

Context-Based Personalisation in Neural Machine Translation of Dialogue

Sebastian T. Vincent



Supervisor: Carolina Scarton

A report submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Computer Science

in the

Department of Computer Science
September 2023

To my parents, Ewa & Marek.

ABSTRACT

Neural machine translation (NMT) has revolutionised automatic translation and has been instrumental in saving costs and improvements in productivity within the translation industry. However, contemporary NMT systems are still primarily designed to translate isolated sentences, disregarding crucial contextual information in the process. This lack of context awareness frequently leads to assumptions about the most likely interpretation of the source text, potentially propagating harmful biases learned from the training data, such as assuming that the average participant in a conversation is male. In the dialogue domain, where the meaning of an utterance may vary depending on what was said before, the environment, the individuals involved, their relationship, and more, translations produced by context-agnostic systems often fall short in capturing the nuances of specific characters or situations.

This thesis expands the understanding of and explores the potential applications of contextual NMT with focus on personalisation. Our methods challenge the prevailing context-agnostic strategy in machine translation and seek to address the aforementioned issues. Our research suggests that by integrating existing information into the translation process we can enhance the quality of translation hypotheses. Additionally, we demonstrate that one type of information can be effectively leveraged to enable manipulation of another. Our experiments involve adapting machine translation systems to individual speakers and productions, focusing on combinations of their individual characteristics rather than relying on discrete labels. We also explore personalisation of language models based on context information expressed in this way: to personalise a model for a particular character, we use a combination of their traits. These personalised language models are then used in an evaluation scenario where the context specificity of machine translation hypotheses is expressed as the pointwise mutual information between the proposed text and its original context. Finally, our best personalised NMT system is thoroughly evaluated in a professional multi-modal setting of translating subtitles for TV series on two language pairs: English-to-German and English-to-French. Throughout the thesis, we report on experiments with various types of context in a setting of translation between English and a range of European languages. Our chosen domain is dialogue extracted from TV series and films, due to the availability of context-rich datasets, as well as the potential practical application of this research to the work of the industrial partner to this PhD, ZOO Digital¹.

¹ <https://www.zoodigital.com/>.

Our research tackles five primary challenges:

1. Direct incorporation of extra-textual information into neural machine translation systems.
2. Zero-shot and few-shot control of this information.
3. Reference-free evaluation and analysis of contextual NMT.
4. Personalisation of language models (LMs) and NMT systems using rich sets of speaker and film metadata annotations.
5. Human evaluation of machine translation in a professional post-editing setting.

By addressing these challenges, this thesis aims to enhance machine translation in dialogue by ensuring translations are better suited to the specific characters, addressees, and contextual factors involved. The research contributes to the advancement of NMT systems that can effectively account for the personalised nature of dialogue.

PUBLICATIONS

Vincent, S., Flynn, R., Scarton, C. (2023), MTCue: Learning Zero-Shot Control of Extra-Textual Attributes by Leveraging Unstructured Context in Neural Machine Translation, in 'Findings of the Association for Computational Linguistics: ACL 2023', Association for Computational Linguistics, Toronto, Canada, pp. 8210-8226.

URL: <https://aclanthology.org/2023.findings-acl.521/>

Vincent, S., Barrault, L. & Scarton, C. (2022a), Controlling formality in low-resource NMT with domain adaptation and re-ranking: SLT-CDT-UoS at IWSLT2022, in 'Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)', Association for Computational Linguistics, Dublin, Ireland (in-person and online), pp. 341–350.

URL: <https://aclanthology.org/2022.iwslt-1.31>

Vincent, S. T., Barrault, L. & Scarton, C. (2022a), Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation, in 'Proceedings of the 23rd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Ghent, Belgium, pp. 121–130.

URL: <https://aclanthology.org/2022.eamt-1.15>

Vincent, S. (2021), Towards Personalised and Document-level Machine Translation of Dialogue, in 'Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop', Association for Computational Linguistics, Online, pp. 137–147.

URL: <https://www.aclweb.org/anthology/2021.eacl-srw.19>

ACKNOWLEDGMENTS

In the first instance I would like to express my deep gratitude to my supervisor Dr Carolina Scarton, who was an outstanding collaborator, advisor and guide throughout my PhD journey. Her contribution played a crucial role in my success, and I am incredibly thankful for her guidance.

I am grateful to the members of my panel: Dr Lexi Birch and Prof Nikos Aletras for taking their time to critically evaluate my work, for our fruitful viva discussion and their guidance on making this thesis the best version of itself. I also thank the members of the intermittent panels held throughout the process: Prof Eleni Vasilaki, Dr Emma Norling, Dr Chenghua Lin and Prof Rob Gaizauskas for their advice and constructive criticism.

I owe an immense debt of gratitude to my family and closest friends. My dearest fiancé Jakub's patience and love has been a blessing on this journey. My parents Ewa and Marek, and my sister Klaudia, have been an endless source of love and support. I cannot forget our lovely greyhound Ronald ([Figure 1](#)) who was by my side during much of the writing process. I am also grateful for the unwavering support of my closest friends Brodie, Meg, Peter, Sylvie, Tasmine, Tom, and Tymon.

Additionally I would like to express my appreciation to my CDT colleagues. First, to my friend, CDT colleague and collaborator Rob, for discussing ideas and concepts with me, and without whose expertise and advice the thesis would not have been what it is today. To Danae, Hussein, Rosa, Will, and many, many others. I am grateful to my industrial collaborators from ZOO Digital: Chris Bayliss, Charlotte Blundell, Alice Dowek, Stuart Green, Mark Matthews, Chris Oakley, Emily Preston, Joshua Rocchi, and Rowanne Sumner for providing a practical scenario for my research, as well as for the collaboration on data collection and human evaluation. I am also thankful to the anonymous reviewers who provided me with helpful feedback regarding my conference and journal submission, strengthening my publications and this thesis. I would like to acknowledge my second supervisor throughout the first half of the thesis, Loïc, and the CDT administrators, Stu, Rachael and Lizzy, the CDT administrators, for their support and guidance. Finally, I am indebted to the CDT directors, Rob and Thomas, for creating the Speech and Language Technologies CDT and made my PhD possible.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of the High Performance Computing Service. This work was also supported by ZOO Digital.



Figure 1: Ronald.

CONTENTS

1	Introduction	1
1.1	Aims and Objectives	4
1.2	Contributions	6
2	Controlling Extra-Textual Attributes in Translation	9
2.1	Chapter Overview	9
2.2	Related Work	10
2.3	Controlling Extra-Textual Attributes about Dialogue Participants	13
2.3.1	Introduction	13
2.3.2	Problem Specification	15
2.3.3	Experimental Setup	16
2.3.4	Results	23
2.3.5	Conclusions	25
2.4	Controlling Formality in Low-Resource NMT	26
2.4.1	Introduction	26
2.4.2	Shared Task Details	27
2.4.3	Proposed Approach	28
2.4.4	Experimental Setup	34
2.4.5	Results	36
2.4.6	Conclusions	38
2.5	MTCUE: Learning Zero-Shot Control of Extra-Textual Attributes in NMT	39
2.5.1	Introduction	39
2.5.2	Proposed Architecture: MTCUE	41
2.5.3	Data: The OpenSubtitles18 Corpus	43
2.5.4	Evaluation	45
2.5.5	Baselines and Implementation	47
2.5.6	Results	49
2.5.7	Examples of Model Outputs (Zero-Shot)	53
2.5.8	Ablation Study	54
2.5.9	Conclusions	56
2.6	Chapter Conclusions	57
3	Reference-less Analysis of Context Specificity in Translation	59
3.1	Introduction	59
3.2	Related Work	61
3.3	Building a Personalised Language Model	64
3.3.1	Datasets	64
3.3.2	Details regarding the data collection campaign	67
3.3.3	Experimental Setup	70

3.3.4	Results	74
3.4	Measuring Personalisation in Professional and Machine Translations . .	78
3.4.1	Datasets	79
3.4.2	Evaluation	82
3.4.3	Machine Translation Systems	82
3.4.4	Results	83
3.5	Cost-benefit Analysis of Human Annotations	88
3.6	Pre-training Strategy: Past Dialogue as Proxy for Metadata	89
3.7	Conclusions	90
4	Assessing Contextual MT in a Professional Scenario of Subtitling	93
4.1	Introduction	93
4.2	Related Work	94
4.3	Experimental Setup	96
4.3.1	Examined System and Baselines	96
4.3.2	Evaluation	96
4.4	Results of Automatic Evaluation	103
4.5	Results of Human Evaluation	103
4.5.1	Error Analysis	104
4.5.2	Analysis of Effort and Quality	110
4.6	Conclusions	116
5	Concluding Remarks	119
5.1	Assessment of Contributions	119
5.2	Limitations	122
5.3	Future Work	123
A	Preliminaries	127
A.1	Introduction to Machine Learning	128
A.1.1	Neural Networks	128
A.1.2	Model Architectures and Components	131
A.1.3	Learning Paradigms	136
A.2	Natural Language Processing	137
A.2.1	Tokenisation	137
A.2.2	Text Embeddings	138
A.2.3	Neural Machine Translation	140
A.2.4	Language Modelling	143
A.2.5	Beam Search	144
A.3	Statistical Concepts Employed Within This Thesis	145
A.3.1	Statistical Significance Testing	147

LIST OF FIGURES

Figure 1.1	Trajectory of the state-of-the-art performance of neural machine translation systems.	1
Figure 2.1	Example of an ambiguous English sentence with all plausible translations to Polish.	14
Figure 2.2	Contributions of each grammatical category to each attribute in the extracted corpus.	18
Figure 2.3	Translation quality (chrF++) for each contextual group.	24
Figure 2.4	Validation accuracy plot showing the effect of applying FORMALITYRERANK to a list of k model hypotheses.	32
Figure 2.5	A high-level overview of MTCue.	40
Figure 2.6	The MTCue architecture.	42
Figure 2.7	UMAP visualisation of how various contexts impact the formality of produced translations when used as input in MTCue.	50
Figure 2.8	Evaluation results from the EAMT22 multi-attribute control task.	51
Figure 3.1	Comparison between CORNELL and our proposed CORNELL-RICH.	66
Figure 3.2	Visualisation of a subset of features of the proposed corpus.	66
Figure 3.3	An illustration of the pre-training and fine-tuning regimens used in the experiments.	71
Figure 3.4	sRR illustrated for one speaker (Hannah).	73
Figure 3.5	Perplexity reduction from training LMCue with individual speaker attributes.	89
Figure 4.1	Snapshots of the ZOOSUBS system in action.	98
Figure 4.2	BLEU, COMET and PMI scores obtained by the evaluated models.	102
Figure 4.3	Effort for each PE for the English-to-French language pair.	111
Figure 4.4	Effort for each PE for the English-to-German language pair.	112
Figure 4.5	A comparison of the effort of translation from scratch and post-editing machine translation outputs.	113
Figure A.1	Example of a feed-forward neural network	129
Figure A.2	The Transformer architecture.	133
Figure A.3	Look-ahead Attention Mask	135
Figure A.4	Word embeddings	139
Figure A.5	The siamese network architecture of SBERT.	139

LIST OF TABLES

Table 2.1	A TV segment along with available metadata.	13
Table 2.2	Attributes and types controlled in the experiment.	16
Table 2.3	Quantities of unique data used for: model pre-training, model fine-tuning and the testing set for calculation of restricted impact.	17
Table 2.4	Training data quantities for all combinations of contexts with examples for each combination.	19
Table 2.5	Comparison of examined approaches to attribute controlling.	20
Table 2.6	Translation performance of all models.	23
Table 2.7	Results of the oracle experiment.	31
Table 2.8	Combinations of formality annotations for the EN-DE-ES triplet extracted from the MuST-C dataset.	33
Table 2.9	Corpora containing training data used in the experiments.	35
Table 2.10	Results on the development sets.	36
Table 2.11	Official results of the automatic evaluation of formality control reported for formal and informal register.	37
Table 2.12	Official results of the automatic evaluation of translation quality.	37
Table 2.13	Percentage of system outputs labelled by professional translators according to the formality level.	38
Table 2.14	Data quantities for the extracted OpenSubtitles18 corpus.	44
Table 2.15	Example of a source-target pair and metadata in OPENSUBTITLES.	45
Table 2.16	Model details for MTCue and baselines.	48
Table 2.17	Translation quality results on the OPENSUBTITLES test set.	49
Table 2.18	Evaluation on the IWSLT 2022 formality control evaluation campaign.	52
Table 2.19	Ablation study on model components and data settings.	55
Table 3.1	A sample from the ZOO-ENGLISH corpus.	65
Table 3.2	Details of annotations compared to data quantities from CORNELL.	67
Table 3.3	A sample from CORNELL-RICH with each type of collected metadata.	68
Table 3.4	Quantities of segments, tokens and unique metadata in the OPENSUBTITLES, CORNELL-RICH and ZOO-ENGLISH datasets.	70
Table 3.5	Model details for LMCUE and BASE-LM.	72
Table 3.6	Perplexity [↓] on different validation and testing sets for the two corpora.	74
Table 3.7	Selected metadata regarding long-term speakers from ZOO-ENGLISH used in the experiment.	75

Table 3.8	Sentences and tokens for which the log likelihood under LMCUE ($\mathcal{S} + \mathcal{P}$) changes the most compared BASE-LM.	75
Table 3.9	Results on the test set for long-term speakers.	76
Table 3.10	Results of evaluation with speaker & film metadata on the test set of unseen speakers.	77
Table 3.11	Details regarding the ZOO-MULTI corpus.	80
Table 3.12	Quantities of segments, tokens and unique metadata in the OPENSUBTITLES and ZOO-ENGLISH datasets.	81
Table 3.13	PMI computed with general and personalised language models on translations	84
Table 3.14	BLEU and COMET scores for the evaluated MT systems.	84
Table 3.15	Results on test_unseen of CORNELL-RICH from different pre-training/fine-tuning setups.	90
Table 4.1	Work assignment to PEs and translators in the human evaluation campaign.	99
Table 4.2	Details regarding PEs who took part in the campaign.	100
Table 4.3	List of errors provided to the human evaluators during the campaign.	101
Table 4.4	Error counts and normalisation coefficients h for each post-editor (PE) in the experiment.	105
Table 4.5	Counts of errors flagged by the PEs for each system.	108

ACRONYMS

NLP natural language processing

NN neural network

FFNN feed-foward neural network

RNN recurrent neural network

SEQ2SEQ sequence-to-sequence

SBERT Sentence-BERT

MT machine translation

NMT neural machine translation

LM language model

MTE machine translation evaluation

BPE Byte-Pair Encoding

PMI pointwise mutual information

MRR mean reciprocal rank

SMRR speaker mean reciprocal rank

EN-DE English-to-German

DE-EN German-to-English

EN-RU English-to-Russian

RU-EN Russian-to-English

EN-PL English-to-Polish

PL-EN Polish-to-English

EN-FR English-to-French

FR-EN French-to-English

EN-IT English-to-Italian

EN-ES English-to-Spanish

PE post-editor

YOE years of experience

INTRODUCTION

Neural machine translation (NMT) has seen a series of rapid advances in the recent years with the advent of the Transformer architecture (Vaswani et al. 2017) along with innovative techniques such as back-translation (Sennrich et al. 2016c) and sub-word tokenisation (Sennrich et al. 2016d). The impressive results achieved with these methods still remain competitive today (Figure 1.1).

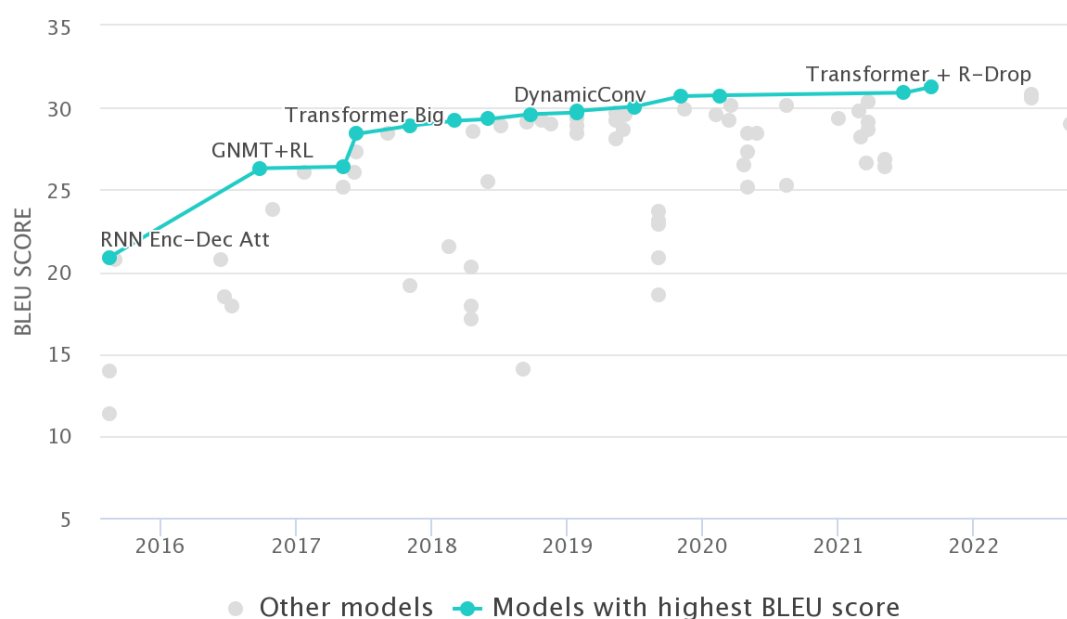


Figure 1.1: Trajectory of the state-of-the-art performance of neural machine translation systems. The benchmark is an English-to-German WMT translation task held by Papers With Code (<https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german>). Taken on 01/03/2023.

Originally, this overnight success of NMT led some researchers to make a claim about human parity (Hassan et al. 2018), one since refuted or challenged by other work (Läubli et al. 2018, Graham et al. 2020, Toral 2020). The opponents argued that the evidence obtained in the evaluation task was too weak for the claim, as the campaign was limited to translation of isolated sentences, failing to capture the complexities of translating full (multi-sentence) texts. As a consequence of this and to address this

limitation, research within NMT started to focus more on incorporating **context** – any textual information related to the input text, whether **intra-** or **extra-textual** – to generate translation hypotheses.

The shift from one paradigm (**sentence-level**) to another (contextual), however, necessitates challenging the long-standing status quo. Dominant in the field for years, the **context-agnostic** framing of machine translation has shaped various systemic aspects in its image. Firstly, the notion of context in translation is unaccounted for in many evaluation benchmarks of NMT, including the WMT 2014 benchmark presented in [Figure 1.1](#). In fact, how well NMT performs is typically measured on context-agnostic test sets; in accordance with this criterion, the best system is the one which can most effectively translate individual sentences divorced from their broader context and the surrounding text. This evaluation approach aligns with the most commonly used metrics, such as BLEU, CHRF++ or COMET, which focus on quantifying the similarity between the reference translation and the model’s hypothesis. Secondly, the majority of training corpora for NMT consist of pairs of sentences expressed in two different languages, devoid of any contextual information. Finally, the architectures of most NMT systems are primarily designed to address the challenge of translating input sequences of limited length into output sequences of similar length. These architectures do not lend themselves easily to accommodating additional inputs, such as contextual information, thus perpetuating the context-agnostic paradigm.

The context-aware framing of NMT has a parallel with everyday life: when we engage in a conversation, we do not process what is said to us in isolation, but rather consider the broader context, including who said it, where and when it is being said, what was said before, and more. While the simplification of disregarding such information has ultimately significantly helped accelerate the progress of NMT, it has also created a simplistic narrative for the task that is difficult to escape. To make the shift towards context-aware NMT, it is imperative that the employed architectures, datasets, and evaluation are adapted to more faithfully reflect the contextual complexity of naturally occurring text.

Recent few years have already seen the first signs of transition towards contextual NMT: a considerable body of work was devoted to the challenge of incorporating **document-level** context (i.e. past and future sentences on the source or target side) into it. Questions regarding this issue have been considered in many publications, including how context affects translation ([Voita et al. 2019a](#)), which parts of the surrounding text are needed to resolve the inter-sentential phenomena (e.g. [Kim et al. 2019](#)), how to efficiently incorporate this information into the translation pipeline (e.g. [Lupo et al. 2022a,b](#)) and how to address the lack of parallel document-level data (e.g. [Yu et al. 2020](#)). In contrast, the problem of incorporating extra-textual information into translation has attracted significantly less focus, with most studies concentrating exclusively on controlling formality (e.g. [Sennrich et al. 2016a](#), [Anastasopoulos et al. 2022](#)) or gender (e.g. [Vanmassenhove et al. 2018](#), [Moryossef et al. 2019a](#)), despite calls for research in

natural language processing (NLP) to become more **personalisation**-focused (Flek 2020, Dudy et al. 2021). More attention ought to be given to how various kinds of such extra-textual information can be incorporated into machine translation, and what effect they have on it. That is precisely the goal of this thesis.

Context plays a special role in interpreting and generating **dialogue**, here defined as the interactive act of communication which relies on verbal and non-verbal cues to maximise the understanding between two or more interlocutors. In contrast with monologue (e.g. news articles, books, blog posts and talks), in dialogue things are often communicated without being said explicitly, as they can be recovered by the addressee from previous utterances or the environment. As the speaker, we possess the knowledge of how we should be addressed, and how we should address our interlocutor (whether formally or informally). Finally, every person has their own unique speaking style, which - though only in part - is determined by their background, gender, age or country of origin. A one-size-fits-all model will ignore all of these factors, creating the *source ambiguity* problem, where the same source sentence, which will occasionally require different interpretations in different contextual scenarios, gets processed the same way in all cases as the context information is disregarded.

As our first contribution, we must now make a novel but necessary distinction between two ways in which extra-textual context can affect text: grammatical and behavioural agreement. **Grammatical agreement** describes the known, well-defined, grammatical way in which context affects its text. A fundamental example of this phenomenon is the use of honorifics in situations requiring formal register in languages such as German (*you* expressed as *Sie* formally or *du* informally). Another example is the expression of the speaker's or addressee's gender in the Polish language via morphosyntactic markers (*I was* expressed as *byłem* by the masculine speaker and *byłam* by the feminine one). **Behavioural agreement**, on the other hand, pertains to the way language *tends to be used* in the given context. For example, the legal jargon is far more likely to be employed by a lawmaker or a student of law than the average person; the phrase *Yes, God!* has a meaning radically different to the queer community than to the Christian community, especially of the older generation; and depending on which anglophone country an English speaker is from, they may use the term *cookie* or *biscuit* to refer to the same sweet snack, or even a different word entirely. The effects of the genre, tone or domain on text, typically studied within machine translation research, all fall within behavioural agreement. None of the above examples are "hard-coded" into language. Furthermore, some context variables influence both types of agreement: Vanmassenhove et al. (2018) highlight that the gender of the speaker influences both the morphology of words describing the subject (grammatical agreement), as well as offer a discussion of the French term *crois* ("believe") which is used more frequently by males than females (behavioural agreement). Within this thesis, we explore context variables which induce either or both types of agreement, and while we offer specific tools for calculating accuracy of grammatical agreement control, we propose that behavioural

agreement be evaluated with a bespoke tandem of systems which calculate context specificity of the given translation, i.e. how likely the given translation is to occur in the provided context when compared to the general case.

Finally, a potential solution to the problem of contextual machine translation may be on the horizon with the recent advent of large language models. With the use of the likes of CHATGPT, one may soon be able to produce a contextualised translation simply by specifying in one's input prompt what kind of adaptations ought to be made to it. However, harnessing the power of such models even for translation alone remains an open task. Furthermore, we envision that in the future where such models become widely available, there are still undeniable advantages to deploying smaller-scale tailored solutions like the ones described in this work. Such models can be trained on modest hardware, and with smaller amounts of hand-selected training data, giving the user total control over what the model is exposed to.

1.1 AIMS AND OBJECTIVES

The majority of this thesis is centred on the domain of scripted dialogue, which encompasses text found in sources such as subtitles, transcriptions or scripts of TV series and film. Several reasons make this domain particularly suitable for our study. First and foremost, when compared to monologue, the interpretation and processing of dialogue is generally more reliant on contextual cues (Halliday & Hasan 1976, Pickering & Garrod 2004, Danescu-Niculescu-Mizil & Lee 2011). Secondly, there is an abundance of parallel dialogue corpora extracted from subtitles. Furthermore, this data can be enriched with various meta-information, including details about the show or film itself (e.g. its plot), the discourse context (e.g. descriptions of scenes) or information about the characters involved (e.g. their ages or countries of origin). In contrast to real-life dialogue, significantly larger datasets exist within this domain, partly due to the ethical concerns surrounding the processing of personal profiles of real individuals. Lastly, it is worth noting that machine translation of subtitles remains an ongoing challenge within the subtitling industry. One of the key objectives of this thesis is to investigate whether contextual systems can improve the efficiency of this automation.

The work presented in this thesis is carried out in a number of language pairs, dictated by the availability of data and the feasibility of application of the models in the future. As such, all explored language pairs involve translation from English (to Polish, French, German, Spanish, Italian, and Russian) and to English (from Polish, French, German, and Russian).

Specifically, we address the following research questions:

RQ1 How can attribute control best be incorporated into neural machine translation in multiple attribute and low-resource scenarios?

Within this research question addressed in [Chapter 2](#), we investigate how

elementary extra-textual phenomena can be effectively controlled in translation, particularly focusing on the English-to-Polish language pair (§ 2.3), as well as exploring low-resource control of formality in four language pairs (§ 2.4). Finally, we present MTC_{UE}, a novel solution which enables control of the same phenomena in a few- and zero-shot fashion, without the need for attribute-annotated translation samples (§ 2.5).

***RQ2** Can language models for film and TV characters be personalised solely relying on their character profiles and information on the discourse environment, and used to evaluate context-specificity in personalised machine translation?*

The second research question (Chapter 3) first explores whether language modelling of dialogue can benefit from rich metadata annotations. We address it by contributing an English-language corpus of film dialogue annotated with rich speaker profiles and film metadata, and showing such annotations effectively enable language model personalisation, even in cases where no prior dialogue data is available for a given character. Secondly, we explore the application of such personalised language models in the evaluation of **context specificity** of machine translations.

***RQ3** How does personalisation affect translation quality and post-editing effort in a real-life scenario of subtitle translation?*

In the final research chapter (Chapter 4) we delve deeper into using contextual information in translation, this time directly focusing on the industrial use case of the thesis. We apply the outputs of MTC_{UE}, as well as several baselines, in a professional multi-modal system for subtitle translation and post-editing. Our analysis suggests that MTC_{UE} makes fewer context, style and fluency errors, especially in the English-to-French (EN-FR) language pair. We also contribute a wealth of findings regarding future human evaluation campaigns in this domain.

This thesis received partial funding from ZOO Digital¹, a Sheffield-based company specialising in subtitling and dubbing services. They also contributed an industrial use case as the objective for the research. Owing to this collaboration, we obtained diverse subtitle and transcribed data from various sub-domains of streaming content, including films, unscripted entertainment, and cartoons. The data came with valuable annotations, such as character profiles and scene descriptions, enabling us to conduct more comprehensive experiments.

¹ <https://www.zoodigital.com/>

1.2 CONTRIBUTIONS

The present work makes the following research contributions:

- We propose a novel distinction between *grammatical* and *behavioural* agreement in contextual translation (Chapter 1).
- We implement a tool for annotating utterances in the Polish language with speaker and interlocutor attributes at over 99% validation accuracy (§2.3).
- We build a machine translation system which produces translations in agreement with the provided gender of the speaker, gender and number of the interlocutor(s) and formality, at over 99% accuracy (§2.3).
- We build a formality-controlling machine translation system for translation from English into German and Spanish in a low-resource scenario (§2.4).
- We propose MTC_{CUE}: a context-aware machine translation model which can utilise any context expressible in natural language to produce translations of better quality (§2.5).
- We show that MTC_{CUE} achieves 100% zero-shot accuracy at controlling the formality of translations in English-to-German on a shared-task test set, and significantly improves on zero-shot and few-shot control of multiple attributes in English-to-Polish translation (§2.5).
- We contribute CORNELL-RICH, a dataset of rich character annotations for a subset of the most featured characters from the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil & Lee 2011) (§3.3.1).
- We propose LMC_{CUE}: an extension of MTC_{CUE} adapted to the task of contextual language modelling and show over a set of comprehensive experiments that LMC_{CUE} significantly outperforms parameter-matched LM baselines (reducing perplexity by up to 6.5%) and performs on par with per-speaker fine-tuning methods, while requiring no such fine-tuning (§3.3).
- We introduce a new metric, speaker mean reciprocal rank (sMRR), which measures how well a personalised language model captures the language patterns of a character (§3.3.3.4).
- We devise a formulation and an experimental analysis regarding the use of personalised LMs to evaluate how specific the given translation hypotheses are to the extra-textual context they arise in, together with a range of examples investigating the evaluation behaviour (§3.4).

- We contribute a cost-benefit study showing which speaker characteristics have contributed the most to perplexity reduction relative to the cost of collecting the information (§3.5).
- We present empirical evidence that pre-training LMCUE on document-level information helps realise personalisation in fine-tuning on considerably smaller corpora with access to extra-textual context (§3.3.3.2).
- We present the results of a human evaluation campaign performed in a professional production setting of subtitle translation across three different types of TV series, showing that leveraging context in machine translation yields significant improvements in context-specific errors marked during post-editing in the English-to-French translation direction (§4.5).
- We contribute a taxonomy of possible errors in post-editing machine-translated subtitles and present a detailed analysis of the types of errors made by non-contextual and contextual machine translation systems (§4.3.2).
- We report the results of a survey conducted among the participating professional translators, gathering their views on the future of machine translation in the subtitle industry, as well as their expectations from the technology (§4.5.2.1).

The rest of the thesis is structured as follows. Immediately following this Introduction are the three research chapters ([Chapter 2](#), [Chapter 3](#), [Chapter 4](#)), each accompanied with a tailored Related Work section providing the necessary background. In the appendix ([Appendix A](#)) we provide the necessary background information and context to enable the reader to understand this research project. It lays the foundation for the rest of the thesis and contains essential information that is necessary for the reader to comprehend the scope and nature of the research. Since many readers will already have been familiar with these concepts, we advise to read the chapter only if the reader is not already familiar with machine learning or Natural Language Processing. Otherwise, if a certain concept is unclear, throughout the thesis we hyperlink the employed concepts to [Appendix A](#) for easy access to a definition. Finally, we conclude the thesis and outline the directions for future work in [Chapter 5](#).

CONTROLLING EXTRA-TEXTUAL ATTRIBUTES IN TRANSLATION

2.1 CHAPTER OVERVIEW

Neural machine translation has made significant progress in the recent years, much owing to the advent of the Transformer architecture (§A.1.2.1), as well as supplementary techniques such as sub-word segmentation (§A.2.1) or back-translation (§A.2.3.1). Nevertheless, much remains to be done in certain aspects of the task. One challenge NMT faces today is source ambiguity. When a source sentence contains the same ordered set of tokens, it is either always translated in the same way, or the set of different hypotheses produced is not related to the context in which the source arises in the original text. This phenomenon is easiest exemplified when the speaker’s grammatical gender is involved, which in some languages can determine the morphological endings to certain words.

In this chapter, we explore the role of *extra-textual* attributes in NMT. In §2.3 (Vincent et al. 2022b), we centre the attention on the English-to-Polish (EN-PL) translation direction and introduce methods which sensitise NMT to four extra-textual variables: formality, the speaker’s gender, and the gender and number of the interlocutor(s). We develop a tool which automatically annotates these four attributes based on morphosyntactic evidence found in the target (Polish) sentence itself, and then use that data to train a model which can control these attributes.

In §2.4 (Vincent et al. 2022a), we present our winning submission to the IWSLT 2022 formality control task (Anastasopoulos et al. 2022). Our approach addresses the problem of low-resource formality control in multiple language pairs, addressing few-shot control (§A.1.3) in English-to-German (EN-DE) and English-to-Spanish (EN-ES), as well as zero-shot control (§A.1.3) in English-to-Russian (EN-RU) and English-to-Italian (EN-IT).

Finally, in §2.5 (Vincent, Flynn & Scarton 2023) we introduce MTC_{CUE}, a contextual translation architecture which consumes as input the source sentence and various contextual information (film metadata, past sentences) and produces translations which are more adapted to the provided context. In our evaluation, we show that MTC_{CUE} offers a novel and robust solution to the two problems described in the preceding sections: by utilising various sources of context, it learns a general representation of context and is then able to scale to new extra-textual variables. We show that MTC_{CUE} trained on metadata and past context achieves over 80% accuracy on the multi-attribute control task described in §2.3 and 100% accuracy at formality control on the IWSLT

2022 shared task test sets in the [EN-DE](#) and [EN-RU](#) pairs, both **in a zero-shot fashion** and requiring no adaptations to those specific downstream attributes.

2.2 RELATED WORK

As a domain, dialogue presents a unique set of challenges within machine translation, owing to the properties of discourse. Dialogue is naturally coherent and cohesive ([Halliday & Matthiessen 2013](#)), and this manifests itself in the text in three ways:

- via *reference*, where the speaker refers to elements with pronouns or synonyms that they judge recoverable from somewhere else in text;
- via *ellipsis and substitution*, where the speaker omits parts of or whole phrases which can be unambiguously recovered by the addressee;
- via *lexical cohesion*, where the speaker chooses words related to those that have been used earlier.

Dialogue is also naturally in agreement with the environment it is spoken in, the dialogue participants and their unique attributes ([Halliday & Matthiessen 2013](#)). It reflects the speaker's chosen tone to effectively convey their message. Ambiguity and polysemy present within the utterance can often be automatically resolved in the presence of context. However, machine translation, typically designed to translate isolated utterances, may fall short in capturing these inherent characteristics, resulting in text that lacks natural properties.

Most studies on incorporating contextual information into [NMT](#) of dialogue have focused on [document-level](#) context, targeting specifically coherence and cohesion phenomena such as ellipsis or reference. Notable approaches include using multiple encoders (e.g. [Miculicich et al. 2018](#)), cache models ([Kuang et al. 2018](#)), automatic post-editing ([Voita et al. 2019a](#)), shallow fusion with a document-level language model ([Sugiyama & Yoshinaga 2021](#)), data engineering ([Lupo et al. 2022a](#)) or simple concatenation models ([Tiedemann & Scherrer 2017](#)). A different branch of contextual models seeks to restrict or guide hypotheses via variable controlling to certain external conditions. One of the first such works focuses on controlling the formality register applied in translation hypotheses. [Sennrich et al. \(2016a\)](#) show that this can be achieved in the [EN-DE](#) translation direction with a **side constraints** approach whereby a meaningful tag is prepended to those training data samples which convey specific formality. The same idea is revisited by numerous works in [NMT](#) and [NLP](#), alternatively under the name of **tagging** or **user embedding**. For example, some researchers have applied it in the context of interlocutors' gender identities: [Vanmassenhove et al. \(2018\)](#) explore the idea of gender identity being a linguistic signal in translation, analysing how a *male/female* tag impacts the vocabulary used by the members of the European parliament. [Moryossef et al. \(2019a\)](#) propose the idea that [NMT](#) models can be primed

to use the correct self-referent gendered words via inference-time prefixes such as “He said:” or “She said:”. Their results suggest that such a priming approach is sufficient to control this aspect of translation; however, their study is evaluated on one specific target scenario (a woman speaking to a plural audience). Findings from both studies suggest that controlling the gender of the speaker improves translation quality in languages which contain a grammatical system for expressing that gender.

The idea of using prefixes or tags in training data was later established as an easily accessible method of imposing constraints in generation tasks, and successfully applied to control the translation length (Lakew et al. 2019a, , using discrete length categories: *short*, *medium* and *long*), vocabulary used (Post & Vilar 2018) or domain and genre (Matusov et al. 2020). In Johnson et al. (2017), the target language itself is treated as extra-textual context in a multilingual setting. Schioppa et al. (2021) notably addresses the problem of controlling attributes such as length, monotonicity (closeness of the word order in the source and target sentences) and formality on a continuous scale by shifting the outputs of each encoder layer by a control vector \mathbf{V} scaled by a continuous weight w which corresponds to the “strength” of the controlled attribute. Their results suggest that this formulation enables more fine-grained adaptations for continuous phenomena than a discrete “bucketing” approach of e.g. Lakew et al. (2019a). Controlling multiple attributes with this approach has not been excessively studied (Schioppa et al. 2021), though works such as Takeno et al. (2017) and Lample et al. (2019) show that this can be facilitated by concatenating the control tokens or averaging the vectors corresponding to their embeddings.

Typically, attribute-controlling models are fully supervised, requiring annotated training data. Such annotations can be obtained directly, e.g. from metadata (Vanmassenhove et al. 2018); although most available corpora are unannotated. Sennrich et al. (2016a) and Elaraby et al. (2018) automatically annotate the data using morphosyntactic parsers based on rules, validating agreement to the attribute in question in target-side sentences. To verify that the rules capture the attribute completely, a precision/recall score is computed against a manually labelled test set. However, this solution is not without its issues: (i) it is time-consuming and difficult to implement (e.g. the annotation rules produced by Elaraby et al. yield a recall value of 50 – 71.42% across markings of different speaker and interlocutor genders), (ii) this solution is only applicable when the effect of the individual attributes on text is fully known, which is not the case for most contextual phenomena in language. Although contextual adaptation in NMT to phenomena that span beyond gender, formality and domain has been discussed theoretically, most empirical research falls back to gender (Rabinovich et al. 2017) or formality control (Niu et al. 2017). Somewhat of an exception, Michel & Neubig (2018a) adapt NMT for each of many speakers by adding a “speaker bias” vector to the decoder outputs. They find that explicitly modelling speaker-related variation has a positive impact on BLEU and slightly improves speaker

classification accuracy (determining the authorship of the translation hypothesis among all speakers).

In conclusion, a tagging approach offers a simple and effective way of introducing a more fine-grained control to an otherwise one-size-fits-all model, however it relies on the user knowing in advance what needs to be controlled, and that data annotation or identification methods exist to collect sufficient training data for the task. In § 2.5 (Vincent, Flynn & Scarton 2023), we address this drawback by proposing a novel NMT architecture which learns from the available context (such as document-level and metadata information) to enable few- and zero-shot control of some of the most popular extra-textual attributes such as formality. Part of that work is motivated by the CUE vectors (Novotney et al. 2022). The CUE approach represents contextual variables as equal-sized vectors computed by passing sentence embeddings of the input context (computed with the DistilBERT model described in Sanh et al. 2019) through a dedicated encoder. Novotney et al. show that incorporating CUE into a language model improves perplexity within the domain of news articles. Their formulation allows the user to train the contextual model on any set of available variables. In contrast, we reformulate CUE for contextual machine translation, provide a detailed analysis of incorporating CUE into the model, emphasise the importance of vectorising the context variables prior to embedding them, and examine the benefits for zero-shot and few-shot performance in contextual NMT tasks. Furthermore, we propose a number of improvements to the original approach, such as altering the way context is combined with the textual information and using a different sentence embedding model, ultimately showing significant improvements over that setup.

2.3 CONTROLLING EXTRA-TEXTUAL ATTRIBUTES OF DIALOGUE PARTICIPANTS

2.3.1 Introduction

In some languages, *dialogue* explicitly expresses certain information about the interlocutors: for example, while in English words describing the speaker “I” and the interlocutor “you” are ambiguous w.r.t. their gender, number and formality, languages such as Polish, German or Spanish will mark for one or more of these attributes. In industrial settings such as dubbing and speech translation, there is an abundance of available metadata about the interlocutors, such as their gender(s), that could be used to help resolve these ambiguities.

Field	Value
source	"Are you blind?"
spoken by (=speaker)	"Anne"
speaker's gender	"feminine"
spoken to (=interlocutor(s))	["Mark", "Colin"]
interlocutor(s)' gender	"masculine"
formality	"informal"

Table 2.1: A TV segment along with available metadata.

Table 2.1 shows an example of such a TV segment: the English sentence ‘*Are you blind?*’, should translate to Polish as ‘*Jesteście ślepi?*’ as the addressee is a group of men and the setting is informal; however, when spoken e.g. formally to a mixed-gender group of people, the correct translation would read ‘*Są państwo ślepi?*’, using a different verb inflection and an honorific (*państwo*). Since the *contextual* information required to resolve the ambiguity in this example does not belong to the text itself, traditional models do not use it. This yields hypotheses which introduce some assumptions about that context, typically reflecting biases present in the (often unbalanced) training data. To avoid this, a better solution is to resolve such ambiguities by using both the available metadata and the source text as translation input. Alternatively, when such information is unavailable, all possible contextual variants could be provided as output, passing the choice from the model to the user (Jacovi et al. 2021, Schioppa et al. 2021).

In the context of the gender of the speaker and interlocutor, prior research has explored two ways in which such information influences a text (Rabinovich et al. 2017, Vanmassenhove et al. 2018). Firstly, naturally occurring texts satisfy grammatical agreement between the gender of the speaker and interlocutor and the utterances which describe them. How this agreement is expressed in speech varies among different languages (Stahlberg et al. 2007). Polish is a *grammatical gender language*: every noun is assigned a gender, and grammatical forms must agree with that noun. In contrast,

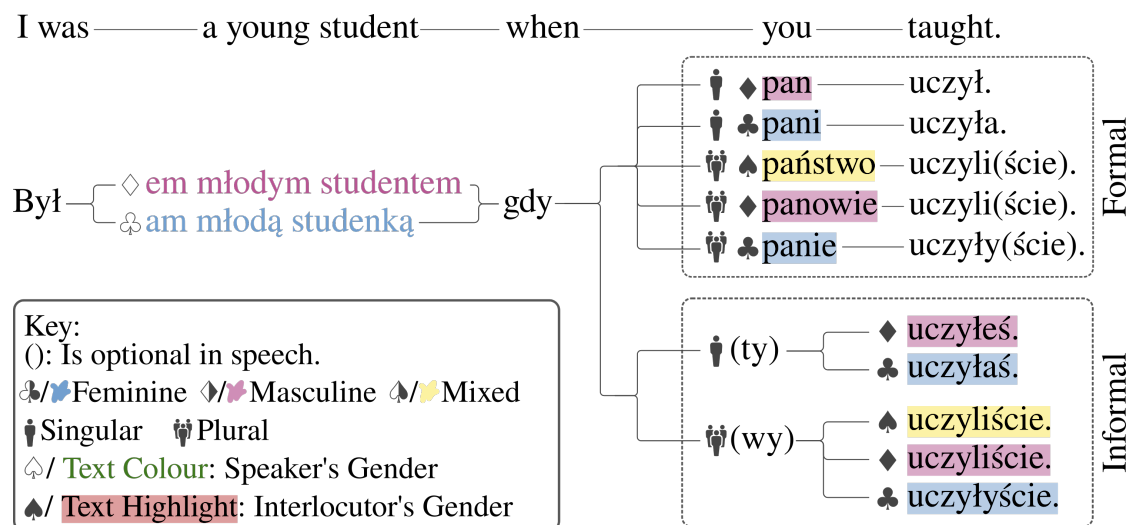


Figure 2.1: Example of an ambiguous English sentence with all plausible translations to Polish. There are a total of 18 equally plausible possible hypotheses based on the combination of contexts.

English is a naturally gender-neutral language, with “no grammatical markings of sex” (Stahlberg et al. 2007, p. 165). Secondly, gender can be seen as a demographic factor that influences the way people express themselves (e.g. word choice). Hereinafter we refer to the former as *grammatical agreement* and the latter as *behavioural agreement*.

In this work, we seek to build neural machine translation (NMT) models that satisfy grammatical agreement. Given an English sentence and a set of attributes (e.g. the gender of the speaker and number of interlocutors), a NMT system must translate this sentence into Polish with a correct grammatical agreement to all attributes but introduce no markings of behavioural agreement. We explore the agreement to one SPEAKER attribute: the gender of the speaker (SPGENDER), and three INTERLOCUTOR attributes: the gender and number of interlocutor(s) (ILGENDER, ILNUMBER), as well as the desired FORMALITY of addressing the interlocutor(s). Figure 2.1 exemplifies the extent of ambiguity these attributes introduce in English-to-Polish translation.

The **main contributions** of the work outlined within this section are:

- (a) A novel English-to-Polish parallel corpus of TV dialogue annotated for SPGENDER, ILGENDER, ILNUMBER and FORMALITY.
- (b) A tool for analysing attributes expressed in Polish utterances.
- (c) The examination of a wide range of approaches to attribute control in NMT, showing that at least four of them can be reliably used for incorporating extra-linguistic information within English-to-Polish translation of dialogue.

This work has been published in the Proceedings of the 23rd Annual Conference of the European Association for Machine Translation (EAMT 2022).¹

2.3.2 Problem Specification

Recognising the small number of studies within machine translation research on the English-to-Polish language direction, as well as our capacity (thanks to the available parsers and native speakers to validate their performance), we decide to focus the study on this language pair. Polish is a West Slavic language spoken by over 50M people over the world (Jassem 2003). It uses an expanded version of the Latin alphabet and is characterised by a complex inflectional morphology (Feldstein 2001). It is a grammatical gender language (Koniuszaniec & Błaszowska 2003) meaning all forms dependent on pronouns must agree to their gender and number. It uses a West Slavic system of honorifics *pani, pan, panie, panowie, państwo* (henceforth *Pan+*) (Stone 1977). Being a null-subject language (Sigurdsson & Egerland 2009), it does not require that pronouns signifying the speaker or the interlocutor are explicit, **unless** they belong to the *Pan+* group (Keown 2003).

English lacks a grammatical gender or a system of honorifics, and the pronoun “you” is used for both plural and singular second person addressees. It is therefore ambiguous w.r.t. some expressions describing the speaker or the interlocutor, which we capture into four attributes, as follows (the attributes are summarised in Table 2.2).

SPEAKER ATTRIBUTES The gender of all forms dependent on the pronoun *ja* (I) must match the gender of the speaker $SP_{GENDER} \in \{feminine, masculine\}$. This includes past and future verbal expressions (e.g. *byłam* ‘I was_{fem}’ vs. *byłem* ‘I was_{masc}’), adjectives (e.g. *piękna* ‘pretty_{fem}’ vs. *piękny* ‘pretty_{masc}’) and nouns (e.g. *wariatka* ‘lunatic_{fem}’ vs. *wariat* ‘lunatic_{masc}’) that describe the speaker.

INTERLOCUTOR ATTRIBUTES All word forms dependent on the pronoun *ty/wy/Pan+* “you”, including the pronoun itself, must match:

- the gender of the interlocutor (IL_{GENDER}); this includes cases analogous to SP_{GENDER}, extended to e.g. vocatives (e.g. *Ty wariatko/cie!* ‘You lunatic_{fem/masc!}’);
- the number of interlocutors (IL_{NUMBER}); this includes verbs and pronouns in second person;

¹ Vincent, S. T., Barrault, L. & Scarton, C. (2022a), Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation, in ‘Proceedings of the 23rd Annual Conference of the European Association for Machine Translation’, European Association for Machine Translation, Ghent, Belgium, pp. 121–130. URL: <https://aclanthology.org/2022.eamt-1.15>

Attribute	Abbreviation	Type
Speaker		
SPGENDER	<sp:feminine>	Feminine speaker
	<sp:masculine>	Masculine speaker
Interlocutor		
ILGENDER	<il:feminine>	Feminine interlocutor(s)
	<il:masculine>	Masculine interlocutor(s)
	<il:mixed>	Mixed-gender interlocutor(s)
ILNUMBER	<singular>	One interlocutor
	<plural>	Multiple interlocutors
FORMALITY	<informal>	Informal
	<formal>	Formal

Table 2.2: Attributes and types controlled in the experiment.

- the formality in addressing the interlocutor (FORMALITY)²; this entails using an inflection of the pronoun Pan+ consistent with ILGENDER and ILNUMBER where applicable, or using polite forms (e.g. *Proszę wejść*. ‘Come in.’).

Throughout this section, when using the term *gender*, we refer to the grammatical gender rendered in text. In the Polish language, the grammatical system of gender in first and second person is a rigid dichotomy of masculine and feminine variants, lacking alternatives for people who identify as neither.

2.3.3 Experimental Setup

2.3.3.1 Data Collection

We collect pre-training data from two corpora: the English-to-Polish part of OpenSubtitles18 (Lison & Tiedemann 2016) and the Europarl (Koehn 2005) corpus. The data quantities can be found in Table 2.3 (column “pretrain”).

Since our studied attributes have a well-understood effect on dialogue, we decide to create a fine-tuning corpus for the task by annotating the pre-training samples with the

² Formality is not necessarily a binary variable, for example Feely et al. (2019) define three stages of formality: formal, polite and informal. In Polish, formality and politeness can be regarded as completely separate phenomena and within this work we focus on the two-stage formality alone.

		pretrain	finetune	amb_test
train	#sents	10.8M	2.9M	—
	#tokens	82.1M	26M	—
valid	#sents	3K	3.5K	—
	#tokens	23.3K	48.7K	—
test	#sents	—	3.5K	1K
	#tokens	—	47.7K	10.3K

Table 2.3: Quantities of unique data used for: model pre-training (pretrain), model fine-tuning (finetune) and the testing set for calculation of restricted impact (amb_test). Values are averaged for source and target text.

desired information; each sample is paired with an annotation of up to four types of attributes. For that purpose, we implement a set of morphosyntactic rules for the Polish SPACY model (Tuora & Kobylinski 2019) which uses the MORFEUSZ2 morphological analyser (Kieras & Wolinski 2017). Since speaker and interlocutor characteristics vary between utterances (as opposed to between words or whole exchanges), we produce one annotation of the four attributes for each utterance. For both speaker and interlocutor gender attributes, the masculine gender makes up over 60% of the corpus. Altogether, a total of 34.33% of the corpus marks at least one of the attributes; Figure 2.2 shows how different linguistic categories contributed to extracting each attribute. In the fine-tuning corpus we only keep samples with at least one attribute marked.

Similarly to Elaraby et al. (2018) and Gonen & Webster (2020), we observe that certain nouns marked as describing the speaker or interlocutor have a fixed gender irrespective of that person’s gender and are therefore inadequate determinants of their gender (e.g. *coward* “tchórz” is always masculine). We could not find a reliable (complete nor heuristic) method to resolve this other than creating a “stopwords” list of all inflexible nouns. The process is now performed in two steps: we first extract a list of sentences containing gender-marked words and then filter out those that were selected based on our “stopwords” list of inflexible nouns.

We extract 223.0K noun-dependent sentences with 9K unique lemmatised nouns in the first pass, build the “stopwords” list of 6.8K words and end up with 67.3K sentences.

ANNOTATION RULES We identify sentences marking for SPGENDER by finding tokens in first person singular and verifying that their head marks feminine or masculine gender. FORMALITY is identified through the use of the inflected pronouns in the *Pan+*

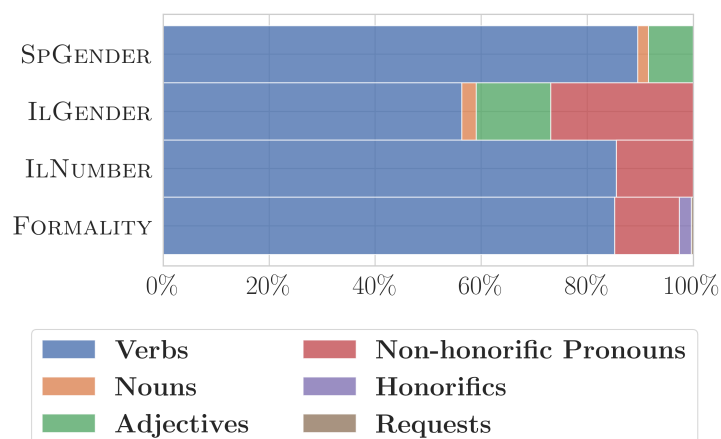


Figure 2.2: Contributions of each grammatical category to each attribute in the extracted corpus.

set (unless it is used as a title, e.g. in ‘*Ms Smith*’). Formal requests are selected by finding *proszę* (‘*please*’) in the target sentence but not in the source. ILGENDER is trivially inferred in formal cases; for informal language, we match structures analogous to those for the SPGENDER and extend them to comparative phrases and vocatives. ILNUMBER follows from the plurality of second-person verbs as well as the use of the pronoun *ty* (‘*you*’, singular) or *wy* (‘*you*’, plural).

To measure the effectiveness of our set of annotation rules, a native Polish speaker with expertise in NLP manually annotated a random sample of 1K sentence pairs from the training corpus for the provided attribute types. Given a sample, the annotator was instructed to identify a type from each attribute, and then highlight a part of the Polish sentence proving its occurrence. Precision and recall (§ A.3) scores were measured between the judgements of the rule set and the annotator. The implemented rule set (hereinafter *Detector*) scored near-perfectly (99.82% precision and 99.17% recall averaged over all attributes) and proved suitable for the tasks of both extracting the corpus and evaluating attribute controlling. Beyond input errors leading to incorrect processing, we observed two consistent cases of failure:

1. when the interlocutor is addressed in plural but is in fact singular (in cases like “Go_{singular} help her. Maybe you [two] will_{plural} figure it out together.” the addressee may be interpreted as *plural* instead of *singular* depending on the majority of grammatical matches for each type);
2. some tag questions (e.g. “prawda?”) or expressions (e.g. the words “kimś” (‘someone_{instr.}’), “czymś” (‘something_{instr.}’)) are consistently incorrectly analysed for dependencies, which sometimes leads to triggering of incorrect rules.

Count			Context				Example	
train	valid	test	SPGENDER	ILGENDER	ILNUMBER	FORMALITY	English	Polish
419.9K	0.8K	0.8K	<i>sp:feminine</i>	*	*	*	I'm an amateur.	Jestem amator ką .
743.6K	0.8K	0.8K	<i>sp:masculine</i>	*	*	*	I'm all alone.	Jestem całkiem sam .
9.3K	0.2K	0.2K	*	<i>il:feminine</i>	<i>plural</i>	<i>informal</i>	You're smitten.	Jeste ście odurzone.
73.8K	0.2K	0.2K	*	<i>il:masculine</i>	<i>plural</i>	<i>informal</i>	Have you met Pete?	Poznali ście Pete'a?
315.9K	0.2K	0.2K	*	×	<i>plural</i>	<i>informal</i>	You need to leave.	Musi cie wyjść.
326.8K	0.2K	0.2K	*	×	<i>singular</i>	<i>informal</i>	I got you something.	Przyniosłem ci coś.
273.0K	0.2K	0.2K	*	<i>il:feminine</i>	<i>singular</i>	<i>informal</i>	Are you sick?	Jeste s chora?
498.7K	0.2K	0.2K	*	<i>il:masculine</i>	<i>singular</i>	<i>informal</i>	Understand?	Zrozumi a łeś?
0.7K	0.1K	0.1K	*	<i>il:feminine</i>	<i>plural</i>	<i>formal</i>	Please, let me explain.	Wyja śn ię paniom .
2.7K	0.2K	0.2K	*	<i>il:masculine</i>	<i>plural</i>	<i>formal</i>	Aren't you?	Panowie nie są?
5.7K	0.2K	0.2K	*	<i>il:mixed</i>	<i>plural</i>	<i>formal</i>	You are wrong.	Mylą się państwo .
63.0K	0.2K	0.2K	*	<i>il:feminine</i>	<i>singular</i>	<i>formal</i>	Martini for you?	Dla pani martini?
144.0K	0.2K	0.2K	*	<i>il:masculine</i>	<i>singular</i>	<i>formal</i>	Let me have your coat.	Wezmę pański płaszcz.
33.5K	0.2K	0.2K	*	×	×	<i>formal</i>	Go ahead.	Proszę kontynuować.

Table 2.4: Training data quantities for all combinations of contexts with examples for each combination, with relevant grammatical expressions highlighted. Since `SPEAKER` and `INTERLOCUTOR` contexts are always independent, the counts include cases where they co-occur. * = this attribute *may* occur in this place; × = this attribute is never expressed within this category.

DATA SELECTION AND ANNOTATION Table 2.4 shows particular groups of contexts, their typical expression, and total count in the corpus³. Similarly to [Sennrich et al. \(2016a\)](#), we mask the annotations of half the training samples every epoch at random and give half of the unannotated sentence pairs a random set of attributes. This helps preserve the translation quality of the model's outputs when insufficient context is given.

We gather a total of 4K unique examples for the validation and testing set, with samples equally distributed across the 14 different context groups (cf. Table 2.4). When evaluating each implemented approach, we provide two results: when *complete context* is given, or when an *isolated attribute* type is provided. Consider a complete-context test case within the `ILNUMBER` group of

<*il:feminine*>,<*plural*>,<*formal*> I like you.

The input for the isolated attribute is as follows:

<*plural*> I like you.

³ Note that `ILGENDER`, `ILNUMBER`, `FORMALITY` are co-dependent, since they all concern the same entity (the interlocutor), and thus different combinations of their types lead to different grammatical expressions.

that is, we omit all types but those belonging to the examined attribute. For the *complete context* case we provide the full input. To evaluate each individual type (e.g. $\langle il:feminine \rangle$ or $\langle formal \rangle$), in the isolated attribute case we gather all validation/testing cases which match the selected type, with a total count of minimum 200 examples (for $\langle il:mixed \rangle$) up to 1,200 (for $\langle plural \rangle$).

Approach	Multi-attribute solution	Embedding size	Input space occupied
<i>Types as Tags</i>			
TAGENC [▲] (Sennrich et al. 2016a)			n_{types}
TAGDEC (Takeno et al. 2017)	++	$n_{types} * d_{model}$	$n_{types} + 1$
TAGENCDEC [▲] (Lakew et al. 2021)			$2 * n_{types} + 1$
<i>Embedded Average</i>			
EMBPWSUM (Lakew et al. 2021)			0
EMBADD (Schioppa et al. 2021)			0
EMBENC (Ours)	$\frac{\sum types}{n_{types}}$	$n_{types} * d_{model}$	1
EMBSOS (Lample et al. 2019)			0
EMBENCOS (Ours)			1
OUTBIAS [▲] (Michel & Neubig 2018a)	$\frac{\sum types}{n_{types}}$	$n_{types} * len_{vocab}$	0

Table 2.5: Comparison of examined approaches. ++ = concatenation. ▲ = Approach originally proposed for single-attribute control and extended by us.

2.3.3.2 Model Settings

In this section, we describe the model architecture and the modifications we apply to adapt it to our problem. We use the Transformer architecture (§A.1.2.1), implemented using PyTorch (Paszke et al. 2019). We draw inspiration from Lakew et al. (2021), where various alterations to the model were tested. Within our setup, these fall into two main categories: *Types as Tags* (TAG*) and *Average Embedding* (EMB*). We extend each approach that was originally proposed as a way of controlling a single attribute to a multi-attribute scenario: for TAG*, we supply multiple tags in a random order, and for EMB* we take the average of embeddings (see Table 2.5 for an overview).

TYPES AS TAGS For the TAG* approach, we associate each attribute type with a special vocabulary token t (e.g. $\langle singular \rangle$, cf. Table 2.2). This token is assigned a unique, trainable embedding ($E(t)$). During fine-tuning, we concatenate a sequence $T = (t_0, \dots, t_k)$ of these *tags* to the source or target sentences⁴. This sequence is treated

⁴ During inference, we supply tags by forcibly decoding the relevant type tokens, followed by a $\langle null \rangle$ token, before the main decoding step commences.

the same as other tokens and is integrated into the training process. We use three settings:

1. TAGENC: appending the tags to the source sentence (Sennrich et al. 2016a).
2. TAGDEC: prepending the tag to the target sentence (Takeno et al. 2017).
3. TAGENCDEC: applying tags to both sentences (Niu & Carpuat 2020).

AVERAGE EMBEDDING As an alternative to supplying the tags as a sequence T , one can average all embeddings in T and use it as a single embedding $\overline{E(T)}$ (Lample et al. 2019). Since the averaging operation is differentiable, the changes in the average can be attributed back to the changes in the individual embeddings. There are a few key differences between this approach and the one above: first, averaging circumvents the problem of tag ordering; second, it assumes that all tag-embedded information contributes equally. We explore five settings for this approach:

1. EMBPWSUM: adding $\overline{E(T)}$ position-wise to each input token (Lakew et al. 2021).
2. EMBADD: adding $\overline{E(T)}$ position-wise to encoder outputs (Schioppa et al. 2021).
3. EMBENC: concatenating $\overline{E(T)}$ to the input (cf. Dai, Liang, Qiu & Huang (2019), but in our approach the embedding is not trained adversarially).
4. EMBEOS: replace the start-of-sequence ($\langle \text{sos} \rangle$) token in the decoder input with $\overline{E(T)}$ (Lample et al. 2019).
5. EMBENCOS: as an additional setting, we test combining EMBENC and EMBEOS.

As a special case, we test OUTBIAS: adding a type embedding as a bias on the final layer of the decoder (Michel & Neubig 2018a). We omit the *black-box injection* method of Moryossef et al. (2019b) as it is not applicable to ILGENDER in plural and to FORMALITY. Our baseline is the pre-trained model without the attribute information.

2.3.3.3 Training Details

We preprocess the corpus with MOSES tools for detokenisation and normalising punctuation⁵, and by applying a set of rules which includes correcting frequent OCR errors and removing start-of-sequence hyphens. We train a joint sub-word segmentation model of 16K tokens using the Byte-Pair Encoding (BPE) algorithm (§ A.2.1) implemented in SentencePiece (Kudo & Richardson 2018) and encode both sides of the corpus. We follow the standard training regimen for a 6-layer Transformer (Vaswani et al. 2017) with an input length limit of 100 tokens; this model has just over 52.3M trainable parameters. All training is done on a single 32GB GPU. As the

⁵ <https://github.com/alvations/sacremoses>

decoding algorithm, we use beam search (§A.2.5) with a beam size of 5. We pre-train the model until a patience criterion of the chrF++ (§A.2.3.2) validation score not increasing for 5 consecutive validation steps, taken roughly four times within every three epochs (§A.1.1). This happens around the 24th epoch, or after 66 hours of training.

Each of the nine architectural upgrades is a copy of the pre-trained model expanded with the relevant component and fine-tuned. The fine-tuning process exposes the model to the fine-tuning corpus in 10 epochs; performance is validated every half epoch. We select the best checkpoint based on the highest chrF++ score on the validation set.

2.3.3.4 Evaluation

We consider the following criteria in evaluation:

1. **Translation Quality.** Attribute-controlled translations should be of quality no worse than translations of the non-specialised model.
2. **Grammatical Agreement.** Attribute-controlled hypotheses should completely agree to the specified type where necessary.
3. **Restricted Impact.** Grammatical agreement should only affect words that explicitly render the attributes. Therefore, if no attribute is to be expressed in the hypotheses, then they should be no different from baseline hypotheses.

We evaluate translation quality with chrF++ (§A.2.3.2)⁶ and BLEU (§A.2.3.2). Grammatical agreement is quantified with the help of the *Detector*. For every attribute, we calculate how many hypotheses agree to the correct type t and to the incorrect type \hat{t} . Let hyp_t be a hypothesis translated using type t as context, and $agree(hyp, t)$ denote that the *Detector* has found evidence of type t expressed in hyp . We express the total agreement score as:

$$Agree = \frac{agree(hyp_t, t)}{agree(hyp_t, t) + agree(hyp_t, \hat{t})}$$

Finally, we quantify restricted impact with a custom metric, which measures that attribute-independent sentences do not carry any attribute-reliant artifacts; we define this metric, AMBID, as:

$$\text{chrF++}(\text{NMT}(src_a, A), \text{NMT}(src_a, \hat{A}))$$

⁶ For clarity, we normalise chrF++ scores to a [0, 100] range.

where A is a set of attribute types and \widehat{A} is the reverse set⁷. We use an attribute-ambivalent testing set of a 1K sentences to calculate this score (Table 2.3, column “amb_test”).

2.3.4 Results

Model	<i>isolated attribute</i>			<i>complete context</i>			
	chrF++ [↑]	BLEU [↑]	Agree [↑] (%)	chrF++ [↑]	BLEU [↑]	Agree [↑] (%)	AMBID [↑]
Baseline	46.60	23.13	74.35	46.60	23.13	74.35	–
TAGENC	48.95	25.52	99.03	52.41	29.16	99.39	95.87
TAGDEC	48.65	<u>25.40</u>	99.21	50.83	27.65	96.84	93.15
TAGENCDEC	48.28	25.26	99.35	51.01	28.15	<u>99.26</u>	82.66
EMBPWSUM	46.03	22.37	100	51.90	28.69	97.90	88.67
EMBADD	47.45	23.61	99.96	51.77	28.56	98.24	87.76
EMBENC	47.72	24.39	83.42	52.23	<u>28.98</u>	<u>99.30</u>	<u>95.58</u>
EMBSOS	48.28	24.90	<u>99.91</u>	<u>52.38</u>	<u>29.09</u>	98.47	92.07
EMBENCOS	48.60	25.08	<u>99.87</u>	51.94	28.77	98.55	92.37
OUTBIAS	48.59	24.98	96.71	49.32	26.11	86.25	94.05

Table 2.6: Translation performance of all models; “*isolated attribute*” means that only one (the investigated) attribute was revealed to the model. Highlighted is the best result in the column; all statistically indistinguishable results (according to a bootstrap resampling method (§A.3.1) with $p \leq 0.05$) are underlined.

We report quantitative results in Table 2.6.

GRAMMATICAL AGREEMENT The *Agree* column in Table 2.6, which shows the agreement scores given by the *Detector*, points to a significant improvement of all tested model variants over the Baseline model in both isolated attribute and complete context scenario, yielding an improvement of between 9.07 and 25.65 percentage points. In the isolated attribute scenario, all methods but OUTBIAS and EMBENC achieve near-perfect (100%) agreement scores. The agreement scores in the *complete context* scenario remain high for other models except TAGDEC, and pick up for EMBENC, suggesting that controlling several attributes generally has no negative impact on individual attributes.

⁷ For the type triplet ILGENDER we assume that $\widehat{il:masculine} = il:feminine$, $\widehat{il:mixed} = il:feminine$, $\widehat{il:feminine} = il:masculine$.

TRANSLATION QUALITY Attribute-controlling models achieve significant gains over baseline for both the isolated attribute and complete context scenarios, and the gains are consistently higher in the latter, suggesting that exposing the models to more context yields better translations. TAGENC achieves the highest improvement over the baseline in terms of chrF++/BLEU for complete context (+5.81 chrF++/+6.03 BLEU). The gains in translation quality are correlated with agreement scores, except for EMBPWSUM, for which the isolated attribute scenario leads to a near-perfect agreement but low quality scores. Further investigation shows that this model learned to overproduce context-sensitive words when given a context of only a subset of types (e.g. translating “you” as “I” to introduce SPGENDER marking), leading to high agreement scores but degradation in quality. This highlights the importance of pairing an accuracy measure with a translation quality metric.

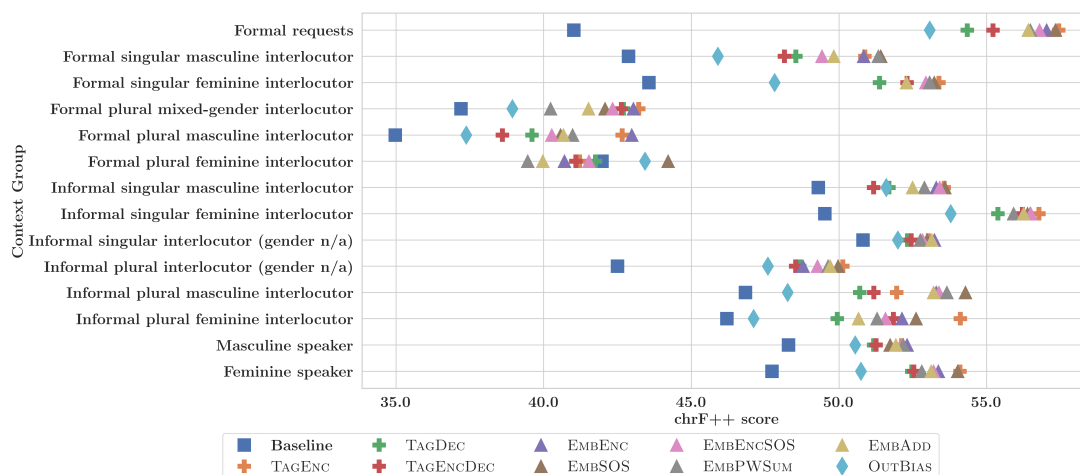


Figure 2.3: Translation quality (chrF++) for each contextual group.

To investigate how successful the models are at modelling each context group individually, we report the mean chrF++ scores obtained for each group’s testing set (Figure 2.3). All contextual models bring significant improvements over the baseline except in the *Formal plural feminine interlocutor* group, for which there was little training data (cf. Table 2.4); improvements are consistently greater for feminine than masculine groups. No single model performs consistently better than others, but TAGDEC, EMBPWSUM and OUTBIAS fall behind on most groups. Finally, we observe no significant gain from including information in both the encoder and the decoder.

RESTRICTED IMPACT The AMBID scores shown in Table 2.6 reveal that TAGENC and EMBENC introduce the least variation in attribute-ambivalent utterances, suggesting that adding contextual information to the encoder input only helps limit creation of

unwanted artifacts. The distance of only 4.13 chrF++ points to the ideal score of 100 for the highest-scoring model suggests good separation of grammatical and behavioural agreement. Some separation-specific modelling may further improve this score, but it was outside the scope of this work.

GENERAL DISCUSSION The results suggest that TAGENC is the most reliable approach to the presented problem, followed by EMBSOS and EMBENC. Notably, we find other methods dubbed as superior to TAGENC in previous work (EMBADD, TAGDEC and TAGENCDEC) to underperform in our case.

2.3.5 *Conclusions*

In this work, we have highlighted the problem of grammatical agreement in translation of TV dialogue in the English-to-Polish language direction. We have created and described a dataset annotated for four speaker and interlocutor attributes that directly influence grammar in dialogue: speaker’s gender, interlocutor’s gender and number and formality relations between them. We have presented a selection of models capable of controlling these attributes in translation, yielding a performance gain of up to +5.81chrF++/+6.03BLEU over the baseline (non-controlling) model. Finally, we have produced a tool that produces an accuracy score for agreement to each type.

Considering all criteria of evaluation, we have identified TAGENC as the best performing approach, with EMBENC, and EMBSOS also achieving competitive performance. TAGENC may be more attractive in scenarios where interventions in the model architecture are impossible as it can be implemented via data preprocessing alone, but the other two have a more scalable design (cf. § 2.2). Finally, contrary to some previous work, we did not find that including the contextual information in the decoder as well as the encoder yielded significant improvements in translation quality or accuracy over including the information in the encoder alone.

2.4 CONTROLLING FORMALITY IN LOW-RESOURCE NMT WITH DOMAIN ADAPTATION AND RERANKING

2.4.1 Introduction

Formality-controlled machine translation enables the user of the translation system to specify the desired formality level of the produced hypothesis at input. Due to discrepancies between different languages in formality expression, it is often the case that the same source sentence has several plausible hypotheses, each aimed at a different audience; leaving this choice to the model may result in an inappropriate translation.

This work describes our team’s submission to the first Special Task on Formality Control in SLT at The International Conference on Spoken Language Translation (IWSLT) 2022 (Anastasopoulos et al. 2022), where the objective was to enable the translation pipeline to generate formal or informal translations depending on user’s input. We participated in the task in four language directions: English-to-German (EN-DE), English-to-Spanish (EN-ES), English-to-Russian (EN-RU) and English-to-Italian (EN-IT). Among these, EN-RU & EN-IT were considered zero-shot (§ A.1.3); for the remaining pairs, small paired formality-annotated corpora were provided.

Our proposed method is language-agnostic and leverages the formality-annotated triplets $(x, y_{\text{formal}}, y_{\text{informal}})$ provided by the task organisers (hereinafter the **IWSLT 2022 corpus**) to pseudo-label a subset of **formality-agnostic** translation corpora (i.e. paired translation datasets containing no explicit formality information). Our translation systems are fine-tuned on this pseudo-labelled data. To further boost the formality control, we implement a *formality-focused hypothesis re-ranking* step. Our zero-shot system uses the re-ranking step alone, i.e. it is not directly fine-tuned to control formality.

More concretely, we extend the provided formality-supervised data by extracting similar samples from the larger unannotated datasets via a language-independent approach of domain adaptation (treating the formality data as “in-domain” sets and the large corpus as an “out-of-domain” set). Our supervised system is fine-tuned on this data, using a *tag* appended to the input of the model. We also re-rank the top n model hypotheses with a formality-focused objective function which uses a relative frequency model built from the provided IWSLT 2022 corpus. To use the same objective in our zero-shot system, we extract samples of particular formality levels for the zero-shot pairs (EN-RU, EN-IT) based on data collected for EN-DE and EN-ES.

Throughout the paper, we use \mathbb{F} to denote the *formal* style and \mathbb{I} to denote the *informal* style. The official evaluation results reveal that, for the supervised pairs, our approach improves formality control by 49.5% accuracy points over the baseline, and for the zero-shot pairs we improve by 33.8%. More specifically, on the test sets for English-to-German and English-to-Spanish, we achieve an average accuracy of 99.5%, and in a zero-shot setting for English-to-Russian and English-to-Italian we obtain 65.9%. Overall,

our submission achieved the best performance in 15 out of 16 directions by automatic evaluation, and according to the human evaluation of the zero-shot English-to-Russian task, our system was able to achieve 85.0%/71.3% control for formal/informal register, the only high-performing zero-shot system in the task. Our work highlights the potential of both data adaptation and re-ranking approaches in attribute control for NMT.

This work has been published in the Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022).⁸

2.4.2 Shared Task Details

The objective of the Special Task on Formality Control in SLT is to train a machine translation system which can control the formality register of the output given some input information. For the supervised language pairs, the participants were provided with short sets of data which could be used for training or validation (400 examples given as triplets of source sentence, informal hypothesis and formal hypothesis). This data (hereinafter the IWSLT 2022 corpus, Nadejde et al. 2022a) comes from two domains, telephone conversations and topical chat (Gopalakrishnan et al. 2019). An associated matched accuracy scoring script was provided by the organisers, and we include its pseudo-code in Algorithm 1. Matched accuracy was the primary metric used to evaluate systems, though translation quality was also measured (via BLEU (§A.2.3.2) and COMET (§A.2.3.2)) to ensure that the systems do not sacrifice quality for formality control. The formality control test set contained 600 paired examples per language pair. The “tst-common” set of the MuST-C corpus (Di Gangi et al. 2019) was used for translation quality testing.

⁸ Vincent, S., Barrault, L. & Scarton, C. (2022a), Controlling formality in low-resource NMT with domain adaptation and re-ranking: SLT-CDT-UoS at IWSLT2022, in ‘Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)’, Association for Computational Linguistics, Dublin, Ireland (in-person and online), pp. 341–350. URL: <https://aclanthology.org/2022.iwslt-1.31>

Algorithm 1: Algorithm for computing the formal & informal matched accuracy.

Input: System *hypotheses*, annotated (\mathbb{F} , \mathbb{I}) *references*

Output: $M\text{-ACC}_{\mathbb{F}}$, $M\text{-ACC}_{\mathbb{I}}$ \triangleright Formal & informal matched accuracy.

```

for hyp  $\in$  hypotheses, (ref $\mathbb{F}$ , ref $\mathbb{I}$ )  $\in$  references do
  for marked_phrase in ref $\mathbb{F}$  do
    if marked_phrase in hyp then
      | match $\mathbb{F}$  + = 1
    end
  end
  for marked_phrase in ref $\mathbb{I}$  do
    if marked_phrase in hyp then
      | match $\mathbb{I}$  + = 1
    end
  end
  if match $\mathbb{F}$  > 0 and match $\mathbb{I}$  = 0 then
    | total $\mathbb{F}$  + = 1
  else if match $\mathbb{I}$  > 0 and match $\mathbb{F}$  = 0 then
    | total $\mathbb{I}$  + = 1
  return total $\mathbb{F}$   $\div$  (total $\mathbb{F}$  + total $\mathbb{I}$ ), total $\mathbb{I}$   $\div$  (total $\mathbb{F}$  + total $\mathbb{I}$ )
end

```

2.4.3 Proposed Approach

At its heart, our method uses an off-the-shelf Transformer model. In the supervised system, formality is controlled by employing a *tagging* approach (Sennrich et al. 2016a), whereby a formality-indicating tag is appended to the source input. This method has been widely used in research in various controlling tasks (e.g. Johnson et al. 2017, Vanmassenhove et al. 2018, Lakew et al. 2019b). The novelty of our approach lies in how the formality-annotated data was collected, which we describe in this section, as well as in our re-ranking step which enables zero-shot control of formality. Throughout this section, we occasionally refer to the MuST-C corpus (Di Gangi et al. 2019); it is a machine translation corpus of transcribed TED talks, translated from English to other languages (e.g. German and Spanish). This is one of the corpora used in our later

experiments (which we describe in §2.4.4) and we used it to develop and test our approach.

AUTOMATIC EXTRACTION OF FORMAL AND INFORMAL DATA As *tagging* requires supervised data to be effective, we seek to enhance the formality-annotated training corpus by annotating samples from formality-agnostic translation corpora. We make the assumption that similar sentences would correspond to a similar formality level, and use a data selection technique to extract from the formality-agnostic samples most similar to the formal and informal sides of the IWSLT 2022 corpus respectively.

Specifically, let $G = (G_x, G_y)$ be the formality-agnostic corpus, and let $S_{\mathbb{F}} = (S_x, S_{y,\mathbb{F}})$ and $S_{\mathbb{I}} = (S_x, S_{y,\mathbb{I}})$ be the formality-annotated corpora (IWSLT 2022). For simplicity, let us focus on adaptation to $S_{\mathbb{F}}$.

Focusing on the sentences on the target side (which explicitly express formality styles), we build a vocabulary of non-singleton tokens from $S_{y,\mathbb{F}}$, then train two language models (§A.2.4): LM_S from $S_{y,\mathbb{F}}$ and LM_G from a random sample of 10K sentences from G_y ; both LMs use the originally extracted vocabulary. Then, we calculate the **sentence-level** perplexity (§A.2.4.1): $\text{PPL}(LM_G, G_y)$ and $\text{PPL}(LM_S, G_y)$. Finally, the sentence pairs within G are ranked by

$$\text{PPL}(LM_S, G_y) - \text{PPL}(LM_G, G_y).$$

The resulting corpora $G_{\text{sorted_by_}\mathbb{F}}$ and $G_{\text{sorted_by_}\mathbb{I}}$ are sorted by the perplexity differences. The intuition behind this approach is that sentences which use a certain formality will naturally rank higher on the ranked list for that formality, due to similarities in the used vocabulary.

Let \mathbb{F}_{pos} and \mathbb{I}_{pos} be the position of a sentence pair in the formal/informal ranking, respectively. We implement a function $Assign_{\alpha}$ which, for an $\alpha \in [0, \mathcal{C})$, assigns a label to the sentence pair (x, y) , using the following rules:

$$Assign_{\alpha} \begin{cases} \mathbb{F}, & \text{if } \mathbb{F}_{pos} - \mathbb{I}_{pos} > \alpha; \\ \mathbb{I}, & \text{if } \mathbb{I}_{pos} - \mathbb{F}_{pos} > \alpha; \\ \text{None}, & \text{otherwise.} \end{cases}$$

where \mathcal{C} is the size of the out-of-domain corpus. We condition assignment on both positional lists since common phrases such as *(Yes! – Ja!)* may rank high on both sides, but should not get included in either corpus.

In other words, we classify sentences as formal or informal based on the *relative position difference* on the formality-ordered lists. We determine α empirically: we test values from range $0.05\mathcal{C}$ and $0.2\mathcal{C}$ by computing a language model from the resulting data and calculating the average perplexity $\text{PPL}(LM_{\text{Corpus}(\alpha)}, \text{IWSLT})$. We select the α

value which minimises this perplexity. We refer to this approach as RD-LABELLING (relative difference labelling).

RELATIVE FREQUENCY MODEL FOR RE-RANKING Sometimes, even when a model gets the formality wrong in its best hypothesis, the correct answer is sometimes found within the n best hypotheses but ranked lower. To address this, we propose a re-ranking approach that uses a formality-specific criterion (distinct from the log probability criterion used in decoding). This method effectively prioritises the hypotheses with the correct formality, moving them to the top of the list.

We conducted an oracle experiment using the provided scoring script to determine the maximum potential improvement achievable by perfectly re-scoring the n -best list. We generated k -best hypotheses for various values of $k \in \{1, \dots, 100\}$ ⁹. From each list of k hypotheses, we selected the first hypothesis (if any) that matched the correct formality according to the M-ACC metric. The results (Table 2.7) demonstrate that as we expand the list of hypotheses, the number of translations with the correct formality increases, reaching an average accuracy of 95.9% (+10.6% compared to the model) for $k = 100$. The column “# Cases” indicates that, on average, in up to 21 cases, a hypothesis of the correct formality could be found with re-ranking. Importantly, regardless of the value of k , selecting the Oracle hypothesis (i.e. the first one on the list with the correct formality) does not compromise translation quality compared to the base Model (column “BLEU”).

To re-rank the hypotheses, we build a simple relative frequency model from the IWSLT 2022 data. For each term $t_i \in \mathcal{T}$ we calculate its occurrence counts \mathbb{F}_{count} in the *formal* set and \mathbb{I}_{count} in the *informal* set. Let $count(t_i) = \mathbb{F}_{count}(t_i) + \mathbb{I}_{count}(t_i)$. Since we wish to focus on terms differentiating the two sets, we calculate the count difference ratio and use it as the weight β :

$$\beta(t_i) = \frac{|\mathbb{F}_{count}(t_i) - \mathbb{I}_{count}(t_i)|}{\max_{t_k \in \mathcal{T}} |\mathbb{F}_{count}(t_k) - \mathbb{I}_{count}(t_k)|}$$

We additionally nullify probabilities for terms for which the difference of the number of occurrences in the formal and informal sets is lower than the third of total occurrences, a value tuned on the validation set:

$$\kappa(t_i) = \begin{cases} 0, & \text{if } \frac{|\mathbb{F}_{count}(t_i) - \mathbb{I}_{count}(t_i)|}{\mathbb{F}_{count}(t_i) + \mathbb{I}_{count}(t_i)} < 0.33; \\ 1, & \text{otherwise} \end{cases}$$

⁹ We capped the search at $k = 100$ due to long inference times for higher k values.

k	Accuracy		δ_{to_best}	# Cases	BLEU	
	Model	Oracle			Model	Oracle
1	83.8%	83.8%	0.00	0.00	25.28	25.28
5	85.8%	89.2%	1.79	7.00	24.80	24.80
10	85.7%	91.3%	2.66	11.50	25.10	25.53
20	85.3%	92.1%	3.46	13.75	24.74	25.15
30	85.1%	93.0%	5.75	16.00	24.68	25.06
40	85.3%	93.6%	7.84	16.75	24.88	25.24
50	85.3%	94.4%	9.64	18.25	24.84	25.20
60	85.2%	95.0%	11.78	19.75	24.71	25.04
70	85.2%	95.0%	12.08	19.75	24.71	25.04
80	85.2%	95.2%	12.78	20.25	24.72	25.04
90	85.2%	95.4%	13.58	20.50	24.72	25.04
100	85.3%	95.9%	14.66	21.25	24.72	25.04

Table 2.7: Results of the oracle experiment. Model was trained in the TINY setting and using the the RD-LABELLING method. Provided values are averaged across the development set. δ_{to_best} describes the average distance to the first hypothesis of correct formality for cases where the most probable hypothesis is incorrect. The column “# Cases” quantifies that phenomenon.

The probabilities are now calculated as

$$p(\mathbb{F}|t_i) = \frac{\mathbb{F}_{count}(t_i)}{count(t_i)} * \beta(t_i) * \kappa(t_i)$$

$$p(\mathbb{I}|t_i) = \frac{\mathbb{I}_{count}(t_i)}{count(t_i)} * \beta(t_i) * \kappa(t_i)$$

For a hypothesis Y , a source sentence S and contexts $c, \hat{c} \in \{\mathbb{F}, \mathbb{I}\}, c \neq \hat{c}$, our objective function in translation thus becomes

$$p(Y|X, c) = p(Y|X) + p(c|Y) - p(\hat{c}|Y)$$

where

$$p(c|Y) = \sum_i p(c|y_i)$$

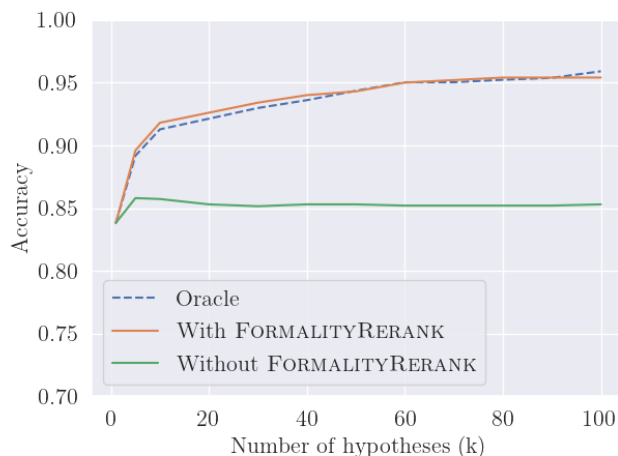


Figure 2.4: Validation accuracy plot showing the effect of applying FORMALITYRERANK to a list of k model hypotheses.

Figure 2.4 shows how validation accuracy increases when this method is used, and that the model is now able to match the oracle accuracy for nearly every k . For $k = 100$ the average improvement in accuracy is 10.2%. The effect of model’s accuracy sometimes surpassing the oracle accuracy (e.g. for $k = 30$) is a by-product of slight sample size variations: the M-ACC metric depends on phrase matches, and a sample is only counted for evaluation if a hypothesis has at least one phrase match against the formality-annotated reference.

GENERALISATION OF RE-RANKING TO ZERO-SHOT LANGUAGE PAIRS Relative frequency re-ranking requires formality-annotated data (i.e. $S_{\mathbb{F}}, S_{\mathbb{I}}$) to be available, which is not the case in our experiments. To enable this re-ranking step, we decide to obtain such a corpus, given supervised training data in other language pairs (in our case, the EN-DE and EN-ES data).

Since the IWSLT 2022 corpus is paired, the formality-agnostic text on the English side does not contain any formality bias that can be leveraged. This points us towards the formality-agnostic corpus G which we can now annotate using RD-LABELLING. However, when applying RD-LABELLING to the MuST-C corpus, we observed that within the set of samples extracted from MuST-C the same source sentences have entirely **different formality expressions** in the German and Spanish corpora, respectively. We confirmed this suspicion by consulting the source sentences and reference translations with native speakers of the respective languages.

Let EN-DE-ES be a corpus of triplets of sentences $(x_{\text{EN}}, y_{\text{DE}}, y_{\text{ES}})$ obtained by identifying English sentences which occur in both the EN-DE and EN-ES parts of MuST-C; due to

the nature of this corpus, EN-DE-ES contains 85.72% of sentence pairs from the EN-DE and 74.13% of pairs from the EN-ES corpus. After marking the target sides of the EN-DE-ES corpus for formality with RD-LABELLING, we quantify in how many cases both languages get the same label (F or I), and in how many cases they get a different label (Table 2.8). Out of all annotated triplets, only 5.8% triplets were annotated in both target languages, significantly less often than expected. Within that group, almost 60% triplets had matching annotations. This implies that - at least in this particular corpus - the same English sentence can sometimes be expressed with different formality in the target language in the same discourse situation. Again, this observation aligns with the intuition of native speakers of the respective languages.

EN-DE	EN-ES	Count	% of annotated
F	F	845	2.85%
I	I	233	0.78%
F	I	381	0.95%
I	F	362	1.22%
F	∅	10851	36.54%
I	∅	7805	26.29%
∅	F	6567	22.12%
∅	I	2749	9.26%

Table 2.8: Combinations of formality annotations for the EN-DE-ES triplet extracted from the MuST-C dataset. “∅” denotes “no annotation”.

Given the non-zero count of triplets with matching formalities, we make another assumption: namely that the English sentences of the triplets with matching formalities may be of “strictly formal” or “strictly informal” nature, meaning the translations of at least some of those sentences to Russian and Italian may express the same formality. To extract F and I sentences for the zero-shot pairs, we adapt the original method, but this time **using English as a pivot** to convey the formality information. We use the English sentences whose German and Spanish translations were both labelled as F or both as I, respectively (columns 1,2 in Table 2.8) and rank the EN-RU and EN-IT corpora by their source sentences’ similarity to that intersection (using the perplexity difference as before).

To infer the final corpora with the RD-LABELLING method, we use the α which yields corpora of similar quantity to the ones for EN-DE & EN-ES, since we could not determine that value empirically.

2.4.4 Experimental Setup

DATA COLLECTION AND PREPROCESSING We collect all datasets permitted by the organisers for our selected language pairs, including:

- **MuST-C (v1.2)** (Di Gangi et al. 2019),
- **Paracrawl (v9)** (Bañón et al. 2020),
- **WMT Corpora** (from the News Translation task) (Barrault et al. 2021):
 - **NewsCommentary (v16)** (Tiedemann 2012),
 - **CommonCrawl** (Smith et al. 2013),
 - **WikiMatrix** (Schwenk et al. 2021),
 - **WikiTitles (v3)** (Barrault et al. 2020),
 - **Europarl (v7, v10)** (Koehn 2005),
 - **UN (v1)** (Ziemski et al. 2016),
 - **Tilde Rapid** (Rozis & Skadiņš 2017),
 - **Yandex**¹⁰.

We list data quantities as well as availability for all language pairs in Table 2.9. We preprocess the WMT and Paracrawl corpora by running a rule-based heuristic of removing sentence pairs with sentences longer than 250 tokens, and with a source-target ratio greater than 1.5, removing non-ASCII characters on the English side and pruning some problematic sentences (e.g. links). We normalise punctuation using the script from Moses (Koehn et al. 2007). After the initial preprocessing, we run the *BiCleaner* tool (Ramírez-Sánchez et al. 2020) on each corpus; the algorithm applies a range of standard preprocessing measures (e.g. removing cases where source and target sentences are identical) and then assigns a confidence score $\in [0, 1]$ to each pair, measuring whether the sentences are good translations of each other, effectively removing potentially noisy sentences. We remove all sentence pairs from the corpora which scored below 0.7 confidence. The final training data quantities are reported in Table 2.9. To train the model on this data, we apply the BPE algorithm (§ A.2.1) implemented in SENTENCEPIECE (Kudo & Richardson 2018) to build a joint vocabulary of 32K tokens across all languages.

Before applying the formality annotation methods we observe that many sentence pairs in our formality-agnostic corpus are not dialogue and too far removed from the domains of our test sets. As the first step, we use the original perplexity-based re-ranking algorithm to prune the corpus. We use the MuST-C corpus as in-domain and all of the data as out-of-domain. We truncate the dataset to the top 5M sentences most

¹⁰ <https://translate.yandex.ru/corpus?lang=en>

Corpus	EN-DE		EN-ES		EN-IT		EN-RU	
MuST-C (v1.2)	0.23M		0.27M		0.25M		0.27M	
Paracrawl (v9)	278.31M		269.39M		96.98M		5.38M	
NewsCommentary v16	0.40M		0.38M		0.09M		0.34M	
CommonCrawl	2.40M		1.85M		–		0.88M	
WikiMatrix	5.47M		–		–		3.78M	
WikiTitles (v3)	1.47M		–		–		1.19M	
Europarl (v7 v10)	1.83M		1.97M		1.91M		–	
UN (v1)	–		11.20M		–		–	
Tilde Rapid	1.03M		–		–		–	
Yandex	–		–		–		1M	
Total								
Raw	291.14M		285.06M		99.23M		12.84M	
Preprocessed	76.99M		91.29M		36.99M		3.86M	
Formality-annotated	F	I	F	I	F	I	F	I
	216.5K	187.2K	111.8K	129.7K	101.0K	172.0K	195.9K	218.4K

Table 2.9: Corpora containing training data used in the experiments. Values indicate number of sentence pairs after preprocessing.

similar the MuST-C data. We then apply RD-LABELLING with α threshold adapted to the data volume. The resulting data quantities can be found in the last row of Table 2.9.

PRE-TRAINING AND FINE-TUNING We train a multilingual 6-layer Transformer model architecture provided within FAIRSEQ (Ott et al. 2019)¹¹. We tie the encoder and decoder weights, use the ADAM optimiser (Kingma & Ba 2015), use a learning rate of $5e - 4$ and a batch size of 2000 tokens. We pre-train for 1.5M iterations (approx. 1.5 epochs) and fine-tune for 0.25M iterations (approx. 47 epochs).

For fine-tuning, we use the MuST-C corpus without formality annotations (to maintain high translation quality), concatenated with the formality-annotated data inferred from the LARGE corpus (to learn formality control). We apply FORMALITYRERANK with $k = 50$. Similarly to pre-training, we average the last 10 checkpoints.

DEVELOPMENT RESULTS The development results (Table 2.10) of our approaches suggest that both RD-LABELLING and FORMALITYRERANK are effective at improving

¹¹ Our choice of the multilingual approach is dictated by the aim to reduce the required computation time rather than the potential benefits of multilingual systems in controlling formality, the exploration of which we leave to future work.

	MuST-C (BLEU)				IWSLT 2022 (M-ACC)				Mean
	EN-DE	EN-ES	EN-RU	EN-IT	EN-DE		EN-ES		
					F	II	F	II	
Pre-trained	28.9	39.5	18.5	29.3	63.4%	36.6%	21.5%	78.5%	50.0%
RD-LABELLING	32.3	40.8	–	–	99.0%	100%	95.2%	99.1%	98.3%
+FORMALITYRERANK	32.3	40.8	20.4	32.0	100%	100%	99.5%	100%	99.9%

Table 2.10: Results on the **development** sets.

formality control without sacrificing translation quality. In particular, RD-LABELLING alone yields a near-perfect for all subcategories except (EN-DE, II); applying FORMALITYRERANK effectively brings the average score up to 99.9%. Note that our pre-trained model for this track achieved lower BLEU scores than for the constrained track, which is explained by the test set coming from the same domain as the constrained training data.

2.4.5 Results

In Table 2.11 and Table 2.12, both taken from the task findings (Anastasopoulos et al. 2022), we report official results of translation quality and formality control, respectively. For our selected language pairs, there was only one other system (UMD) which took part in the task. These results show that by the automatic metrics our methods surpassed those of the competitor for every language pair and register direction. The exception is formal in EN-RU where our system is worse by 0.5%, but our averaged accuracy for this language pair is overall better by 42.6% points.

Overall, our submitted system achieved a near-ideal accuracy of 99.2%. The zero-shot system achieved an impressive average accuracy of 83.8%, an improvement of 33.8% over the baseline and significantly better than the competitors for the EN-RU pair. The human evaluation findings (reported in Table 2.13) confirm the effectiveness of our zero-shot system for EN-RU. In disagreement with the automatic metric, the human evaluation found that our EN-IT system produces mostly formality-neutral hypotheses when asked for formal style, indicating an area for improvement in the future.

Our results, particularly for the zero-shot pairs (Table 2.11, column *Zero-shot*) suggest an interesting phenomenon: for our employed datasets, for any language pair there is a **dominant** formality type, which is the formality type that the baseline model learns to express in translation the majority of the time, as if by default. This dominant type varies across languages, and for the zero-shot pairs in our evaluation it is particularly strong (e.g. 94.5% II vs 5.5% IF in EN-IT). The dominant formalities were controllable to a much higher extent by our submitted model (98.6% and 99.5% respectively) than the

System	Supervised				Zero-shot			
	EN-DE		EN-ES		EN-IT		EN-RU	
	F	I	F	I	F	I	F	I
Baseline	45.8%	54.2%	36.6%	63.4%	5.5%	94.5%	93.4%	6.6%
UMD	99.4%	96.5%	99.5%	93.2%	32.8%	97.9%	100%	0.1%
UoS (Ours)	100%	100%	98.1%	100%	51.2%	98.6%	99.5%	85.8%

Table 2.11: Official results of the automatic evaluation of formality control (matched accuracy) reported for formal (F) and informal (I) register. Baseline scores provided from task organisers. Scores in bold indicate highest in column.

System	EN-DE		EN-ES		EN-IT		EN-RU	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Baseline	32.6	55.0	37.4	70.0	32.2	64.0	19.5	32.0
UMD	22.4	16.1	27.8	34.4	22.9	24.7	14.4	7.5
UoS (Ours)	32.5	49.7	37.0	63.5	33.1	56.2	21.5	35.7

Table 2.12: Official results of the automatic evaluation of translation quality. Baseline scores provided from task organisers. Scores in bold are highest in column.

non-dominant ones (51.2% and 85.8% respectively); this result is also consistent with that of our competitor. This suggests that it is easy for a pre-trained translation model to learn controlled expression of the dominant type within a dichotomous phenomenon, while learning to render the less-expressed type is significantly harder, especially in a low-resource scenario like the present one.

Table 2.12 shows that both ours and our competitor’s submissions sometimes slightly degrade translation quality scores. This could possibly be caused by the models selecting terms in the hypotheses which are of correct formality but less fitting translation candidates in general. Nevertheless, our models exhibit this effect to a significantly lower effect than our competitors: our models degrade COMET by 3.95 on average, whereas UMD’s by as much as 34.58 points. Similarly with BLEU, the competitor’s models degrade it by 8.55 whereas ours actually improve by 0.6. To summarise, our models achieved near full supervised formality control and over 83% zero-shot control while maintaining competitive translation quality scores.

Pair	System	Register	F	I	Neutral	Other	IAA
EN-IT	UMD	F	13.7	25.2	47.0	14.2	0.91
	UMD	I	1.0	78.3	11.5	9.2	
	UoS (Ours)	F	6.0	7.2	81.3	5.5	
	UoS (Ours)	I	0.3	81.0	13.2	5.5	
EN-RU	UMD	F	77.2	0.2	7.0	15.7	0.85
	UMD	I	74.3	0.7	7.8	17.2	
	UoS (Ours)	F	85.0	0.3	6.0	8.7	
	UoS (Ours)	I	10.3	71.3	3.2	15.2	

Table 2.13: Percentage of system outputs (with a given formality level (Register) and track (Track)) labelled by professional translators according to the formality level: F(formal), I(informal), Neutral or Other. IAA was computed using the Krippendorff’s α coefficient. Values in green indicate scores which should be as high as possible (correct register) and values in red indicate scores which should be as low as possible (incorrect register). We highlight best scores between the two competing systems.

2.4.6 Conclusions

Overall results suggest that with sufficient training data formality control can be easily facilitated either via direct supervision or re-ranking, and collection of data necessary to facilitate these methods is possible given small initial samples for all formality types. Our methods applied to the supervised language pairs (EN-DE, EN-ES) worked near-unfailingly. Using English as a pivot language to propagate formality information from one language to another helped achieve impressive results for zero-shot pairs, but the results were not as good as for the supervised pairs.

We suspect that the significant accuracy gains from FORMALITYRERANKING may have been partially due to formality in the studied language pairs itself being expressed primarily via certain token words such as the honorific *Sie* in German creating a *pivot* effect (Fu et al. 2019). As such, it may be of interest for future research to study such methods applied to more complex phenomena, such as grammatical expression of gender.

2.5 MTCUE: LEARNING ZERO-SHOT CONTROL OF EXTRA-TEXTUAL ATTRIBUTES IN NEURAL MACHINE TRANSLATION

2.5.1 Introduction

Although NMT has progressed at a fascinating pace in recent years, contemporary methods focus on translating isolated sentences and overlook the importance of adapting to broader context, such as the description of the discourse situation. Conversely, some researchers have suggested that incorporating fine-grained adaptations based on extra-textual context could benefit conversational machine translation (van der Wees et al. 2016). Most existing work that does consider context in translation has focused on document-level context only, aiming to enhance the coherence and cohesion of the translated document (e.g. Tiedemann & Scherrer 2017). Only a limited amount of research has successfully adapted NMT to extra-textual context variables using supervised learning frameworks on labelled datasets, targeting individual aspects such as gender (Vanmassenhove et al. 2018, Moryossef et al. 2019a, Vincent et al. 2022b), formality (Sennrich et al. 2016a, Nadejde et al. 2022b), translators' or speakers' style (Michel & Neubig 2018b, Wang, Hoang & Federico 2021) and translation length (Lakew et al. 2019a), sometimes controlling multiple attributes simultaneously (Schioppa et al. 2021, Vincent et al. 2022b). However, to our knowledge, no prior work has attempted to model the impact of continuous extra-textual contexts in translation or combined the intra- and extra-textual contexts within a robust framework. This is problematic since translating sentences without or with incomplete context is akin to a human translator working with incomplete information. Similarly, only a handful of earlier studies have contemplated the idea of controlling these extra-textual attributes in a zero-shot or few-shot fashion (Moryossef et al. 2019a, Anastasopoulos et al. 2022); such approaches are essential given the difficulty of obtaining the labels required for training fully supervised models.

In some domains, extra-textual context is paramount and NMT systems oblivious to this information are expected to under-perform. For instance, for the dubbing and subtitling domain, where translated shows can span different decades, genres, countries of origin, etc., a one-size-fits-all model is limited by treating all input sentences alike. In this domain, there is an abundance of various metadata (not just document-level data) that could be used to overcome this limitation. However, such adaptation is not trivial: (i) the metadata often comes in quantities too small for training and with missing labels; (ii) it is expressed in various formats and types, being difficult to use in a standard pipeline; (iii) it is difficult to quantify its exact (positive) effect.

In this paper, we address (i) and (ii) by proposing MTCUE (Machine Translation with Contextual universal embeddings), a novel NMT framework that bridges the gap between training on discrete control variables and intra-textual context as well as allows the user to utilise metadata of various lengths in training, easing the need for

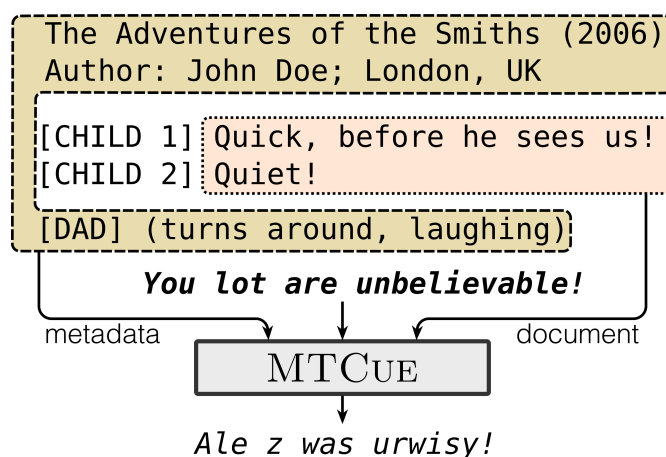


Figure 2.5: A high-level overview of MTCUE (EN-PL).

laborious data editing and manual annotation (Figure 2.5). During inference, when context is provided verbatim, MTCUE falls back to a code-controlled translation model; by vectorising the inputs, it exhibits competitive performance for noisy phrases and learns transferrability across contextual tasks. While (iii) is not directly addressed, our evaluation encompasses two translation quality metrics and two external test sets of attribute control, showing the impact on both translation quality and capturing relevant contextual attributes.

MTCUE can generalise to unseen context variables, achieving 100% accuracy at a zero-shot formality controlling task; it learns to map embeddings of input contexts to discrete phenomena (e.g. formality), increasing explainability; and it exhibits more robust few-shot performance at multi-attribute control tasks than a “tagging” baseline.

The main contributions of this work are:

1. MTCUE (§ 2.5.2): a novel framework for **combining (un)structured intra- and extra-textual context in NMT** that significantly improves translation quality for four language pairs in both directions: English-to-German (EN-DE), German-to-English (DE-EN), English-to-French (EN-FR), French-to-English (FR-EN), English-to-Polish (EN-PL), Polish-to-English (PL-EN), English-to-Russian (EN-RU) and Russian-to-English (RU-EN).
2. A comprehensive evaluation, showing that MTCUE can be primed to exhibit **excellent zero-shot and few-shot performance** at downstream contextual translation tasks (§ 2.5.6 and § 2.5.8).
3. Pre-trained models, code, and an organised version of the OpenSubtitles18 (Lison et al. 2018) dataset **with the annotation of six metadata** are made available.

We also present the experimental settings including the data used (§2.5.3), evaluation methods (§2.5.4) and implementation details (§2.5.5), as well as conclusions (§2.5.9). This work has been published in Findings of the Association for Computational Linguistics (ACL 2023).^{12 13}

2.5.2 Proposed Architecture: MTCUE

MTCUE is an encoder-decoder Transformer (§A.1.2.1) architecture with two encoders: one dedicated for contextual signals and one for inputting the source text. The signals from both encoders are combined using parallel cross-attention in the decoder. Below we describe how context inputs are treated in detail, and later we describe the context encoder and context incorporation.

VECTORISING CONTEXTS Context comes in various formats: for example, the speaker’s gender or the genre of a film are often supplied in corpora as belonging to sets of pre-determined discrete classes, whereas plot descriptions are usually provided as plain text (and could not be treated as discrete without significant loss of information). To leverage discrete variables as well as short and long textual contexts in a unified framework, we define a **vectorisation function** that maps each context to a single meaningful vector, yielding a matrix $E_{c \times r}$, where c is the number of contexts and r is the embedding dimension. The function is deterministic (the same input is always embedded in the same way) and semantically coherent (semantically similar inputs receive similar embeddings). We use a sentence embedding model (§A.2.2; Reimers & Gurevych 2019a) for vectorisation, which produces embeddings both deterministic and semantically coherent. Motivated by Khandelwal et al. (2018) and O’Connor & Andreas (2021) who report that generation models mostly use general topical information from past context, ignoring manipulations such as shuffling or removing non-noun words, we hypothesise that sentence embeddings can effectively compress the relevant context information into a set of vectors, which, when processed together within a framework, will formulate an abstract representation of the dialogue context. We select the MINILMv2 sentence embedding model (Wang, Bao, Huang, Dong & Wei 2021), which we access via the sentence-transformers library¹⁴. In the experiments, we also refer to DISTILBERT (Sanh et al. 2019) which is used by one of our baselines, and a discrete

¹² Vincent, S., Flynn, R., Scarton, C. (2023), MTCue: Learning Zero-Shot Control of Extra-Textual Attributes by Leveraging Unstructured Context in Neural Machine Translation, in ‘Findings of the Association for Computational Linguistics: ACL 2023’, Association for Computational Linguistics, Toronto, Canada, pp. 8210-8226. URL: <https://aclanthology.org/2023.findings-acl.521/>

¹³ The work presented in this section was carried out in collaboration with Robert Flynn who contributed the ideas of using positional embeddings for document-level information and using QK-NORM, participated in project discussions, reviewed drafts of the paper and assisted in responding to reviewers.

¹⁴ <https://sbert.net/>

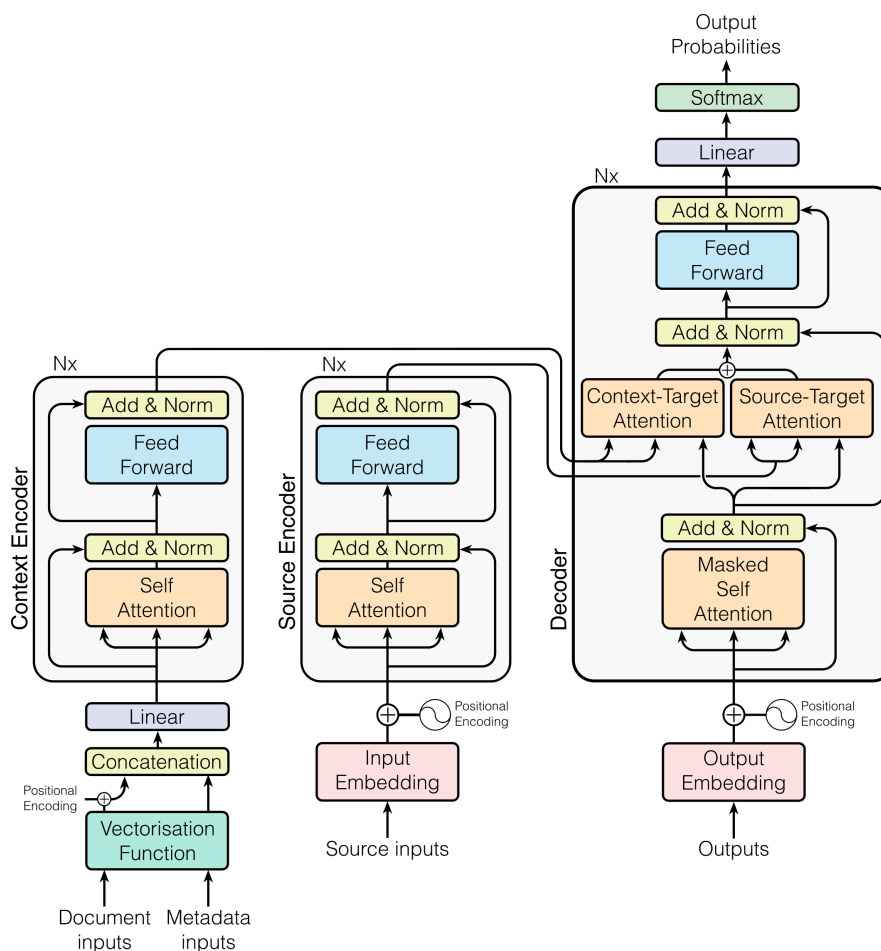


Figure 2.6: The MTCUE architecture. Stylised after the Transformer architecture figure in (Vaswani et al. 2017).

embedding function which maps unique contexts to the same embeddings but has no built-in similarity feature.

For any sample, given a set of its k textual contexts $C = [c_1, \dots, c_k]$, we vectorise each one separately using the method described above. The resulting array of vectors is the input we supply to the context encoder in MTCUE.

CONTEXT ENCODER The context encoder of MTCUE is a standard self-attention encoder with a custom input initialisation. Its inputs are sentence embeddings of context (§2.5.2) projected to the model’s dimensions with a linear layer ($384 \rightarrow d_{model}$). In preliminary experiments, we observe that the first layer of the context encoder receives abnormally large input values, which sometimes leads to the explosion of the

query (**Q**) and key (**K**) dot product \mathbf{QK}^T . We prevent this by replacing the scaled dot product attention with query-key normalisation (§A.1.2.2; Henry et al. 2020)¹⁵.

We apply positional embeddings (§A.1.2.2) to context, in order to (a) indicate the distance of a past utterance to the source sentence and (b) to distinguish metadata inputs from document information. In particular, when translating the source sentence s_i at position i in the document, a sentence distance positional embedding (*POS*) is added to the embedding representations of each past sentence s_{i-j} , with $j \in [0, t]$ where t is the maximum allowed context distance: $e'(s_{i-j}, j) = e(s_{i-j}) + POS(j)$. Metadata contexts (m_0, \dots, m_n) do not receive positional embeddings since their order is irrelevant. The final vectorised input of the context encoder is: $e'(s_i, 0), e'(s_{i-1}, 1), \dots, e'(s_{i-t}, t), e(m_0), \dots, e(m_n)$.

CONTEXT INCORPORATION The outputs of the context and source encoders (respectively \mathcal{C} and \mathcal{S}) are combined in the decoder using **parallel attention** (Libovický et al. 2018). Let the output of the decoder self-attention be \mathcal{T} . Let $\mathcal{T}_{out} = \text{FFN}(\mathcal{T}') + \mathcal{T}'$, where \mathcal{T}' is the multi-head attention output; i.e. \mathcal{T}_{out} is \mathcal{T}' with the feed-forward layer and the residual connection applied. In a non-contextual Transformer, source and target representations are combined with cross-attention:

$$\mathcal{T}' = \text{mAttn}(kv = \mathcal{S}, q = \mathcal{T})$$

In contrast, parallel attention computes individual cross-attention of \mathcal{T} with \mathcal{S} and \mathcal{C} and then adds them together:

$$\mathcal{S}' = \text{mAttn}(kv = \mathcal{S}, q = \mathcal{T})$$

$$\mathcal{C}' = \text{mAttn}(kv = \mathcal{C}, q = \mathcal{T})$$

$$\mathcal{T}' = \mathcal{C}' + \mathcal{S}'$$

Parallel attention is only one of many combination strategies which can be used, and in preliminary experiments we found the choice of the strategy to have a minor impact on performance.

2.5.3 Data: The OpenSubtitles18 Corpus

The publicly available OpenSubtitles18¹⁶ corpus (Lison et al. 2018), hereinafter **OPENSUBTITLES**, is a subtitle dataset in .xml format with IMDb ID attribution and timestamps. It is a mix of original and user-submitted subtitles for movies and TV content. Focusing on four language pairs ($\text{EN} \leftrightarrow \{\text{DE}, \text{FR}, \text{PL}, \text{RU}\}$), we extract parallel

¹⁵ An alternative solution applies layer normalisation to the input of the first layer, but we found that this degraded performance w.r.t. QK-NORM.

¹⁶ Created from data from <https://opensubtitles.org/>.

Data type	EN↔DE	EN↔FR	EN↔PL	EN↔RU
Source & target	5.3M	14.7M	12.9M	12.4M
<i>metadata</i>				
Genre	45.3%	57.8%	60.5%	73.4%
PG rating	35.9%	46.9%	48.8%	62.3%
Writer(s)	45.3%	57.1%	58.9%	71.7%
Year	45.3%	57.8%	60.5%	73.7%
Country	37.7%	42.9%	45.7%	42.7%
Plot description	43.4%	57.1%	59.7%	72.6%
<i>previous dialogue</i>				
$n - 1$	60.1%	68.0%	63.7%	73.6%
$n - 2$	42.0%	51.2%	46.4%	57.9%
$n - 3$	31.2%	40.1%	35.5%	46.9%
$n - 4$	23.9%	32.2%	28.0%	38.6%
$n - 5$	18.7%	26.2%	22.4%	32.2%

Table 2.14: Data quantities for the extracted OpenSubtitles18 corpus. An average of 81% samples has at least one other context than the current sentence.

[sentence-level](#) data with source and target document-level features (up to 5 previous sentences), following [Voita et al. \(2019b\)](#). There are timestamps and overlap values for each source-target sample in the corpus; we only take into account pairs with overlap ≥ 0.9 and we use two criteria to build any continuous document: (1) no omitted pairs (due to poor overlap) and (2) no distance greater than seven seconds between any two consecutive pairs. To generate train/validation/test splits, we use generated lists of held-out IMDB IDs based on various published test sets ([Müller et al. 2018](#), [Lopes et al. 2020](#), [Vincent et al. 2022b](#)) to promote reproducibility. We also extract a range of metadata by matching the IMDb ID against the Open Movie Database (OMDb) API.¹⁷ [Table 2.14](#) shows training data quantities and portions of annotated samples per context while [Table 2.15](#) shows an example of the extracted data. We select six metadata types that we hypothesise to convey useful extra-textual information: *plot description* (which may contain useful topical information), *genre* (which can have an impact on the language used), *year of release* (to account for the temporal dimension of language), *country of release* (to account for regional differences in expression of English), *writers* (to consider writers’ style), *PG rating* (which may be associated with

¹⁷ <https://ombapi.com/>

e.g. the use of adult language). For validation and testing, we randomly sample 10K sentence pairs each from the corpus, based on held-out IMDb IDs.

Key	Value
Source (EN)	This is the Angel of Death, big daddy reaper.
Target (PL)	To anioł śmierci. Kosiarz przez wielkie "k".
PG rating	PG rating: TV-14
Released	Released in 2009
Writers	Writers: Eric Kripke, Ben Edlund, Julie Siege
Plot	Dean and Sam get to know the whereabouts of Lucifer and want to hunt him down. But Lucifer is well prepared and is working his own plans.
Genre	Drama, Fantasy, Horror
Country	United States, Canada

Table 2.15: Example of a source-target pair and metadata in `OPENSUBTITLES`.

The corpus is first detokenised and has punctuation normalised (using Moses scripts, [Koehn et al. 2007](#)). Then a custom cleaning script is applied, which removes trailing dashes, unmatched brackets and quotation marks, and fixes common OCR spelling errors. Finally, we perform sub-word tokenisation ([§A.2.1](#)) via the BPE algorithm with Sentencepiece ([Kudo & Richardson 2018](#)).

Film metadata (which comes from OMDb) is left intact except when the fields contain non-values such as “N/A”, “Not rated”, or if a particular field is not sufficiently descriptive (e.g. a PG rating field represented as a single letter “R”), in which case we enrich it with a disambiguating prefix (e.g. “R” → “PG rating: R”). Regardless of the trained language pair, metadata context is provided in English (which here is either the source or target language). document-level context is limited to source-side context. Since for language pairs into English the context input comes in two languages (e.g. English metadata and French dialogue), we use multilingual models to embed the context in these pairs.

2.5.4 Evaluation

We evaluate the presented approach with the general in-domain test set as well as two external contextual tasks described in this section.

TRANSLATION QUALITY The approaches are evaluated against an in-domain held-out test set of 10K sentence pairs taken from OPENSUBTITLES. As metrics, we use BLEU¹⁸ (§A.2.3.2) and COMET¹⁹ (§A.2.3.2).

CONTROL OF MULTIPLE ATTRIBUTES ABOUT DIALOGUE PARTICIPANTS (EAMT 2022) The EAMT 2022 task, which we introduced in §2.3 (see also Vincent et al. 2022b), evaluates a model’s capability to control information about dialogue participants in English-to-Polish translation. The task requires generating hypotheses that align with four attributes: gender of the speaker and interlocutor(s) (masculine/feminine/mixed), number of interlocutors (one/many), and formality (formal/informal). These attributes can occur in a total of 38 unique combinations. We investigate whether MTCUE can learn this task through zero-shot learning (pre-training on other contexts) or through few-shot learning (when additionally fine-tuned on a constrained number of samples).

To prepare the dataset, we use scripts associated with §2.3 (see also Vincent et al. 2022b) to annotate OPENSUBTITLES with the relevant attributes, resulting in a corpus of 5.1M annotated samples. To leverage the context representation in MTCUE, we transcribe the discrete attributes to natural language by creating three sentences that represent the context. For example, if the annotation indicates that the speaker is male, the interlocutor is a mixed-gender group, and the register is formal, we create the following context: (1) “I am a man”, (2) “I’m talking to a group of people” and (3) “Formal”.

We train seven separate instances of MTCUE using different artificial data settings. Each setting contains the same number of samples (5.1M) but a varying number of **annotated** samples. To address class imbalances in the dataset (e.g. *masculine speaker* occurring more often than *feminine speaker*) and ensure equal representation of the 38 attribute combinations, we collect multiples of these combinations. We select sample numbers to achieve roughly equal logarithmic distances: 1, 5, 30, 300, 3K and 30K supervised samples per each of 38 combinations, yielding exactly 38, 180, 1,127, 10,261, 81,953 and 510,683 samples respectively. Including the zero-shot and full supervision (5.1M cases), this results in a total of eight settings. Each model is trained with the same hyperparameters as MTCUE, and on the same set of 5.1M samples, with only the relevant number of samples annotated (non-annotated samples are given as source-target pairs without contexts). We compare our results against our TAGGING approach which achieved the best performance in §2.3). We train the TAGGING model in replicas of the eight settings above.

ZERO-SHOT CONTROL OF FORMALITY (IWSLT 2022) We experiment with the generalisation of MTCUE to an unseen type of context: formality. In the IWSLT 2022 formality control task (Anastasopoulos et al. 2022), the model’s challenge is to produce

¹⁸ Computed with SacreBLEU (Post 2018).

¹⁹ Computed using the wmt20-comet-da model.

hypotheses agreeing with the desired formality (formal/informal). For the English-to-German language pair, the task provides a set of paired examples (each source sentence is paired with a formal reference and an informal one), to a total of 400 validation and 600 test examples; for the English-to-Russian pair, only the 600 test examples are provided. We test the capacity of MTCUE to control formality zero-shot, given a textual cue as context input.

To evaluate the performance of any tested model, we need a fair method of choosing a context to condition on, since in a zero-shot setting the model organically learns the tested attributes from various contexts rather than specific cherry-picked sentences.

To do so, we sample some metadata from the validation set of the OPENSUBTITLES data and pick eight contexts (four for the *formal* case and four for the *informal* case) which either used formal or informal language themselves or represented a domain where such language would be used. We also add two generic prompts: *Formal conversation* and *Informal chit-chat*. The full list of prompts is as follows:

- Formal:
 1. *Formal conversation*
 2. *Hannah Larsen, meet Sonia Jimenez. One of my favourite nurses.*
 3. *In case anything goes down we need all the manpower alert, not comfortably numb.*
 4. *Biography, Drama*
 5. *A musician travels a great distance to return an instrument to his elderly teacher*
- Informal:
 1. *Informal chit-chat*
 2. *I'm gay for Jamie.*
 3. *What else can a pathetic loser do?*
 4. *Drama, Family, Romance*
 5. *Animation, Adventure, Comedy*

We then run the evaluation as normal with each context separately, and select the highest returned score for each attribute.

2.5.5 Baselines and Implementation

In our experiments, we compare MTCUE with three types of baselines:

1. BASE and BASE-PM. These are pre-trained translation models that match MTCUE either in the shape of the encoder-decoder architecture (BASE) or in terms of the total number of parameters (BASE-PM). For BASE-PM, the extra parameters are obtained from enhancing the source encoder, increasing the number of layers (6 \rightarrow 10) and doubling the feed-forward dimension (2048 \rightarrow 4096).

Model	Params	d_{model}	Layers			h	FFN dim.			GPU Hour/Epoch	Epochs to best
			Cxt	Src	Dec		Cxt	Src	Dec		
BASE	66M	512	–	6	6	8	–	2048	2048	–	–
BASE-PM	107M	512	–	10	6	8	–	4096	2048	–	–
TAGGING	107M	512	–	10	6	8	–	4096	2048	0.74 ± 0.35	6.13 ± 4.09
NOVOTNEY-CUE	99M	512	6	6	6	8	2048	2048	2048	1.29 ± 0.56	9.13 ± 3.60
MTCUE	105M	512	6	6	6	8	2048	2048	2048	0.81 ± 0.39	9.38 ± 4.57

Table 2.16: Model details for MTCUE and baselines. Timings and epochs are averaged across all language directions.

2. TAGGING. Following previous work (e.g. [Schioppa et al. 2021](#), [Vincent et al. 2022b](#)), we implement a model that assigns a discrete embedding to each unique context value. Architecturally, the model matches BASE-PM. The tags are prepended to feature vectors from the source context and then together fed to the decoder.
3. NOVOTNEY-CUE. This baseline is a re-implementation of the CUE vectors architecture ([Novotney et al. 2022](#)) for NMT. It utilises DISTILBERT for vectorisation and averages the context feature vectors to obtain the decoder input. In contrast, MTCUE employs a parallel attention strategy.

In experiments on formality control, we also report results from the two submissions to the IWSLT 2022 task, both implementing a supervised and a zero-shot approach:

1. [Vincent et al. \(2022a\)](#) (see also § 2.4). This (winning) submission combines the TAGGING approach with formality-aware re-ranking and data augmentation. The authors augment the original formality-labelled training samples by matching sentence pairs from larger corpora against samples of specific formality (akin to the Moore-Lewis algorithm described in [Moore & Lewis 2010](#)). Their zero-shot approach relies on heuristically finding a suitable sample of formality-annotated data similar to the provided set and performing the same algorithm above.
2. [Rippeth et al. \(2022\)](#) who fine-tune large pre-trained multilingual MT models with additive control ([Schioppa et al. 2021](#)) on data with synthetic formality labels obtained via rule-based parsers and classifiers.

IMPLEMENTATION We implement MTCUE and all its components in FAIRSEQ, and use HuggingFace ([Wolf et al. 2020](#)) for vectorising contexts. We use hyperparameters recommended by FAIRSEQ, plus optimise the learning rate and the batch size in a grid search. We found that a learning rate of 0.0003 and a batch size of simulated 200K tokens worked best globally. Table 2.16 presents the architecture details and runtimes for the models. All training is done on a single A100 80GB GPU, one run per model. We use early stopping based on validation loss with a patience of 5 (§A.1.1).

2.5.6 Results

Model	EN-DE		EN-FR		EN-PL		EN-RU		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
<i>Baselines</i>										
*BASE	33.60	45.90	34.54	46.92	28.08	58.52	31.37	62.94	31.90	53.57
*BASE-PM	34.36	46.77	35.31	48.87	28.66	60.97	32.40	64.55	32.68	55.29
TAGGING	34.88	49.21	36.74	<u>51.57</u>	29.08	64.29	32.32	<u>65.12</u>	33.26	57.55
NOVOTNEY-CUE	35.30	49.83	36.75	50.52	29.09	62.69	32.36	<u>64.90</u>	33.38	56.99
<i>Proposed</i>										
MTCUE	36.02	50.91	37.54	52.19	29.36	63.46	33.21	65.21	34.03	57.94

Model	DE-EN		FR-EN		PL-EN		RU-EN		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
<i>Baselines</i>										
*BASE	39.53	59.56	35.46	55.10	34.42	50.38	39.37	55.99	37.20	55.26
*BASE-PM	40.32	60.88	36.16	56.28	35.03	51.77	40.04	<u>56.86</u>	37.89	56.45
TAGGING	41.52	62.63	37.10	57.41	36.19	53.46	40.33	57.14	38.79	57.66
NOVOTNEY-CUE	40.86	61.91	36.51	56.21	35.28	52.17	39.44	56.08	38.03	56.59
<i>Proposed</i>										
MTCUE	40.95	61.58	36.57	<u>56.87</u>	35.68	52.48	39.97	<u>56.92</u>	38.29	56.96

Table 2.17: Translation quality results on the OPENSUBTITLES test set. *Model trained without access to any context. We highlight the best result in each column and underline all statistically indistinguishable results, $p \leq 0.05$ (except the Average column).

TRANSLATION QUALITY Results in Table 2.17 show that MTCUE beats all non-contextual baselines in translation quality, achieving an average improvement of +1.51 BLEU/+3.04 COMET over BASE and +0.88/+1.58 over BASE-PM. It is also significantly better than NOVOTNEY-CUE (+0.46/+0.66). MTCUE achieves comparable results to the parameter-matched TAGGING model, consistently outperforming it on all language directions from English, and being outperformed by it on directions into English. Since the primary difference between the two models is that MTCUE sacrifices more parameters to process context, and TAGGING uses these parameters for additional processing of source text, we hypothesise that the difference in scores is due to the extent to which context is a valuable signal for the given language pair: it is less important in translation into English. This is supported by findings from literature:

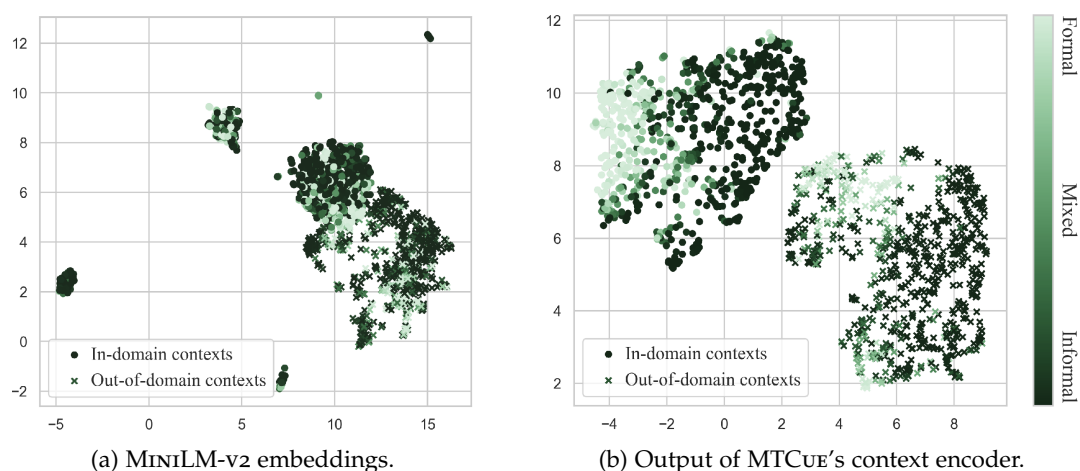


Figure 2.7: UMAP visualisation of how various contexts impact the formality of produced translations when used as input in MTCUE.

English is a language that does not grammatically mark phenomena such as gender (Stahlberg et al. 2007).

The largest quality improvements with MTCUE are obtained on EN-DE (+1.66/+4.14 vs BASE-PM and +1.14/+1.70 vs TAGGING) and EN-FR (+2.23/+3.32 vs BASE-PM and +0.80/+0.62 vs TAGGING) language pairs. Contrastively, the smallest improvements against BASE-PM are obtained on the RU-EN pair. MTCUE is outperformed by TAGGING the most on PL-EN (−0.51/−0.98). As far as training efficiency, MTCUE trains significantly faster than NOVOTNEY-CUE, converging in a similar number of epochs but using significantly less GPU time, on par with TAGGING (Table 2.16). Finally, all contextual models considered in this evaluation significantly outperform the parameter-matched translation model (BASE-PM), clearly signalling that metadata and document context are an important input in machine translation within this domain, regardless of the chosen approach.

CONTROL OF MULTIPLE ATTRIBUTES ABOUT DIALOGUE PARTICIPANTS (EAMT 2022) MTCUE achieves 80.25 zero-shot accuracy at correctly translating the speaker and interlocutor attributes, an improvement of 12.08 over the non-contextual baseline, also expressed in increased translation quality (25.22 vs 23.36 BLEU). Furthermore, it bests TAGGING at few-shot performance by 5 to 8 accuracy points, reaching above 90% accuracy with only 190 of the 5.1M annotated samples (Figure 2.8). Both TAGGING and MTCUE perform similarly with more supervised data. The TAGGING model achieves +2 to +3 accuracy points in the 1K to 100K range, while BLEU remains comparable. We hypothesise that this happens because MTCUE relies strongly on its pre-training prior

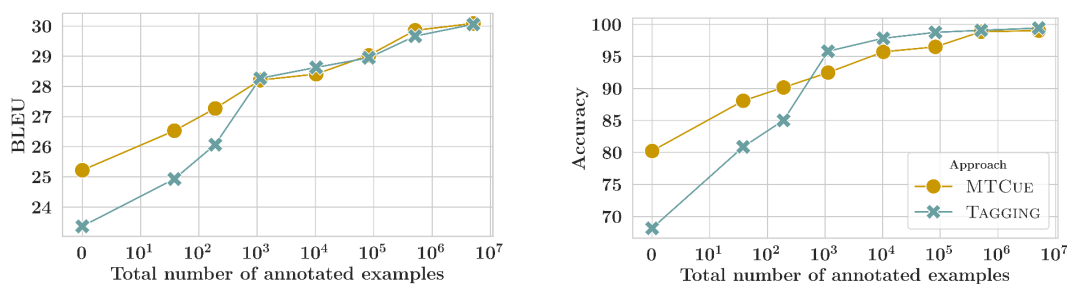


Figure 2.8: Evaluation results from the EAMT22 multi-attribute control task.

	Model	Accuracy	BLEU
0-shot	TAGGING	67.53	23.35
	MTCUE	80.35	25.13
100-shot	TAGGING	84.97	26.06
	MTCUE	90.11	27.22
5M-shot	TAGGING	99.35	30.03
	MTCUE	99.07	30.11

when context is scarce: this proves useful with little data, but becomes less relevant as more explicitly labelled samples are added. Finally, with full supervision, both models achieve above 99% accuracy.

ZERO-SHOT CONTROL OF FORMALITY (IWSLT 2022) MTCUE appears to successfully control the formality of translations in a zero-shot fashion, achieving nearly 100% accuracy on the IWSLT 2022 test sets across two language pairs, beating all zero-shot models on the EN-RU pair and performing on par with the best supervised model for EN-DE. Notably, both baselines presented in Table 2.18 were built to target formality specifically, unlike MTCUE which is a general-purpose model.

Following MTCUE’s success at controlling formality with sample contexts, we investigate the relationship between context embeddings and their corresponding formality control scores. We consider all 394 unique contexts from the OPENSUBTITLES validation data, and another 394 document contexts (individual past sentences) at random (in-domain). We also use an in-house dataset from a similar domain (dubbing of reality cooking shows with custom annotations of scene contents) and select another 394 metadata and 394 document contexts from there (out-of-domain). We run inference on the IWSLT 2022 test set with each context individually (1,576 runs), and use UMAP (McInnes et al. 2018) to visualise (i) the input embedding from MINILM-v2, (ii) the

	Model	Supervision	Formal	Informal	Average
EN-DE	Non-context baseline	–	74.5	25.5	50.0
	Rippeth et al. (2022)	Supervised	99.4	96.5	98.0
	Vincent et al. (2022a)	Supervised	100.0	100.0	100.0
	MTC _{CUE}	Zero-shot	100.0	100.0	100.0
EN-RU	Non-context baseline	–	96.4	3.6	50.0
	Rippeth et al. (2022)	Zero-shot	100.0	1.1	50.5
	Vincent et al. (2022a)	Zero-shot	99.5	85.8	92.7
	MTC _{CUE}	Zero-shot	100.0	99.4	99.7

Table 2.18: Evaluation on the IWSLT 2022 formality control evaluation campaign. Baseline systems were trained on different corpora.

output vector of the context encoder and (iii) the corresponding formality score (Figure 2.7).

We invite the reader to pay attention to the separation of dark and light points in Figure 2.7b that is not present in Figure 2.7a. There is a spatial property that arises in the context encoder and is shown by Figure 2.7b, namely a relationship between the feature vectors from context encoder and formality scores across both domains: contexts yielding translations of the same register tend to be clustered together. This is true for both in-domain data (circles) and out-of-domain data (crosses), suggesting that after training this effect generalises to unseen contexts.

For further investigation, we sample a few contexts at random which yield 100% zero-shot accuracy (from the “ends” of the colour scale) and find that these contexts tend to have semantic relationships with the type of formality they induce in translations. For example, contexts like “What’s wrong with you?”, “Wh-what’s he doing now?” yield all-informal translations while “Then why are you still in my office?” or “I can see you’re very interested.” result in all-formal ones. This confirms our hypothesis: MTC_{CUE}’s context encoder aligns the semantic representation of the input context to the most likely formality it would produce, akin to a human translator deducing such information from available data. Outside of an evaluation scenario like the present one, MTC_{CUE} may therefore be able to predict from the given context what formality style should be used: an effect only facilitated by the context encoder.

2.5.7 Examples of Model Outputs (Zero-Shot)

To exemplify how the zero-shot performance of MTCUE manifests in practice, we present some examples of outputs: [Example 1](#) and [Example 2](#) related to the EAMT 2022 multi-attribute control task, and [Example 3](#) related to the IWSLT 2022 formality control task.

Example 1	EAMT 2022
Source	I just didn't want you to think you had to marry me.
Context	<i>I am a woman. I am talking to a man</i>
Reference	Bo nie chcia ^{am} , żebyś myśla ^ł , że cię zmuszam do ślubu. (<i>'Because I didn't want_{feminine} you to think_{masculine} I am forcing you into a wedding.'</i>)
BASE-PM	Po prostu nie chcia ^{em} , żebyś myśla ^a , że musisz <u>za mnie wyjść</u> . (<i>'I just didn't want_{masculine} you to think_{feminine} you had to marry_{feminine} me.'</i>)
MTCUE	Nie chcia ^{am} , żebyś myśla ^ł , że musisz <u>się ze mną ożenić</u> . (<i>'I didn't want_{feminine} you to think_{masculine} you had to marry_{masculine} me.'</i>)

The phrase *to marry someone* can be translated to Polish in several ways, indicating that the addressee is to be a wife (*ożenić się z kimś*), a husband (*wyjść za kogoś [za męża]*) or neutral (*wziąć ślub*). While the reference in [Example 1](#) uses the neutral version, both BASE-PM and MTCUE opted for feminine/masculine variants. However, the gender of the speaker is feminine, so the phrase "... had to marry me" should use either the neutral version (*wziąć ślub*) or the feminine one (*ożenić się*). The baseline model incorrectly picks the masculine version while MTCUE is able to pick the correct one based on the context given. MTCUE also correctly translates the gender of the interlocutor: both in the top example (*myślał* vs *myślała*) and the bottom one (*aś* vs *eś*, even though a synonymous expression is used in translation, agreement remains correct).

Example 2

EAMT 2022

Source	So then you confronted Derek.
Context	<i>I am talking to a woman</i>
Reference	A więc doprowadziłaś do konfrontacji z Derekiem. (‘So then you led _{feminine} to a confrontation with Derek.’)
BASE-PM	Więc wtedy skonfrontowałeś się z Derekiem. (‘So then you confronted _{masculine} Derek.’)
MTCUE	Więc skonfrontowałaś się z Derekiem. (‘So then you confronted _{feminine} Derek.’)

Example 2 highlights a case where the morphological ending to the verb *confronted* depends on the gender of the interlocutor. MTCUE correctly opts for the *aś* ending associated with feminine grammatical gender as opposed to the masculine *eś* used by the BASE-PM model.

Example 3

IWSLT 2022

Source	I got a hundred colours in your city.
MTCUE (formal)	Ich habe 100 Farben in <u>Ihrer</u> Stadt.
MTCUE (informal)	Ich hab 100 Farben in <u>deiner</u> Stadt.

Finally, **Example 3** shows MTCUE produces correct possessive adjectives for each desired formality.

2.5.8 Ablation Study

We discuss the robustness of MTCUE with an ablation study on the model components as well as a complementary ablation on types of context (metadata vs document). We evaluate three language pairs (EN→DE,FR,PL) and report results from single runs (Table 2.19): COMET score on the OpenSubtitles18 data and zero-shot accuracy at the two contextual tasks (on the **validation** sets in all cases).

Removing the context encoder (output of the linear layer is combined with source straight away) or the position embeddings has only a minor effect on the COMET score; replacing MINILM-v2 with a discrete embedding function hurts performance the most. Positional embeddings seem more important to the EAMT 2022 task than IWSLT 2022

<i>Ablation</i>	COMET			ZERO-SHOT ACCURACY	
	EN→DE	EN→FR	EN→PL	IWSLT 2022 (DE)	EAMT 2022
Full MTCUE	46.89	54.06	62.67	100.0	81.35
no context encoder	46.76	53.73	63.26	89.10	77.42
no pos. embeddings	46.68	53.81	62.47	91.65	70.91
no MINILM-v2	45.32	53.42	62.55	50.00	70.16
no metadata	45.23	53.64	62.64	89.70	83.41
no doc.-level data	46.23	53.49	61.67	68.80	74.64
random context	42.17	51.94	61.74	49.90	68.44
no context*	41.22	50.07	58.94	50.00	67.53

Table 2.19: Ablation study on model components and data settings. *Corresponds to non-contextual Transformer.

- possibly because EAMT 2022 focuses on sentence-level phenomena, so the order of past context matters.

Replacing MINILM-v2 with a discrete embedding function removes the zero-shot effect in both tasks. An interesting finding is that between metadata and document-level data, it is the latter that brings more improvements to contextual tasks; this means that our model potentially scales to domains without metadata. Finally, using random context degrades performance w.r.t. full model implying that the gains come from signals in data rather than an increase in parameters or training time.

Finally, while we do not ablate the choice of representing context as equal-sized vectors, in this paragraph we offer a discussion of the alternative choices. Vectorisation of extra-textual context is only one of many ways in which it can be incorporated into the model. Alternatively, the textual fields of the context variables could be used as input verbatim, each individually in their respective context encoders or collectively, separated by a custom token. Another possibility is using the represented context as a prefix to the source or target sentence, eliminating the need for a separate encoder. However, these approaches come with challenges. First, Transformer models, particularly at the scale characteristic of NMT applications, notoriously struggle with longer sequences (Dai, Yang, Yang, Carbonell, Le & Salakhutdinov 2019). Adding extra tokens in the source or target input could potentially harm the model’s quality. Moreover, even in a separate encoder, context inputs with lengthy sequences, such as those containing plot or character descriptions, may not be processed optimally, and the complete pipeline may be slow to train. One notable advantage of vectorisation is its equal treatment of each context variable, whether it is *genre* or *plot description*, at the beginning of

training. This helps prevent information overload and ensures that longer contexts do not dominate the model’s attention. Lastly, the cosine similarity objective used to train `MINILM-v2` increases the interchangeability of context inputs, which is paramount in a scenario where not all variables are guaranteed to always be represented, but there is often a significant overlap between some of them.

2.5.9 Conclusions

We have presented `MTCUE`, a new `NMT` architecture that enables zero- and few-shot control of contextual variables, leading to superior translation quality compared to strong baselines across multiple language pairs (English to others, cf. [Table 2.17](#)). We demonstrated that using sentence embedding-based vectorisation functions over discrete embeddings and leveraging a context encoder significantly enhances zero- and few-shot performance on contextual translation tasks. `MTCUE` outperforms the winning submission to the IWSLT 2022 formality control task for two language pairs, with zero-shot accuracies of 100.0 and 99.7 accuracy respectively, without relying on any data or modelling procedures for formality specifically. It also improves by 12.08 accuracy points over the non-contextual baseline in zero-shot control of interlocutor attributes in translation at the EAMT 2022 English-to-Polish task. Our ablation study and experiments on formality in English-to-German demonstrated that the context encoder is an integral part of our solution. The context embeddings produced by the context encoder of the trained `MTCUE` can be mapped to specific effects in translation outputs, partially explaining the model’s improved translation quality. Our approach emphasises the potential of learning from diverse contexts to achieve desired effects in translation, as evidenced by successful improvements in formality and gender tasks using film metadata and document-level information in the dialogue domain.

2.6 CHAPTER CONCLUSIONS

Within this chapter we strived to answer **RQ1**, i.e. how attribute control can best be incorporated into neural machine translation in multiple attribute and low-resource scenarios. We have drawn the following conclusions:

1. **Interlocutor attributes can be completely controlled in translation provided adequate training data.** Our work on English-to-Polish translation with the use of extra-textual attributes highlights that, as far as the effect of **grammatical agreement** is concerned, the availability of good quality training data in a sufficient quantity should be the primary concern.
2. **When considering a fully supervised multi-attribute machine translation scenario, various methods have been suggested for integrating context embeddings, and their performances tend to be quite similar.** As shown in our assessment discussed in §2.3, approaches that utilise additional embeddings of context annotations achieve a comparable level of success when trained on the same dataset. This observation implies that forthcoming fully supervised approaches to similar problems should prioritise ensuring the availability and quality of data, rather than solely concentrating on modelling efficiency.
3. **Formality can be fully controlled in a low-resource scenario.** In §2.4 we showed how formality can be controlled given only a sample amount of training data. Following conclusion 2., we predominantly focused on the quality and quantity of the training data rather than modelling. We also implemented a hypothesis re-ranking approach which further boosted the formality accuracy in our submissions from 98.3% to 99.9% without impacting translation quality. This re-ranking component in particular may offer an alternative solution to grammatical agreement in translation in general: if a reward model can be built to assess the accuracy of a hypothesis, then the hypotheses which express the given phenomenon correctly can be preferred over others.
4. **Context has a greater impact on translation from English compared to translation into English.** In the evaluation campaign utilising our proposed MTCUE architecture, we found that the models trained to translate *from* English experienced greater advantages from our approach than models trained on the same datasets for translation *into* English. While this observation is based on empirical evidence, it aligns with the prevailing pattern where translating from English tends to derive more benefits when context is utilised. This could be attributed, at the very least in part, to English being a naturally gender-neutral language, and one that does not explicitly differentiate between various levels of formality.
5. **Information for desirable attributes can be contained in the underlying representation of context information.** Through training MTCUE on document-level

information and film metadata, we successfully managed to manipulate a range of speaker and interlocutor attributes in zero-shot translation. This outcome can be attributed to how our architecture encodes and handles context. Importantly, this strategy might pave the way for future transfer learning applications, like training on one set of contexts and assessment on another.

REFERENCE-LESS ANALYSIS OF CONTEXT SPECIFICITY IN TRANSLATION WITH PERSONALISED LANGUAGE MODELS

3.1 INTRODUCTION

In [Chapter 2](#), we have presented several ways in which context can be used to improve both translation quality and context appropriateness, exploring scenarios with varying access to supervised data. In [§2.5](#) specifically, we presented the MTCUE system which is trained to accept as input a range of various contextual variables. We have shown that the model leads to improvements in translation quality compared to its non-contextual counterpart. Thanks to the evaluation tools described in [§2.3](#) and [§2.4](#) we were also able to quantify how well the model performs at satisfying grammatical agreement in the given tasks.

What is missing from [Chapter 2](#) – and from literature hitherto – is a method of quantifying how well the given contextual translation method actually captures the effect of *behavioural* agreement. While metrics such as BLEU and COMET highlight potential improvements in translation quality, they would not be helpful in distinguishing between two contextual systems: just because one achieves a higher BLEU score, it does not mean that it is better at capturing the context-specific features of text. The present Chapter takes a step back to focus on evaluating *context specificity* directly from dialogue sentences, divorced from their source counterparts. We arrive at a reference-free evaluation formulation which focuses on the generated translations and, in a two-model setting, calculates how specific the translations are in the given extra-textual context.

Studies of sociolinguistics have long accepted that spoken language is not universal ([Milburn 2004](#)). Contrary to this, conventional approaches to generation tasks in natural language processing (NLP) build models in a one-size-fits-all fashion, and most often for a particular language and domain, disregarding the context of the processed text. In practice, this leads to assuming the most likely scenario as context, sometimes resulting in harmful predictions (e.g. the “masculine default” in [Schiebinger 2014](#)). **Personalisation** – adapting model predictions to the unique [dialogues](#) of individuals – offers clear benefits in generation tasks ([Flek 2020](#), [Dudy et al. 2021](#)), where context information can help disambiguate the input text, aiding correct interpretation and minimising sample bias in training data ([Dudy et al. 2021](#)). Personalised systems, by more effectively capturing the speaking patterns of individuals with specific characteristics or in particular environments, can be used to generate text adapted to the individual or more accurately estimate the likelihood of their authorship of a sentence.

Demographic factors have been shown to improve the performance in various NLP tasks, such as classification (Hovy 2015), generation (Zeng et al. 2019), and translation (§2.3, Vincent et al. 2022b). This impact can be *grammatical*, which is well-defined and of a morphosyntactic nature, and *behavioural*, which is more fluid and pertains to the way language is used by certain demographics or in certain situations. Grammatical agreement can be exemplified by gender, which, in some languages, determines the morphological ending of self-referent verbs. Behavioural agreement is more subtle. For example, the utterance “They’re done!” has a different meaning when said by a baker about a batch of cookies than when exclaimed by a frenzied king about his treacherous subjects. Contextual language generation methods frequently concentrate exclusively on grammatical agreement (e.g. document-level translation in Voita et al. 2019b), and unsurprisingly so, as it is the more well-understood and easier to evaluate of the two types of adaptation. In practice, however, both grammatical and behavioural agreement are required in the language generation process and a robust framework accommodates one as well as the other.

This Chapter focuses on context-based personalisation of LMs and NMT systems for speakers and productions in TV series and film (which we refer to as *productions*). The way language is used in this domain can vary greatly; for example, TV writers will construct characters who mimic the way of speaking of a certain group they represent; productions from a certain decade, country or within a specific genre will capture the discourse nuances of that group. We demonstrate how speaker and production metadata can be used to create context-based personalised LMs that model the language of a specific speaker or production more effectively than a one-size-fits-all model. We then apply these LMs in practice to measure the **context specificity** of the tokens in translation hypotheses – the extent to which tokens occurring in the text are specific to the given speaker and metadata context – both in professional and machine translations, in the domain of translating dialogue from TV series.

To define our metric, we borrow from the statistical concept of pointwise mutual information (PMI) which measures how likely a particular utterance is to occur in the provided extra-textual context. The metric is intuitive in interpretation: a positive PMI suggests that the utterance is more likely in the given context than in the general case (analogously for the negative score). Our results suggest that the degree to which professional translations in our domain are context-specific (PMI of 0.073) can be preserved to a better extent by a contextual machine translation model (PMI of 0.051) than a non-contextual model (PMI of 0.028). This is also reflected in the contextual model’s superior BLEU and COMET scores.

Our selected domain presents an additional challenge: models must be robust to the scenario where there are no prior dialogue samples for the given speakers or productions, i.e. when new content arrives and only metadata is available. This is known as the cold start problem (e.g. Schein et al. 2002, Huang et al. 2014) where there is insufficient content to characterise the subjects of a given system. Models adapted

solely on past dialogue are not robust for this case, and we argue that a context-based approach is more effective, mimicking the benefits of personalisation by estimating token distributions for similar character/production profiles.

Since the datasets in our selected domain contain identifiable information such as titles and character names, we were able to collect a rich set of metadata annotations for the selected corpus, allowing us to perform experiments on up to 14 unique metadata variables at once, to our knowledge the richest set of metadata information for personalisation. In contrast, metadata-based approaches to personalisation reported in previous work in different domains were small-scale, leveraging a few simple and mostly categorical variables (Huang et al. 2014, Lynn et al. 2017, King & Cook 2020, Welch et al. 2020, Guo et al. 2021).

Our work is presented in three parts: first, we consider whether rich character profiles can be used to model the characters' speaking styles (§3.3.4.1), including for characters which did not appear in the training data, by learning from data for characters with similar profiles (§3.3.4.2). Then, we explore how such personalised LMs can be used to estimate the context specificity (or extent of personalisation) of professional and machine translations (§3.4). Finally, in §3.5 we report on a cost-benefit analysis of the manually collected annotations, encouraging future effort in collecting annotations which proved most cost effective in our experiments. Additionally, we contribute CORNELL-RICH (§3.3.1), a corpus of rich character and film annotations for the Cornell Movie Dialogue Corpus (Danescu-Niculescu-Mizil & Lee 2011) (CORNELL) and sMRR, a bespoke evaluation metric for personalised language models. As an addendum, in §3.3.3.2 we also discuss our pre-training strategy which contributed to the success of the approach to personalising LMs. This paper also presents the related work (§3.2), and conclusions (§3.7).¹

3.2 RELATED WORK

PERSONALISATION IN NLP Personalisation in NLP can generally be split into three groups with respect to how much data is available for a speaker:

1. *full supervision*, where there is sufficient training data to fine-tune a model for a particular speaker,

¹ The present Chapter has been submitted to a conference and is awaiting review at the time of thesis submission. It lists eight authors, of which the first is the author of this thesis, the last is the PhD supervisor and the remaining six are employees at ZOO Digital Group PLC, the industrial partner to this thesis. The two top credited authors from the company - Alice Dowek and Rowanne Sumner - are the two data annotators mentioned in the paper, directly supervised by Charlotte Blundell and receiving instructions from the first author. Emily Preston, Chris Bayliss and Chris Oakley were involved in supervising the data collection process for the ZOO corpus discussed in the paper. Finally, the majority of the authors contributed by revising the paper before submission to a venue. An earlier draft of this work has been made available on arXiv (Vincent, Sumner, Dowek, Blundell, Preston, Bayliss, Oakley & Scarton 2023).

2. *few-shot*, where some supervised data exists but not in quantities sufficient for supervised training,
3. *zero-shot*, where no samples of text exist for the speaker used in evaluation.

Full supervision is usually facilitated through some form of a *user embedding* or *tagging* approach (e.g. [Sennrich et al. 2016b](#), [Keskar et al. 2019](#), [Miresghallah et al. 2022](#)): in abstract, a unique token is assigned to each speaker and provided together with samples of dialogue of that speaker during training and inference. Among few-shot approaches, [King & Cook \(2020\)](#) examine several personalisation methods for language modelling of blog posts with sample adaptation data for new users: fine-tuning, interpolation (averaging the fine-tuned speaker model with a general model), priming (updating the cell state of the language model; only possible within selected architectures), and demographic-based adaptation (fine-tuning on text from users with the same age and gender). They find that interpolation works best when a relatively large amount of data is available for the speaker, while priming outperforms all other methods when little data is available. [Welch et al. \(2022\)](#) consider a scenario where models built for “anchor users” (who boast a large history of posting) are leveraged to build models for new users (with a small number of posts), focusing on the similarity between samples of users’ posts. They find that interpolating fine-tuned models of several anchor users based on the similarity between their and the new user’s user embedding performed better than weighted fine-tuning and interpolation based on authorship attribution and perplexity-based methods.

Zero-shot approaches typically leverage background data available for the new speakers, like demographic factors or metadata. [Huang et al. \(2014\)](#) rely on the social network of a user to model their language; [Lynn et al. \(2017\)](#) use age, gender, and personality traits to improve user modelling in multiple NLP tasks; [Zeng et al. \(2019\)](#) leverage user profiles to improve comment generation on a social media corpus; [Welch et al. \(2020\)](#) produce compositional demographic word embeddings by learning demographic-specific vectors for same training data as the NMT models to measure each word in the vocabulary. Demographic-based adaptation was found inferior to interpolation and priming in the few-shot scenario by [King & Cook](#), but their study only used two factors: age and gender. More recently, substantial strides have been made in leveraging large language models (LLMs) as a potential cornerstone for personalised NLP applications. [Salemi et al. \(2023\)](#) introduce the LaMP benchmark, specifically tailored for various personalised LLM tasks. Their discussion delves into strategies for personalising LLM outputs at inference time. Aware of the trade-off between input length and quality in LLMs, they opt for a method involving a retrieval module that selects the most useful features from the user’s profile for the given input.

Our work is also positioned in the zero-shot category as we rely on rich metadata annotations to model the dialogue of individual screen characters appearing in particular productions. Unlike [King & Cook](#), we leverage textual (real-valued) metadata

annotations, which in personalisation are preferable to categorical values (Lynn et al. 2017), and a significantly higher count of them (up to 14). Ultimately, our approach, though not reliant on potentially potent LLMs, boasts full trainability and deployability on commonplace hardware. In contrast with Salemi et al., our method not only scales effectively to the zero-shot scenario but also capitalises on supervised dialogue samples for users with accessible records of past dialogues or responses. Importantly, we also leverage our personalised LMs to quantify context specificity in professional and machine translation.

A few works have explored the idea that context in LMs can be summarised with pre-trained models. Ippolito et al. (2020) propose a language model which selects the best continuation to a story from a finite set of pre-trained embeddings. Novotney et al. (2022) introduce the notion of CUE (contextual universal embedding) vectors, representing individual context variables as pre-trained vectors. They use DISTILBERT to obtain the vectors, pass them through a context encoder and average the result. Novotney et al. demonstrate that including article metadata in the form of CUE improves perplexity in language modelling of the articles. Vincent, Flynn & Scarton (2023) explore this idea further, applying it to machine translation of dialogue and showing that pre-training on film metadata helps zero- and few-shot performance in some contextual MT tasks. In this paper, we leverage context in the same way as Vincent, Flynn & Scarton (2023), but focus on contextual language modelling, and specifically on personalisation for individual characters and films. We also explore a practical exploration of such personalised LMs in evaluation of contextual MT (§3.4), and contribute an evaluation metric for personalised LMs (§3.3.3.4).

EVALUATION OF CONTEXTUAL MACHINE TRANSLATION Traditional measures of MT quality (§ A.2.3.2) are based on sentence-level matching to references, and offer little insight into performance at maintaining or introducing context-specific features of the source text. Alternative evaluation methods of contextual MT have been explored to address this. When contextual phenomena are directly observable and necessitate grammatical agreement (e.g. in formality transfer or document-level translation), evaluation usually involves parsing tools (Sennrich et al. 2016b, Vincent et al. 2022a) or contrastive evaluation on bespoke test suites (Bawden et al. 2018a, Müller et al. 2018, Voita et al. 2019b, Lopes et al. 2020). However, the creation of such tools and test sets is expensive, and as argued in Post & Junczys-Dowmunt (2023), strong performance at contrastive evaluation does not necessarily entail the ability to generate contextual translations in practice. Evaluation of behavioural agreement (e.g. preserving individual style of a character or production), has mostly been limited to classification systems (e.g. Michel & Neubig 2018b) which attribute the input text as belonging to one of a list of speakers. However, such systems depend on sufficient quantities of training data for each considered speaker, which is usually not readily available.

LANGUAGE MODELS IN MACHINE TRANSLATION Language models have been utilised to improve machine translation as means of improving fluency (Stahlberg et al. 2018), boosting document-level performance (e.g. Sugiyama & Yoshinaga 2021) or evaluation (Edunov et al. 2020). In contrast, we build a tandem of language models from the same training data as the NMT models to measure to what extent the NMT generations are context-specific. Our metric is pointwise mutual information (PMI) (§ A.3), computed in a vein similar to Sugiyama & Yoshinaga (2021) who used PMI between document context and the target utterance to boost document-level performance of MT.

3.3 BUILDING A PERSONALISED LANGUAGE MODEL

The first stage of our work delves into building a personalised language model for dialogue associated with rich contextual annotations. We create two metadata-rich datasets (§3.3.1) and train a contextual language model to capture the distribution of the tokens in the dialogue given a set of contextual variables (§3.3.3.1). This section addresses the following research questions (RQs):

RQA How can rich character profiles be used to model the characters’ speaking styles? (§3.3.4.1)

RQB How can a LM be personalised for a specific character solely by learning from data for characters with similar profiles? (§3.3.4.2)

3.3.1 Datasets

Since context-annotated dialogue data is hard to come by, we use a combination of manual and automatic annotation to create two English-language corpora: ZOO-ENGLISH and CORNELL-RICH, which we describe in this section. The domain of both corpora is TV series and film dialogue respectively, and samples within each corpus consist of: an utterance in English and a set of up to 14 textual metadata annotations for the speaking character (age bracket, country of origin, description, gender, profession, religion and characteristic quote) and for the production (country, genre, PG rating, plot description, writers, year). Below we summarise the descriptions for each corpus.

THE ZOO-ENGLISH CORPUS The ZOO-ENGLISH corpus is a private in-house collection of subtitles for nine anglophone TV series. It totals 157K dialogue lines and annotations for 159 speakers of 101K lines. It is divided into traditional test, valid and train, but features an additional test set of metadata and dialogue from 11 **held-out** speakers who do not appear in the remaining sets. Quantitative details are reported in Table 3.4 (rows 1-5) and a sample from the corpus is presented in Table 3.1. The corpus was created from production-ready subtitle files from which dialogue with character and TV series attributions was extracted. This data was subsequently annotated with

Metadata type (% annotated)	Value
Speaker metadata (50.3%)	
Age bracket	Adult
Description	Chris Kraft is an American aerospace and NASA engineer who was instrumental in establishing the agency's Mission Control Center (...)
Characteristic quote	He just kited a damn check.
Country of origin	United States
Gender	I am a man
Profession	Flight Director
Religion	Christian
Production metadata (87.1%)	
Genre	Drama, History
PG rating	PG rating: TV-14
Names of writers	Written by: Mark Lafferty
Country of production	United States
Year of release	Released in 2020
Plot description	U.S. fighter pilots are recruited to test experimental aircraft and rockets to become first Mercury astronauts.

Table 3.1: A sample of metadata from the ZOO-ENGLISH corpus.

production metadata (automatically, via the OMDb API²) and character metadata for the most frequently speaking characters. The annotation process is detailed in §3.3.2.

THE CORNELL-RICH CORPUS Much like ZOO-ENGLISH, CORNELL-RICH is a dataset of rich character and production annotations, albeit for film dialogue extracted from scripts. It includes 14 distinct metadata variables captured as text. The collected annotations can be linked to the entries of CORNELL (Danescu-Niculescu-Mizil & Lee 2011) which is a corpus of exchanges from a set of film scripts, with character dialogue attributions (Figure 3.1 illustrates how CORNELL-RICH enriches the original corpus). Both dialogue data and annotations are in English. CORNELL-RICH comprises annotations for 863 speakers (speaker *profiles*), covering 135.7K utterances; nearly half of the annotated

² <https://www.omdbapi.com/>

speakers have 150+ lines of dialogue and about 25% have 200+. At least 64.1% of conversational exchanges feature at least one annotated character and as much as 95.5% of the featured films are annotated with film metadata (Table 3.2). We provide a full list of collected metadata with examples in Table 3.3.³

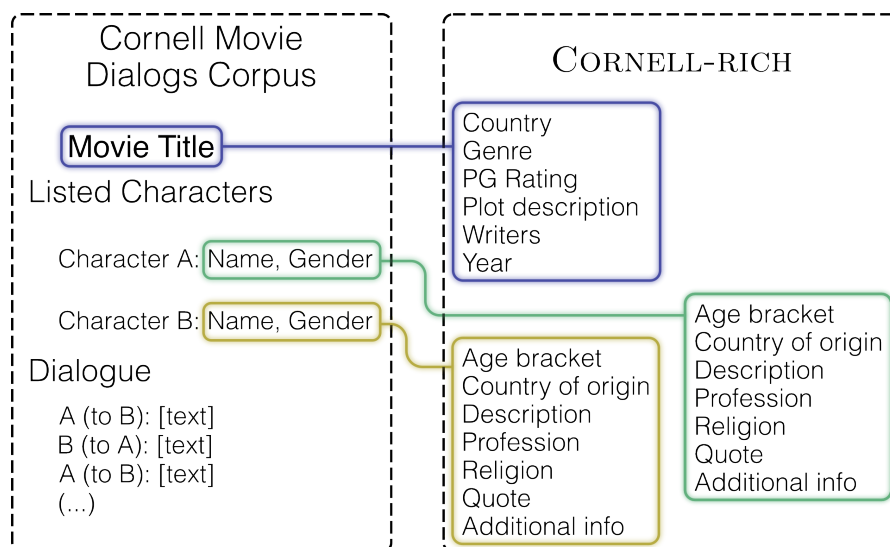


Figure 3.1: Comparison between CORNELL and our proposed CORNELL-RICH.

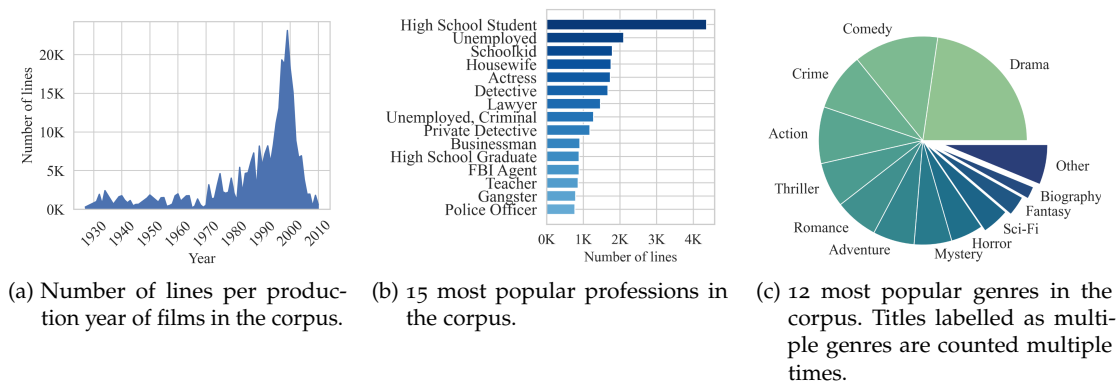


Figure 3.2: Visualisation of a subset of features of the proposed corpus.

As per Figure 3.2, the annotated films span nearly a century, with most lines coming from between the 1990s and 2005; the distribution of professions is significantly flatter,

³ The annotations are available to download at <https://github.com/st-vincent1/cornell-rich>












	Count	% of all	Utterances	% of all
speakers	9.0K	–	304.3K	–
	863	9.5%	135.7K	44.6%
	8.2K	90.5%	170.1K	55.9%
exchanges	83.1K	–	304.3K	–
 ⇔ ( or 	53.3K	64.1%	202.4K	66.5%
 ⇔ 	36.8K	44.3%	134.4K	44.2%
 ⇔ 	16.5K	19.8%	68.0K	22.3%
films	617	–	304.3K	–
annotated	589	95.5%	291.0K	95.6%

Table 3.2: Details of annotations compared to data quantities from CORNELL.  = speaker with rich annotations.  = speaker without rich annotations.

with the dominant field (“High School Student”) only making up about 3% of the corpus. Finally, the most popular genres include drama, comedy, crime, and action.

THE OPENSUBTITLES CORPUS We use the OpenSubtitles18⁴ corpus (Lison et al. 2018) (OPENSUBTITLES) to pre-train the language models. It is a large collection of subtitles with timestamps that facilitate the extraction of document-level information. Focusing on past context with no loss of generality, we extract up to 3 past sentences based on the timestamps (Table 3.4, rows 11-12). Roughly 68% samples contain at least one past sentence. We detail how the models are pretrained in §3.3.2.

3.3.2 Details regarding the data collection campaign

The data collection process of the CORNELL-RICH and ZOO-ENGLISH annotations was carried out by two annotators, both native English speakers and experts in the dubbing and subtitling industry, formally employed by ZOO Digital. After parsing CORNELL⁵ and ZOO-ENGLISH, a spreadsheet of characters was generated that included the names of the characters, the names of their source films, and the number of lines attributed to each character.

From previous work (e.g. Johannsen et al. 2015) and hypotheses made based on experts’ experience, we pre-defined a number of categories of information to collect

⁴ Based on <https://opensubtitles.org/>

⁵ <https://convokit.cornell.edu/documentation/movie.html>

Metadata type (% annotated)	Value
Speaker metadata (44.6%)	
Gender	A man
Age bracket	Adult
Profession	Attorney
Description	Galvin graduated from Boston College’s law school. Galvin had a promising legal career ahead of him at an elite Boston law firm until he was framed for jury tampering by a partner due to his plans to expose the firm’s underhanded activities. (...)
Quote	Your honor, with all due respect: if you’re going to try my case for me, I wish you wouldn’t lose it.
Country of origin	USA
Religion	Christian
Film metadata (95.5%)	
Genre	Comedy, Drama
PG Rating	PG Rating: R
Names of writers	Written by: Paul Andréota, André Cayatte, Henri Coupon
Country of production	France, Italy
Year of release	Released in 1974
Plot description	A French judge try to acquit a man who is accused of murdering his lover.

Table 3.3: A sample from CORNELL-RICH with each type of collected metadata.

about each character. Specifically, we selected categories that we hypothesised to (i) be identifiable from the available sources and (ii) influence a person’s speaking style or vocabulary used. They were: their **age bracket** \in {child, teen, young adult, adult, elderly}, **profession**, **character description** (a few sentences summarising their personality or character arc), **religion** and a **characteristic quote**: a typical or quotable phrase the character might say. Additionally, the **gender** annotations from the original corpus were re-used, and an optional column “**additional information**” was included

to collect comments from experts⁶. The characters with the most lines spoken were prioritised. This resulted in 863 characters being annotated for CORNELL-RICH and 159 for ZOO-ENGLISH.

ANNOTATION SOURCES Annotations are based on publicly available pages from Wikipedia⁷ for individual films, as well as fan-made Fandom⁸ pages for both films and characters. Where information was unavailable from these sources, the annotators either referred back to the corpus itself or skipped the given field altogether. The film metadata was obtained via the OMDb API⁹.

ANNOTATION DECISIONS The annotation process involved matching every script's name against an IMDb entry, which did not always yield a match as some scripts had been scrapped or rewritten or characters' names had been changed. Unidentifiable films and characters were not considered for annotation. Some information, especially *religion*, was occasionally difficult to find, in which case it would be skipped or labelled as *Unknown*. It was challenging to produce annotations for characters based on real people, or for a real person played by themselves. Where characters were based on historical figures, the annotators focused on the production interpretation of the person; when dealing with a characterisation of the person at a specific point in time, the focus was on their behaviour at that point in time. Finally, some characteristics were unsuitable for selected character information: e.g. when a character was immortal, it did not fit into set age brackets, and for some characters there were limited clues to determine their age bracket. In both cases, the final annotations were based on the annotators' expertise.

PREPROCESSING Since both ZOO-ENGLISH and CORNELL are of high quality as is, our preprocessing only involves normalising punctuation, removing tokenisation using the `sacremoses` package¹⁰, fixing leftover punctuation issues (e.g. ensuring all multi-dots use three dots) and removing HTML tags. We also preprocess all (original and added) annotations so that: (i) all empty fields are expressed as an empty string; (ii) there are no multiple expressions of the same discrete type (e.g. *m* and *M* to denote masculine gender); (iii) all attributes are expressed in unambiguous natural language (e.g. a PG rating of "R" is rewritten as "PG Rating: R"). OPENSUBTITLES is preprocessed following Vincent, Flynn & Scarton (2023), using the scripts provided by the authors. For subword tokenisation, we use SentencePiece to train a BPE model of 8K tokens on the train split of CORNELL-RICH; it is then used to tokenise all datasets.

6 Upon inspection: the annotators predominantly used this field to provide the actor's name, an interesting fact about the character (e.g. "Plays a caricature of himself"), or trivia.

7 <https://wikipedia.org/>

8 <https://fandom.com/>

9 <https://omdbapi.com/>

10 <https://pytorch.org/project/mosestokenizer/>

Row	Dataset & split	segments	tokens	Total number of metadata types
(1)	ZOO-ENGLISH			
(2)	train	140.4K	1.1M	
(3)	valid	4K	31.3K	13
(4)	test	6K	47.1K	
(5)	test_unseen	6.7K	51.5K	
(6)	CORNELL-RICH			
(7)	train	289.0K	3.1M	
(8)	valid	5K	51.2K	14
(9)	test	5K	54.4K	
(10)	test_unseen	5.2K	54.6K	
(11)	OPENSUBTITLES			
(12)	train	14.7M	109.6M	3*

Table 3.4: Quantities of segments, tokens (pre-tokenisation) and unique metadata (speaker and production) in datasets. *OPENSUBTITLES uses three past sentences as proxy metadata.

3.3.3 Experimental Setup

3.3.3.1 LMCUE Architecture

Our selected language model architecture is adapted from the MTCUE model (Vincent, Flynn & Scarton 2023), which is a Transformer-based multi-encoder contextual machine translation system. MTCUE processes the source text with a source encoder and the context information with an additional context encoder. We convert MTCUE to a language model by removing the source encoder, resulting in a conditional encoder-decoder LM where context is treated as the input to the encoder. The sequence of context information is converted to a sequence of equal-sized vectors with a sentence embedding model (MINILM-v2). This approach has the advantage of treating both discrete and continuous (text) inputs in the same way, potentially utilising the semantic information of the discrete labels, as well as allowing longer spans of context as input without issues of long-range dependencies. The target sequences are contextualised via standard encoder-decoder attention which maps queries (target) to keys and values (context). We select this approach as MTCUE can process large sets of contextual information and has the potential to scale well to few- and zero-shot scenarios, which

in our case are explored when we consider test sets with completely new speakers. Hereinafter we refer to this architecture as LMCUE¹¹.

3.3.3.2 Pre-training

Our preliminary experiments showed that training LMCUE from scratch on CORNELL-RICH lead to results inferior to a non-contextual language model trained on the same data (discussed in §3.6). We therefore experimented with pre-training the model first. Since a larger corpus of dialogue with character metadata is typically unavailable, we used a corpus with document-level information and treated the **past dialogue** for any sentence as the **metadata context**. We hypothesised that at a larger scale, the effect of metadata embeddings on text generation will be similar to the effect of embeddings of past dialogue (Figure 3.3), meaning the pre-training procedure allows the model to learn dependencies between the context and the text.

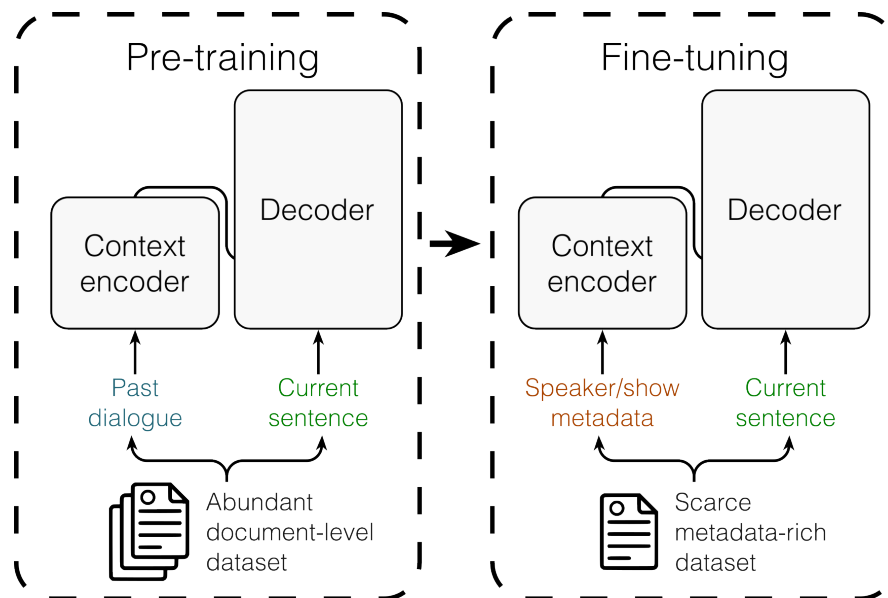


Figure 3.3: An illustration of the pre-training and fine-tuning regimens used in the experiments.

This approach has the advantage that metadata-rich corpora are likely to be too small to train a model from scratch, but document-level information is abundant. In our case, pre-training on past dialogue proved successful; consequently, all models considered in our experiments are pre-trained. For this purpose, we use the OPENSUBTITLES corpus (§ 3.3.1).

¹¹ We make the implementation of LMCUE available at <https://github.com/st-vincent1/LMCue>.

3.3.3.3 Baselines and Implementation

We consider three baselines: a non-contextual LM (BASE-LM), a speaker-wise fine-tuning baseline (SPFINE TUNING) and a linear interpolation method (LERP) which ensembles SPFINE TUNING with the general model BASE-LM at test time; both baselines are modelled after King & Cook (2020).

We implement LMCUE by modifying the code provided by (Vincent, Flynn & Scarton 2023). The model has 159M parameters and comprises a context encoder (38M) and a decoder (121M). 25% of the decoder’s parameters are used by the encoder-decoder attention; a non-contextual decoder of this shape would have 91M parameters. To make the comparison fair, BASE-LM matches the total number of parameters in LMCUE (159M) and is therefore wider than the decoder in LMCUE (for details see Table 3.5). This strong baseline removes the possibility that the model improves simply because of a higher parameter count¹². All other baselines (SPFINE TUNE, LERP) share the architecture and size of BASE-LM.

		Params	d_{model}	n_{layers}	h	FFN dim.
(1)	LMCUE (Enc.)	38M	512	6	8	2048
(2)	LMCUE (Dec.)	121M	768	12	12	3072
(3)	LMCUE (total)	159M	—	—	—	—
(4)	BASE-LM	159M	1024	12	16	4096

Table 3.5: Model details for LMCUE and BASE-LM.

The LMCUE models are pre-trained on OPENSUBTITLES (using past dialogue as context), while BASE-LM is pre-trained on the text part of the corpus, one sentence at a time. We use off-the-shelf model architectures with pre-defined hyperparameters in FAIRSEQ and only tune on three values each for batch size (simulated 200K to 400K tokens) and learning rate ($3e - 4$ to $1e - 3$) based on validation performance on valid in CORNELL-RICH. For fine-tuning, we separately adapt these parameters for each dataset and metadata combination: learning rate ($5e - 5$ to $1e - 3$) and batch size (0.25K to 20K tokens). The best fine-tuning set of learning rate and batch size was $5e - 5$ and 1.5K for LMCUE and $2e - 3$ and 3K for BASE-LM. Each model was trained on a single 32GB V100 GPU with an early stopping condition of validation loss not improving for 5 epochs. Pre-training LMCUE and BASE-LM took 35 and 17.5 GPU hours respectively while fine-tuning these models took respectively 0.78 and 0.32 GPU hours on average.

¹² Results from using the smaller baseline LM (91M params) were consistently inferior to 159M by up to 0.75 perplexity.

3.3.3.4 Evaluation

For evaluation, we use perplexity (PPL) as well as sMRR, which we define as follows: let M_j be a model personalised for a speaker s_j and U_i be a set of utterances by a speaker s_i . We calculate speaker reciprocal rank sRR for any speaker k by scoring the U_k with M_1, \dots, M_n (expressed with log likelihood), then ranking the models best to worst by this score¹³ and taking the reciprocal rank ($1/\text{rank}$) of M_k , the model for speaker k (see Figure 3.4). sMRR is sRR averaged for all speakers; $1/\text{sMRR}$ is the average rank of the

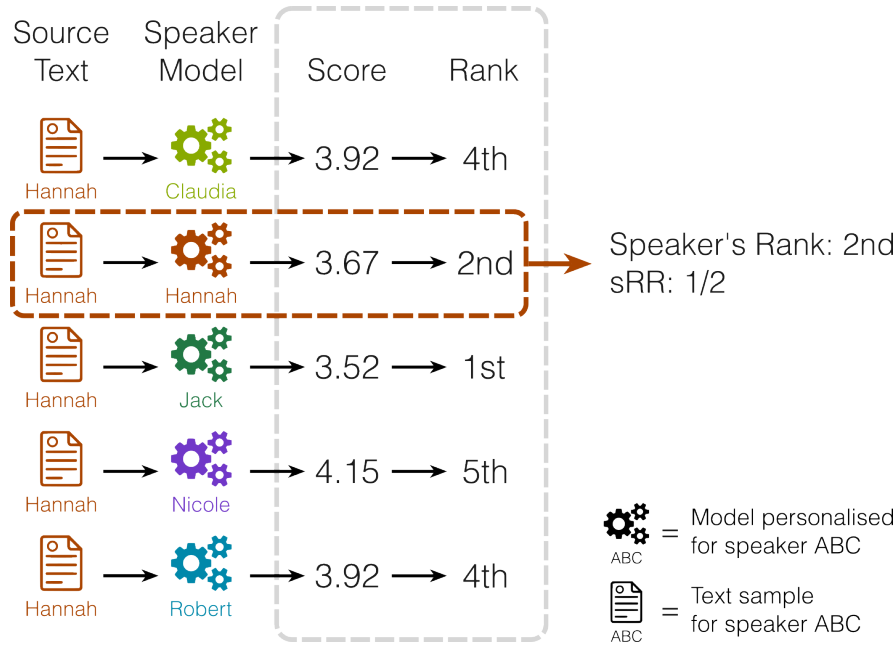


Figure 3.4: sRR illustrated for one speaker (Hannah).

correct speaker model. Intuitively, this metric captures the strength of the association between dialogue and the speaker model: sMRR of 1.0 indicates that for any speaker j , the model M_j produces the best score for U_j .

Unless otherwise specified, all results are calculated from five runs with different random initialisation, and the reported value is the **mean** result. We **highlight** the best overall result. Unless another result is underlined, it is significantly worse (indicating a less effective model) than the best result in bold, with a confidence interval of 95% (computed with a one-tailed t-test, $t(4) = 1.65$, $p = 0.05$).

¹³ Ties are resolved pessimistically.

3.3.4 Results

This section presents the results of training LMCUE on the CORNELL-RICH and ZOO-ENGLISH corpora. As before, we use the umbrella term **production** to refer to the different types of media (film and TV series) in the corpora. Furthermore, we use \mathcal{S} and \mathcal{P} to denote that *Speaker* or *Production* metadata was used in training (or both, i.e. $\mathcal{S} + \mathcal{P}$).

3.3.4.1 Are speaker profiles helpful?

We examine whether including speaker profiles as a supplementary input in language modelling can result in significant quantitative improvements. For this, we train models on the train splits and evaluate on the test splits of both corpora, with overlapping speakers (= unique combinations of speaker profiles) between them. As demonstrated in Table 3.6, context-based personalisation with LMCUE results in substantial reductions in perplexity compared to the best baseline, with a decrease of 5.4% for CORNELL-RICH and 6.5% for ZOO-ENGLISH, respectively.

	CORNELL-RICH		ZOO-ENGLISH	
	valid	test	valid	test
<i>baselines</i>				
BASE-LM	22.35	23.38	18.42	18.41
<i>proposed</i>				
LMCUE (\mathcal{S})	21.37	22.37	17.52	17.55
LMCUE (\mathcal{P})	21.07	22.04	17.18	17.29
LMCUE ($\mathcal{S} + \mathcal{P}$)	21.14	22.13	17.13	17.21

Table 3.6: Perplexity[↓] on different validation and testing sets for the two corpora.

IS SPEAKER-BASED ADAPTATION BETTER THAN DIRECT FINE-TUNING? To determine the effectiveness of our context-based adaptation approach when compared to LMs fine-tuned on the available speaker dialogue, we focus our setup on five long-term (multi-episode TV) speakers with at least 3,000 lines of dialogue sources from the ZOO-ENGLISH corpus (Table 3.7). For each speaker, we use 400 and 600 of these lines for validation and testing, respectively. Within this experiment, we use SPFINETUNE and LERP as baselines. We obtain SPFINETUNE by fine-tuning the LM on all ZOO-ENGLISH data initially (FT₁), and then once more on speaker data alone. LERP is a

<i>ID</i>	<i>#Lines</i>	<i>Age</i>	<i>Profession</i>	<i>Country</i>	<i>Genre</i>	<i>Characteristic quote</i>
sp01	7.5K	Teen	Student, Spy	United States	Comedy	"Look, I love you! I have loved you since the moment I saw you. Please! I'll never get a chance to tell you."
sp02	3.9K	Young Adult	Unemployed, Community Service	United Kingdom	Comedy, Drama	"In the words of the great Lionel Richie...hello."
sp03	3.2K	Adult	Actor	United States	Docuseries	"So be present, be spontaneous. Enjoy the moment, enjoy yourself and learn."
sp04	3.1K	Adult	Criminal Profiler	United States	Crime, Drama, Horror	"It isn't very smart to piss off a guy who thinks about killing people for a living."
sp05	3.1K	Adult	Psychiatrist	United States	Crime, Drama, Horror	"Before we begin, I must warn you... nothing here is vegetarian."

Table 3.7: Selected metadata regarding long-term speakers from ZOO-ENGLISH used in the experiment.

<i>ID</i>	<i>Top-gaining sentence (4+ words)</i>	<i>Five top-gaining tokens</i>	<i>Top-losing sentence (4+ words)</i>
sp01	"Paranoid and can fit into small spaces."	Okay, Wait, spy, Mom, mission	"To teach and to lend a guiding hand."
sp02	"Fucking nuns! Fucking shit!"	Fuck, Shit, Fucking, fucking, fuck	"English, Math and French."
sp03	"I love this car."	Wow, ital, coffee, brain, b	"I'm not opposed to doing things to my teeth."
sp04	"One missing kid's a boy."	killer, Jack, kill, close, life	"She was a slim and delicate pig."
sp05	"Is your conscience clear?"	got, killer, Will, Jack, Ab	"Simpler times in boatyards with dad."

Table 3.8: Sentences and tokens for which the log likelihood under LMCUE ($\mathcal{S} + \mathcal{P}$) changes the most compared BASE-LM.

mean interpolation of SPFINE-TUNE with FT1. We do this individually for each speaker $\in \{\mathbf{sp01}, \dots, \mathbf{sp05}\}$.

LMCUE achieves results comparable to all speaker-fine-tuned models (Table 3.9). When using speaker metadata (\mathcal{S}), LMCUE achieves sMRR of 1.0 just like fine-tuned models, suggesting the perplexity improvements come from the model's context-based predictions. LMCUE (\mathcal{P}) achieves lower sMRR (0.8): its predictions are based only on production metadata, not considering that two different characters may come from the same production. Speaker profiles \mathcal{S} are necessary for full speaker adaptation.

Any adapted model, whether fine-tuned or metadata-based, yields a reduction in perplexity between 5.1% and 6.8% which is comparable to results on test. SPFINE-TUNE achieves the best overall perplexity reduction of 1.32 and 1.0 sMRR, with LMCUE ($\mathcal{S} + \mathcal{P}$) yielding a statistically comparable reduction of 1.29 and the same sMRR while requiring (i) no fine-tuning and (ii) the maintenance of only one model for all speakers.

	sMRR [↑]	PPL [↓]					Mean
		sp01	sp02	sp03	sp04	sp05	
<i>baselines</i>							
<i>non-context</i>							
BASE-LM	0.2	15.24	17.39	23.53	18.64	21.14	19.19
<i>fine-tuning</i>							
SPFINE-TUNE	1.0	14.54	16.01	21.76	17.36	<u>19.50</u>	17.84
LERP	1.0	14.35	16.31	22.25	17.66	19.91	18.10
<i>proposed</i>							
<i>metadata-based</i>							
LMCUE (\mathcal{S})	1.0	14.99	16.75	21.86	17.54	19.89	18.21
LMCUE (\mathcal{P})	0.8	14.68	17.17	<u>21.26</u>	<u>17.12</u>	19.45	17.94
LMCUE ($\mathcal{S} + \mathcal{P}$)	1.0	14.77	16.77	21.22	17.10	<u>19.47</u>	<u>17.87</u>

Table 3.9: Results on the test set for long-term speakers. Underlined results are on par with results in **bold**.

To illustrate how personalisation manifests in practice, we identify the predictions of LMCUE ($\mathcal{S} + \mathcal{P}$) with the most increased and decreased log likelihood compared to BASE-LM (compare Table 3.7 and Table 3.8). Top-gaining tokens have strong associations with certain categories, like *profession* (sp01 “Student, Spy” → *spy, Mom, mission*), *age* (sp02 “Young Adult” → expletives) or *genre* (sp04, sp05 “Crime, Drama, Horror” → *killer*). Similarly, top-gaining sentences for sp01 and sp02 have a comedic overtone (matching the genre), while the top-losing sentences do not fit these characters’ demographic profiles.

3.3.4.2 Zero-shot Transfer to Unseen Speakers

In this section, we assess the effectiveness of speaker adaptation for completely **new test speakers** featured in the test_unseen sets of both corpora. To reiterate, these speakers’ dialogue is excluded from training and validation data (although there are overlaps in production metadata). As before, we fine-tune the pre-trained LMCUE on the train splits. We compare the performance only to BASE-LM since other baselines are not equipped to work well in this zero-shot scenario.

Results for this scenario reported in Table 3.10 show that LMCUE (\mathcal{S}) still improves perplexity over a parameter-matched LM. Though these improvements are smaller than in the supervised scenario, they are still significant, especially for ZOO-ENGLISH

Approach	CORNELL-RICH		ZOO-ENGLISH	
	test_unseen		test_unseen	
	PPL \downarrow	sMRR \uparrow	PPL \downarrow	sMRR \uparrow
<i>baselines</i>				
BASE-LM	23.62	0.03	17.11	0.09
<i>proposed</i>				
LMCUE (\mathcal{S})	23.28	0.70	16.49	0.39
LMCUE (\mathcal{P})	22.00	0.80	16.21	0.19
LMCUE ($\mathcal{S} + \mathcal{P}$)	22.31	0.96	16.35	0.32

Table 3.10: Results of evaluation with speaker & film metadata on the test set of unseen speakers.

(-0.62). More importantly, for both corpora \mathcal{S} is strongly beneficial towards high speaker separation (i.e. the model assigns the highest probability to dialogue which matches the given speaker’s profile), as measured by sMRR. Perplexity does improve more when \mathcal{P} is also used ($1.4 \rightarrow 5.6\%$ for CORNELL-RICH, $3.6 \rightarrow 4.4\%$ for ZOO-ENGLISH), though in this scenario we are evaluating the easier task of modelling new speakers in seen or unseen productions. \mathcal{P} production metadata alone yields the best reduction of 6.9/5.3%. Using it has a different effect on the two test sets: first, in CORNELL-RICH it induces a stronger boost in sMRR than \mathcal{S} ($+0.08$), while in ZOO-ENGLISH it decreases it considerably (-0.20); second, using it in conjunction with \mathcal{S} results in best sMRR in CORNELL (0.94), but not so for ZOO-ENGLISH. This can be explained by the fact that ZOO-ENGLISH uses a pool of only nine productions (vs 595 in CORNELL-RICH), so adding \mathcal{P} on top of \mathcal{S} is unlikely to increase speaker separation. In contrast, CORNELL-RICH uses a rich pool of films, so film metadata is more likely to be unique between any two speakers, thus introducing it separates the two speakers even more, increasing sMRR. This is also why sMRR is so high for LMCUE (\mathcal{P}): with 24 unique films between the 30 speakers the film metadata is rarely shared between any two speakers, making their context inputs more dissimilar. The magnitude of improvements in sMRR is also different for the two corpora, which again could be attributed to scale (863 vs 159 speakers, 595 vs 9 productions). Increasing the number of annotated entities can therefore improve the personalisation effect. Nevertheless, a score of 0.39 still suggests that LMCUE ranks an unseen character on the 2.56th position with a model built from their demographic profile, on average.

Using LMCUE ($\mathcal{S} + \mathcal{P}$), we queried the words for which log likelihood increased the most w.r.t. BASE-LM in the test set of CORNELL-RICH and obtained a list of the following fifteen tokens:

crew shark ship azz birds casino space leads
power ocean camp boat cops baby ace

Many of these tokens are context-specific and would only appear in certain scenarios or domains. For example, *casino* or *space* are unlikely to appear in a sentence unless they represent locations within the film. A subset of the provided tokens (*crew, shark, ship, ocean, birds*) may also collectively describe a single scenario, such as an adventure or thriller film set on a ship in the middle of an ocean. We hypothesise that a few such films appeared in the training set CORNELL-RICH, allowing LMCUE to develop a strong prior for predicting these tokens when metadata of similar films is provided as input. Finally, these tokens are notably more generic than those in Table 3.9: we observe that the effect of biasing speaker-specific vocabulary may be limited for some tokens compared to the supervised scenario (e.g. tokens representing names of the character’s co-stars are not related to demographic features so would not be affected in a zero-shot scenario).

3.4 MEASURING PERSONALISATION IN PROFESSIONAL AND MACHINE TRANSLATIONS

In §3.3 we have established empirically that LMCUE exhibits effects of context-based personalisation, acting as a person- and production-specific language model when provided with their metadata, and is comparable with speaker-specific fine-tuning approaches (§ 3.3.4.1). Compared to a general language model, it assigns higher probability to tokens which are more likely to occur in the given character and production context. Within this section, we use this model as a “contextual oracle”, applying it to various streams of dialogue to obtain judgements on how likely the dialogue is to be said in the given context. We also use a non-contextual language model as a “non-contextual oracle”, to measure the extent to which the given text co-occurs specifically with the provided context. We are interested in the following research question:

RQC Can MT offer personalisation benefits proportional to professional translations?

We operate on four iterations of the same text: original version in English (ORIGINAL), professional translations of the original text to French, German or Polish (REFERENCE), and several versions of machine-translated text, which we describe below. Our goal is to establish to what extent the effect of personalisation (context-specificity to particular character and production descriptions) is found in professional and machine translations, and whether hypotheses generated by a contextual machine translation system exhibit stronger personalisation effects compared to non-contextual.

3.4.1 Datasets

THE ZOO-MULTI CORPUS For our experiments, we use the ZOO-MULTI. It consists of a subset of the episodes featured in the ZOO-ENGLISH corpus, but is extended in two ways: (i) it includes professional translations to French, German and Polish, and (ii) it contains three additional groups of contextual annotations. In total, the following context types are included in this dataset:

1. **document-level context:** the previous five utterances in the source language.
2. **production metadata:** country, genre, PG rating, writers, plot description and year; this metadata was obtained via the OMDb API¹⁴ and is the same kind of information used in OPENSUBTITLES.
3. **scene metadata:** location, description of activity performed by the characters and the topic of their conversation.
4. **speaker profiles:** age bracket, character description, country of origin, gender identity, profession, religion and a characteristic quote.
5. **addressee characteristics:** number of interlocutors, their gender identities¹⁵ and the formality register.

We report an example from the corpus in [Table 3.11](#). The slight discrepancies between quantities for different language pair are the result of some series not being translated to the given language. This corpus is split differently into training, validation and testing subsets than ZOO-ENGLISH.

This corpus is split differently into training, validation and testing subsets than ZOO-ENGLISH. Specifically, we pre-select 14 episodes of three unique TV series:

- four episodes (240 min) of **The Big Family Cooking Showdown**¹⁶ (BIGFAM; 2017-2018), an unscripted British family team cooking competition.
- four episodes (240 min) of **The Right Stuff**¹⁷ (RIGHTSTUFF; 2020), a scripted American historical drama series about the United States' space programme.
- six episodes (180 min) of **The World According to Jeff Goldblum**¹⁸ (WORLDJEFF; 2019-2022), an American documentary series which follows the actor and musician Jeff Goldblum¹⁹ on his exploration of various subjects through conversations with experts and enthusiasts.

¹⁴ <https://www.omdbapi.com/>

¹⁵ For multiple addressees, the gender identity of the group is used, e.g. "All female".

¹⁶ https://en.wikipedia.org/wiki/The_Big_Family_Cooking_Showdown

¹⁷ [https://en.wikipedia.org/wiki/The_Right_Stuff_\(TV_series\)](https://en.wikipedia.org/wiki/The_Right_Stuff_(TV_series))

¹⁸ https://en.wikipedia.org/wiki/The_World_According_to_Jeff_Goldblum

¹⁹ https://en.wikipedia.org/wiki/Jeff_Goldblum

Type of text field (% populated)	Example
source sentence	That will take 12 hours.
target sentence	Cela prendra 12 heures.
past dialogue (83.4% ± 6.9)	
<i>n</i> – 5	Unless it explodes when you shoot it.
<i>n</i> – 4	I suppose that could happen. What do you suggest?
<i>n</i> – 3	Nothing.
<i>n</i> – 2	Nothing. We do nothing.
<i>n</i> – 1	We wait for the battery to run down and let the oxidizer boil off.
production (59.1% ± 28.6)	
country	United States
genre	Drama, History
PG rating	PG rating: TV-14
plot description	U.S. fighter pilots are recruited to test experimental aircraft and rockets to become first Mercury astronauts.
writers' names	Written by: Mark Lafferty
year	Released in 2020
scene (81.6% ± 0.6)	
activity	Talking, (...) arguing
conversation topic	Rocket launch malfunction
location	Blockhouse
speaker (45.9% ± 22.5)	
age bracket	Adult
description	Chris Kraft is an American aerospace and NASA engineer who was instrumental in establishing the agency's Mission Control Center (...)
characteristic quote	He just kited a damn check.
country of origin	United States
gender	I am a man
profession	Flight Director
religion	Christian
addressee (90% ± 9.6)	
number & gender	I am talking to a man
register (formality)	Informal chit-chat

Table 3.11: Details regarding the ZOO-MULTI corpus (EN-FR sample). Certain examples have been shortened for brevity.

The remaining data is split as follows, in two different data settings: **DISJOINT**, where we use the remaining episodes as validation data, and all other TV series as training data (so that there is no overlap between training and validation/testing data) and

OVERLAP, where we include the remaining episodes in the training data, and select random utterances from other series for validation. For example, if RIGHTSTUFF is used for testing, in the DISJOINT setting no samples from that show are allowed in training, but the OVERLAP setting will use all non-testing episodes as training data.

Row	Dataset & split	Number of samples		
		EN-FR	EN-DE	EN-PL
(1)	ZOO-MULTI			
(2)	DISJOINT			
(3)	train*	58.5K	59.0K	107.1K
(4)	valid*	4.0K	3.8K	4.1K
(5)	OVERLAP			
(6)	train*	60.3K	60.8K	106.1K
(7)	valid*	2.3K	2.3K	2.3K
(8)	test	7.8K	7.8K	7.6K
(9)	OPENSUBTITLES			
(10)	train	14.7M	5.3M	12.4M

Table 3.12: Quantities of segments in ZOO-ENGLISH and OPENSUBTITLES. *Values are averaged over dataset iterations generated for each of the three series.

The motivation for this dual setup is to represent the real-life scenarios of the subtitle translation task: when no past episodes are available for the considered series (DISJOINT), and when there are some already completed translations that can be leveraged, e.g. for past seasons of a series, or a prequel to the film (OVERLAP). Since we operate in a low-resource scenario, to maximise the usage of our available data, when evaluating on any of the three testing series we generate a new training and validation dataset which uses the remaining testing sets for training; see Table 3.12, rows 1-8 for quantitative details).

OPENSUBTITLES CORPUS For pre-training, we re-use the version of OPENSUBTITLES corpus described in Vincent, Flynn & Scarton (2023). The dataset comprises sentence pairs annotated with six production metadata (analogous to e.g. ZOO-MULTI) and document-level data. Data quantities are listed in Table 3.12, row 10.

3.4.2 Evaluation

We evaluate how well adapted to context individual version of the text are by finding out the degree of co-occurrence between individual sentences or translations and their specific *extra-textual* context. We express results as the PMI (§A.3) between the context \mathcal{C} and the target utterance or hypothesis \mathcal{H} , which is computed as:

$$\text{PMI}(\mathcal{C}, \mathcal{H}) = \log \frac{p(\mathcal{H} | \mathcal{C})}{p(\mathcal{H})} \quad (3.1)$$

$$= \log p(\mathcal{H} | \mathcal{C}) - \log p(\mathcal{H}) \quad (3.2)$$

PMI rewards positively those tokens which occur more frequently in the context \mathcal{C} than in the general distribution termed with the prior $p(\mathcal{H})$. In practice, both terms of Equation 3.2 are computed with language models: $\log p(\mathcal{H} | \mathcal{C})$ with LMCUE ($\mathcal{S} + \mathcal{P}$) and $\log p(\mathcal{H})$ with BASE-LM. Both LMs are pre-trained according to the strategy described in §3.3.3.2 and fine-tuned on the context and target-side dialogue from ZOO-MULTI corpus (OVERLAP setting). We train a separate tandem of language models for each language pair, and for statistical significance we train five distinct instances of each model (each with a different random seed).

3.4.3 Machine Translation Systems

We use the MTCUE architecture (§2.5, Vincent, Flynn & Scarton 2023) for the experiments, leveraging all available context listed in §3.4.1. We scale the original model up from 106M to 386M parameters by switching the underlying encoder-decoder model from Transformer *base* to *big* (c.f. Vaswani et al. 2017). We also implement a corresponding parameter-matched baseline (BASE-NMT), which is a standard encoder-decoder Transformer with double the number of encoder and decoder layers of MTCUE. We pre-train the systems on OPENSUBTITLES with metadata and document-level information (as per the original paper) and fine-tune them on the ZOO-MULTI corpus for each language pair separately.

Both models are trained in a pipeline process described in three steps:

1. The source encoder and decoder (standard NMT architecture) are pre-trained on the translation objective on OPENSUBTITLES (§2.5.3).
2. MTCUE with pre-loaded weights from step 1. is fine-tuned on OPENSUBTITLES with document-level information and production metadata²⁰;

²⁰ In preliminary experiments we tested a range of pre-training settings, including the strategy of pre-training on document data alone outlined in §3.6, but found that pre-training on both document information (with position information) and production metadata consistently yielded the best performance in this set of experiments.

3. MTCUE is further fine-tuned on ZOO-MULTI (§3.4.1).

BASE-NMT is trained in the same three steps, albeit without the contextual information. For hyperparameter search, we adapt the learning rate (η) and batch size (β) values for pre-training (OPENSUBTITLES) and fine-tuning (ZOO-MULTI) models separately, both times using grid search (§A.1.1), choosing from six values from a range between 0.00005 and 0.002 for (η) and five values from a range between 20,000 and 400,000 tokens for β . We find that for pre-training, the best-performing values are an η of 0.0005 and a β of 80,000. For fine-tuning, we trained 4×5 copies of models for each test set and data strategy, for 4 values of $\alpha \in \{1e-4, \dots, 1e-3\}$ and 5 values of $\beta \in \{4,000, \dots, 40,000\}$. For each data and test setting, we select the model with the best validation BLEU (§A.2.3.2).

For this set of experiments, personalised LMs implemented via LMCUE are fine-tuned on the ZOO-MULTI corpus, each time on the given target-side text and speaker and production metadata only (the same set of context values used to train the models examined in §3.3). In other words, while we train the contextual models on a larger set of context variables, we only test context specificity to the speaker and production context. The hyperparameter search for pre-training and fine-tuning these language models is conducted following the procedure described in §3.6 and §3.3.3.3.

3.4.4 Results

A positive value of PMI for REFERENCE (mean score of 0.073; Table 3.13) suggests the presence of a co-occurrence effect between professional translations of the test set and their extra-textual context. We calculated that the ORIGINAL text in English obtained a score of 0.087. The two values are not directly comparable, but they are of similar magnitude which suggests that the context-specific traits of the original text are well preserved in the professional translations for this test set.

The non-contextual machine translation model, BASE-NMT, achieves significantly lower PMI on average (Table 3.13, 0.026 for DISJOINT, 0.028 for OVERLAP): the absence of context at generation time results in translations less adapted to the specific characters and productions. The nevertheless positive values can be explained by the presence of domain-specific terms such as “pan” which do not need context to be translated correctly but will occur more often in specific contexts (e.g. cooking shows), yielding a positive PMI. The contextual MTCUE achieves higher PMI on average than BASE-NMT (+0.015 for DISJOINT, +0.023 for OVERLAP), meaning that using the relevant context does make the hypotheses more personalised, and the greater improvement for OVERLAP suggests that context can be utilised even better when previous samples from the same speakers or series are given. Interestingly, while personalisation is stronger in OVERLAP systems than DISJOINT, the MTCUE (DISJOINT) still performs better than either BASE-NMT system, signifying the robustness of CUE vectors to this zero-shot adaptation to new series and speakers. Among all three target languages, reference translations to

	EN-FR	EN-DE	EN-PL	Average
REFERENCE	0.101	0.037	0.081	0.073
DISJOINT				
BASE-NMT	0.042	-0.004	0.041	0.026
MTCUE	0.066	0.007	0.049	0.041
OVERLAP				
BASE-NMT	0.040	-0.006	0.049	0.028
MTCUE	0.069	0.012	0.072	0.051

Table 3.13: PMI computed with general and personalised language models on translations. Results computed from five different runs (instances of the language models). Highlighted results are statistically significantly better than those from other systems.

	EN-FR		EN-DE		EN-PL	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
DISJOINT						
BASE-NMT	34.89	23.69	35.71	28.59	31.13	31.50
MTCUE	35.73	25.81	36.22	29.29	31.62	<u>32.66</u>
OVERLAP						
BASE-NMT	35.06	23.75	36.15	29.75	31.59	32.83
MTCUE	36.14	27.04	36.90	30.53	32.18	31.95

Table 3.14: BLEU and COMET scores for the evaluated MT systems. COMET score was computed using the wmt20-comet-da model. Highlighted results are the best in the column, and all underlined results are statistically indistinguishable from them ($p = 0.05$).

German are correlated with context the least ($PMI = 0.037$) and translations to French the most ($PMI = 0.101$). The MT systems' results follow a similar trend.

For completeness, we provide the BLEU and COMET scores, comparing the MT to the human references (Table 3.14): in both data settings, MTCUE matches the references to a significantly higher extent than the baseline.

ANALYSIS OF EXAMPLES In this section, we delve into a detailed analysis of examples where the hypotheses of the contextual MT model (MTCUE) were assigned a stronger

PMI score compared to the hypotheses from BASE-NMT, quantified with the **diff** score defined below:

$$\mathbf{diff} = \text{PMI}_{\text{MTCUE}(\text{Source})} - \text{PMI}_{\text{BASE-NMT}(\text{Source})}$$

All the examples presented herein are derived from systems trained in the OVERLAP data setting. We have deliberately chosen examples that admit multiple translations, depending on the context they occur in. Across the seven examples, the overarching context can be summarised as *A British family cooking competition show*. Each of our chosen cases in some way relies on this context.

Example 1	diff = 0.26
Source	We're okay, we're doing just fine, just...
Reference	Wszystko jest w porządku. (<i>'Everything is alright.'</i>)
BASE-NMT (✗)	Nic nam nie jest. (<i>'We are fine.'</i>)
MTCUE (✓)	Radzimy sobie. (<i>'We're coping.'</i>)

Our first example exhibits a behavioural agreement adaptation. During cooking show walkthroughs, there are often discussions about managing stress and working under pressure. This discourse is sincere and allows the contestants to share about their struggles. In **Example 1**, the non-contextual translation *Nic nam nie jest* would be far less likely to occur in this walkthrough setting than *Radzimy sobie*. The former phrase is more suited for immediate impact situations (e.g. falls) rather than stressful ongoing situations like a competition show. It also carries a more defensive tone which makes it ill-fitted for a lighthearted family series.

Example 2	diff = 0.20
Source	Well deserved.
Reference	Zasłużone zwycięstwo. (<i>'Well deserved victory.'</i>)
BASE-NMT (✗)	Zasłużyłeś. (<i>'You_{masculine} deserve this.'</i>)
MTCUE (✓)	Zasłużyliście. (<i>'You_{plural} deserve this.'</i>)

In [Example 2](#), MTCUE discerns the number of addressees correctly: in this family cooking competition show, congratulations are typically extended to a family, i.e. a group of people. Our internal case-by-case analysis continually found that the personalised exhibits an inclination towards the correct gender and plurality.

Examples [3](#) and [4](#) illustrate two cases where significantly stronger PMI scores are assigned to hypotheses that align better with the context, even though they may not be the correct translations.

Example 3	diff = 1.46
Source	Try the balls.
Reference	Spróbuj kulkę. (‘Try a ball.’)
BASE-NMT (✗)	Spróbuj piłeczek. (‘Try the <i>footballs</i> .’)
MTCUE (✗)	Spróbuj jajek. (‘Try the <i>eggs</i> .’)

In [Example 3](#), MTCUE translates *balls* as *eggs* and receives a PMI score higher by 1.46 points compared to BASE-NMT, which generated *footballs*). While neither hypothesis is entirely accurate (the original *balls* likely referred to meatballs or dough balls), the term *eggs* is highly specific to the context of cooking, leading to a higher rating from the personalised LM.

Example 4	diff = 0.35
Source	Spice girls.
Reference	Spice Girls.
BASE-NMT (✓)	Spice girls.
MTCUE (✗)	Dziewczyny z przyprawami. (‘Girls with spices.’)

In [Example 4](#), MTCUE opts for a literal translation *girls with spices* for the band *Spice Girls*. This choice is deemed more contextually appropriate, likely due to the use of *przyprawy* (en. *spices*), which is very specific to cooking. Both examples show that our method must be used in tandem with a standard translation quality metric such as BLEU or COMET, as our approach focuses on monolingual evaluation and does not prioritise source sentence faithfulness.

[Example 5](#) highlights a scenario where the correct translation of the ambiguous verb *make* is generated by the contextual model, interpreting it as *prepare*. This translation is

Example 5		diff = 1.19
Source	I am now making my...	
Reference	Przygotuje... (<i>'I will make [= prepare] my...'</i>)	
BASE-NMT (✗)	Teraz robię... (<i>'I am now making [=do]...'</i>)	
MTCUE (✓)	Teraz przygotowuję... (<i>'I am now making [= prepare] my...'</i>)	

awarded a significantly higher PMI score than the incorrect translation (*do*) generated by the baseline model.

Example 6		diff = 1.45
Source	Can somebody get a pan of simmering water on, please?	
Reference	Czy ktoś może zagotować wodę? (<i>'Can somebody boil the water please?'</i>)	
BASE-NMT (✗)	Może ktoś nałożyć wodę na patelnię? (<i>'Could someone put some water on the pan?'</i>)	
MTCUE (✗)	Podajcie patelnię. (<i>'Pass me the pan.'</i>)	

Example 7		diff = 0.62
Source	What is this, cake dough?	
Reference	To ciasto? (<i>'Is this cake dough?'</i>)	
BASE-NMT (✓)	Co to? Ciasto? (<i>'What's this? Cake dough?'</i>)	
MTCUE (✗)	Ciasto? (<i>'Cake dough?'</i>)	

In both [Example 6](#) and [Example 7](#), the translation of MTCUE is rewarded more positively than that of BASE-NMT even though neither fits the context more than the other. On top of that, MTCUE's hypotheses are inferior translations of the source sentence: in [Example 6](#) the translation of MTCUE completely changes the meaning of the source sentence, while in [Example 7](#) it omits the translation of *What's this?*. These examples spotlight potential challenges in our evaluation model, particularly

in low-resource scenarios. The limited quantity of samples in our fine-tuning corpus (approximately 50 to 100K depending on the target language) may have contributed to the model being more prone to learning spurious correlations due to an insufficient population of tokens. If a common token or phrase occurs disproportionately often in certain contexts, it may be considered more context-specific during evaluation (and skew the PMI score towards a positive co-occurrence factor), even if this context specificity is not reflective of real-world language use. Conversely, tokens or phrases occurring disproportionately infrequently in certain contexts may unfairly lower the score. Effectively addressing this challenge requires the collection and augmentation of context-annotated data. However, this process must be conducted thoughtfully, striving to inclusively represent diverse social groups and actively avoiding the emergence of harmful correlations.

3.5 COST-BENEFIT ANALYSIS OF HUMAN ANNOTATIONS

Granular manual annotations are costly to obtain. Cost-benefit analysis helps avoid the misallocation of limited annotation funding and resources. This section presents the results of the cost-benefit analysis we conducted to show which individual speaker attributes produce the most **benefit** (reduction in perplexity) w.r.t. the perceived **cost** of producing them.

We asked the two human annotators to assess the effort required for the annotation task using three metrics on a Likert scale of 1 to 10: *access* (how difficult it was to find information), *credibility* (how confident they were in the accuracy/usefulness of the information), and *time* (how much time was needed relative to other fields). We took the mean of both annotators' scores after reversing credibility ($C = 10 - C + 1$). We then conducted a simple experiment to measure the **benefit** of each metadata type by fine-tuning the pre-trained LMCUE on each speaker metadata type evaluated individually. Finally, we measured the reduction in perplexity from including this information (Figure 3.5) compared to the 91M parameter decoder in LMCUE, since that is the decoder we are trying to improve with context.

The figure suggests that *description*, *profession* and *quote* yield the greatest perplexity reduction in both datasets, around 5 to 6%. *Description*, the best-performing attribute, alone achieves 88.7/91.9% of the perplexity reduction of LMCUE (S). On the other hand, *age bracket*, *religion* and *country of origin* yield the smallest improvements, and a better improvement can be achieved with the parameter-matched BASE-LM. For CORNELL-RICH, they still help marginally (1 to 2%), while for ZOO improvements from *age* and *religion* are negligible. This analysis suggests why King & Cook (2020), who implemented context-based adaptation using only *age* and *gender*, found it inferior to other methods; we found other variables such as *description* to be significantly more useful.

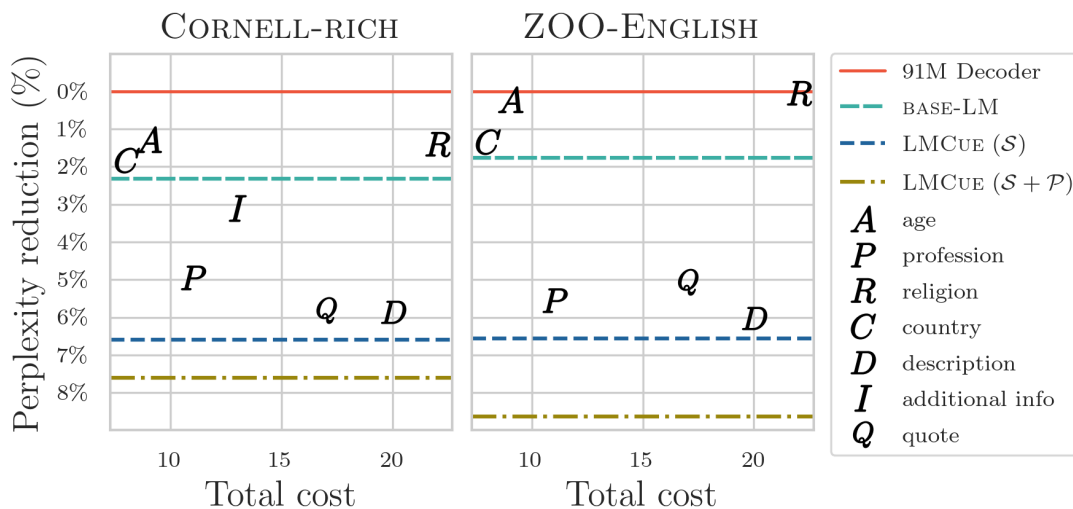


Figure 3.5: Perplexity reduction from training LMCUE with individual speaker attributes.

Other findings of the analysis are consistent among the two corpora. *Profession*, collected at a relatively small cost, is one of the top-3 attributes for both, hence the most cost-effective. *Religion* is the least cost-effective attribute, requiring the most effort but producing the least benefit. Finally, experimental attributes such as characteristic *quotes* and *additional information*²¹ have been shown to be useful, the latter placing in the middle of the ranking whilst the former is on par with the best attribute for CORNELL-RICH.

3.6 PRE-TRAINING STRATEGY: PAST DIALOGUE AS PROXY FOR METADATA

This supplementary section presents empirical evidence that past dialogue can be used as a proxy for fine-tuning LMCUE on speaker or production metadata. When fine-tuning, we use both *Speaker* and *Production* metadata. We report performance on `test_unseen` to also present sMRR scores.

As per Table 3.15, pre-training on OPENSUBTITLES leads to best perplexity when no context is used (BASE-LM), however using context yields improvements in sMRR, and these are stronger when metadata is used instead of dialogue. Similarly, without pre-training we also obtain the best perplexity with BASE-LM; here even sMRR remains at a baseline level, i.e. the contextual model fails to learn contextual dependencies correctly. Metadata only leads to superior results when both pre-training and fine-tuning are included in the pipeline. Interestingly, pre-training on dialogue yielded the

²¹ Since *additional information* was not collected for ZOO, it is not present on the ZOO plot of Figure 3.5.

	Pre-training ✗ / ✓	Context	Fine-tune	PPL↓	sMRR↑
BASE-LM	✓	–	✗	28.78	0.03
LMCUE	✓	dialogue	✗	37.19	0.29
LMCUE	✓	metadata	✗	30.95	0.43
BASE-LM	✗	–	✓	39.60	0.03
LMCUE	✗	–	✓	51.14	0.03
BASE-LM	✓	–	✓	23.62	0.03
LMCUE	✓	dialogue	✓	22.31	0.96
LMCUE	✓	metadata	✓	22.71	0.89

Table 3.15: Results on `test_unseen` of CORNELL-RICH from different pre-training/fine-tuning setups. New results (top 5 rows) come from single runs.

best results, though pre-training on metadata is not far behind (+0.4 PPL, -0.07 sMRR). We hypothesise that since past dialogue is much more diverse than film metadata (which contains many repeated fields), it is overall the better pre-training proxy for fine-tuning on new types of metadata, such as speaker profiles. For applications on other datasets, we therefore recommend pre-training on a similar dataset (domain-wise) with access to document-level information.

3.7 CONCLUSIONS

We have argued for context-based personalisation of language models by training a conditional generation architecture on dialogue accompanied by rich contextual annotations. Our approach performs on par with expensive speaker-specific fine-tuning methods. We have also explored using such models to evaluate the context-specificity of professional and machine translations, providing insight into how well the generated translations are specific to the extra-textual context, without direct comparison to the human references. Finally, we have contributed CORNELL-RICH, a set of rich speaker and production annotations for a publicly available dialogue dataset, which can be used to reproduce our results in personalised language modelling. Below we summarise the findings specific to each research question.

RQA How can rich character profiles be used to model the characters’ speaking styles? (§3.3.4.1)

The LMCUE architecture can be trained to exhibit personalisation based on Speaker context: LMCUE (\mathcal{S}) reduces perplexity by 4.3/4.7% compared to a parameter-matched general LM, or by 5.4/6.5% when \mathcal{P} roduction data is also used. In a few-shot scenario, LMCUE ($\mathcal{S} + \mathcal{P}$) is better than linear interpolation LERP and comparable with speaker-specific fine-tuning (SPFINE-TUNE). Since LMCUE requires no fine-tuning to specific speakers, it is favourable if metadata is available.

RQB How can a LM be personalised for a specific character solely by learning from data for characters with similar profiles? (§3.3.4.2)

On a test set with unseen speakers, context-based personalisation yields a high speaker separation effect (models assign the highest probability to dialogue which matches the given speaker’s profile). Using both \mathcal{S} and \mathcal{P} metadata on this test set reduces perplexity by 5.6/4.4%, a similar magnitude to that for seen speakers (5.4/6.5%), suggesting that our method scales robustly to this scenario, unlike speaker-specific fine-tuning which cannot be applied on new characters. Having a varied pool of speakers and productions in training data correlates positively with sMRR.

RQC Can MT offer personalisation benefits proportional to professional translations? (§3.4.4)

Utilising speaker and metadata annotations in MT makes the language used in hypotheses more context-specific, as measured by the PMI score between such context and the generated text, when compared to a context-agnostic system. However, this context specificity is still stronger in gold standard (professional) translations. Our findings suggest that contextual language models could be paired with automatic metrics for a more well-rounded evaluation of machine translation as they bring the aspect of the translations fitting the specific extra-textual context.

RQD Which character metadata are the most cost-effective for personalisation? (§3.5)

Textual metadata (descriptions, quotes, professions) is significantly more useful for personalisation with LMCUE than discrete metadata (e.g. age bracket). Particularly in our evaluation, descriptions alone achieve results on par with using the entire speaker profile. Furthermore, the utility of individual attributes seems to be positively correlated with the diversity in their representation.

ASSESSING CONTEXTUAL MACHINE TRANSLATION IN A PROFESSIONAL SCENARIO OF SUBTITLING

4.1 INTRODUCTION

Interlingual subtitling of videos involves a two-step process. Initially, the video is transcribed in its native language, and the transcribed text is then transformed into concise subtitle blocks that match the video’s timing. These subtitles are then manually translated into the desired language, while staying within the subtitle constraints, such as reading speed, maximum number of lines and characters in a line, length proportions of the top and bottom lines, and additional considerations such as distinguishing when dialogue from two or more speakers is displayed. As a heavily involved process with multiple guidelines, the task necessitates manual quality checks after each major step.

The challenges involved in the above process individually fall within what the field of speech and language technologies commonly aims to tackle. Speech recognition research focuses on automatic transcription of text, while machine translation research considers the problem of automatic translation. Subtitle segmentation is sometimes conceived of as an individual task (e.g. [Ponce et al. 2023](#)), though predominantly it is incorporated into the guidelines of the main task. For example, [Papi et al. \(2022\)](#) unify speech transcription and segmentation of the resulting text.

Within this work, we replace the step of translating the native subtitles to the desired languages from scratch with post-editing machine translations of the source text. Such a formulation is far from new: machine translation has consistently demonstrated its potential to reduce effort in the subtitling domain, to a varying degree (e.g. [C. M. de Sousa et al. 2011](#), [Huang & Wang 2023](#)). Nevertheless, these studies have often relied on off-the-shelf general-purpose [NMT](#) engines such as Google Translate¹. Our work challenges this setup: using [GOOGLE](#) as one of our baselines, we present that by just tailoring an [NMT](#) engine specifically to the target domain we can not only significantly enhance the accuracy of translation hypotheses (as measured by automatic metrics), but also significantly reduce the human effort required to post-edit them. On top of that, we compare two such in-domain systems: a [context-agnostic](#) one and one which leverages a wide range of [context](#) information.

Our prior investigations outlined in [Chapter 2](#) and [Chapter 3](#) have illustrated the advantages of incorporating [extra-textual](#) information in both [NMT](#) and language modelling contexts. In this Chapter, we show the impact such incorporation of context

¹ <https://translate.google.com/>

has on post-editing effort, compared to a non-contextual domain-specific translation system and a general-purpose commercial system like Google Translate.

We perform a thorough evaluation with the assistance of professionals with expertise in translation, post-editing and quality checking. Hereinafter we refer as PEs to those who were tasked with post-editing work, and as translators to those who were tasked with translation from scratch. The campaign takes place in a full-context multi-modal environment where the professionals have access to the video material and are able to directly jump to the segment corresponding to the utterance they are reviewing, as well as see the preceding and succeeding segments. We employ a total of eight post-editors (PEs), four for English-to-German and four for English-to-French translation, and four translators, two per language pair. In our evaluation, we measure both the effort it takes to post-edit or translate the TV series content, as well as take note of specific translation errors observed by the post-editors. We find that the contextual MTCUE makes consistently fewer errors related to context and style in EN-FR translation, while performing on par with its non-contextual counterpart in general translation quality. Furthermore, both of these domain-adapted systems make fewer total errors than Google Translate and their outputs are easier to post-edit. Finally, between MTCUE and BASE-NMT, our experiment did not find that either system's outputs require significantly less effort in post-editing than the other. However, our survey among the professional post-editors revealed that errors related to style and context in particular often necessitate complete rewrites of machine translation (MT) outputs over corrections. This finding motivates future research within contextual NMT.

The rest of the Chapter is structured as follows. §4.2 presents the work related to this subject. §4.3 describes the experimental setup. In §4.4 and §4.5 we report on the results of the automatic and human evaluation respectively. Finally, §4.6 concludes the Chapter.²

4.2 RELATED WORK

Over the last few years the problem of automatic translation of video subtitles has been given a volume of attention. Among many others, C. M. de Sousa et al. (2011), Koponen et al. (2020) and Huang & Wang (2023) observe that post-editing the outputs of an NMT system is a promising alternative to translation *ex novo*. Such an approach can reduce the temporal, technical and cognitive effort of both novice and professional translators

² The present Chapter is to be submitted as a conference paper to the Annual Conference of the European Association for Machine Translation (EAMT 2024). It is going to list five authors, of which the first is the author of this thesis, the last is the PhD supervisor and the remaining three are employees at ZOO Digital Group PLC, the industrial partner to this thesis. Chris Bayliss was in charge of development work within the ZOOSUBS system, enabling the human evaluation campaign. Charlotte Blundell acted as the project manager on the company side, facilitating communication with the production team performing the evaluation work. Both Chris and Charlotte took part in team meeting during which consultations regarding the work setup took place. Chris Oakley was the industry-side supervisor of the project.

and subtitlers. Moreover, as highlighted by the findings of a survey among professional subtitlers detailed by [Karakanta et al. \(2022\)](#), professionals generally have a positive outlook on including automatic components (such as speech recognition, translation, and subtitle segmentation) into their workflow, reporting that these components serve as starting templates, reduce effort and sometimes provide useful suggestions.

However, automatic translation presents a range of challenges that remain unsolved. [Gupta et al. \(2019\)](#) list a set of issues encountered with this specific problem in their practical setting. The most common errors include: (i) cases where machine-translated text disregards the subtitle block limitations, and a shorter and often paraphrased translation is required, (ii) contextual errors, where words used are lexically correct but morphologically inconsistent with the surrounding text or video material, and (iii) lexical inconsistency errors, where the employed vocabulary does not comply with the standard language or industry usage, or is inconsistent with the video material or surrounding text. The surveyed subtitlers in [Karakanta et al.](#) note two main issues: lexical errors, including the translation of idioms and figurative language (“automatic translation still tends to be a bit too literal”), and context issues, such as inconsistent translation of the same term across multiple segments. Context issues have also been pointed out as the culprit in automatic translation of text in many works that leveraged the OpenSubtitles corpus ([Lison et al. 2018](#)), a dataset of user-submitted subtitles and their translations. Specifically, recent work highlights that many translation errors found in this domain are related to the use of context, which includes document-level information ([Tiedemann & Scherrer 2017](#), [Bawden et al. 2018b](#)), extra-textual information contained implicitly in the text such as the speaker’s gender identity ([Vincent et al. 2022b](#)) and explicit extra-textual information ([Vincent, Flynn & Scarton 2023](#)). The contextualised translation of our selected model enables improvements in both of these areas.

When considering the task of post-editing machine-translated subtitles, it appears that the environment setup plays an important role: [Huang & Wang \(2023\)](#) show that post-editing in a multi-modal scenario decreases the cognitive load of student translators compared to a mono-modal (text-only) scenario, and argue that this could be explained by the dual coding theory, according to which the interactions between the verbal and non-verbal information enhances the translators’ understanding of the material. Within our multi-modal human evaluation study, we measure the technical and temporal effort. While we do not directly measure cognitive effort (due to lack of appropriate measuring equipment), we conduct a survey among the PEs and translators and report their perception of the study in §4.5.2.1.

Finally, the experimental work presented in this Chapter employs the MTCUE architecture introduced in §2.5 and [Vincent, Flynn & Scarton \(2023\)](#), trained according to the regimen and on the datasets described in Chapter 3. MTCUE is a multi-encoder Transformer designed for contextual NMT; in [Vincent, Flynn & Scarton](#) we show that it is capable of leveraging contextual signals such as film metadata and document-level

information to improve translation quality, as well as enable better control of fundamental extra-textual phenomena in translation, such as speaker’s gender and formality register.

4.3 EXPERIMENTAL SETUP

In this section, we outline the experimental setup for our human evaluation experiment. Our primary objective is to compare the translation quality of various machine translation models, both in terms of similarity to the reference text, as well as in terms of the number of errors and the post-editing effort required to achieve sufficient quality, measured in a human evaluation campaign. Specifically, we employ these models to generate translations for two language pairs: EN-DE and EN-FR. As our test set, we use the three TV series: BIGFAM, WORLDJEFF and RIGHTSTUFF, as described in §3.4.1.

4.3.1 Examined System and Baselines

The objective of the experiment is to compare the contextual MTCUE system (§2.5, Vincent, Flynn & Scarton 2023), trained to translate dialogue with regards to the extra-textual context it arises in, to non-contextual machine translation. The MTCUE instance used in this experiment is the one already described in detail in §3.4.3 within the previous chapter (Chapter 3). We compare it to three baselines:

1. GOOGLE³, a readily available general purpose NMT engine used extensively in prior work on automated translation of subtitles.
2. BASE-NMT §3.4.3, a non-contextual baseline translation model matched to MTCUE in terms of the total number of parameters. Similarly to MTCUE, we re-use the model described in §3.4.3.
3. REF the production-approved human translations of the test set. This baseline is omitted during automatic evaluation (in fact, it is used as the reference text to calculate the translation metrics), but is used as a baseline in the human evaluation, where the professionals are asked to post-edit this text (unaware that it is of already sufficient quality).

Additionally, for the BASE-NMT and the MTCUE models, two instances are trained of each, in each of the distinct data settings DISJOINT and OVERLAP (§3.4.1).

4.3.2 Evaluation

We conduct automatic and human evaluation. For automatic evaluation, we use BLEU (§A.2.3.2) and COMET (§A.2.3.2) as metrics and compare the outputs of the machine

³ <https://translate.google.com/>

translation systems (BASE-NMT, GOOGLE, MTCUE) against the reference (REF). The human evaluation campaign is conducted in a professional environment, in collaboration with ZOO Digital. The objective of this evaluation is to highlight whether MTCUE indeed makes fewer context-related errors, and additionally to shed some light on whether such a contextual MT model may help decrease the effort required to post-edit subtitles within this domain.

The task is implemented and performed by professional translators and PEs using ZOOSUBS, an in-house software belonging to ZOO Digital, built to facilitate manual translation of video material (Figure 4.1). This specialised software offers a user interface that displays the video material along with its associated subtitles in the original language. Additionally, it provides a set of windows where the translator can input translations in the desired target language (Figure 4.1a). When a specific workflow (such as the one employed in this study) involves pre-translated text, the boxes are initially populated with draft translations, which the post-editors may edit, divide or combine as they see fit.

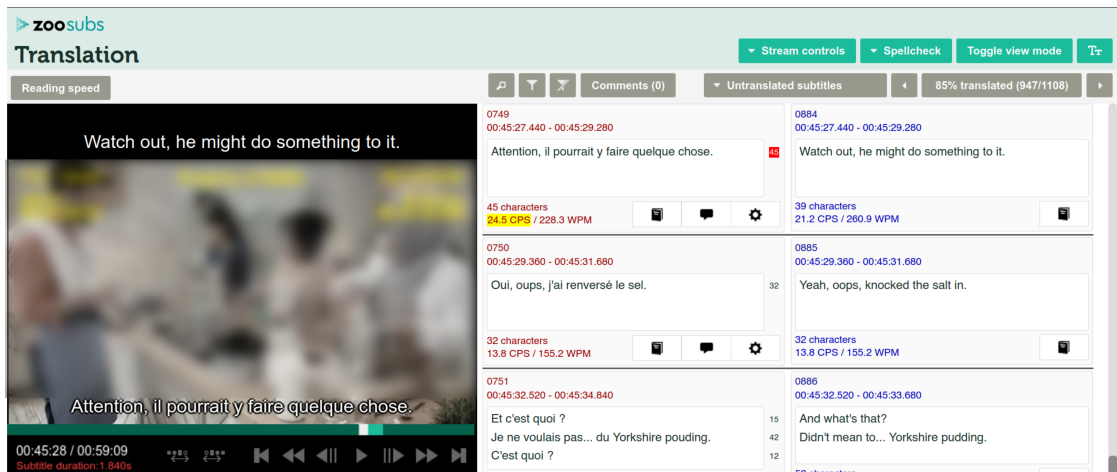
When post-editing the subtitles, the PE can click on any of the target-language text boxes to edit the text within it. The system automatically tracks the time a worker spends editing the given box, as well as the number of keystrokes made. These metrics are recorded for each window separately and are taken into account only if actual changes were made to the text. Once changes are made, the worker is prompted to enter the reason for making a change, choosing from a pre-existing list of errors or optionally providing their own custom description (Figure 4.1b). Multiple errors can be marked at once.

In our campaign, we leverage this functionality to measure the total and average time and number of keystrokes made by (a) translators, (b) PEs given some pre-existing translations. We also measure the total number of boxes edited. Finally, for the purpose of this project we create a custom list of errors that the PEs are prompted to select from.

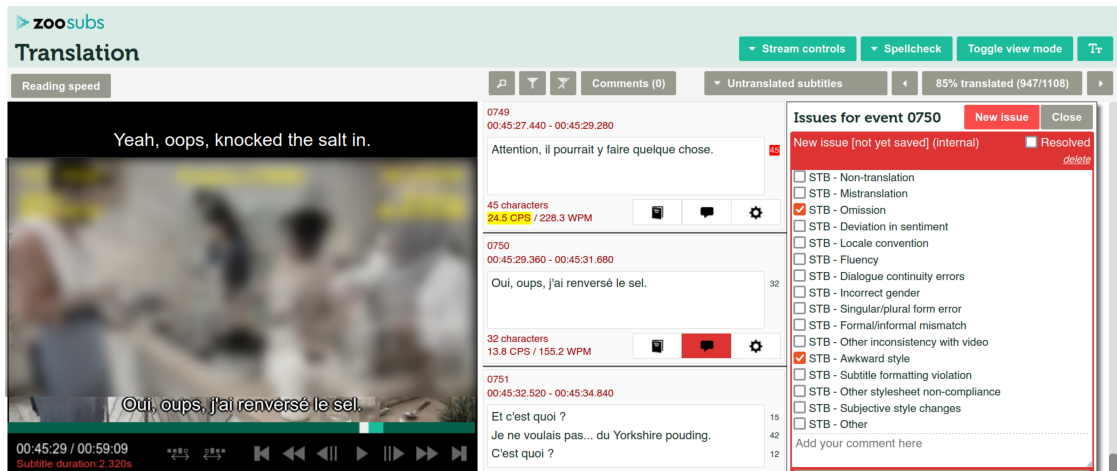
WORKER SETUP The PEs operate on **seven episodes** total (half the episodes of every show featured in §3.4.1), comprising 54% of the original test set’s segments. They are unaware that some of the text they work with is machine translated, but are told that it is for a research project and asked to relax some constraints such as adhering to the reading speed constraints. In addition, we ask four translators (two per language pair) to translate one 60-minute episode of BIGFAM from scratch and record their effort.

For each of the seven episodes, the PEs are asked to post-edit one out of four versions of the text:

1. GOOGLE: a baseline machine translation from Google Translate.
2. BASE-NMT: a baseline machine translation from BASE-NMT (OVERLAP).
3. MTCUE: translation hypotheses from MTCUE (OVERLAP).



(a) A snapshot of the ZOOSUBS system. The original subtitles (far-right column) are translated to their counterparts in the target language. These translated subtitles are displayed immediately in the video on the right.



(b) A snapshot of the ZOOSUBS system when the error selection window is prompted.

Figure 4.1: Snapshots of the ZOOSUBS system in action.

4. REF: the reference human-written and production-ready translations (to account for the fact that PEs can sometimes post-edit a translation even when the original one is valid).

Our setup ensures that the same PE evaluates the output for each episode exactly once (i.e. does not see two different versions of the same text) (Table 4.1). When referring to

individual PEs, we use the notation PE.[L][i], where L ∈ {G (German), F (French)}, and i denotes the PE ID ∈ [1, 4].

Title	BigFam		RightStuff		WorldJeff		
Episode no	9th	11th	5th	8th	8th	5th	10th
PE.1	REF	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE
PE.2	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE
PE.3	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT	REF
PE.4	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT
Translator 1	<i>From Scratch</i>						
Translator 2	<i>From Scratch</i>						

Table 4.1: Work assignment to PEs and translators in the human evaluation campaign. Within both language pairs the work assignment is the same.

DETAILS REGARDING THE PEs The recruited PEs and translators are professionals within the subtitle domain and freelance employees of ZOO DIGITAL. All recruited workers were informed that the undertaken work is carried out for a research project, but nevertheless they were paid for their effort at competitive PE and translator rates, standard within the company for this type of work. All work conducted for this human evaluation campaign was led and managed by a project manager employed by ZOO Digital. This occurred while the author of this thesis was involved in an internship with the company (as part of their sponsorship of the PhD), gaining access to the ZOOSUBS system and on-site data.

The ZOO project manager also had information about the PEs and translators background and, as part of this work, they also answered a short survey about their views regarding machine translation:

1. Basic information

- a) Years of experience (YOE) as a translator.
- b) YOE in the domain of subtitle translation.
- c) YOE in post-editing.
- d) Did your professional training as a translator comprise training in post-editing specifically?

2. Views on machine translation

- a) Which one would you prefer: translating a stream from scratch or doing a quality check (post-editing) a stream? Why?

- b) What are your views on the use of machine translation in the industry?
- c) In your opinion, are there any benefits to post-editing translations rather than translating from scratch?

We report the *Basic information* in Table 4.2, while devoting a later section (§4.5.2.1) to discuss the PEs views on machine translation. All French translators have training in post-editing, and three out of four prefer it to translating from scratch, while no German translators have received such training in the past, and all but one strictly prefer translation from scratch. All PEs have at least one YOE in post-editing and one and a half in the subtitle domain. Although the translators within both pairs have a similar amount of experience in translation in general and in the subtitle domain (11.5 ± 6.5 for French vs 12.5 ± 5.0 for German), the French translators have the advantage in terms of YOE in both subtitling (a mean difference of 2.1 YOE) and post-editing (a mean difference of 3.3 YOE).

	English-to-French				English-to-German			
	PE.F1	PE.F2	PE.F3	PE.F4	PE.G1	PE.G2	PE.G3	PE.G4
Translation YOE	15	8	3	20	7	18	8	17
YOE in subtitles	8	6	1.5	20	7	5	8	7
YOE in post-editing	8	6	3	10	5	5	1	3
Post-editing training?	✓	✓	✓	✓	✗	✗	✗	✗
Prefer post-editing?	✓	✓	✗	✓	✓/✗	✗	✗	✗

Table 4.2: Details regarding PEs who took part in the campaign.

ERROR TAXONOMY Upon correcting an individual translation, the PE is prompted to select a reason for the correction from a fixed list of possible errors. For this project, we customise the list to involve not only standard translation errors such as *Mistranslation* or *Omission* but also focus on the context-specific errors (e.g. *formality mismatch*) as well as task-specific errors (*subtitle formatting violation*). To build the taxonomy, we first compiled a list of candidate errors from three sources:

- general machine translation errors reported in previous work (Freitag et al. 2021, Sharou & Specia 2022),
- the original list of issues already present in the ZOOSUBS system,
- errors deemed relevant based on previous work on post-editing machine-translated subtitles (§4.2).

We then narrowed down the list of candidates to errors relevant to the study, and removed duplicates. At that point the taxonomy was uploaded to the system and the thesis' author undertook a test evaluation against a stream with 446 segments to validate the reliability of the list. As a result, some errors were split into more granular categories, some were renamed and some generalised. Table 4.3 presents the final list of errors compiled.

Type	Description
Translation quality	
<i>Catastrophic translation</i>	Would be impossible to post-edit; the text must be translated from scratch.
<i>Mistranslation</i>	Incorrect and does not preserve the meaning or function of the source text.
<i>Omission</i>	Part of the source text was left untranslated.
<i>Deviation in sentiment</i>	Does not preserve the sentiment of the source (e.g. does not match the expressed excitement), or negates the sentiment (e.g. from positive to negative).
<i>Locale convention</i>	Violates locale convention, such as currency and date format.
<i>Fluency</i>	Contains punctuation, spelling and grammar errors.
Context	
<i>Incorrect gender</i>	Misgenders the speaker or the addressed person(s).
<i>Incorrect plurality</i>	Incorrectly refers to a single person when a group is addressed, or vice versa.
<i>Wrong formality</i>	Expressed in informal style or uses informal addressing when should use formal, or vice versa.
<i>Other inconsistency with video</i>	Contains inconsistencies with the video material not falling within any of the above.
Style	
<i>Subtitle formatting violation</i>	Violation of the subtitle blocking guidelines.
<i>Other style sheet non-compliance</i>	Does not conform to the provided style sheet.
<i>Awkward style</i>	The style of the translation does not reflect the style of the source sentence and/or the context.
<i>Subjective style changes</i>	The translation is acceptable but the editor suggests improvements in style.
Other	All other error types; the evaluators are invited to describe the errors in a text box provided.

Table 4.3: List of errors provided to the human evaluators during the campaign.

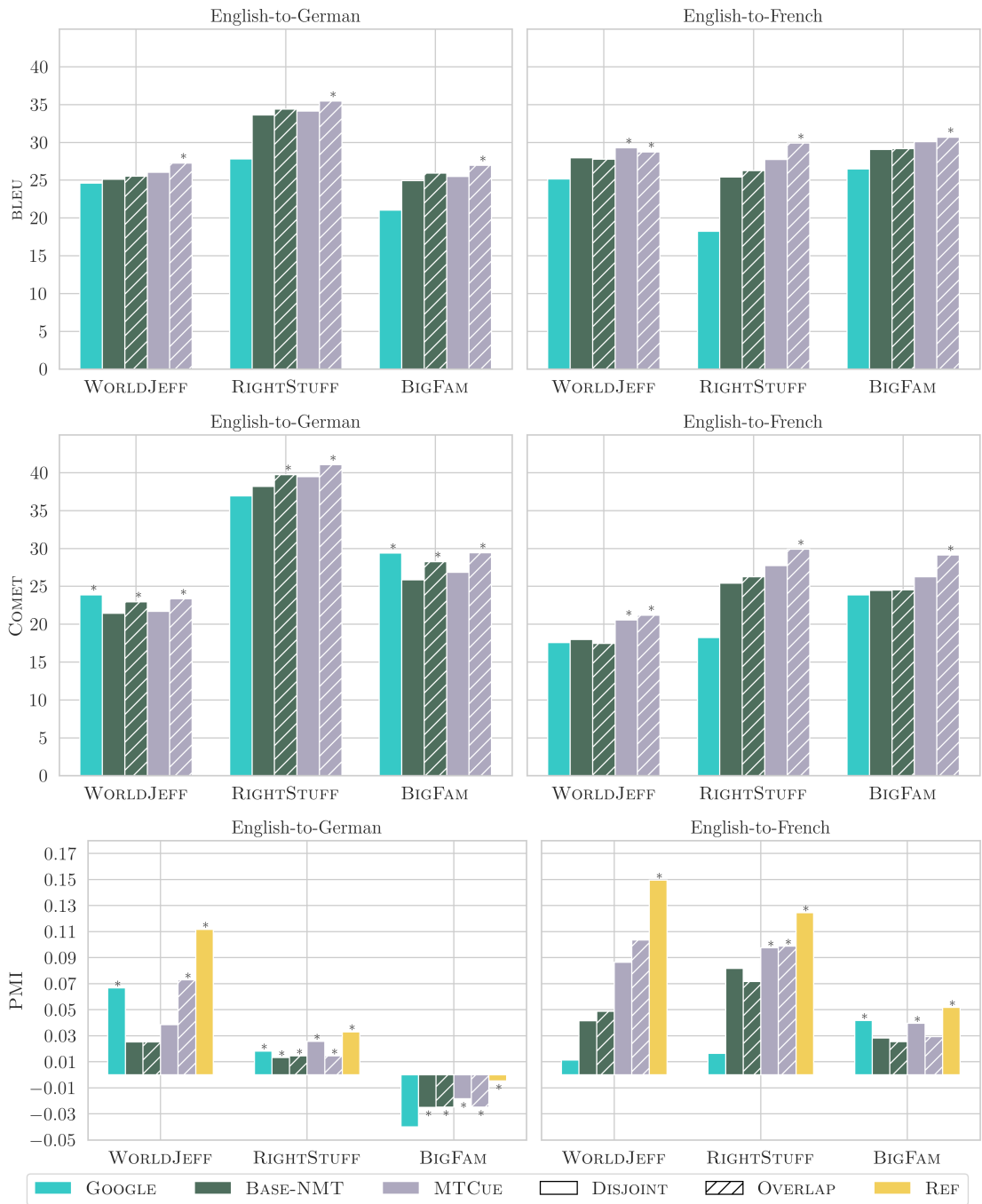


Figure 4.2: BLEU, COMET and PMI scores obtained by the evaluated models. Asterisks (*) over bars indicate the best result along with all statistically indistinguishable results computed either via bootstrap resampling (§A.3.1) or t-test for PMI (§A.3.1), $p = 0.05$.

4.4 RESULTS OF AUTOMATIC EVALUATION

In [Figure 4.2](#) we report the results of the evaluation campaign, as measured by BLEU ([§ A.2.3.2](#)), COMET ([§ A.2.3.2](#)) and in addition by our context specificity metric introduced in [§ 3.4](#). Our first observation is that, especially for the EN-DE pair, the BLEU and COMET scores of the different models vary considerably, with COMET generally indicating smaller gains from using MTCUE compared to the baseline, and even finding GOOGLE to be the best model for two out of three series, unlike BLEU which always finds it to be the worst. This discrepancy could be explained by the fact that the hypotheses from BASE-NMT and MTCUE have similar lengths to the reference translations (7.06 words per segment vs 7.04 in references), and the hypotheses from GOOGLE are considerably longer (8.29 words per segment). This may inflate the COMET scores which are computed via pooled sentence embeddings ([§ A.2.2](#)) as more information is expressed in the average segment. BLEU circumvents this inflation via the built-in brevity penalty ([§ A.2.3.2](#)).

The BLEU scores suggest that the BASE-NMT and MTCUE systems significantly outperform GOOGLE, even in the DISJOINT data setting, with the discrepancy particularly high for RIGHTSTUFF, the fictional series about a mission in space. Secondly, for both BASE-NMT and MTCUE, fine-tuning on prior episodes of the testing show (OVERLAP) generally results in higher scores. MTCUE is consistently the strongest-performing model, highlighting that context information may be particularly useful in this scenario. According to the PMI scores, the professional translations (REF) consistently exhibit the highest context specificity, however the MTCUE system (whether in the DISJOINT or the OVERLAP setting) is on par with this reference score 5 out of 6 times. The GOOGLE system achieves low scores especially in the EN-FR pair, but manages to achieve results comparable to best in half of the cases. Finally, MTCUE remains on par or better at context specificity than BASE-NMT in all cases.

4.5 RESULTS OF HUMAN EVALUATION

This section presents the results of human evaluation. First, [§ 4.5.1](#) discusses the analysis of the specific errors marked by the PEs. Then, in [§ 4.5.2](#), we analyse the effort required to post-edit the outputs of the individual systems (as well as the professional references, i.e. REF). Finally, in [§ 4.5.2.1](#) we analyse the results of a post-campaign survey regarding the PEs' views on machine translation and post-editing in the subtitling industry, which helps us incorporate the human feedback to shed light on potential future directions for this work.

JOB COMPLETION As part of the professionals' contract with the company, they were allowed to withdraw from completing the work at any point if they deemed the compensation inadequate for the required job. At the midpoint of the campaign, two PEs (PE.G1 and PE.G3) contacted the project manager on the company side. They expressed

concerns regarding the machine translation output, indicating that they believed the quality was subpar and the task potentially required more effort than translation from scratch. As an alternative solution, they proposed narrowing the scope of the remaining work to error identification and marking, without making the necessary corrections. These two PEs had post-edited two and three (out of seven) episodes respectively before making this request. Consequently, 32% of the episodes in the EN-DE language pair lack post-editing results. In our setting without repeated measurements, where each PE post-edited a distinct version of each episode, any missing results negatively impact result reliability. Given that we could still gather complete error data, we present our error analysis in both language pairs in §4.5.1. However, our effort analysis in §4.5.2 is centred on the EN-FR pair, with results for PE.G2 and PE.G4 shown when per-PE effort or comparison to translation from scratch is discussed.

4.5.1 Error Analysis

In this section we analyse the errors marked by the PEs. Our initial inspection of the results indicated that each PE marked a significantly different total number of errors. For example, within the EN-FR pair, PE.F1 marked a total of 232 errors across all episodes while PE.F4 marked 878. This makes direct comparison of the error counts across systems unreliable as each PE also post-edited a different number of segments for each system (cf. Table 4.1). Since each PE post-edits seven episodes total, each the output of one out of four examined systems, one of the systems is post-edited only once by any PE. For example, in Table 4.1, PE.1 is assigned two episodes for REF, MTCUE and GOOGLE, but only one for BASE-NMT. In this example, if PE.1 generally marks fewer errors than others, BASE-NMT will be disproportionately rewarded.

To make them comparable, we normalise each of the measurements by computing a *normalisation constant* h for each PE and then multiplying the error count for the given category by the PE's h . Let $\text{ERR}_{PE_i,x}$ denote the number of errors within the category x for the i -th PE. We compute the normalised count $\text{norm}(\text{ERR}_{PE_i,x})$ as described by Equation 4.1.

$$\begin{aligned} \text{norm}(\text{ERR}_{PE_i,x}) &= \text{ERR}_{PE_i,x} \times h_i \\ \text{where } h_i &= \frac{\max(\text{ERR}_{PE_1,\text{total}}, \text{ERR}_{PE_2,\text{total}}, \text{ERR}_{PE_3,\text{total}}, \text{ERR}_{PE_4,\text{total}})}{\text{ERR}_{PE_i,\text{total}}} \end{aligned} \quad (4.1)$$

We report the total error counts as well as the normalisation multipliers in Table 4.4.

ERROR POST-PROCESSING Our evaluation took place in a multi-modal subtitling environment. However, the model outputs, whether from GOOGLE or one of our own systems, were formatted as lists of sentences, lacking adaptation to the constraints typically associated with subtitles. In standard quality checks conducted by the PEs, the

English-to-German			English-to-French		
<i>PE ID</i>	<i>Error count</i>	<i>h</i>	<i>PE ID</i>	<i>Error count</i>	<i>h</i>
PE.G1	1526	1.76	PE.F1	232	14.68
PE.G2	2452	1.10	PE.F2	182	18.71
PE.G3	2690	1.0	PE.F3	3406	1.0
PE.G4	1832	1.47	PE.F4	878	3.88

Table 4.4: Error counts and normalisation coefficients h for each PE in the experiment.

task encompasses not only correcting translation errors but also ensuring the subtitles comply with strict guidelines. This includes adhering to reading speed and length limits, balancing the length of the top and bottom subtitle, disambiguating speaker turns through the use of colours or dashes, and applying appropriate HTML formatting such as italics where necessary, as specified by a style sheet. Given the focus of this project is on contextual machine translation, our systems were not designed to create translations conforming to these stringent guidelines. Consequently, our primary interest was in identifying the translation errors alone.

To faithfully replicate the work environment of the PEs, we applied a greedy reformatting tool (built into ZOOSUBS) to reformat our translations as subtitles. We made it clear to the PEs that this work is conducted for research purposes, and that standard subtitle formatting and reading speed guidelines are relaxed for this project.

To ensure that the translation and non-translation errors are kept separate during the process, we included several environment-specific errors for the workers to select from:

- *Subtitle formatting violation*: This category addressed cases where the subtitle is not optimally split across multiple blocks or where the top and bottom lines are not of similar lengths.
- *Other style sheet non-compliance*: As the task sought to mimic a production environment, the workers were provided with a style sheet guide from a professional streaming company, which covered rules such as translation of names and titles, the usage of italics or punctuation conventions.

There were also instances where a PE encountered both translation and non-translation errors within the same segment, as exemplified below.

Example	Series: WORLDJEFF
Source	Hi. Can I take a look at what you're doing by any chance?
BASE-NMT (X)	Hi. Kann ich mir zufällig ansehen, was du [BREAK] machst?
Post-ed.	Hi. Kann ich mir vielleicht ansehen, [BREAK] was Sie da machen?
Errors	<i>Mistranslation</i> <i>Subtitle formatting violation</i> <i>Formal/informal mismatch</i>

In the above **Example**, both translation errors (*Mistranslation* of *by any chance* and *Formal/informal mismatch* of *you're doing*) and non-translation errors (*Subtitle formatting violation* of the position of the subtitle break) are present. While disregarding the non-translation error counts in such cases is straightforward, correcting the effort rates (editing time and keystrokes) is more challenging. To precisely gauge the effort required solely for addressing translation-related errors, we employ a correction method. Specifically, let $ERR_{non-translation}$ and $ERR_{translation}$ be the total effort expended by a PE on a segment that has only non-translation and only translation errors marked, respectively. We calculate a translation share as follows:

$$\text{Translation Share (TS)} = \frac{ERR_{translation}}{ERR_{translation} + ERR_{non-translation}}$$

We then use this share to determine the share of the effort spent on translation in segments that had both errors marked:

$$\text{Effort on Translation Errors in Mixed Segment} = \text{Effort in Mixed Segment} \times \text{TS}$$

For example, if a PE takes three seconds for translation errors and two seconds for non-translation errors on average, where they marked both types we multiply their total effort for that segment by $\frac{3}{3+2}$.

We also noted that the **Other** category was used substantially and decided to parse the contents of the optional description text box, to verify whether some of them fit already pre-defined categories. Indeed, most commonly reported **Other** errors were "Grammar", "Punctuation", "Timing", "SGP" (spelling, grammar, punctuation) and "Literal translation". Such errors were removed from the **Other** category and pigeonholed as appropriate (e.g. "Grammar" as *Fluency*). More complex comments

such as “wissen Sie should not be in the translation” were left categorised as *Other*. In total, 69.3% of **Other** errors were re-categorised.

RESULTS The calculated normalised counts of errors within each category (Table 4.5) suggest that MTCUE performs no worse than both non-contextual MT systems overall (row **Total**), while performing significantly better in the **Context** and **Style** categories in EN-FR, pointing to gains related to the use of context information.

The most frequently flagged errors in both language pairs were consistently *Mistranslation* and *Fluency*. *Mistranslation* was reported a similar number of times for all three machine translation systems in EN-DE and three times less frequently for post-editing REF. In EN-FR, this gap between the machine and human translation was similar, though within the MT systems themselves, the GOOGLE system had a significantly higher error rate for *Mistranslation* errors (38.80 mean) than the next best system, i.e. BASE-NMT (22.73); the contextual MTCUE achieved an even lower rate of 20.10. Interestingly, MTCUE also produced outputs of higher *Fluency* than other systems, even surpassing REF for EN-FR, though at our confidence interval of 80% this difference is insignificant.

In both language pairs, the *Omission* error was consistently marked the fewest times in GOOGLE-generated text (see rows labelled *Omission* within the **Translation quality** category). In both cases, REF scored significantly above the mean. This is unsurprising: translations authored by the general-purpose GOOGLE engine tend to be overly literal and faithfully preserve the contents of the input sentence, whereas *dialogue* translation often necessitates that the translator let go of individual features of the source text, or opt for alternative expressions, to maintain the brevity and dynamics of the source dialogue, leading to spontaneous omissions in the reference translations. One such example is the notorious use of *wissen Sie* in place of the English *you know* by GOOGLE in translations to German

As shown in Example 1, the filler phrase *you know* is translated literally by GOOGLE – and necessitates post-editing – but is ignored by MTCUE. BASE-NMT and MTCUE, trained on dialogue, are characterised both by the preference of brevity and dynamics expression in translations while also maintaining a closer link with the source text. As a result, the number of times both systems were marked with *Omission* was near average. However, hypotheses generated by MTCUE prompted the PEs to mark *Omission* more times than BASE-NMT, suggesting that MTCUE’s behaviour more closely matches that of professional translators. Other **Translation quality** errors were relatively infrequent and with insignificant differences between systems.

While we did not provide explicit ways for the PEs to mark errors to do with speaker style so as not to bias them towards seeking out contextual issues, we instead provided categories for most frequent contextual errors: *Incorrect gender*, *Plural/singular form* and *Formal/informal mismatch*, as well as loose categories for **Style**, in the intention to collect measurements of how often the PEs feel the need to alter the style of the translations.

Error type	Normalised count			
	GOOGLE	BASE-NMT	MTCUE	REF
Translation quality	13.12 ± 14.46	<u>8.70 ± 11.67</u>	8.49 ± 10.90	4.56 ± 5.14
<i>Catastrophic translation</i>	<u>0.50 ± 0.27</u>	0.46 ± 0.18	<u>0.88 ± 0.95</u>	0.72 ± 0.68
<i>Mistranslation</i>	<u>26.99 ± 8.58</u>	25.69 ± 7.67	<u>26.74 ± 6.15</u>	8.76 ± 5.51
<i>Omission</i>	0.26 ± 0.15	2.32 ± 2.20	3.54 ± 2.79	5.38 ± 6.75
<i>Deviation in sentiment</i>	<u>1.11 ± 0.66</u>	0.83 ± 0.30	<u>1.25 ± 0.88</u>	5.23 ± 4.40
<i>Locale convention</i>	2.04 ± 0.00	<u>0.94 ± 0.46</u>	0.61 ± 0.30	0.91 ± 1.03
<i>Fluency</i>	16.88 ± 15.22	<u>9.54 ± 11.17</u>	7.10 ± 6.52	4.18 ± 3.65
Context	5.34 ± 5.68	<u>2.64 ± 3.45</u>	2.21 ± 2.55	1.18 ± 1.13
<i>Incorrect gender</i>	<u>2.20 ± 1.58</u>	<u>1.69 ± 1.90</u>	1.43 ± 1.17	1.60 ± 1.19
<i>Plural/singular form error</i>	<u>0.99 ± 0.81</u>	0.80 ± 0.63	<u>1.19 ± 1.24</u>	0.33 ± 0.00
<i>Formal/informal mismatch</i>	11.31 ± 4.55	<u>5.29 ± 4.60</u>	3.86 ± 3.60	1.19 ± 1.31
Style	<u>12.19 ± 9.79</u>	8.12 ± 6.59	<u>9.88 ± 7.83</u>	3.77 ± 3.86
<i>Awkward style</i>	17.70 ± 7.76	11.82 ± 5.21	<u>13.11 ± 7.04</u>	4.70 ± 4.34
<i>Subjective style changes</i>	<u>2.55 ± 2.09</u>	1.65 ± 1.59	<u>2.33 ± 2.28</u>	2.13 ± 2.52
Other	<u>2.12 ± 3.43</u>	<u>3.26 ± 4.48</u>	2.10 ± 2.46	3.39 ± 5.88
Total	9.58 ± 11.35	<u>6.44 ± 9.05</u>	6.41 ± 8.82	3.86 ± 4.70
Translation quality	20.01 ± 23.05	9.27 ± 9.52	<u>10.21 ± 8.88</u>	6.60 ± 5.08
<i>Catastrophic translation</i>	<u>3.41 ± 1.38</u>	2.25 ± 2.39	<u>2.86 ± 3.03</u>	2.51 ± 3.26
<i>Mistranslation</i>	38.80 ± 14.35	<u>22.73 ± 8.49</u>	20.10 ± 7.34	7.24 ± 3.61
<i>Omission</i>	2.40 ± 2.40	3.91 ± 1.49	5.56 ± 4.09	7.48 ± 5.13
<i>Deviation in sentiment</i>	5.93 ± 5.90	<u>7.82 ± 6.09</u>	11.59 ± 0.00	6.74 ± 3.03
<i>Locale convention</i>	4.29 ± 2.49	0.73 ± 0.51	0.21 ± 0.00	0.63 ± 0.00
<i>Fluency</i>	30.83 ± 31.77	<u>7.28 ± 3.75</u>	5.92 ± 4.18	7.82 ± 7.35
Context	<u>5.41 ± 3.64</u>	6.09 ± 4.26	3.86 ± 3.11	1.29 ± 1.07
<i>Incorrect gender</i>	3.49 ± 2.59	6.96 ± 5.57	<u>4.77 ± 3.98</u>	0.49 ± 0.44
<i>Plural/singular form error</i>	4.50 ± 1.92	5.84 ± 4.60	1.97 ± 0.62	0.00 ± 0.00
<i>Formal/informal mismatch</i>	<u>7.44 ± 4.63</u>	<u>5.58 ± 3.76</u>	4.23 ± 2.93	1.69 ± 1.10
Style	11.05 ± 7.07	10.35 ± 3.69	3.41 ± 2.53	5.55 ± 3.41
<i>Awkward style</i>	11.13 ± 7.46	9.55 ± 1.27	2.89 ± 2.76	4.10 ± 1.28
<i>Subjective style changes</i>	<u>10.94 ± 8.16</u>	11.15 ± 5.52	4.18 ± 2.87	6.28 ± 4.09
Other	37.20 ± 52.68	11.19 ± 16.44	<u>23.67 ± 29.23</u>	27.05 ± 24.68
Total	17.02 ± 25.78	8.84 ± 9.20	<u>9.63 ± 13.85</u>	8.83 ± 12.84

Table 4.5: Counts of errors flagged by the PEs for each system. The best (i.e. lowest mean) result in each row is highlighted and all statistically indistinguishable results underlined (one-tailed t-test, confidence interval of 80%, $p = 0.2$). REF scores are excluded from statistical significance. Error rates for categories in bold (e.g. **Style**) are calculated based on all errors within the category.

Example 1		Target: German; series: WORLDJEFF
Source	Things catch my eye, you know, and I get a little fascinated.	
Reference	Die Dinge fallen mir auf, und ich bin etwas fasziniert.	
GOOGLE (✗)	Die Dinge fallen mir ins Auge, wissen Sie, und ich bin ein wenig fasziniert.	
Post-ed.	Die Dinge fallen mir ins Auge, wissen Sie , und ich bin ein wenig fasziniert.	
Error	<i>Awkward style</i>	
BASE-NMT (✓)	Die Dinge fallen mir auf, und ich bin etwas fasziniert.	
MTCUE (✓)	Die Dinge fallen mir ins Auge und ich bin etwas fasziniert.	

Since all of the post-edited content is dialogue, the style of the translation can be directly associated with the style of the speaker’s expression, without biasing the PE towards thinking in terms of what is a characteristic way of expression for the given speaker. Our findings regarding some **Context** categories (*Incorrect gender, Formal/informal mismatch*) are consistent between the two language pairs, and MTCUE was found to be superior in most categories in both cases, with the overall score for the **Context** category being significant at 80% confidence for EN-FR. The *Plural/singular* form error required few corrections in EN-DE (and BASE-NMT was found superior to MTCUE) and more in EN-FR (where MTCUE was found superior).

The findings from the **Style** category also work in favour of contextual MT, where it was found comparable to non-contextual systems for the EN-DE pair and significantly better than them for the EN-FR pair, requiring the fewest style-based adjustments, even fewer than REF. Within the EN-DE pair, *Subjective style changes* were flagged only up to 4 – 5 times per 100 segments for any system, and a consistent number of times between systems, and *Awkward style* was flagged the fewest times for REF (4.68 on average), much less frequently than for the other systems, among which GOOGLE required the most edits and BASE-NMT the fewest.

Overall, our error count analysis suggests that within the EN-FR pair, MTCUE has significantly reduced the number of errors marked for contextual and stylistic reasons compared to non-contextual systems, while not degrading overall translation quality. The findings within the EN-DE pair are too variable to yield definitive conclusions, but entail no degradation of quality leading from the inclusion of context, a significant improvement for contextual phenomena compared to GOOGLE, and highlight that MTCUE makes the fewest contextual errors overall.

4.5.2 *Analysis of Effort and Quality*

This section delves into the analysis of per-PE effort spent post-editing or translating the outputs of each system. Based on the observation that some measurements of editing time and keystrokes were out of the distribution, we normalise these by first computing the 97.5th percentile for the given language pair and task (translation or post-editing) and set all per-segment measurements to be capped at that percentile. Our obtained percentiles were: 37 seconds and 69 keystrokes for translation, and 45 seconds and 54 keystrokes for post-editing.

EFFORT PER PE Analysis of effort metrics for the individual PEs (Figure 4.3) reveals a significant discrepancy between the total effort put in by PE.F3 compared to the other three. With the exception of PE.F4, post-editing the outputs of GOOGLE generally took the longest, required the most keystrokes and resulted in the most changed translations (as measured by HTER), even if the error rate per 100 segments was not the highest (see PE.F1, PE.F2). Between BASE-NMT and MTCUE, Judgements of PE.F1 and PE.F2 suggest that the outputs of these systems required a similar amount of post-editing effort (though PE.F2 found significantly more errors in BASE-NMT's outputs). According to all four metrics, PE.F3 found the outputs of MTCUE to require less post-editing work than BASE-NMT. Finally, PE.F4 identified a similar number of errors in both, but made more significant alterations to the contextual outputs of MTCUE.

Results for the EN-DE language pair (Figure 4.4) suggest that each PEs contributed similar effort. Interestingly, the error rate and effort measures of these PEs are closer in magnitude to the outlier PE.F3 within the EN-FR pair. Putting PEs from both pairs together we find an interesting correlation: those PEs who expressed a preference for post-editing marked significantly fewer errors overall. This data suggests that those professionals who prefer translation opted for the approach of spending any effort necessary to match the quality of the resulting text to what they would have produced from scratch, while those who prefer post-editing have contributed a fixed amount of effort, possibly characteristic of their typical post-editing assignment.

Error rate per 100 segments for this pair suggests that GOOGLE consistently requires the most edits overall, and REF the least, though only PE.G4 made drastically fewer edits to this already production-ready text. Between BASE-NMT and MTCUE, PE.G2 and PE.G3 found MTCUE to be less erroneous (and PE.G3 found it to be on par with REF), while PE.G1 and PE.G4 identified fewer errors in BASE-NMT.

As we are missing effort measurements for PE.G1 and PE.G3, the following analysis is based only on the other two PEs. Results from PE.G2 indicate that the quality of translations from GOOGLE and BASE-NMT is comparable, requiring the most complex and laborious edits. MTCUE's hypotheses required less work from this PE, and REF text still less. Results obtained from PE.G4's edits are somewhat different. This PE made next to no edits to the REF text, which could be interpreted as them being the least

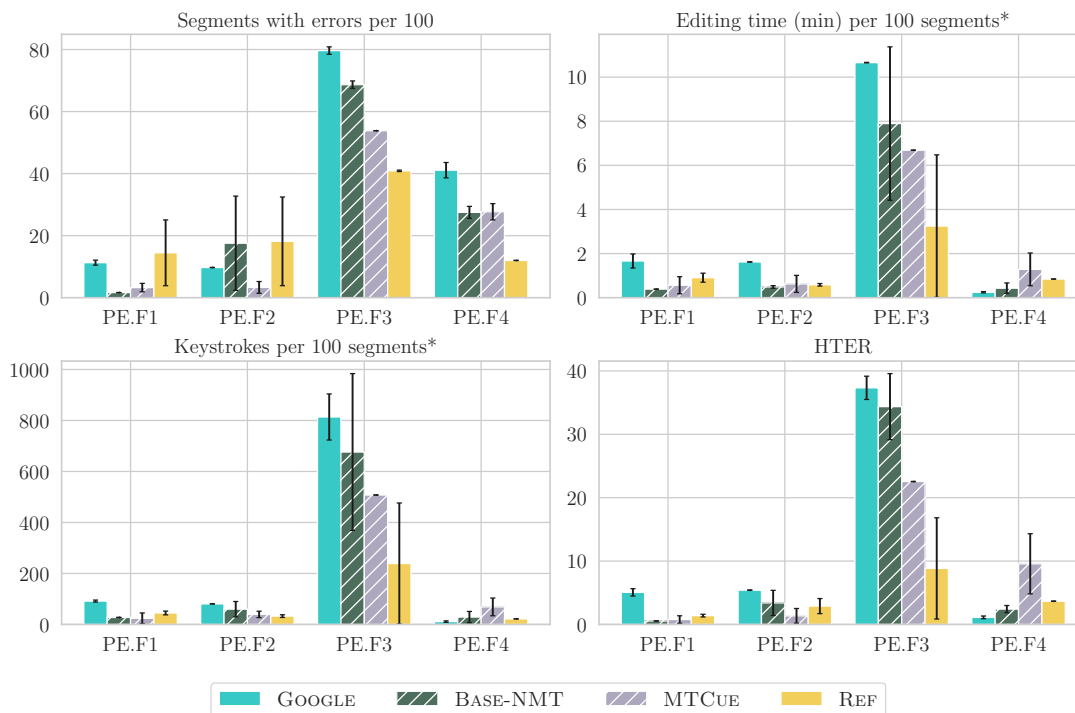


Figure 4.3: Effort for each PE for the English-to-French language pair.

subjective of the PEs, only making edits when they are necessary. This PE found MTCUE to require more edits than BASE-NMT and on par with GOOGLE. Interestingly, even though editing MTCUE’s outputs took more time and keystrokes, GOOGLE’s outputs yielded a HTER value about 10 points higher than MTCUE. Since GOOGLE is the more literal MT system, and MTCUE produces more dialogue-like responses, these findings suggest that, other things being equal, a literal and overly long translation of dialogue may take less effort to post-edit than an incorrect platonic (dialogue-like) response, even if more profound edits are required.

COMPARISON WITH TRANSLATION EFFORT In Figure 4.5 we compare the unnormalised post-editing effort (gold bars) to the translation effort (turquoise bars) for the 9th episode of BIGFAM. For PE, we excluded measurements from post-editing reference translations (REF). For both language pairs, translating segments from scratch requires between 4 and 6 times effort, both technical and temporal.

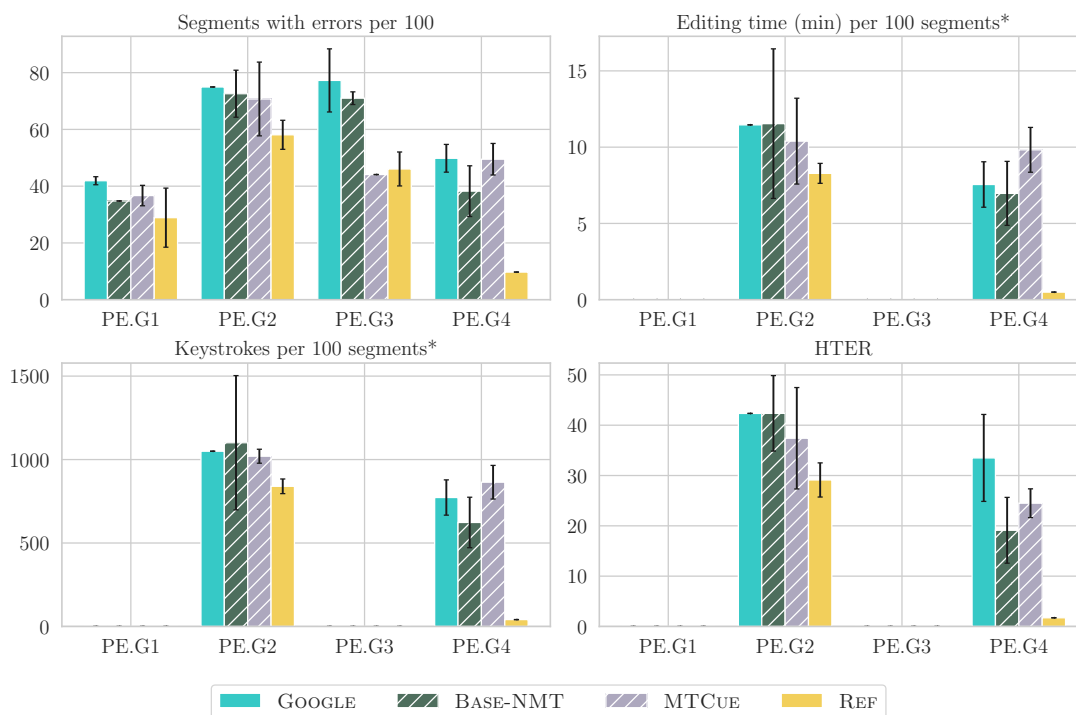


Figure 4.4: Effort for each PE for the English-to-German language pair.

4.5.2.1 Analysis of the professionals' views on post-editing and MT

In our final analysis, we present the PEs' responses to a survey regarding views on post-editing and machine translation.

When asked to express their preference between post-editing and translation from scratch, most of the German PEs indicated a preference for translating from scratch, whether for the domain of subtitles or in general. In particular, three out of four, expressed frustration with machine translations, highlighting their stiffness and literal nature. They pointed out that MT omits many aspects of the original text, such as slang, gender agreement, references to the video and people's speaking styles. In contrast, they viewed translation as a more creative process, which yields more idiomatic and fluent translations. Moreover, they noted that post-editing currently demands significant effort, sometimes even surpassing that of translating from scratch, yet it is compensated at a considerably lower rate than translation. One PE remarked that post-editing often feels like damage control rather than effort to deliver the best possible translation.

Conversely, most of the French PEs (three out of four) expressed a preference for post-editing. Two out of the three French translators who favoured post-editing cited their specialisation in quality checking as the reason for their preference. The one

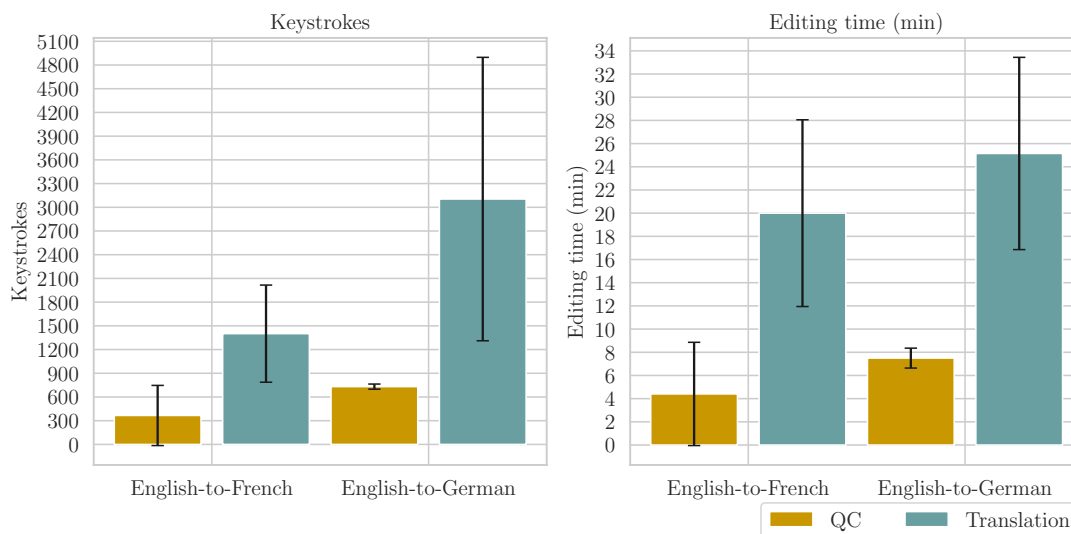


Figure 4.5: A comparison of the effort of translation from scratch and post-editing machine translation outputs, captured per 100 segments.

translator who did not prefer post-editing expressed that, due to recurring issues with subtitle formatting within the project, post-editing was taking considerably longer than anticipated. They believed that translating from scratch would have focused more on content and therefore be less tedious.

Regarding their views on machine translation, *PEs* in both languages agreed that *MT* can be a helpful tool, and one *PE* even noting significant recent improvements in *MT* quality. However, most *PEs* concurred that there is still a substantial gap in quality, rendering *MT* insufficiently competent to replace translation from scratch. Despite this, the *PEs* expressed optimism that *MT* could bridge this gap in the future, potentially resulting in a post-editing workflow that saves effort.

Finally, the majority of *PEs* recognized the advantages of post-editing, including the reduction of temporal effort in some cases and the potential to improve consistency in translating terminology (through a form of translation memory). A French *PE* pointed out that post-editing enables greater attention to detail. For some, these benefits are anticipated in the future, when the technology reaches a sufficient level, requiring edits only for a fraction of segments.

4.5.2.2 Examples of challenging translations for MTCUE

We present several examples of corrections made in the post-editing process to reflect what kind of corrections required attention as well as what mistakes need to be improved upon in the future.

Example 2	Target: French; series: BIGFAM
<hr/>	
Source	Testing your creativity as home cooks.
Addressee	<i>I am talking to a group of people</i>
MTCUE (X)	Vous testez votre créativité de cuisinière familiale.
Post-ed.	Vous testez votre créativité de cuisiniers familiale.
Error	<i>Incorrect gender</i>

Despite being given the sufficient gender and number context for the interlocutor in [Example 2](#), MTCUE still generated the incorrect noun form *cuisinière* indicating a singular female family cook as opposed to the masculine plural form *cuisiniers* (which can also be used to refer to a mixed-gender group). We have shown in [§2.5](#) that generating correct gender and number forms is a capacity which improves in MTCUE with more training data, and while the error counts suggest that the model makes fewer such errors than its non-contextual counterpart (BASE-NMT), there is still room for improvement which can be facilitated through inclusion of more annotated samples. In data settings similar to ours, where parallel translations to multiple languages are available, future such efforts could explore leveraging annotations obtained automatically for one language such as Polish in other languages such as French.

Example 3	Target: German; series: BIGFAM
<hr/>	
Source	I really want to stir it.
Previous sentence	<i>I really want to stir that sugar.</i>
MTCUE (X)	Ich möchte es wirklich umrühren.
Post-ed.	Ich möchte ihn wirklich umrühren.
Error	<i>Incorrect gender</i>

[Example 3](#) highlights another gender-related error, albeit one unrelated to the speaker or the interlocutor of the conversation. The source sentence uses the pronoun *it* to refer to sugar. The information what the pronoun refers to is recoverable from one of the provided context sentences, namely the immediate previous sentence. However, MTCUE incorrectly translates *it* as *es*, when the correct pronoun is the masculine *ihn* – *sugar* in the German language translates as the masculine *der Zucker*. This type of gender error can be categorised as a document-level problem as the information is recoverable from intra-textual context. While this work has not focused on document-level performance of MTCUE, it is crucial that a contextual model performs well in this capacity, and improving this performance is a suitable direction for future work.

[Example 4](#) presents a scenario where MTCUE incorrectly interprets the exclamation *No way* as *Under no circumstance*, which fails to account for the sense of disbelief and amazement that the victorious family is experiencing. Such an interpretation

Example 4	Target: German; series: BIGFAM
Source	No way, no way.
Video context	<i>The victorious family is in disbelief about their triumph.</i>
MTC _{UE} (✗)	Auf keinen Fall. (‘Under no circumstance.’)
Post-ed.	Unmöglich. (‘Unbelievable.’)
Error	Other: (<i>inconsistency with video</i>)

relies strongly on the visual context, of which effective incorporation into the machine translation process in a multi-modal framework is an area for future work.

Example 5	Target: German; series: BIGFAM
Video context	<i>Two cooks in front of a chopping board.</i>
Source N	Get that Welly on that board.
Reference N	Leg das Welly auf das Brett.
MTC _{UE} (✗)	Stell die Welly auf das Brett.
Post-ed.	Legt das Wellington auf das Brett.
Error	<i>Awkward style</i>
Source N+1	She’s on.
Reference N+1	Es ist drauf.
MTC _{UE} (✗)	Sie ist dran.
Post-ed.	Ist drauf.
Error	Other: (<i>inconsistency with video</i>)

Example 5 presents a difficult scenario. On the one hand, MTC_{UE} uses the incorrect German preposition *an* to translate the English *on*, instead of the correct *auf* (*on that board = auf der Tafel*). On the other hand, there is a more interesting error, and it comes from mistranslating *She* as *Sie*. The pronoun is a reference to pork Wellington, abbreviated to *Welly* by the speaker, and incorrectly assigned the feminine article *sie*, instead of the neuter *das*. The error is difficult not to make since in sentence N+1, the English speaker personifies the object by referring to it as *She* - consequently, even a document-level system could take this into account and incorrectly interpret what *Welly* is. And the correct interpretation is crucial to selecting the right verb *legen* over *stellen* which should be used to translate *get* when referring to meat. While the PE described this as an *inconsistency with video* error, it is challenging to outline the minimal set

of context information sufficient for the correct interpretation and translation of this example. The context of cooking, the light-hearted, casual character of the show and the manner of British speech in this scenario, as well as more concrete information, such as what meal is being made and what the cooks are doing in the moment, all could aid this process. An important challenge for future contextual systems is going to be to discern which type of information is necessary and when.

4.6 CONCLUSIONS

In this Chapter we have presented the results of automatic and human evaluation campaign on the use of machine translation in post-editing translations of subtitles for TV series in a multi-modal scenario, with the focus on contextual *MT*. We have drawn the following conclusions:

1. **Rich contextual annotations benefit machine translation, even in a scenario of translating unseen series and speakers.** Our automatic evaluation results (§ 4.4) suggest that *MTCUE* surpasses the translation quality of *BASE-NMT* in both *DISJOINT* and *OVERLAP* data settings, despite only being fine-tuned on 50 – 100K samples of annotated dialogue. The scores are nevertheless consistently better when there is overlap between training and testing series and speakers.
2. **Custom machine translation models for dialogue are more helpful in post-editing than general systems such as Google Translate.** Our results suggest that the *GOOGLE* system consistently makes the most errors and requires the most effort to obtain translations of sufficient quality, and we have argued that overbearing literalness and stiffness of the subtitles may be the root cause.
3. **Translations may receive edits even if they are correct.** One of the texts post-edited in our evaluation was a production quality human translation (*REF*). We found that the *PEs* consistently applied changes to this text despite it needing no changes, with three of them doing so at a rate of over 40 errors per 100 segments. In workflows which already include a post-editing step to human translations, this finding should be taken into account when estimating the efficiency of using *MT* instead of human translations.
4. **Rich contextual annotations may benefit a post-editing workflow.** Between the two systems trained on the *ZOO-MULTI* corpus (*BASE-NMT* and *MTCUE*), neither model's outputs are categorically quicker or easier to post-edit. However, our results regarding error counts indicate that *MTCUE* – the contextual model – makes fewer errors related to **Style**, **Context** and *Fluency*, especially in the *EN-FR* language pair.
5. **Post-editing machine translation requires significantly less technical and temporal effort compared to translation from scratch.** Post-editing of any machine

translation output (even GOOGLE) required between four to six times less technical and temporal effort than translation from scratch, though the post-campaign survey revealed that some PEs considered the job to sometimes be harder and less interesting than translation from scratch.

6. **Machine translation is viewed positively, though sceptically, by the PEs.** The post-campaign survey revealed the general consensus among PEs in the subtitling industry: machine translation can be a useful tool, however its usefulness depends on how it is implemented into the system, and how good is the average translation. Unfortunately, it also points to a general pessimism towards MT ever becoming capable of handling colloquial language and behavioural agreement. Since each PE in our experiment was given outputs of both non-contextual and contextual systems, as well as professional translations, it is unclear which systems' outputs the PEs had in mind when writing this feedback. In the direction of improving machine translation with context, future human evaluation campaign could involve sufficiently large groups of PEs to devote individual groups to post-editing contextual MT outputs exclusively. This way, more clear feedback could be collected as to whether MT is improving at expressing behavioural agreement.
7. **In a professional human evaluation, the recruited PEs may contribute a varying amount of effort.** We found that different PEs put in a different amount of effort, with a correlation between the amount of effort contributed and the post-editing training and preference of the PE. However, there may be other reasons for this variability. Since the PEs were told about the research nature of the project, they may have approached this project with less vigilance than if the work was undertaken for actual clients. On the flip side, some PEs may have eventually realised they were dealing with some MT outputs – they were not told this explicitly – and became generally more scrutinous as a result, expecting to make many more corrections than in a typical PEs task. This would perhaps explain why some PEs took to post-editing REF at rates sometimes matching the outputs of the MT systems. In future campaigns, robustness to such variables can be ensured by e.g. having a more varied pool of workers and assigning significantly fewer segments to each. However, such a solution was inaccessible to us as our pool of workers was very limited.
8. **The capability of MT to handle subtitle formatting is of the utmost importance to the PEs.** While generating translations which adhere to subtitle constraints was out of scope for this project, it is clear that the lack of regard for subtitle formatting when MT systems are used is a serious concern for the PEs who must devote their time to manually reformatting the input text. Future systems implemented for this task should therefore take into account such constraints, possibly as an additional set of contextual variables. Similarly, style sheet adherence could also be seen as a contextual translation problem and addressed accordingly.

Finally, as per the suggestion of one of the [PEs](#), the ZOOSUBS interface could be developed further to add machine translation as an assistance tool for translators, letting them choose when they wish to use the machine-generated outputs, e.g. when dealing with basic or repetitive segments, while letting them maintain the potential for creative control, ultimately yielding a more efficient and satisfactory workflow.

CONCLUDING REMARKS

5.1 ASSESSMENT OF CONTRIBUTIONS

This thesis has investigated context-based personalisation in neural machine translation. Our work has explored the issue of grammatical agreement to gender and formality, proposed MTCUE, a novel model architecture for leveraging contextual information in NMT, explored reference-free evaluation of context specificity in NMT with the use of personalised language models, and conducted a human evaluation campaign in a real-world task of subtitle translation, offering insight into the benefits of using context in this domain, as well as collecting the views of translation professionals on the current state of machine translation in their domain. In addition, we have contributed a morphosyntactic annotation tool for gender, number and formality phenomena in the Polish language, and CORNELL-RICH, a publicly available corpus of rich metadata annotations for films and characters featured in the Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil & Lee 2011). Below we detail contributions made in effort to answer each research question.

RQ1 How can attribute control best be incorporated into neural machine translation in multiple attribute and low-resource scenarios? (Chapter 2)

We have concluded that attribute control can be incorporated in several ways. If adequate training data is available, then interlocutor attributes such as gender and formality can be fully controlled. However, even in a low-resource scenario the attributes can be partially and fully controlled, by implementing methods such as attribute-specific hypothesis re-ranking and data augmentation. Finally, more complex attributes such as plot descriptions can be used to improve translation quality, and when provided alongside other metadata can be used to create a representation space for context which then enables few- and zero-shot control of attributes such as formality and gender. This approach is implemented via MTCUE proposed in §2.5. To summarise, we have made the following contributions:

- An annotation tool for dialogue expressed in Polish which leverages MORFEUSZ2 (Kieras & Wolinski 2017) to detect the presence of grammatical markers of speaker's and interlocutor(s)' genders, the number of interlocutors and the formality relation between the speaker and the interlocutor(s), their genders and the number of interlocutors (§2.3).

- Experiments and performance analysis of nine attribute-controlling approaches utilising the above corpus, including consideration for translation quality, accuracy of controlling the phenomena, and the impact of this control on context-ambivalent examples (§2.3).
- An exploration of low-resource and zero-shot approaches to formality control in four language directions: English-to-German, English-to-Spanish, English-to-Italian, and English-to-Russian, showing that both perplexity-based data augmentation and phenomenon-specific hypothesis re-ranking methods are an effective tool in improving formality control in NMT (§2.4).
- MTCUE: a novel model architecture which utilises unstructured context to improve neural machine translation (§2.5).
- A comprehensive set of experiments showing that MTCUE significantly improves the quality of translation from English to four other languages (as measured by BLEU (§A.2.3.2) and COMET (§A.2.3.2)), and achieves excellent few-shot and zero-shot performance at attribute-controlling tasks such as formality and gender (§2.5).

RQ2 Can language models for film and TV characters be personalised solely relying on their character profiles and information on the discourse environment, and used to evaluate context-specificity in personalised machine translation? (Chapter 3)

Our work conducted within Chapter 3 has highlighted that language models for film and TV characters can indeed be personalised by leveraging the characters' and production profiles. Our evaluation suggests that this way of personalisation leads to performance comparable with speaker-specific fine-tuning methods, but requires no such fine-tuning and is also scalable to a scenario with little to no annotated data for specific speakers, mimicking the personalisation effect based on data from similar speakers. Finally, we managed to successfully leverage such personalised LMs to evaluate context specificity in machine translation, and showed that MTCUE, the previously introduced contextual MT model, achieves higher context specificity in its translations compared to a non-contextual baseline, but lower context specificity compared to the professional human translations.

To support this conclusion, the following contributions have been made:

- CORNELL-RICH: a publicly available corpus of rich metadata annotations for films and characters featured in the Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil & Lee 2011), and consisting of seven metadata types for 863 characters and six types for 595 films (§3.3.1).

- LMCUE: a robust implementation of a conditional Transformer-based language model built for the use case of personalisation (§3.3).
- sMRR: an evaluation metric for speaker-oriented personalisation, which signifies expresses the correlation between the speaker models and the speakers (§3.3.3.4).
- A profound set of experiments and analysis of LMCUE’s performance against strong baselines and on two corpora (§3.3).
- A formulation and experimental analysis regarding the use of personalised LMs to evaluate how specific the given translation hypotheses are to the extra-textual context they arise in (§3.4).
- A cost-benefit analysis revealing which types of manual metadata annotations are the most useful when personalising LMs (§3.5).
- Empirical evidence showing that pre-training language models on document-level data helps realise personalisation in fine-tuning on significantly smaller corpora (§3.6).

RQ3 *How does personalisation affect translation quality and post-editing effort in a real-life scenario of subtitle translation?* (Chapter 4)

Finally, in Chapter 4 we conducted a human evaluation campaign in two language pairs: English-to-German and English-to-French, asking professional translators to post-edit human and machine translations of seven episodes of three different TV series. Our automatic evaluation revealed that MTCUE exhibits the highest translation quality among the surveyed systems, and within the human evaluation the outputs of MTCUE were corrected the fewest times for context-related errors among all systems, confirming that that the improvements are likely context-related. In our analysis of post-editing effort we did not find significant reductions stemming from these improvements. However, a survey conducted among the participants revealed that context- and style-related errors are among the most disruptive ones, motivating future work within this area. In total, we made the following contributions:

- a taxonomy of errors which includes translation, context, style and subtitle formatting issues (§4.3.2).
- a comprehensive human evaluation campaign conducted in a practical multi-modal scenario of subtitle translation, suggesting that contextual machine translation may have a positive effect on the number of context, style and translation errors marked in the English-to-French translation direction (§4.5).
- an analysis of the views of translators and post-editors on the use of machine translation in subtitling, both currently and in the future (§4.5.2.1).

5.2 LIMITATIONS

This section outlines the limitations encountered during the course of our research, shedding light on the factors that have constrained the scope and applicability of our study.

LANGUAGE PAIR LIMITATION While the research presented within this thesis was carried out in six language pairs (sometimes in both directions), we recognise that these are mainly European languages and that English is a common denominator. Translation from English into other languages is a scenario typically required by the industrial partner to this thesis, ZOO DIGITAL: English content often needs to be translated or dubbed so that the streaming services can make it available in other countries. The choice of language pairs was also limited by the data and evaluation tools we had access to (and in the case of the shared task participation, the setting considered in the shared task). Furthermore, the human evaluation presented in [Chapter 4](#) was conducted only in two language pairs ([EN-DE](#) and [EN-FR](#)), attributed to the availability of data as well as the scarcity of qualified human post-editors capable of conducting meaningful evaluations. As such, the generalisability of our human evaluation findings may be restricted. However, our employed methods are language-independent, meaning the presented research could be expanded to other pairs in the future.

DOMAIN LIMITATION Our study mainly focuses on translations in the domain of scripted dialogue, which aligns with the domain of interest of the industrial partner to this thesis. Although previous work has shown that certain accommodative characteristics of dialogue are inherently present in scripted dialogue ([Danescu-Niculescu-Mizil & Lee 2011](#)), it goes without saying that TV dialogue is not the same as real dialogue. The cardinal difference between scripted and real dialogue is the avoidance in scripts of “redundant” elements of everyday speech such as hedges, stutter, interruptions (unless they serve a specific purpose), to maintain focus on the aim of the conversation and to speed up the story line ([Remael 2003](#)). As such, one must be cautious when drawing conclusions about real-life dialogue from the presented work. Instead, the implications of this work are relevant to the industry of automatic subtitling, dubbing, translation and automatic speech recognition of TV and film content. Additional validation may need to be performed in order to generalise our findings to other domains.

SMALL DATASET LIMITATION [Chapter 3](#) and [Chapter 4](#) present conclusions based on relatively small fine-tuning datasets. This limitation arises from the constraints of data collection and availability. Consequently, the results obtained with these datasets may not entirely capture the actual potential of contextual machine translation and personalised language modelling. We expect these benefits to strengthen with larger and more diverse datasets. As stated in the previous paragraph, we advocate for

conscious effort in creating such datasets, taking into account social biases and equal representation of different ethnicities and races, genders, sexual preferences, ages, disabilities, religions, backgrounds, etc.

HUMAN EVALUATION LIMITATION In our human evaluation in [Chapter 4](#) we collected one measurement per system per episode due to constraints in the availability of post-editors with expertise in the surveyed languages. It must be recognised that different [PEs](#) may have different preferences for post-editing, which could impact the robustness of the results. Additionally, our use of [PEs](#) with varying levels of post-editing training and preferences may limit the direct transferability of our findings to specific locales.

ETHICAL CONSIDERATIONS We acknowledge the ethical considerations surrounding personalisation and the use of sensitive data such as gender in the context of machine translation. While we do not foresee any direct application of our work in an unethical manner, it is crucial to recognise that, like all research employing generative models, our work is susceptible to inheriting the unintended biases already present in these models, including social biases. Therefore, when controlling contextual attributes, researchers must exercise consciousness of the biases in their data to fully understand the models' behaviour.

In our research, our ability to explore gender distinctions was constrained by the limited availability of data, constraining our focus to binary gender categories. Nevertheless, we strongly advocate for the development of datasets that encompass a broader spectrum of gender identities. Such an initiative not only serves to mitigate potentially harmful biases but also fosters diversity and inclusivity within machine translation and language modelling systems. By acknowledging and addressing these ethical concerns, researchers can work collectively towards more unbiased AI technologies.

5.3 FUTURE WORK

This thesis opens up several compelling avenues for future research, spanning from exploring alternative data sources to investigating the impact of the presented architectures in other domains and in settings with larger datasets. In this section we describe each direction in detail.

BUILDING RELEVANT LINGUISTIC RESOURCES FOR MORE LANGUAGES. In [§2.3](#) we have shown that the availability of training data which adequately captures a specific phenomenon in translation is the primary criterion for controlling that phenomenon. To that end, future work could explore building such datasets or annotation tools. Since the process may require advanced linguistic knowledge and be time-consuming,

an interesting avenue to consider would be (i) leveraging the resources for existing languages to build them for new ones, and (ii) exploring multilingual solutions.

APPLYING MTCUE TO OTHER DATA DOMAINS AND AT SCALE. The present work has applied MTCUE to the domain of TV dialogue, and it would be an interesting future direction to explore whether similar benefits can be obtained in other metadata-dependent domains such as the biomedical or news. In these domains, context could play the role of a “sub-domain indicator”; similarly to tagging systems, the prior calculated based on the article abstract, publication year or journal could better inform the system of what kind of article is being processed, improving the modelling of subject-specific terms.

EXPANDING TO DOCUMENT-LEVEL LANGUAGE MODELLING. Our proposed LM-CUE model (Chapter 3) processes text utterance by utterance, which is a setting necessitated by our objective of using the model in evaluating machine translation. However, a clear increment of this work could scale the model to processing entire dialogues. The primary challenge with this kind of extension is efficient leveraging of speaker metadata, which is subject to change many times within a conversation. Future work could explore utilising gating mechanisms or other ways of tying the metadata of the correct speakers to their utterances.

LEVERAGING INTERLOCUTORS’ CHARACTERISTICS. In Chapter 3 and Chapter 4 we have focused on speaker profiles but not the profiles of the other interlocutors (with the exception of [grammatical agreement](#) attributes, i.e. their gender, number and formality). In preliminary experiments, we found no benefit to utilising their metadata, however this could be because the impact of this metadata on language modelling and translation is too small to learn from corpora of the sizes which were explored in our sections. Future work could explore efficient ways of incorporating profiles for both the speaker and the interlocutor, perhaps in a document-level framework (as suggested in the previous direction).

PSEUDO-LABELLING. The primary setback to a wider adaptation of our model is the lack of useful metadata in some domains. In Chapter 3 we reported on a manual annotation campaign for collecting rich metadata of TV characters and performed an ablation study to highlight the cost-benefit trade-off for annotations made in this domain. Future work could explore semi-automatic ways of collecting metadata, perhaps utilising a human-in-the-loop approach with a large language model, where the bulk of the metadata collection is automated, and human annotators perform the less expensive task of verifying the genuinity of the collected information.

FURTHER USE OF PERSONALISED LANGUAGE MODELS FOR EVALUATION OF CONTEXT-SPECIFICITY IN TRANSLATION. In [Chapter 3](#) we outlined an idea of using personalised language models to evaluate [context specificity](#) in translations. Future work could expand on this idea, conducting a thorough investigation across multiple domains and datasets, as well as other problems of contextual translation (such as [document-level](#) translation).

FURTHER HUMAN EVALUATION INVESTIGATIONS IN PROFESSIONAL SETTINGS. Future human evaluation campaigns in setups similar to our ([Chapter 4](#)) could (i) expand to different language pairs, (ii) see the participants grouped into post-editors of contextual and non-contextual machine translation, so that their post-campaign views can reflect on the quality of the different types of [MT](#). Furthermore, the current translation workflow within ZOO DIGITAL includes either translation from scratch or post-editing (which is paid at a lower rate), and several [PEs](#) have suggested that post-editing machine translation involves more work than standard quality checks. As part of the expansion with machine translation, post-editing [MT](#) could be considered a third tier, compensated adequately to the effort it takes compared to translation *ex novo* and quality checking.

PRELIMINARIES

In this supplementary chapter, we delineate the essential concepts that constitute the foundation of knowledge necessary for understanding the research chapters of the thesis. The reader unfamiliar with any of the foundational concepts may use the references provided in text to learn it by reading the explanation provided in this chapter. This chapter is organised into three sections:

§A.1 (*Introduction to Machine Learning*) explains the fundamental principles of machine learning essential for understanding the research experiments. This section delves into the architectures and mechanisms employed in the experiments, including the Transformer architecture and the attention mechanism.

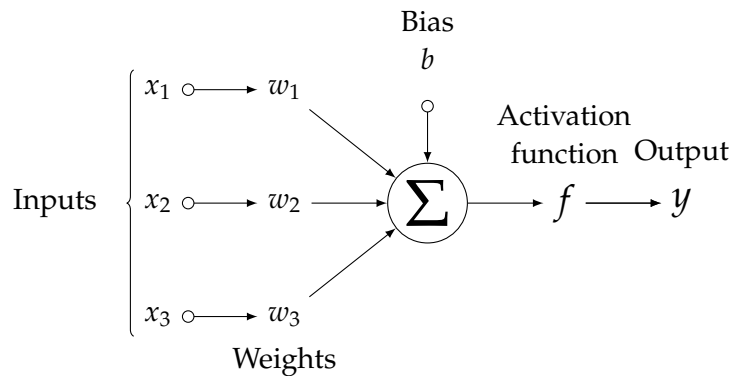
§A.2 (*Natural Language Processing*) provides the reader with an understanding of how the machine learning concepts covered in §A.1 are applied to text processing. We explore two specific tasks, namely machine translation and language modelling, which are extensively experimented with in later chapters.

Finally, the supplementary §A.3 (*Statistical Concepts Employed Within This Thesis*) covers a subset of mathematical and statistical concepts that will arise at least once later in the thesis. This section provides the necessary background knowledge to comprehend them effectively.

A.1 INTRODUCTION TO MACHINE LEARNING

A.1.1 Neural Networks

A neural network (NN) is a simplified model of interconnected artificial neurons, as proposed in 1943 by McCulloch & Pitts. A neuron with three inputs is presented below.



Weights w_1, w_2, w_3 are numerical values associated with the connections between the input layer (inputs x_1, x_2, x_3) and the output y . The first step to calculating y is multiplying the input values and the corresponding weights. A bias term b is added to the result, and finally the **activation function** f is applied, yielding the following equation for y :

$$y = f\left(\sum_{i=1}^N x_i w_i + b\right)$$

An example of a commonly used activation function is rectified linear unit (ReLU), which nullifies negative outputs: $f(x) = \frac{x+|x|}{2}$.

A feed-forward neural network (FFNN), the simplest kind, is a network of multiple layers of artificial neurons (Figure A.1). It consists of an input layer, an output layer, and at least one hidden layer¹ in between. It can be seen as a directed acyclic graph: the connections always flow forward from layer l_i to the next layer l_{i+1} . The network is fully connected, which means that all neurons in layer l_i are connected to all neurons in layer l_{i+1} .

SOFTMAX

Softmax is a type of activation function, albeit applicable to vectors rather than individual values, and in machine learning it is used to interpret the outputs of a network as a

¹ A neural network is called deep if it consists of multiple hidden layers.

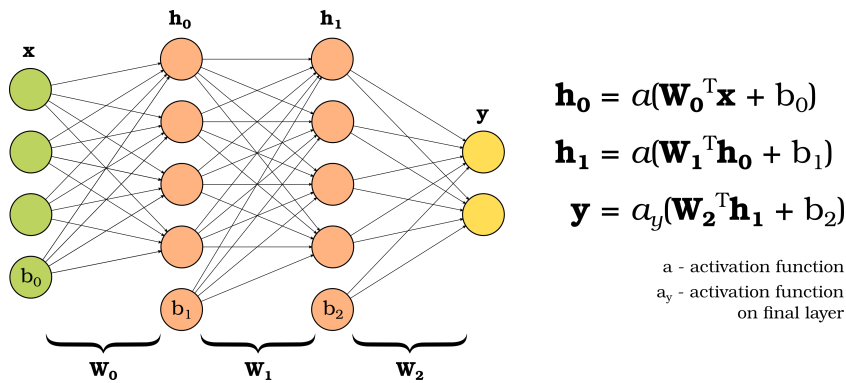


Figure A.1: Example of a feed-forward neural network with $k = 2$ hidden layers. x represents the input, W_0, W_1, W_2 represent the weights between the input layer and the first hidden layer, the hidden layers and the last hidden layer and the output layer, consecutively. A bias term b is added during each computation. The activation function a_y is different from a as it is applied before the last layer (see later).

probability distribution. Given a vector $\mathbf{v} \in \mathbb{R}$, $\text{softmax}(\mathbf{v})$ converts \mathbf{v} into a probability distribution:

$$\text{softmax}(\mathbf{v}) = \frac{\exp(v_i)}{\sum_{j=0}^{|\mathbf{v}|-1} \exp(v_j)} \text{ for } i \text{ in } 1, 2, \dots, |\mathbf{v}|$$

For example,

$$\text{softmax} \left(\begin{bmatrix} 0.3 \\ 2.4 \\ -1.2 \\ 0.5 \\ 1.5 \end{bmatrix} \right) = \begin{bmatrix} 0.07 \\ 0.59 \\ 0.01 \\ 0.09 \\ 0.24 \end{bmatrix}$$

With softmax, we can produce a probability distribution over categories (for classification tasks), words (for most textual applications like translation) and many other types of information. In [Figure A.1](#), softmax could be used as f_y .

TRAINING A NEURAL NETWORK An NN is trained² to learn parameters \mathbf{W}_j and \mathbf{b}_j for each layer j such that the generated probability distribution $\hat{\mathbf{y}}$ is as close to the true probability distribution \mathbf{y} as possible. In order to achieve that, a loss function is defined which expresses the error of the model's predictions w.r.t. the gold standard answer.

² We focus on supervised learning only.

An optimiser algorithm, such as Adam (Kingma & Ba 2015) is then used to minimise that loss function. The cross-entropy loss (§A.3) lends itself especially well to textual applications of neural networks as it is an easy and reliable way of comparing two probability distributions.

The goal of training a network is to enable it to produce outputs for unseen examples by learning from data. Crucially, a well-trained network should generalise effectively to unseen data. Merely performing well on the training data is insufficient, as overfitting can lead to poor performance on new samples. Techniques such as dropout (§A.1.1) can be used to minimise this generalisation error.

To measure the generalisability of a network, one typically splits the available data into three sets: a **training** set (which the model directly adapts to), a **validation** set (which the model does not see during training, but performance on which can be computed periodically to verify whether the model's performance on a held-out set improves), and a **testing** set (which is not used at all during training and only afterwards to verify final performance). When training a neural model, one typically seeks to achieve a set of weights which:

1. performs well on the training data (i.e. over time reduces the error on training batches);
2. performs as well possible on the held-out validation data;
3. generalises well to the held-out testing data.

The point at which a network is considered to be optimally trained and the procedure should cease is described by a stopping criterion (§A.1.1).

STOPPING CRITERIA

A neural network is trained to approximate the function given by the set of the training examples, leading to continuous improvement in its performance on that training set. To maximise the generalisation power of the network, it is ideal to halt the training process when the performance on the held-out validation set no longer shows improvement. However, a decline in performance during one validation step does not necessarily indicate full optimisation, since this performance can fluctuate. Common stopping criteria then involve either halting the training after a specific number of epochs without improvement (referred to as *patience*) or when the improvement fails to meet a minimum threshold. Alternatively, the stopping criterion can be based on a predetermined number of training updates, irrespective of the model's performance. The choice of the stopping criterion depends on factors such as network size, network type, dataset characteristics, and the specific task at hand.

HYPERPARAMETER SEARCH

NNs training is governed by several tunable parameters, such as learning rate (the coefficient for changing the model weights in response to the calculated loss), batch size (number of examples to process in one step) and dropout (§A.1.1). For a network to train optimally, a suitable values for these parameters must be found, as there is no particular set of values which performs optimally for any problem. Such values can be obtained via a **hyperparameter search**. Let A, B, C be the parameters we wish to tune and $[A_0, A_k], [B_0, B_k], [C_0, C_k]$ be feasible ranges for candidate values for each hyperparameter. Two typical strategies for the search are:

- **grid search**: selecting (n_A, n_B, n_C) values from the ranges for each parameter in order, then training a model copy with each parameter combination ($n_A \times n_B \times n_C$ combinations total);
- **randomised search**: selecting m random combinations of randomly selected parameters from the given ranges (m combinations total).

In both cases, the best combination results in a model which performs best on the validation data according to a pre-selected metric, such as the validation loss or a downstream performance metric like BLEU (§A.2.3.2).

DROPOUT

By now a default tactic in many models, the dropout regularisation technique (Srivastava et al. 2014) randomly sets the output of certain nodes to zero during each forward and backward pass, which can be seen as a way of simulating multiple sub-models within a single model and approximating their average output.

A.1.2 Model Architectures and Components

A.1.2.1 Model architectures

RECURRENT NEURAL NETWORK

Recurrent neural networks (RNNs) were the foundation of most state-of-the-art architectures for neural machine translation until the release of the Transformer (§A.1.2.1). The RNN can be thought of as a neural network spread through time: at each timestep t , the hidden state \mathbf{h}_t is re-computed from the previous hidden state \mathbf{h}_{t-1} and based on current input \mathbf{x}_t (Lipton et al. 2015):

$$\mathbf{h}_t = a(\mathbf{W}_{hx}\mathbf{x}_t, \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h)$$

where a is the activation function. the output \hat{y}_t is produced as a function of the hidden state \mathbf{h}_t :

$$\hat{y}_t = \text{softmax}(\mathbf{W}_{y/h}\mathbf{h}_t + \mathbf{b}_y)$$

This initial formulation of RNNs imposed a fixed-size limitation on the input and output of the network; in contrast, for applications like NMT it is desirable that the length of the output is not constrained by the length of the input. The encoder-decoder model mentioned earlier (Sutskever et al. 2014) was originally proposed precisely to address this shortcoming of RNNs: they used one RNN as the encoder and another as the decoder.

Despite their success and widespread use, RNNs were inherently non-parallelisable and could not handle long inputs reliably. The Transformer (§A.1.2.1) addresses these two shortcomings: rather than process the input token by token like RNNs, this non-recurrential sequence-to-sequence (seq2seq) (§A.1.2.1) model processes it simultaneously via a parallelisable self-attention mechanism.

THE TRANSFORMER

The Transformer (Figure A.2, Vaswani et al. 2017) is the first model completely based on the attention mechanism (§A.1.2.2) and variants of it are at the foundation of most contemporary state-of-the-art models in seq2seq and language modelling tasks. It consists of an encoder and a decoder, each of which is a stack of N layers³. Typically, each layer consists of at least one multi-head attention mechanism (§A.1.2.2): a Transformer for NMT uses one self-attention in each encoder layer and a self-attention + an encoder-decoder attention in each decoder layer. The source information flows through all the encoder layers sequentially, and results in a feature matrix \mathcal{C} which can be seen as an encoded, self-contextualised version of the input. The target information flows through the N decoder layers similarly: its self-attention self-contextualises this information, while the encoder-decoder attention contextualises it with \mathcal{C} . The output of the last decoder layer is subjected to a linear transformation and finally, the softmax operation (§A.1.1) is applied to the resulting values in order to obtain a probability distribution.

Since the only interaction between tokens in a Transformer is implemented via attention layers, which are permutation equivariant, the model has no inherent notion of token positions in sequences. To address this, positional encoding (§A.1.2.2) is used. The original Transformer uses a cosine position embedding.

A.1.2.2 The attention mechanism

At the core of the attention mechanism are *keys* (\mathbf{K}), *queries* (\mathbf{Q}) and *values* (\mathbf{V}). \mathbf{K} is the encoding of the data (e.g. word embeddings), on which we wish to compute the

³ The original publication, Vaswani et al. (2017), uses $N = 6$ for base models and $N = 12$ for big models.

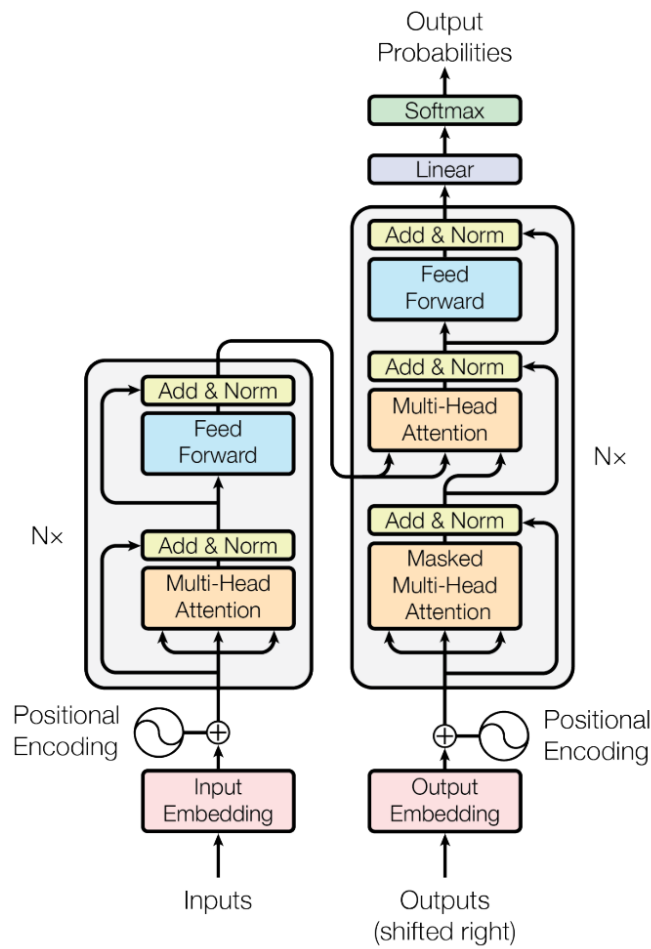


Figure A.2: The Transformer architecture (Vaswani et al. 2017, p. 3).

attention vector. \mathbf{V} can be seen as a different representation of the data in \mathbf{K}^4 . \mathbf{Q} is a reference used to compute the attention. For example, in Vaswani et al. (2017), \mathbf{Q} corresponds to a particular word when \mathbf{K} corresponds to all words in the sequence. Note that \mathbf{K} , \mathbf{Q} and \mathbf{V} are all matrices.

We use \mathbf{Q} and \mathbf{K} to compute the energy scores e using a compatibility function f :

$$e = f(\mathbf{Q}, \mathbf{K})$$

In Vaswani et al. (2017), this amounts to computing the importance of the word denoted by \mathbf{Q} with the rest of the sequence (denoted by \mathbf{K}). This result is obtained with the *dot*

⁴ Galassi et al. (2020) note that in some architectures $\mathbf{V} = \mathbf{K}$.

product of \mathbf{K} and \mathbf{Q} , scaled by $\frac{1}{\sqrt{d_k}}$ to achieve more stable gradients. Together, this yields a *scaled multiplicative* compatibility function.

We transform the energy scores into attention weights using a *distribution function*, g :

$$\mathbf{a} = g(e)$$

In the Transformer architecture, the distribution function used is the softmax (§A.1.1). Finally, we apply the attention weights to values \mathbf{V} by multiplying the values with \mathbf{a} :

$$\mathbf{z} = \mathbf{aV}$$

Later on we will explore how the resulting vector \mathbf{z} is parsed further by the Transformer.

Together, within Vaswani et al. (2017), the application of attention results in the following vector \mathbf{Z} :

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (\text{A.1})$$

The calculation of \mathbf{K} , \mathbf{V} and \mathbf{Q} itself is the result of the following operations on the input matrix \mathbf{X} :

$$\mathbf{K} = \mathbf{W}^K\mathbf{X}$$

$$\mathbf{V} = \mathbf{W}^V\mathbf{X}$$

$$\mathbf{Q} = \mathbf{W}^Q\mathbf{X}$$

where $\mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^Q$ are weight matrices which are initialised to random values and trained along with the model.

MULTI-HEAD ATTENTION

In Vaswani et al. (2017), the context vectors are calculated in multiple distinct heads (*multi-head* attention), and then merged together and multiplied with a weight matrix to arrive at the final embedding. The reasoning behind using multiple heads is to enlarge the representation space. Since all heads are initialised randomly, they can impact the sequence differently, producing more diverse hypotheses.

The self-attention described above is computed in h separate heads (within this thesis, usually either 8 or 16), in a setting called *multi-head attention*. Each head has its own $\mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^Q$ matrices, resulting in different attention matrices. The final vectors \mathbf{z} are then concatenated and multiplied with matrix \mathbf{W}^O into a single matrix \mathbf{z} (Vaswani et al. 2017):

$$\begin{aligned} \text{MultiHead}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O \\ \text{where } \text{head}_i &= \text{Attention}(\mathbf{KW}^K_i, \mathbf{VW}^V_i, \mathbf{QW}^Q_i) \end{aligned}$$

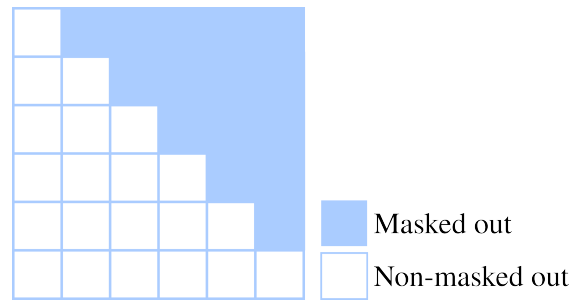


Figure A.3: Example of a mask used to nullify attention given to future words ($n = 6$). The mask adds negative infinities to the attention matrix in places which correspond to future tokens.

Multi-head attention is the basic implementation of attention used in the model. Attention blocks appear in the model under three labels:

- **self-attention in the encoder:** \mathbf{K} , \mathbf{V} and \mathbf{Q} come from the output of the previous encoder. In the first encoder they come from the input embedding.
- **cross-attention in the decoder:** \mathbf{K} , \mathbf{V} come from the output of the encoder while \mathbf{Q} comes from the previous decoder layer.
- **masked self-attention in the decoder:** \mathbf{K} , \mathbf{V} and \mathbf{Q} similarly come from the output of the previous decoder and in the first decoder they come from the output embedding. Values to the right of each token are masked out to prevent illegal flow of information. During inference, the translation hypothesis is produced one token at a time. When token t_i is generated, the model only has access to tokens $T_{<i} \in \{t_1, t_2, \dots, t_{i-1}\}$. The decoder can only make use of links between t_i and the words that occurred before it (and itself). But the original self-attention mechanism relates every word to all other words in the sequence. In order to re-use the same mechanism in the decoder, succeeding tokens for each token t_i are masked out (Figure A.3).

The FFNN connection in the encoder is calculated for each word separately, and in the following way:

$$\begin{aligned}\mathbf{h} &= \mathbf{W}_0^T \mathbf{x} + \mathbf{b}_0 \\ \mathbf{z} &= \text{ReLU}(\mathbf{h}) \\ \mathbf{f} &= \mathbf{W}_1^T \mathbf{z} + \mathbf{b}_1\end{aligned}$$

Let the FFNN and Attention blocks described above be sub-layers. The authors employ a residual connection around each sub-layer in the architecture. The residual connection

facilitates the addition of the input to the output of the sub-layer ($x + \text{Sublayer}(x)$). The result of the addition is also normalised (yielding $\text{LayerNorm}(x + \text{Sublayer}(x))$).

The purpose of residual connections is to retain information from earlier layers in later layers. For example, it helps to efficiently propagate the positional encoding information to further layers.

QUERY-KEY NORMALISATION

Query-key normalisation [Henry et al. \(2020\)](#) applies l^2 normalisation to \mathbf{Q} and \mathbf{K} matrices and scales them up by a learnable parameter g (initialised to the 97.5 percentile of training sequence lengths); the attention equation shown in [Equation A.1](#) becomes

$$\mathbf{Z} = \text{softmax}(g \times l_2(\mathbf{Q})l_2(\mathbf{K})^T)\mathbf{V}$$

POSITIONAL EMBEDDING

Because the attention mechanism does not inherently encode position information which is a crucial feature of text, **positional encoding** is necessary. Originally, [Vaswani et al. \(2017\)](#) implement absolute position embeddings (APE) by adding sine and cosine functions of varying frequencies to token embeddings; later approaches such as BERT ([Devlin et al. 2019](#)) replace the sin/cos approach with learned embeddings: a randomly-initialised vector for any integer position k up until the limit K , $k \in \{0, 1, \dots, K\}$ is adapted during training.

A.1.3 Learning Paradigms

ZERO-SHOT LEARNING

Zero-shot learning refers to the paradigm in machine learning where a model is enabled to exhibit behaviour unseen during training. Typically the training and test data in an experimental setup will share some common specifications. For example, in machine translation the underlying assumption could be that the source language is always Ukrainian and the target language is always Polish. If either or both languages changed during testing, then the model would be queried for zero-shot adaptation to other languages.

FEW-SHOT LEARNING

The few-shot learning paradigm is analogous to zero-shot except the model is trained on only a **few** examples of the downstream task, as opposed to a standard paradigm where the model is adapted with many examples.

A.2 NATURAL LANGUAGE PROCESSING

A.2.1 Tokenisation

Tokenisation is the process of converting a string S consisting of characters into an array T of individual *tokens* based on a delimiter character d or a tokenisation algorithm in such a way that S can be unambiguously recovered from T . The primary function of tokenisation is enabling the conversion of input text S to an array of units interpretable by text processing systems, such as FFNNs. Since any text is described by a finite number of tokens, a complete vocabulary of these tokens V can be collected. For processing with neural networks, the tokens in V can each be assigned a unique embedding. The most common delimiter d is the space character (" "), resulting in V of all words in the text. However, for large datasets V will become too big for standard networks, surpassing the memory capacity of systems or resulting in training and inference slowdowns. There are two solutions to this problem:

1. selecting a threshold t_V for the capacity of V and only storing the most frequent t_V tokens, treating the remaining tokens as unknown;
2. employing an alternative tokenisation method, such as a sub-word algorithm.

SUB-WORD TOKENISATION

Due to the constraints on the vocabulary size, models in NLP in the past have struggled with rare words (Sennrich et al. 2016d). This issue is especially prominent in languages such as German where words are often formed by compounding. For example, the phrase *suggestion for improvement* can be translated to German as *der Verbesserungsvorschlag*. Intuitively, a more successful tokenisation method would recognise such compounds and split them into the individual items, and optionally the connecting element (e.g. tokenise the string "der Verbesserungsvorschlag" as ["der", " ", "Verbesserung", "s", "vorschlag"]). This is the motivation behind **sub-word tokenisation** which treats sub-word units as tokens, meaning one word can be tokenised as one or more tokens. There are three main sub-word tokenisation algorithms: **Byte-Pair Encoding (BPE)** (Sennrich et al. 2016d), **unigram** (Kudo 2018) and **WordPiece** (Wu et al. 2016). Algorithm 2 describes the BPE algorithm which is used as the tokenisation method throughout the thesis.

Algorithm 2: Byte-Pair Encoding

Input: C, n_V ▷ Training corpus C ; Vocabulary size n_V .
Output: V ▷ Vocabulary V of size n_V built from C .

Function BuildVocab(n_V, C):

```

 $V \leftarrow$  vocabulary of all characters in  $C$ 
while  $|V| < n_V$  do
   $a, b \leftarrow$  two most frequent consecutive tokens from  $V$  in  $C$ 
   $a \cdot b \leftarrow$  concatenation of  $a$  and  $b$ 
   $V = V \cup a \cdot b$ 
end
return  $V$ 

```

A.2.2 Text Embeddings

In order to apply NNs in practice we need a method for realising our input as a vector of numerical values, and for making sense of the numerical values in the output. Textual applications of NNs in particular require that a mapping is created between text numerical values on which the network operates. This is commonly done with **word embeddings**, where each *word* (or sub-word) is treated as a **token** and mapped to a vector, or **sentence embeddings**, where each *sentence* is mapped to one vector.

WORD EMBEDDINGS

Given the input sequence of words $T = t_1, t_2, \dots, t_k$, each word t_i is represented as a vector of real numbers \mathbf{x}_i . This can be seen as a preliminary neural transformation where T becomes \mathbf{X} through multiplication by embedding matrix W_e . Given a set of vocabulary words V , \mathbf{X} is obtained by first creating a **one-hot encoding** of token t_i and multiplying it with the embedding matrix to obtain the embedding \mathbf{x}_i (Figure A.4). The embedding matrix can either be trained with the rest of the network, or pre-trained (e.g. GLoVe, Pennington et al. 2014) and frozen (kept fixed during training).

SENTENCE EMBEDDINGS

Sentence embeddings map sentences to real-valued vectors. The definition of *sentence* here is loose and the embeddings can be successfully applied to words, sentences and paragraphs alike, so long as the text does not exceed the token limit of a particular model. Sentence embeddings are useful when one needs to compare how similar two sentences are, in tasks such as semantic search.

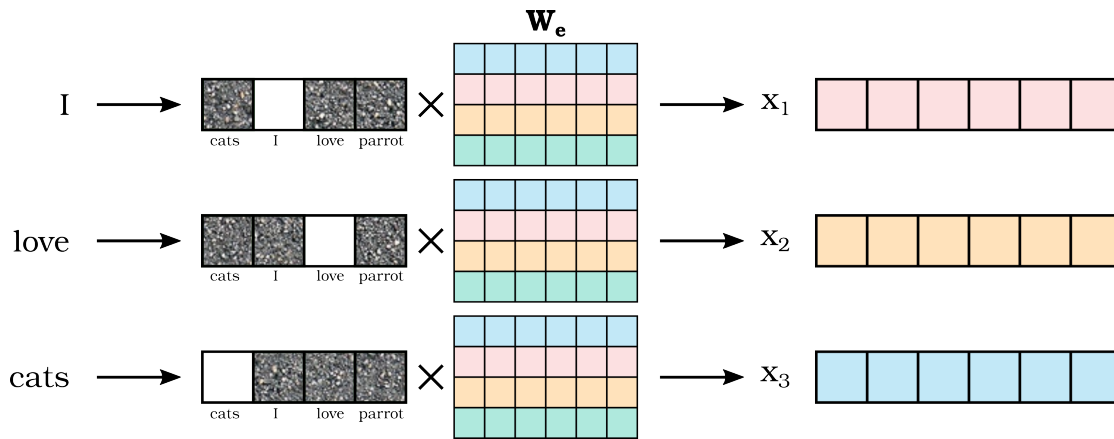


Figure A.4: Example of parsing input tokens to numerical values with word embeddings, where input = “I love cats” and vocabulary = [cats I love parrot]. First, input tokens are one-hot encoded according to their positions in the vocabulary. Then, they are multiplied by the embedding matrix W_e .

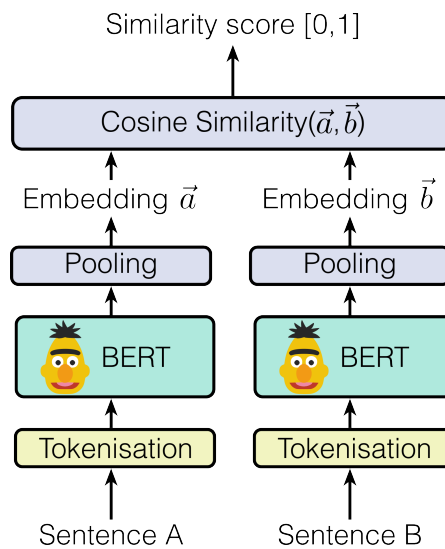


Figure A.5: The siamese network architecture of Sentence-BERT (SBERT). The parameters of the BERT models are shared.

The state-of-the-art approach to computing sentence embeddings was proposed by Reimers & Gurevych (2019b) who introduced the SBERT (Sentence-BERT) architecture, which finetunes BERT (Devlin et al. 2019) in a siamese or triplet network configuration (Figure A.5). Sentence A and Sentence B are tokenised and the representations of their

tokens is computed using BERT. Then those representations are pooled (converted to a single vector via a pre-defined operation such as averaging), resulting in embeddings \vec{a} and \vec{b} respectively. As the output of the siamese network, cosine similarity is computed between the two vectors, and compared to the gold standard answer, updating the parameters of the BERT model according to the training criterion. The training data comprises pairs A, B which are either identical in meaning (with a cosine similarity of 1) or completely dissimilar (with a cosine similarity of 0).

A.2.3 Neural Machine Translation

Neural machine translation (NMT) involves translating a sequence of tokens in the source language, $\mathbf{x} = \{x_1, \dots, x_t\}$ to a sequence of tokens in the target language, $\mathbf{y} = \{y_1, \dots, y'_t\}$. Specifically, we aim to find $\hat{\mathbf{y}}$ such that

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} P_{\theta}(\mathbf{y}|\mathbf{x})$$

where θ is the matrix of weights in the NMT model. We find the translation by first training a neural model to learn the conditional probability distribution, and then given the source sentence \mathbf{x} , we search for the target sentence \mathbf{y} which maximises the conditional probability (Bahdanau et al. 2015). NMT is based on seq2seq models which use the **encoder-decoder** architecture. The original idea behind seq2seq was to encode the input sequence \mathbf{x} into a fixed-length vector \mathbf{c} , and then decode the output sequence \mathbf{y} from \mathbf{c} .

$$\log(p_w(y_i|x_i)) = f_D(f_E(x)) = f_D(c) = y$$

In such a setting, the encoder and the decoder can be trained together to maximise the probability of a correct translation given a source sentence:

$$\max_w \frac{1}{N} \sum_{i=1}^N \log(p_w(y_i|x_i))$$

where w are the weights (parameters) of the model, and (y_i, x_i) pairs respectively represent the correct translation and source sentence examples from a sentence-aligned (parallel) corpus C , $|C| = N$.

Bahdanau et al. (2015) observe that the encoder-decoder architecture forces too strict a compression of information with its single vector \mathbf{c} . They instead propose the **attention** architecture, which holds one entry per word and, when processing a particular word, allows to compute a relevance weight for all the other words. At its heart, attention computed for a vector $\mathbf{x}^{1 \times d}$ is a vector $\mathbf{a}^{1 \times d}$ of weights, $a_i \in [0, 1]$. This idea is based on

the observation that when humans translate, they pay attention to specific elements of the sequence rather than the whole sequence equally. Since its formulation, attention has led to the development of many groundbreaking architectures, among which is the foundation of the current state-of-the-art approach to neural machine translation, i.e. the Transformer architecture (§A.1.2.1).

A.2.3.1 Back-translation

Back-translation (Sennrich et al. 2016c, Edunov et al. 2018) is a data augmentation technique used to obtain a parallel corpus $(X_{btr(Y_{mono})}, Y_{mono})$ given a monolingual corpus Y_{mono} and a parallel corpus (X_p, Y_p) . Let src, tgt be the source and target languages respectively; let $MT_{(src \rightarrow tgt)}$ denote a machine translation model trained to translate text from src to tgt . Assuming the availability of The process of back-translation to augment (X_p, Y_p) with $(X_{btr(Y_{mono})}, Y_{mono})$ occurs in two steps:

1. A back-translation system $MT_{btr} = MT_{(tgt \rightarrow src)}$ is trained on (X_p, Y_p) .
2. MT_{btr} is used to translate Y_{mono} , yielding $X_{btr(Y_{mono})}$.

A forward translation model $MT_{(src \rightarrow tgt)}$ can then be trained on the concatenation of (X_p, Y_p) and $(X_{btr(Y_{mono})}, Y_{mono})$.

A.2.3.2 Evaluation

Machine translation evaluation (MTE) serves to produce a judgement on MT quality, either automatically (with algorithms or pre-trained models) or through manual human judgements. Traditionally this is done by comparing the translation hypothesis to a human-written reference translation, though some methods (like quality estimation) are reference-free, relying only on the source sentence and the hypothesis.

Some automatic MTE metrics like BLEU and COMET, being particularly popular, serve as the cornerstone of measuring progress in the field, by tracking results obtained on the same test sets over time. Automatic MTE can also aid the development of MT systems, indicating the efficiency of prototypes. Human assessment on the other hand is too expensive to use in development, but irreplaceable when it comes to evaluation. Automatic metrics can be faulty: n-gram-based metrics (e.g. BLEU) are notorious for punishing hypotheses which use synonymous translation alternatives, being based on lexical identity between tokens in the hypothesis and the reference, while learned metrics (e.g. COMET) can exhibit social biases, a characteristic shared by most trained neural networks.

BLEU

Given a reference translation and a translation hypothesis, BLEU calculates the n-gram overlap between them:

$$\text{BLEU} = \text{BP} \times \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

where BP is the brevity penalty, defined as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

and p_n is the modified precision score, calculated as follows:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')}$$

where *Candidates* are the sentences to be evaluated, $\text{Count}(\text{n-gram})$ is the number of times the n-gram appears in C, and $\text{Count}_{\text{clip}}$ is the same but clipped to the maximum number of occurrences of the same n-gram in the reference (extra counts are treated as errors).

BLEU has attracted attention of researchers as one of the first automatic MTE which correlated well with human judgements. Unfortunately, several scholars have pointed out that this metric is not without its problems (Callison-Burch et al. 2006, Reiter 2018). For example, since it is based entirely on lexical matching, it will fail to recognise a word's synonym as its viable translation.

CHRF++

Character n-gram F-score (chrF) (Popovi 2017) compute the character n-gram overlap between hypothesis and reference. First, we compute the precision and reference scores:

$$\begin{aligned} \text{CHRP} &= \frac{\# \text{ of character n-grams in hyp. \& ref.}}{\# \text{ of character n-grams in hyp.}} \\ \text{CHRR} &= \frac{\# \text{ of character n-grams in hyp. \& ref.}}{\# \text{ of character n-grams in ref.}} \end{aligned}$$

The chrF score is defined as follows:

$$\text{CHRF}(\beta) = (1 + \beta^2) \frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}$$

β balances the importance of recall w.r.t. precision ($\beta = 1$ makes them equally important). Finally, chrF++ (Popovi 2017) is a better-performing variation on CHRf in which both character and word n-gram scores are computed and averaged together. Popovi comment that $n = 6$ yields the best correlation to human judgement for character n-grams and $n = 1$ or $n = 2$ works best for word n-grams (and throughout this work, a value of $n = 2$ is used).

COMET

COMET (Rei et al. 2020) combine the ideas from two metrics: it uses contextual embeddings like BERTSCORE (Zhang et al. 2019) and is trained to optimize human correlation, like RUSE (Shimanaka et al. 2018). Let s, r be the source sentence and the reference, and h be the translation hypothesis. Using contextual embeddings, COMET calculates the following distances between h and s or r : element-wise products ($h \odot s$ and $h \odot r$) and absolute element-wise differences ($|h - r|$ and $|h - s|$). They append the distances to reference and hypothesis embeddings and train a feed forward neural network to minimise the error between the scores produced and human quality assessments.

Despite more sophisticated metrics achieving better human correlation as measured by the annual Metrics task at WMT (e.g. Mathur et al. 2020), BLEU remains the most commonly used one.

A.2.4 Language Modelling

Language modelling refers to the task of determining the probability of a set of words (or tokens) occurring in a sequence. The most fundamental language models (LMs) are **n-gram** models which operate on the assumption of the Markov property, where the probability of a word depends on a limited (typically fixed) number n of preceding words. Mathematically, the probability distribution of an n-gram language model is expressed as:

$$p(w_k | w_{k-1}, w_{k-2}, \dots, w_1) \approx p(w_k | w_{k-1}, w_{k-2}, \dots, w_{k-(n-1)})$$

While n-gram models are computationally efficient and can capture local dependencies, they struggle to handle long-range dependencies and suffer from the sparsity of data when faced with unseen n-grams. Contemporary language models address these shortcomings by increasing the **context window** (n) and leveraging powerful mechanisms such as self-attention to capture complex relationships between words. While these models still use the Markov property, increasing n to several thousand tokens yields significantly more contextually grounded models. However, these improvements come at the cost of training and inference efficiency when compared to n-gram models.

A.2.4.1 Evaluation

Language models can be evaluated *intrinsically* (i.e. on how well they capture a given text) and *extrinsically* on downstream tasks they are later used on, such as language understanding (e.g. GLUE, Wang et al. 2018). In this section, we focus only on metrics used later in the thesis: perplexity (§ A.2.4.1) and mean reciprocal rank (§ A.2.4.1). Perplexity is an intrinsic measure while mean reciprocal rank is an extrinsic one.

PERPLEXITY

Perplexity is an automatic measure of how well the learned probability distribution of words matches the distribution of the given text. Given a causal LM with parameters θ , the perplexity for a sequence $S = (s_0, s_1, \dots, s_n)$ is given by Equation A.2:

$$\text{PERPLEXITY} = \sqrt[n]{\frac{1}{\sum_i^n \log p_\theta(s_i | s_{<i})}} \quad (\text{A.2})$$

MEAN RECIPROCAL RANK

Mean reciprocal rank (MRR) captures the effectiveness of a system which returns a list of ranked results to a query. Let q^i be a query and a^i be a list of n possible answers $a^i \in (a_1^i, a_2^i, \dots, a_n^i)$ ordered descendingly from highest-ranked. Let rank_{q^i, a^i} denote the position of the (first) correct item in a^i . Then, the reciprocal rank is the multiplicative inverse of that position, and, when q belongs to a list of queries Q , the **mean reciprocal rank** is the average of the reciprocal ranks computed for all queries in Q :

$$\text{MRR} = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{\text{rank}_{q^i, a^i}}$$

Given a list of predicted tokens (t_0, t_1, \dots, t_k) , an language model can be queried to produce a probability distribution over the vocabulary for the next word t_{k+1} . If there exists a gold-standard answer for t_{k+1} , then MRR can be computed on the probability-ranked list of predicted tokens.

A.2.5 Beam Search

Textual output of a neural network is traditionally generated in an *autoregressive* way, i.e. the text is decoded in a pre-determined direction (usually left-to-right) and one token at a time. The simplest approach to decoding a sequence of tokens, referred to as greedy decoding, simply takes the most conditionally probable token at each step. An alternative and more robust strategy (used particularly in NMT) is beam search, which can be seen as a width-limited breadth-first graph search algorithm, where width is

usually referred to as the *beam size*. When the beam size is equal to one, beam search falls back to greedy decoding.

Let k be the beam size and v be the vocabulary size. At each step of the decoding process the k hypotheses with the highest cumulative probability are kept. Given the k hypotheses obtained from step i , at step $i + 1$ a token is conditionally predicted from each of the top hypotheses, yielding $k \times v$ new hypotheses from which the most probable k are selected as the result of step $i + 1$.

A.3 STATISTICAL CONCEPTS EMPLOYED WITHIN THIS THESIS

CLASSIFICATION PERFORMANCE METRICS

We consider four performance metrics applicable to tasks with a binary outcome (e.g. correct or incorrect classification) for each sample from a collection.

Let TP , TN , FP , and FN denote the true positive, true negative, false positive, and false negative outcomes of an experiment, respectively. Then,

- **ACCURACY** is defined as the sum of correct outcomes (TP , TN) divided by the sum of all outcomes:

$$\text{ACCURACY} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **PRECISION** measures the proportion of correctly classified positive samples out of all samples predicted as positive. It quantifies the accuracy of positive predictions. Precision is computed using the formula:

$$\text{PRECISION} = \frac{TP}{TP + FP}$$

A high precision indicates a low rate of false positives, meaning that the model is reliable when it predicts a positive outcome.

- **RECALL** measures the proportion of correctly classified positive samples out of all actual positive samples. It quantifies the ability of the model to identify positive samples correctly. Recall is calculated using the formula:

$$\text{RECALL} = \frac{TP}{TP + FN}$$

A high recall indicates a low rate of false negatives, indicating that the model is effective in capturing positive samples.

- **F1 SCORE** is the harmonic mean of precision and recall, providing a single metric that balances both measures. It combines precision and recall into a single value that summarises the model's performance. The F_1 score $\in [0, 1]$ is calculated as:

$$F_1 \text{ SCORE} = 2 \times \frac{\text{PRECISION} \times \text{RECALL}}{\text{PRECISION} + \text{RECALL}}$$

CROSS-ENTROPY

Cross-entropy is a measure of the differences between two probability distributions. It quantifies the average number of bits needed to represent the true distribution p compared to the predicted distribution q :

$$H(p, q) = - \sum_{i=1}^n p(x_i) \log(q(x_i))$$

where the probability distributions $p(i)$ and $q(i)$ are obtained over n possible outcomes.

LABEL-SMOOTHED CROSS-ENTROPY

Label-smoothed cross-entropy, often used when training the Transformer (§A.1.2.1) for the NMT task, is a variation on the cross-entropy function. It adds a small amount of uncertainty to the true predictions, preventing the overconfidence of the model in what it has predicted. This also effectively encourages it to learn more robust and calibrated probabilities. While the standard cross-entropy uses one-hot encoded labels for each sample, the label-smoothed variant assigns a positive value $\epsilon \in [0, 1]$ to the true class and the remaining probability mass $(1 - \epsilon)$ is distributed equally among the remaining classes.

POINTWISE MUTUAL INFORMATION

PMI measures the level of association between two events. Given two events **A** and **B**, **PMI** is computed as the logarithmic probability of the joint occurrence of **A** and **B**, divided by the product of their individual probabilities:

$$PMI(\mathbf{A}, \mathbf{B}) = \log \frac{P(\mathbf{A}, \mathbf{B})}{P(\mathbf{A}) \times P(\mathbf{B})}$$

PMI indicates how much more likely it is for **A** and **B** to occur together than if they were independent; positive **PMI** indicates positive correlation, zero implies independence and negative **PMI** indicates negative correlation.

A.3.1 Statistical Significance Testing

BOOTSTRAP RESAMPLING

When comparing two machine translation systems, \mathcal{M}_1 and \mathcal{M}_2 on a given test set with a quantitative metric like BLEU (§A.2.3.2), an important thing to consider is how to ensure that the difference in scores is sufficient to claim one system significantly better than the other. Koehn (2004) propose that the statistical significance of this difference can be computed with bootstrap resampling: given a test set of n sentences, we sample (with replacement) a subset of k sentences from this collection, and compute the quality score for \mathcal{M}_1 and \mathcal{M}_2 . We repeat this procedure a large number of times (e.g. 1000). If e.g. 95% of the time \mathcal{M}_1 is better (i.e. obtains a higher score) than \mathcal{M}_2 , then we can conclude with 95% certainty that it is the better system.

T-TEST

A t-test⁵ is a statistical test which compares the means of two groups of measurements, typically to determine whether there is a statistically significant difference between them. In the context of results from evaluating a model, given the baseline model M and the tested model M' and the results from n separate runs for each, we can calculate a t -value based on an array of differences of model scores δ as follows:

$$t = \frac{\bar{\delta}}{\sigma(\delta)/\sqrt{n}}$$

where $\bar{\delta}$ corresponds to the mean and $\sigma(\delta)$ to the standard deviation of the differences. The obtained t -value is then compared to the critical t -value at the selected level of significance (expressed as a confidence interval or a p -value) and degrees of freedom (i.e. $n - 1$). The result is statistically significant if the observed t -value is greater than the critical t -value.

⁵ Within this thesis we only consider a one-tailed t-test, which assumes a specific direction of change in results, for example when we compare an improved model to a baseline one.

BIBLIOGRAPHY

Anastasopoulos, A., Barrault, L., Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., Currey, A., Dinu, G., Duh, K., Elbayad, M., Emmanuel, C., Estève, Y., Federico, M., Federmann, C., Gahbiche, S., Gong, H., Grundkiewicz, R., Haddow, B., Hsu, B., Javorský, D., Kloudová, V., Lakew, S., Ma, X., Mathur, P., McNamee, P., Murray, K., Nădejde, M., Nakamura, S., Negri, M., Niehues, J., Niu, X., Ortega, J., Pino, J., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Turchi, M., Virkar, Y., Waibel, A., Wang, C. & Watanabe, S. (2022), Findings of the IWSLT 2022 evaluation campaign, in 'Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)', Association for Computational Linguistics, Dublin, Ireland (in-person and online), pp. 98–157.

URL: <https://aclanthology.org/2022.iwslt-1.10>

Bahdanau, D., Cho, K. H. & Bengio, Y. (2015), 'Neural machine translation by jointly learning to align and translate', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* pp. 1–15.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D. & Zaragoza, J. (2020), ParaCrawl: Web-scale acquisition of parallel corpora, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 4555–4567.

URL: <https://aclanthology.org/2020.acl-main.417>

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M. & Zampieri, M. (2020), Findings of the 2020 conference on machine translation (WMT20), in 'Proceedings of the Fifth Conference on Machine Translation', Association for Computational Linguistics, Online, pp. 1–55.

URL: <https://aclanthology.org/2020.wmt-1.1>

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M. & Monz, C., eds (2021), *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online.

URL: <https://aclanthology.org/2021.wmt-1.0>

- Bawden, R., Sennrich, R., Birch, A. & Haddow, B. (2018a), 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>
- Bawden, R., Sennrich, R., Birch, A. & Haddow, B. (2018b), 'Evaluating discourse phenomena in neural machine translation', *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1, 1304–1313.
- C. M. de Sousa, S., Aziz, W. & Specia, L. (2011), Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles, *in* 'Proceedings of the International Conference Recent Advances in Natural Language Processing 2011', Association for Computational Linguistics, Hissar, Bulgaria, pp. 97–103.
URL: <https://aclanthology.org/R11-1014>
- Callison-Burch, C., Osborne, M. & Koehn, P. (2006), Re-evaluating the role of Bleu in machine translation research, *in* '11th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Trento, Italy, pp. 249–256.
URL: <https://aclanthology.org/E06-1032>
- Dai, N., Liang, J., Qiu, X. & Huang, X. (2019), Style transformer: Unpaired text style transfer without disentangled latent representation, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 5997–6007.
URL: <https://aclanthology.org/P19-1601>
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. & Salakhutdinov, R. (2019), Transformer-XL: Attentive language models beyond a fixed-length context, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 2978–2988.
URL: <https://aclanthology.org/P19-1285>
- Danescu-Niculescu-Mizil, C. & Lee, L. (2011), Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, *in* 'Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics', pp. 76–87.
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

- Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: <https://aclanthology.org/N19-1423>
- Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M. & Turchi, M. (2019), MuST-C: a Multilingual Speech Translation Corpus, in 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 2012–2017.
URL: <https://aclanthology.org/N19-1202>
- Dudy, S., Bedrick, S. & Webber, B. (2021), Refocusing on relevance: Personalization in NLG, in 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 5190–5202.
URL: <https://aclanthology.org/2021.emnlp-main.421>
- Edunov, S., Ott, M., Auli, M. & Grangier, D. (2018), 'Understanding back-translation at scale', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* pp. 489–500.
URL: <https://www.aclweb.org/anthology/D18-1045>
- Edunov, S., Ott, M., Ranzato, M. & Auli, M. (2020), On the evaluation of machine translation systems trained with back-translation, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 2836–2846.
URL: <https://aclanthology.org/2020.acl-main.253>
- Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H. & Osama, A. (2018), 'Gender aware spoken language translation applied to English-Arabic', *2nd International Conference on Natural Language and Speech Processing, ICNLSP 2018* pp. 1–6.
- Feely, W., Hasler, E. & de Gispert, A. (2019), Controlling Japanese honorifics in English-to-Japanese neural machine translation, in 'Proceedings of the 6th Workshop on Asian Translation', Association for Computational Linguistics, Hong Kong, China, pp. 45–53.
URL: <https://aclanthology.org/D19-5203>
- Feldstein, R. F. (2001), *A Concise Polish Grammar*, Slavic and East European Language Research Center (SEELRC), Duke University, 2001.
- Flek, L. (2020), Returning the N to NLP: Towards contextually personalized classification models, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online,

- pp. 7828–7838.
 URL: <https://aclanthology.org/2020.acl-main.700>
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q. & Macherey, W. (2021), ‘Experts, errors, and context: A large-scale study of human evaluation for machine translation’, *Transactions of the Association for Computational Linguistics* **9**, 1460–1474.
 URL: <https://aclanthology.org/2021.tacl-1.87>
- Fu, Y., Zhou, H., Chen, J. & Li, L. (2019), Rethinking text attribute transfer: A lexical analysis, in ‘Proceedings of the 12th International Conference on Natural Language Generation’, Association for Computational Linguistics, Tokyo, Japan, pp. 24–33.
 URL: <https://aclanthology.org/W19-8604>
- Galassi, A., Lippi, M. & Torroni, P. (2020), ‘Attention in Natural Language Processing’, *IEEE Transactions on Neural Networks and Learning Systems* pp. 1–18.
- Gonen, H. & Webster, K. (2020), Automatically identifying gender issues in machine translation using perturbations, in ‘Findings of the Association for Computational Linguistics: EMNLP 2020’, Association for Computational Linguistics, Online, pp. 1991–1995.
 URL: <https://aclanthology.org/2020.findings-emnlp.180>
- Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R. & Hakkani-Tür, D. (2019), Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations, in ‘Proc. Interspeech 2019’, pp. 1891–1895.
 URL: <http://dx.doi.org/10.21437/Interspeech.2019-3079>
- Graham, Y., Federmann, C., Eskevich, M. & Haddow, B. (2020), Assessing human-parity in machine translation on the segment level, in ‘Findings of the Association for Computational Linguistics: EMNLP 2020’, Association for Computational Linguistics, Online, pp. 4199–4207.
 URL: <https://aclanthology.org/2020.findings-emnlp.375>
- Guo, D., Zhang, Z., Fan, P., Zhang, J. & Yang, B. (2021), ‘A context-aware language model to improve the speech recognition in air traffic control’, *Aerospace* **8**(11).
 URL: <https://www.mdpi.com/2226-4310/8/11/348>
- Gupta, P., Sharma, M., Pitale, K. & Kumar, K. (2019), ‘Problems with automating translation of movie/tv show subtitles’, *CoRR* **abs/1909.05362**.
 URL: <http://arxiv.org/abs/1909.05362>
- Halliday, M. A. K. & Hasan, R. (1976), *Cohesion in English*, Longman, London.
- Halliday, M. A. & Matthiessen, C. M. (2013), *Halliday’s introduction to functional grammar: Fourth edition*, Routledge.

- 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.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z. & Zhou, M. (2018), 'Achieving Human Parity on Automatic Chinese to English News Translation', *arXiv* .
- Henry, A., Dachapally, P. R., Pawar, S. S. & Chen, Y. (2020), Query-key normalization for transformers, in 'Findings of the Association for Computational Linguistics: EMNLP 2020', Association for Computational Linguistics, Online, pp. 4246–4253.
URL: <https://aclanthology.org/2020.findings-emnlp.379>
- Hovy, D. (2015), Demographic factors improve classification performance, in 'Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Beijing, China, pp. 752–762.
URL: <https://aclanthology.org/P15-1073>
- Huang, J. & Wang, J. (2023), 'Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators' effort', *Perspectives* 31(4), 620–640.
URL: <https://doi.org/10.1080/0907676X.2022.2026424>
- Huang, Y.-Y., Yan, R., Kuo, T.-T. & Lin, S.-D. (2014), Enriching cold start personalized language model using social network information, in 'Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Baltimore, Maryland, pp. 611–617.
URL: <https://aclanthology.org/P14-2100>
- Ippolito, D., Grangier, D., Eck, D. & Callison-Burch, C. (2020), Toward better storylines with sentence-level language models, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 7472–7478.
URL: <https://aclanthology.org/2020.acl-main.666>
- Jacovi, A., Marasović, A., Miller, T. & Goldberg, Y. (2021), Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI, in 'Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency', FAccT '21, Association for Computing Machinery, New York, NY, USA, p. 624–635.
URL: <https://doi.org/10.1145/3442188.3445923>
- Jassem, W. (2003), 'Polish', *Journal of the International Phonetic Association* 33(1), 103–107.
- Johannsen, A., Hovy, D. & Søgaard, A. (2015), Cross-lingual syntactic variation over age and gender, in 'Proceedings of the Nineteenth Conference on Computational

- Natural Language Learning', Association for Computational Linguistics, Beijing, China, pp. 103–112.
URL: <https://aclanthology.org/K15-1011>
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. & Dean, J. (2017), 'Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation', *Transactions of the Association for Computational Linguistics* 5, 339–351.
- Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M. & Turchi, M. (2022), Post-editing in automatic subtitling: A subtitlers' perspective, in 'Proceedings of the 23rd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Ghent, Belgium, pp. 261–270.
URL: <https://aclanthology.org/2022.eamt-1.29>
- Keown, A. (2003), 'Motivations for Polish pronouns of address', *Glossos* 4(4).
- Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C. & Socher, R. (2019), 'CTRL: A conditional transformer language model for controllable generation', *arXiv* pp. 1–18.
- Khandelwal, U., He, H., Qi, P. & Jurafsky, D. (2018), Sharp nearby, fuzzy far away: How neural language models use context, in 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 284–294.
URL: <https://aclanthology.org/P18-1027>
- Kieras, W. & Wolinski, M. (2017), 'Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego', *Jezyk Polski* 97(1), 75–83.
- Kim, Y., Tran, D. T. & 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 Computational Linguistics, Hong Kong, China, pp. 24–34.
URL: <https://aclanthology.org/D19-6503>
- King, M. & Cook, P. (2020), Evaluating approaches to personalizing language models, in 'Proceedings of the Twelfth Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 2461–2469.
URL: <https://aclanthology.org/2020.lrec-1.299>
- Kingma, D. P. & Ba, J. (2015), Adam: A method for stochastic optimization, in Y. Bengio & Y. LeCun, eds, '3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings'.
URL: <http://arxiv.org/abs/1412.6980>

- Koehn, P. (2004), Statistical significance tests for machine translation evaluation, in 'Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Barcelona, Spain, pp. 388–395.
URL: <https://aclanthology.org/W04-3250>
- Koehn, P. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, in 'Conference Proceedings: the tenth Machine Translation Summit', AAMT, AAMT, Phuket, Thailand, pp. 79–86.
URL: <http://mt-archive.info/MTS-2005-Koehn.pdf>
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007), Moses: Open source toolkit for statistical machine translation, in 'Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions', Association for Computational Linguistics, Prague, Czech Republic, pp. 177–180.
URL: <https://aclanthology.org/P07-2045>
- Koniuszaniec, G. & Błaszczowska, H. (2003), 'Language and gender in Polish', *Gender across Languages* 3, 259–285.
- Koponen, M., Sulubacak, U., Vitikainen, K. & Tiedemann, J. (2020), MT for subtitling: User evaluation of post-editing productivity, in 'Proceedings of the 22nd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Lisboa, Portugal, pp. 115–124.
URL: <https://aclanthology.org/2020.eamt-1.13>
- Kuang, S., Xiong, D., Luo, W. & Zhou, G. (2018), Modeling coherence for neural machine translation with dynamic and topic caches, in 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 596–606.
URL: <https://aclanthology.org/C18-1050>
- Kudo, T. (2018), Subword regularization: Improving neural network translation models with multiple subword candidates, in 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 66–75.
URL: <https://aclanthology.org/P18-1007>
- Kudo, T. & Richardson, J. (2018), SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Brussels, Belgium,

- pp. 66–71.
URL: <https://aclanthology.org/D18-2012>
- Lakew, S. M., Di Gangi, M. & Federico, M. (2019a), Controlling the output length of neural machine translation, in 'Proceedings of the 16th International Conference on Spoken Language Translation', Association for Computational Linguistics, Hong Kong.
URL: <https://aclanthology.org/2019.iwslt-1.31>
- Lakew, S. M., Di Gangi, M. & Federico, M. (2019b), 'Controlling the Output Length of Neural Machine Translation', *arXiv* .
- Lakew, S. M., Federico, M., Wang, Y., Hoang, C., Virkar, Y., Barra-Chicote, R. & Enyedi, R. (2021), 'Machine translation verbosity control for automatic dubbing', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2021-June*, 7538–7542.
- Lample, G., Subramanian, S., Smith, E. M., Denoyer, L., Ranzato, M. & Lan Boureau, Y. (2019), 'Multiple-attribute text rewriting', *7th International Conference on Learning Representations, ICLR 2019* pp. 1–20.
URL: <https://openreview.net/pdf?id=H1g2NhC5KQ>
- Läubli, S., Sennrich, R. & Volk, M. (2018), Has machine translation achieved human parity? a case for document-level evaluation, in 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Brussels, Belgium, pp. 4791–4796.
URL: <https://aclanthology.org/D18-1512>
- Libovický, J., Helcl, J. & Mareček, D. (2018), Input combination strategies for multi-source transformer decoder, in 'Proceedings of the Third Conference on Machine Translation: Research Papers', Association for Computational Linguistics, Brussels, Belgium, pp. 253–260.
URL: <https://aclanthology.org/W18-6326>
- Lipton, Z. C., Berkowitz, J. & Elkan, C. (2015), 'A Critical Review of Recurrent Neural Networks for Sequence Learning', pp. 1–38.
URL: <http://arxiv.org/abs/1506.00019>
- Lison, P. & Tiedemann, J. (2016), OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles, in 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', European Language Resources Association (ELRA), Portorož, Slovenia, pp. 923–929.
URL: <https://aclanthology.org/L16-1147>

- Lison, P., Tiedemann, J. & Kouylekov, M. (2018), OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)', European Language Resources Association (ELRA), Miyazaki, Japan.
URL: <https://aclanthology.org/L18-1275>
- Lopes, A., Farajian, M. A., Bawden, R., Zhang, M. & 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>
- Lupo, L., Dinarelli, M. & Besacier, L. (2022a), Divide and rule: Effective pre-training for context-aware multi-encoder translation models, *in* 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 4557–4572.
URL: <https://aclanthology.org/2022.acl-long.312>
- Lupo, L., Dinarelli, M. & Besacier, L. (2022b), Focused concatenation for context-aware neural machine translation, *in* 'Proceedings of the Seventh Conference on Machine Translation (WMT)', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp. 830–842.
URL: <https://aclanthology.org/2022.wmt-1.77>
- Lynn, V., Son, Y., Kulkarni, V., Balasubramanian, N. & Schwartz, H. A. (2017), Human centered NLP with user-factor adaptation, *in* 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Copenhagen, Denmark, pp. 1146–1155.
URL: <https://aclanthology.org/D17-1119>
- Mathur, N., Wei, J., Freitag, M., Ma, Q. & Bojar, O. (2020), Results of the WMT20 metrics shared task, *in* 'Proceedings of the Fifth Conference on Machine Translation', Association for Computational Linguistics, Online, pp. 688–725.
URL: <https://aclanthology.org/2020.wmt-1.77>
- Matusov, E., Wilken, P. & Herold, C. (2020), Flexible customization of a single neural machine translation system with multi-dimensional metadata inputs, *in* 'Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)', Association for Machine Translation in the Americas, Virtual, pp. 204–216.
URL: <https://aclanthology.org/2020.amta-user.10>
- Mcculloch, W. & Pitts, W. (1943), 'A Logical Calculus of Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics* 5, 127–147.

- McInnes, L., Healy, J. & Melville, J. (2018), 'Umap: Uniform manifold approximation and projection for dimension reduction'.
URL: <https://arxiv.org/abs/1802.03426>
- Michel, P. & Neubig, G. (2018a), Extreme adaptation for personalized neural machine translation, in 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 312–318.
URL: <https://aclanthology.org/P18-2050>
- Michel, P. & Neubig, G. (2018b), 'Extreme adaptation for personalized neural machine translation', *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 2*, 312–318.
- Miculicich, L., Ram, D., Pappas, N. & Henderson, J. (2018), Document-level neural machine translation with hierarchical attention networks, in 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Brussels, Belgium, pp. 2947–2954.
URL: <https://aclanthology.org/D18-1325>
- Milburn, T. (2004), 'Speech community: Reflections upon communication', *Annals of the International Communication Association* 28(1), 411–441.
URL: <https://doi.org/10.1080/23808985.2004.11679041>
- Miresghallah, F., Shrivastava, V., Shokouhi, M., Berg-Kirkpatrick, T., Sim, R. & Dimitriadis, D. (2022), UserIdentifier: Implicit user representations for simple and effective personalized sentiment analysis, 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. 3449–3456.
URL: <https://aclanthology.org/2022.naacl-main.252>
- Moore, R. C. & Lewis, W. (2010), Intelligent selection of language model training data, in 'Proceedings of the ACL 2010 Conference Short Papers', Association for Computational Linguistics, Uppsala, Sweden, pp. 220–224.
URL: <https://aclanthology.org/P10-2041>
- Moryossef, A., Aharoni, R. & Goldberg, Y. (2019a), 'Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection', pp. 49–54.
- Moryossef, A., Aharoni, R. & Goldberg, Y. (2019b), Filling gender & number gaps in neural machine translation with black-box context injection, in 'Proceedings of the First Workshop on Gender Bias in Natural Language Processing', Association for Computational Linguistics, Florence, Italy, pp. 49–54.
URL: <https://aclanthology.org/W19-3807>

- Müller, M., Rios, A., Voita, E. & Sennrich, R. (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>
- Nadejde, M., Currey, A., Hsu, B., Niu, X., Federico, M. & Dinu, G. (2022a), CoCoA-MT: A Dataset and Benchmark for Contrastive Controlled MT with Application to Formality, *in* 'Findings of the Association for Computational Linguistics: NAACL 2022', Association for Computational Linguistics, Seattle, USA.
- Nadejde, M., Currey, A., Hsu, B., Niu, X., Federico, M. & Dinu, G. (2022b), CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality, *in* 'Findings of the Association for Computational Linguistics: NAACL 2022', Association for Computational Linguistics, Seattle, United States, pp. 616–632.
URL: <https://aclanthology.org/2022.findings-naacl.47>
- Niu, X. & Carpuat, M. (2020), 'Controlling neural machine translation formality with synthetic supervision', *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* 2(1), 8568–8575.
- Niu, X., Martindale, M. & Carpuat, M. (2017), A study of style in machine translation: Controlling the formality of machine translation output, *in* 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Copenhagen, Denmark, pp. 2814–2819.
URL: <https://aclanthology.org/D17-1299>
- Novotney, S., Mukherjee, S., Ahmed, Z. & Stolcke, A. (2022), CUE vectors: Modular training of language models conditioned on diverse contextual signals, *in* 'Findings of the Association for Computational Linguistics: ACL 2022', Association for Computational Linguistics, Dublin, Ireland, pp. 3368–3379.
URL: <https://aclanthology.org/2022.findings-acl.265>
- O'Connor, J. & Andreas, J. (2021), What context features can transformer language models use?, *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. 851–864.
URL: <https://aclanthology.org/2021.acl-long.70>
- Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. & Auli, M. (2019), fairseq: A fast, extensible toolkit for sequence modeling, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)', Association for Computational

- Linguistics, Minneapolis, Minnesota, pp. 48–53.
URL: <https://aclanthology.org/N19-4009>
- Papi, S., Karakanta, A., Negri, M. & Turchi, M. (2022), Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora, *in* ‘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 2: Short Papers)’, Association for Computational Linguistics, Online only, pp. 480–487.
URL: <https://aclanthology.org/2022.aacl-short.59>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019), ‘PyTorch: An imperative style, high-performance deep learning library’, *Advances in Neural Information Processing Systems* 32(NeurIPS).
- Pennington, J., Socher, R. & Manning, C. D. (2014), GloVe: Global Vectors for Word Representation, *in* ‘Empirical Methods in Natural Language Processing (EMNLP)’, pp. 1532–1543.
URL: <http://www.aclweb.org/anthology/D14-1162>
- Pickering, M. J. & Garrod, S. (2004), ‘Toward a mechanistic psychology of dialogue’, *Behavioral and Brain Sciences* 27(02).
- Ponce, D., Etchegoyhen, T. & Ruiz, V. (2023), Unsupervised subtitle segmentation with masked language models, *in* ‘Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)’, Association for Computational Linguistics, Toronto, Canada, pp. 771–781.
URL: <https://aclanthology.org/2023.acl-short.67>
- Popovi, M. (2017), ‘CHR F ++ : words helping character n-grams’, 2(1), 612–618.
- Post, M. (2018), A call for clarity in reporting BLEU scores, *in* ‘Proceedings of the Third Conference on Machine Translation: Research Papers’, Association for Computational Linguistics, Brussels, Belgium, pp. 186–191.
URL: <https://aclanthology.org/W18-6319>
- Post, M. & Junczys-Dowmunt, M. (2023), ‘Escaping the sentence-level paradigm in machine translation’.
- Post, M. & Vilar, D. (2018), Fast lexically constrained decoding with dynamic beam allocation for 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. 1314–1324.
URL: <https://aclanthology.org/N18-1119>
- Rabinovich, E., Mirkin, S., Patel, R. N., Specia, L. & Wintner, S. (2017), ‘Personalized machine translation: Preserving original author traits’, *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference 1*, 1074–1084.
- Ramírez-Sánchez, G., Zaragoza-Bernabeu, J., Bañón, M. & Rojas, S. O. (2020), Bifixer and bicleaner: two open-source tools to clean your parallel data, *in* ‘Proceedings of the 22nd Annual Conference of the European Association for Machine Translation’, European Association for Machine Translation, Lisboa, Portugal, pp. 291–298.
URL: <https://aclanthology.org/2020.eamt-1.31>
- Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. (2020), COMET: A neural framework for MT evaluation, *in* ‘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>
- Reimers, N. & Gurevych, I. (2019a), Sentence-bert: Sentence embeddings using siamese bert-networks, *in* ‘Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing’, Association for Computational Linguistics.
URL: <https://arxiv.org/abs/1908.10084>
- Reimers, N. & Gurevych, I. (2019b), Sentence-BERT: Sentence embeddings using Siamese BERT-networks, *in* ‘Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)’, Association for Computational Linguistics, Hong Kong, China, pp. 3982–3992.
URL: <https://aclanthology.org/D19-1410>
- Reiter, E. (2018), ‘A structured review of the validity of BLEU’, *Computational Linguistics* 44(3), 393–401.
URL: <https://aclanthology.org/J18-3002>
- Remael, A. (2003), ‘Mainstream Narrative Film Dialogue and Subtitling’, *The Translator* 9(2), 225–247.
- Rippeth, E., Agrawal, S. & Carpuat, M. (2022), Controlling translation formality using pre-trained multilingual language models, *in* ‘Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)’, Association for Computational Linguistics, Dublin, Ireland (in-person and online), pp. 327–340.
URL: <https://aclanthology.org/2022.iwslt-1.30>

- Rozis, R. & Skadiņš, R. (2017), Tilde MODEL - multilingual open data for EU languages, *in* 'Proceedings of the 21st Nordic Conference on Computational Linguistics', Association for Computational Linguistics, Gothenburg, Sweden, pp. 263–265.
URL: <https://aclanthology.org/W17-0235>
- Salemi, A., Mysore, S., Bendersky, M. & Zamani, H. (2023), 'Lamp: When large language models meet personalization'.
- Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019), DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *in* '5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019'.
URL: <http://arxiv.org/abs/1910.01108>
- Schein, A. I., Popescul, A., Ungar, L. H. & Pennock, D. M. (2002), Methods and metrics for cold-start recommendations, *in* 'Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '02, Association for Computing Machinery, New York, NY, USA, p. 253–260.
URL: <https://doi.org/10.1145/564376.564421>
- Schiebinger, L. (2014), 'Scientific research must take gender into account', *Nature* 507(7490), 9.
- Schioppa, A., Vilar, D., Sokolov, A. & Filippova, K. (2021), Controlling machine translation for multiple attributes with additive interventions, *in* 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 6676–6696.
URL: <https://aclanthology.org/2021.emnlp-main.535>
- Schwenk, H., Chaudhary, V., Sun, S., Gong, H. & Guzmán, F. (2021), WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia, *in* 'Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume', Association for Computational Linguistics, Online, pp. 1351–1361.
URL: <https://aclanthology.org/2021.eacl-main.115>
- Sennrich, R., Haddow, B. & Birch, A. (2016a), 'Controlling politeness in neural machine translation via side constraints', *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference* pp. 35–40.
- Sennrich, R., Haddow, B. & Birch, A. (2016b), Controlling politeness in neural machine translation via side constraints, *in* 'Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human

- Language Technologies', Association for Computational Linguistics, San Diego, California, pp. 35–40.
URL: <https://aclanthology.org/N16-1005>
- Sennrich, R., Haddow, B. & Birch, A. (2016c), 'Improving neural machine translation models with monolingual data', *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers 1*, 86–96.
- Sennrich, R., Haddow, B. & Birch, A. (2016d), 'Neural machine translation of rare words with subword units', *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers 3*, 1715–1725.
- Sharou, K. A. & Specia, L. (2022), A taxonomy and study of critical errors in machine translation, in 'Proceedings of the 23rd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Ghent, Belgium, pp. 171–180.
URL: <https://aclanthology.org/2022.eamt-1.20>
- Shimanaka, H., Kajiwara, T. & Komachi, M. (2018), RUSE: Regressor using sentence embeddings for automatic machine translation evaluation, in 'Proceedings of the Third Conference on Machine Translation: Shared Task Papers', Association for Computational Linguistics, Belgium, Brussels, pp. 751–758.
URL: <https://aclanthology.org/W18-6456>
- Sigurdsson, H. & Egerland, V. (2009), 'Impersonal null-subjects in Icelandic and elsewhere*', *Studia Linguistica* **63**(1), 158–185.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9582.2008.01157.x>
- Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C. & Lopez, A. (2013), Dirt cheap web-scale parallel text from the Common Crawl, in 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Sofia, Bulgaria, pp. 1374–1383.
URL: <https://aclanthology.org/P13-1135>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014), 'Dropout: A simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research* **15**(56), 1929–1958.
URL: <http://jmlr.org/papers/v15/srivastava14a.html>
- Stahlberg, D., Braun, F., Irmen, L. & Sczesny, S. (2007), 'Representation of the sexes in language', *Social Communication* pp. 163–187.
- Stahlberg, F., Cross, J. & Stoyanov, V. (2018), Simple fusion: Return of the language model, in 'Proceedings of the Third Conference on Machine Translation: Research

- Papers', Association for Computational Linguistics, Brussels, Belgium, pp. 204–211.
URL: <https://aclanthology.org/W18-6321>
- Stone, G. (1977), 'Address in the Slavonic Languages', *The Slavonic and East European Review* 55(4), 491–505.
- Sugiyama, A. & Yoshinaga, N. (2021), Context-aware decoder for neural machine translation using a target-side document-level language model, 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. 5781–5791.
URL: <https://aclanthology.org/2021.naacl-main.461>
- Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *Advances in Neural Information Processing Systems* 4(January), 3104–3112.
- Takeno, S., Nagata, M. & Yamamoto, K. (2017), 'Controlling Target Features in Neural Machine Translation via Prefix Constraints', *Afnlp* pp. 55–63.
- 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. & 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. (2020), Reassessing claims of human parity and super-human performance in machine translation at WMT 2019, in 'Proceedings of the 22nd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Lisboa, Portugal, pp. 185–194.
URL: <https://aclanthology.org/2020.eamt-1.20>
- Tuora, R. & Kobylinski, L. (2019), Integrating Polish Language Tools and Resources in Spacy, in 'Proceedings of PP-RAI 2019 Conference', pp. 210–214.
- van der Wees, M., Bisazza, A. & Monz, C. (2016), Measuring the effect of conversational aspects on machine translation quality, in 'Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers', The COLING 2016 Organizing Committee, Osaka, Japan, pp. 2571–2581.
URL: <https://aclanthology.org/C16-1242>

- Vanmassenhove, E., Hardmeier, C. & Way, A. (2018), 'Getting gender right in neural machine translation', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* pp. 3003–3008.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in Neural Information Processing Systems* pp. 5999–6009.
URL: <https://arxiv.org/pdf/1706.03762.pdf>
- Vincent, S., Barrault, L. & Scarton, C. (2022a), Controlling formality in low-resource NMT with domain adaptation and re-ranking: SLT-CDT-UoS at IWSLT2022, in 'Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)', Association for Computational Linguistics, Dublin, Ireland (in-person and online), pp. 341–350.
URL: <https://aclanthology.org/2022.iwslt-1.31>
- Vincent, S., Flynn, R. & Scarton, C. (2023), MTCue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation, in 'Findings of the Association for Computational Linguistics: ACL 2023', Association for Computational Linguistics, Toronto, Canada, pp. 8210–8226.
URL: <https://aclanthology.org/2023.findings-acl.521>
- Vincent, S., Sumner, R., Dowek, A., Blundell, C., Preston, E., Bayliss, C., Oakley, C. & Scarton, C. (2023), 'Personalised language modelling of screen characters using rich metadata annotations'.
- Vincent, S. T., Barrault, L. & Scarton, C. (2022b), Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation, in 'Proceedings of the 23rd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Ghent, Belgium, pp. 121–130.
URL: <https://aclanthology.org/2022.eamt-1.15>
- Voita, E., Sennrich, R. & Titov, I. (2019a), Context-aware monolingual repair for neural machine translation, in 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 877–886.
URL: <https://aclanthology.org/D19-1081>
- Voita, E., Sennrich, R. & Titov, I. (2019b), 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, Florence, Italy, pp. 1198–1212.
URL: <https://aclanthology.org/P19-1116>

- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2018), GLUE: A multi-task benchmark and analysis platform for natural language understanding, *in* 'Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP', Association for Computational Linguistics, Brussels, Belgium, pp. 353–355.
URL: <https://aclanthology.org/W18-5446>
- Wang, W., Bao, H., Huang, S., Dong, L. & Wei, F. (2021), MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers, *in* 'Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021', Association for Computational Linguistics, Online, pp. 2140–2151.
URL: <https://aclanthology.org/2021.findings-acl.188>
- Wang, Y., Hoang, C. & Federico, M. (2021), Towards modeling the style of translators 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. 1193–1199.
URL: <https://aclanthology.org/2021.naacl-main.94>
- Welch, C., Gu, C., Kummerfeld, J. K., Perez-Rosas, V. & Mihalcea, R. (2022), Leveraging similar users for personalized language modeling with limited data, *in* 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 1742–1752.
URL: <https://aclanthology.org/2022.acl-long.122>
- Welch, C., Kummerfeld, J. K., Pérez-Rosas, V. & Mihalcea, R. (2020), Compositional demographic word embeddings, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 4076–4089.
URL: <https://aclanthology.org/2020.emnlp-main.334>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020), Transformers: State-of-the-art natural language processing, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 38–45.
URL: <https://aclanthology.org/2020.emnlp-demos.6>
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young,

- C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. & Dean, J. (2016), 'Google's neural machine translation system: Bridging the gap between human and machine translation', *CoRR* **abs/1609.08144**.
URL: <http://arxiv.org/abs/1609.08144>
- Yu, L., Sartran, L., Stokowiec, W., Ling, W., Kong, L., Blunsom, P. & Dyer, C. (2020), 'Better document-level machine translation with Bayes' rule', *Transactions of the Association for Computational Linguistics* **8**, 346–360.
URL: <https://aclanthology.org/2020.tacl-1.23>
- Zeng, W., Abuduweili, A., Li, L. & Yang, P. (2019), Automatic generation of personalized comment based on user profile, in 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop', Association for Computational Linguistics, Florence, Italy, pp. 229–235.
URL: <https://aclanthology.org/P19-2032>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2019), 'BERTScore: Evaluating Text Generation with BERT', *arXiv* pp. 1–43.
URL: <http://arxiv.org/abs/1904.09675>
- Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. (2016), The United Nations parallel corpus v1.0, in 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3530–3534.
URL: <https://aclanthology.org/L16-1561>

DECLARATION

I, Sebastian T. Vincent, declare that the work in this dissertation was carried out in accordance with the requirements of the University of Sheffield's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Sheffield, United Kingdom, September 2023