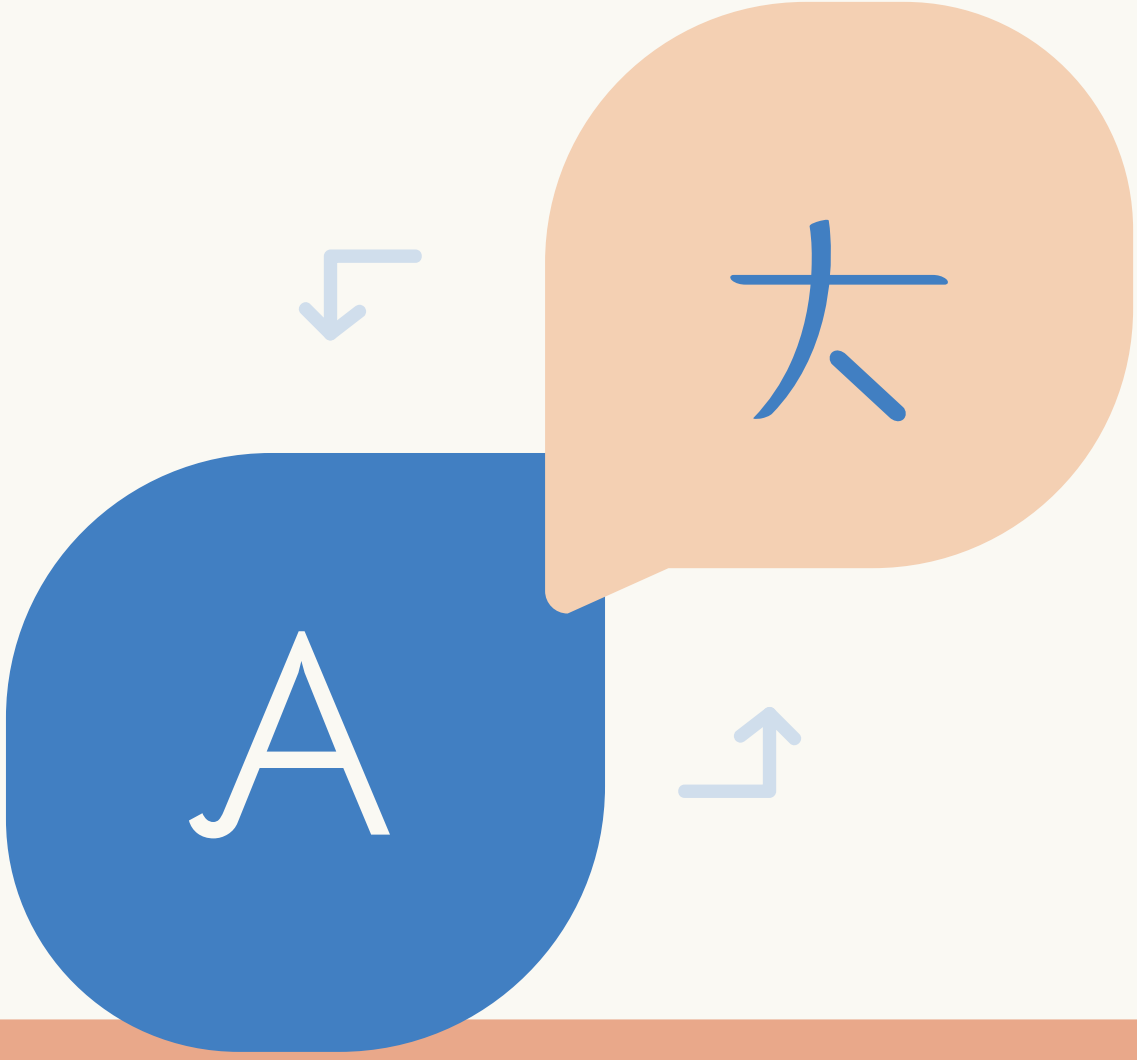


Journal of Computational & Applied Linguistics

Volume 2
2024

NEW BULGARIAN UNIVERSITY

Address
New Bulgarian University
Research Center for Computational
and Applied Linguistics
21, Montevideo Str., 309/ Building 2
1618 Sofia
jcal@nbu.bg

ISSN (online): 2815-4967

Journal of Computational & Applied Linguistics

Volume 2
2024

NEW BULGARIAN UNIVERSITY

Journal of Computational and Applied Linguistics (JCAL)

The Journal of Computational and Applied Linguistics **(JCAL) is a yearly open access peer-reviewed journal publishing articles which contribute new results in computational linguistics, applied linguistics and related areas. The goal of the journal is to bring together researchers and practitioners from academia and industry, to establish collaborations and to focus on advanced concepts in these fast-developing fields.**

The Journal has the following structure:

Articles
Reviews
Announcements
Bibliographies

Authors are invited to submit papers which illustrate research results, projects, surveying works and industrial experiences describing significant advances in the following fields:

Formal Grammars
Corpus Linguistics and Applications
Computational Linguistics
Contrastive Studies
Languages Technologies and Translation
Language for Specific Purposes
Multilingualism

Article Submission Guidelines:

File format: Microsoft DOCX;
Paper size: A4;
Orientation: Portrait;
Font: Book Antiqua;
Margins (Top, Bottom, Left, Right): 0.98 inch;
Line spacing: Single;
Paragraph style: Block format;
Title: block caps, bold, centred, size 14;
Author(s): Academic title, first name, last name, affiliation; font size 12, italics, right;
Abstract: font size 11;
Keywords: font size 10, italics;
Body of the article: font size 12, single line spacing, justified;
Spaces between paragraphs;
no header, footer or page numbers;
Footnotes: font size 10;
References: Harvard style, size 11, single line spacing.

Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated for this Journal. They should be made via e-mail: jcal@nbu.bg.

Contents

https://doi.org/10.33919/JCAL.24.2.1

LANGUAGE CORPORA AS A RESOURCE FOR MACHINE TRANSLATION

Todor Lazarov

New Bulgarian University,
Computational and Applied Linguistics Research Centre
tlazarov@nbu.bg

Abstract

The article discusses the importance of linguistic corpora in both statistical and neural machine translation. It explains that these models require large, representative parallel corpora to capture language phenomena and determine translation probabilities between source and target languages. Corpus linguistics, the field dedicated to creating and structuring such corpora, is essential for developing linguistic resources that can support language learning, lexicography, and linguistic analysis.

Keywords: *Corpora, language resources, representativeness, balance, sampling*

Nowadays, with the advancements of translation technologies, machine translation (MT) and artificial intelligence (AI) linguistic resources are becoming increasingly pivotal for research in these areas and, more importantly, improvements in the available technology and its practical applications. Language resources are also important for courses related to language education, translation technologies and natural language processing (NLP) since they provide the fundamental understanding of how advanced language systems work. These resources, which include large, structured collections of text and annotated data, provide the foundation for training statistical and neural machine translation models.

In recent years, the study of corpora is incorporated into university curricula, especially in courses related to translation technologies translation studies, and linguistics, since through the understanding of corpora, students gain an understanding of the empirical basis of language technologies, including their design and limitations. This theoretical and practical knowledge is essential for grasping the principles behind MT and the broader field of NLP. As corpus linguistics plays a pivotal role in language technologies, familiarity with its tools and methodologies empowers students to contribute meaningfully to technological innovations and linguistic research.

Both traditional statistical translation systems and neural machine translation systems, together with state-of-the-art AI systems, rely primarily on sufficiently large and reliable language corpora to be trained and evaluated. Manually crafted and annotated corpora play an essential role in the design of these types of systems. Manually created corpora have an advantage over automatically compiled corpora because the data in them (which can be automatically collected) is processed by humans and the probability of errors is minimised. This characteristic suggests their greater reliability for language learning and adequate data extraction in comparison to the automatic collection of text.

Machine translation – both statistical and neural – relies on representative parallel corpora of sufficient volume to represent linguistic phenomena in their various manifestations. In order to create a statistical translation model, it is necessary to know what is the probability that a language unit (word, phrase) in the target language will be a translation of a language unit from the source language – $p(t|s)$. In order to achieve this, it is necessary to develop algorithms and systems that process information and language resources to extract linguistic data. The creation, structuring and annotation of linguistic corpora is the main subject of study in corpus linguistics. Corpus linguistics is a scientific branch of linguistics that studies the development and creation of linguistic corpora that are sufficiently representative in terms

of their size and that are composed of natural language examples. Globally, corpus linguistics has made significant achievements and has traditionally been applied mainly in the creation of different types of dictionaries for certain linguistic purposes, as well as in foreign language teaching. In corpus linguistics, large collections of annotated and unannotated texts are used to extract statistical patterns that demonstrate a linguistic phenomenon. Corpus linguistics can also be helpful in the effective use of different types of corpus data in the formulation of grammatical rules, different types of grammars, etc. The corpus-based approach can be useful in the study of many different aspects and areas of linguistic knowledge at different linguistic levels (phonetics, morphology, syntax, etc.), as well as in various practical and theoretical goals and tasks (Tisheva, 2014). The corpus approach helps linguists in the study of large volumes of linguistic material composed of different texts in the study and description of natural language and the formulation of theoretical propositions about the functioning of language. Contemporary corpora are used as empirical linguistic material, serving as a basis for the objectivity of relevant research and conclusions that do not rely solely on the individual linguistic sense of the linguist.

According to the definition given by the European Linguistic Resources Association, the term *linguistic resource* means a system of linguistic data or descriptions of such data that is available for machine processing for various purposes. 'Language resources range from simple lists to complex resources in which different types of linguistic information are organised: word lists, text collections, corpora, electronic dictionaries, thesauri, ontologies, subject dictionaries, databases, sound recordings, videos, etc.' (Koeva, 2009, pp. 59–60). In general, language resources may be classified according to the structure in which the language data is presented or according to their specific content. Depending on how they are structured, language resources can be divided into thesauri, corpora, dictionaries, thematic dictionaries, terminology bases or databases (Tisheva, 2014). Resources may contain only written texts or may have complex content if they also include images, audio or video. A more detailed classification of language resources takes into account the format, language and subject area of the included data. According to the format in which the data is presented, three types of resources can be defined: speech – which includes recordings of dialectal or colloquial speech, radio or television broadcasts; written and multimodal (multimodal, multimedia) (Tisheva, 2014). According to the languages represented, the resources are divided into monolingual, bilingual or multilingual, and according to the thematic area to which the language data are related – into general and specialised.

There is some ambiguity in defining relatively similar concepts such as a textual archive, textual library, a collection of texts and corpora (Koeva, 2009, p. 60). Corpora are distinctive in terms of their structuring and usage. Corpora are collections of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2005). Representativeness is a feature that determines the objectivity of corpora. When excerpting them, the maximum scope and variety of materials are chosen in order to allow for the fullest description of the objective picture of linguistic practice. This is the main characteristic that distinguishes a linguistic corpus from a random set of texts (Baeva, 2011, p. 49). The pre-processing of the text encompasses different levels of the linguistic information embedded in the text: tokenisation (dividing the text into a sequence of symbols), parsing (dividing the text into certain units – phonemes, morphemes, lexemes, word combinations, sentences, etc.); annotation (attributing linguistic information about the characteristics of each individual unit – morphological, syntactic, morphosyntactic, semantic features, etc.); resolution of various language-specific phenomena such as pronominal and non-pronominal anaphora, ellipses, etc. and of various types of linguistic polysemy. The different levels of analysis use different software applications: morphological analysers (taggers), syntactic analysers (parsers), anaphora extenders, text structure analysers, etc. From a contemporary point of view, the theoretical definition of a language corpus is descriptive rather than definitional. The main criteria for defining a set of texts as a corpus are usually the computer-accessible, electronic form, subordinated to a given format (e.g. txt, XML); the purposeful and well-documented structural organisation; the representativeness of linguistic variation for predefined, specific linguistic phenomena; the balance in the distribution of texts in the overall structure relative to usage in the language community or time span under study. Definitions for *language corpus* usually emphasise the criteria for selection of corpora units, the importance of the corpora for various linguistic studies and the opportunities that corpora provide for computer processing (Koeva, 2014, p. 29). The analysis and processing of the different types of corpora are a scientific subject within the field of natural language processing. A corpus is a set of texts or fragments of linguistic material in written or spoken format, mainly in electronic form, organised in a systematic and structured way so that they can be used for the study of different linguistic usages. The definition adopted in this text is 'A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language' (EAGLES 1996: 4). Textual corpora are usually used for

statistical analysis, for studying the frequency of certain usages or linguistic rules specific to a given linguistic or literary domain; they are also used to study historical documents, ancient manuscripts, etc.

Characteristics and criteria in defining linguistic corpora are also subject to debate. The characteristics of a linguistic corpus can be its size, textual typology, the realisation of linguistic phenomena (general or specialised), internal structure, annotation structure, etc. It should be clarified that these characteristics also do not have a uniform interpretation and principles of definition, their use is rather dictated by a desire for a conceptualised application of a different set of methods in the construction of language corpora.

In this article, we characterise corpora by traditionally accepted features, which are largely an amalgamation of the broader range of characteristics for a linguistic corpus – balance, representativeness and sampling (Leech, 1991, p. 10). The representativeness of a language corpus is determined by the extent to which it can cover the full range of different manifestations of a given language phenomenon. A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalised to the said language variety (Leech, 1991, p. 27). The aim is that each language corpus provides a representative sample of the examined linguistic phenomenon. The selection of representative texts is an essential part of building a corpus of linguistic data. In this sense, the representativeness of most corpora is largely determined by two factors: the range of genres included in the corpus (balance) and the criteria for selecting texts from each genre (sampling) (Sinclair, 2005, p. 12). 'The representativeness of a corpus relates to its ability to provide reliable linguistic facts about a particular state of language, a specialised domain of language use, and/or a set of linguistic phenomena [...]. Representativeness is determined by the variety of text types that illustrate a given state of language or its specialised domain of use, more precisely it depends on the definition of the set of representative texts, the hierarchical organisation of their types and the relationship between them (balance), and the techniques that are used to obtain representative samples (sampling)' (Koeva, 2014, p. 30). The balance of a corpus depends on its text typology – the classification and the characteristics of the text types involved. A balanced corpus should cover in some proportion as many text types as possible in order to be representative of the language as a whole or of a given linguistic phenomenon (Biber, 1993, p. 243). Various ways of achieving balance are known, based on both external and internal criteria. Text selection criteria 'which are derived from an examination of the communicative function of the text are called external criteria, and those which reflect the linguistic characteristics of the text are called internal cri-

teria. A corpus / Corpora should be designed and built exclusively on the basis of external criteria' (Sinclair, 2005, p. 6). 'Representativeness and balance are closely related to sampling' (Koeva, 2014, p. 33). Determining the text types to be included in corpus construction implies that the corpus units used should demonstrate as far as possible all the linguistic properties for the text type to which they belong, since 'different linguistic features are distributed differently (within texts, between texts, between different text types), and the representative corpus should enable analysis of these different occurrences' (Biber, 1993, p. 243). Achieving sampling is dependent on the choice of representative texts, the number of which for each text category must be appropriate to the frequency or importance of that category for the linguistic phenomenon being illustrated.

In many ways, progress in natural language research depends directly on the availability of linguistic data. This is especially applicable to the fields of statistical and neural machine translation, which rely on large amounts of parallel text (Koehn, 2010, p. 79). Since both approaches need sufficient representative data, the quality of the parallel corpora used to train this type of systems is an important factor for the feasibility and reliability of the systems. For this reason, the selection and preparation of the parallel corpora determine the quality of the results obtained. This can be demonstrated with the widely available machine translation systems that offer good translation for languages with sufficient available resources. It follows that modern machine translation systems depend directly on the quality and quantity of available language resources. Languages or certain linguistic phenomena for which there are insufficient representative corpora are poorly studied, and the capabilities that modern machine translation systems provide are in most cases inapplicable to them. Not all languages have easily accessible and sufficiently large and representative language corpora. Due to the lack of such corpora, it is difficult to use statistical and neural machine translation approaches to create automatic translation systems or for empirical studies. The resources available on the Internet cannot be used directly to train machine translation systems due to the fact that any freely available text must be processed before it can become a reliable language corpus. The steps involved in this processing are: collection of texts; correlation of texts; preparation for statistical processing (normalisation, tokenisation); alignment at sentence level of originals with their translations (Koehn, 2010, p. 79). The presented characteristics of language resources show that the possibilities offered by statistical and neural approaches for the study of languages depend primarily on the availability of sufficiently representative data about them.

References

Baeva, D. (2011). 'Za korpusite kato ezikovi resursi i tyahnoto prilojenie' (About corpora as language resources and their application). In: *Nauchi trudove na Rusenskia universitet*, t. 50, 48–53.

Biber, D. (1993). 'Representativeness in Corpus Design'. In: *Literary and Linguistic Computing*, vol. 8, 243–257.

EAGLES (1996). 'Expert Advisory Group for Language Engineering Standards, Preliminary Recommendations on Corpus Typology'. *Instituto di Linguistica Computazionale*. Available at: https://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koeva, S. (2009). 'Ezikovi resursi i kompyutarni programi s prilojenie v lingvistichnite izsledvania' (Language Resources and Computer Software Applied to Linguistic Research). In: Barkalova, P. (ed.) *Prilojenie na informatsionnite tehnologii v rabotata na filologa i pri izgrajdaneto na ezikovi resursi.* Sofia: Arhimed, 57–74.

Koeva, S. (2014). 'Balgarskiyat natsionalen korpus v konteksta na svetovnata teoria i praktika'. (The Bulgarian National Corpus in the Context of World Theory and Practice). In: *Ezikovi resursi i tehnologii za balgarski ezik*. Sofia: Akademitshno izdatelstvo 'Prof. Marin Drinov', 29–53.

Leech, G. (1991). 'The State of the Art in Corpus Linguistics'. In: *English Corpus Linguistics*. London: Longman, 8–29.

Sinclair, J. (2005). 'Corpus and Text: Basic Principles'. In: *Developing Linguistic Corpora: A guide to Good Practice*. Available at: https://pdfs.semanticscholar.org/07fe/f4d946b51c7045c6968066f222f31192a389.pdf.

Tisheva, Y. (2014). 'Ezikovi bazi danni. Korpusi i elektronni resursi za balgarskata ustna rech' (Linguistic databases. Corpora and electronic resources for Bulgarian oral speech). *Littera et Lingua*, vol. 11(1–2). Available at: https://naum.slav.uni-sofia.bg/en/lilijournal/2014/11/1-2/ytisheva.

https://doi.org/10.33919/JCAL.24.2.2

EMOTIONALTURK, A NEW STANDARD IN EVALUATING AUTOMATIC TURKISH EMOTION DETECTION: A CASE STUDY IN A CROSS-LINGUAL TRANSFER SETTING

Egemen Ipek[1], Pranaydeep Singh[1], Els Lefever[1]

[1]*Ghent University*

Abstract

Current Turkish emotion detection from text (EDFT) research, and EDFT research in general, does not justify the rationale behind the emotion framework used in their EDFT datasets. Adapting emotion frameworks directly from psychology research, the most popular ones based on physiology, creates a gap between what is latently available in textual data and what the model output is, regardless of its performance. Therefore, the article introduces a dataset, EmotionalTurk, built with an emotion framework empirically grounded in textual data. While EmotionalTurk is currently limited to a test set in size, the article seizes the opportunity to explore cross-lingual transfer capabilities of multilingual transformers – an area that has not been previously investigated – using mBERT-based and XLM-RoBERTaBASE-based models, and establishes the SOTA score, which also serves as the baseline in this study in cross-lingual Turkish EDFT.

Keywords: *emotion detection from text in Turkish, cross-lingual transfer, lower-resource languages, multilingual transformers, classifiers*

Introduction

Text is the primary medium through which people express themselves online (Sailunaz *et al.,* 2018). Through text, people share their opinions, achievements, sentiments etc. and make them publicly available. This vast amount of publicly available textual data offers significant potential for gathering emotion insights. However, manually analysing and gathering insights from such a large volume of data is not feasible. Therefore, the act of gathering emotion insights from textual data has become a machine learning, and more specifically, a natural language processing (NLP) task. To address the task of EDFT, research in NLP has mainly focused on developing classifiers. These classifiers are typically trained using supervised machine learning methods. Because supervised machine learning methods require labelled data, EDFT researchers have adopted emotion frameworks from psychology research to create label sets for EDFT.

EDFT has been applied in research to gather insights from various impactful areas. One significant application is in suicide prevention (Desmet and Hoste, 2013). Another critical application area of EDFT is used in understanding public sentiment towards events such as earthquakes or pandemics (Sosea *et al*., 2022). Additionally, EDFT shares areas of application with sentiment analysis, including the collection of customer insights (De Bruyne *et al*., 2022). In these cases, EDFT provides a more fine-grained view of public sentiment, going beyond the polarity perspective sentiment analysis provides.

Currently, one of the major issues in EDFT is the lack of consistency in emotion frameworks adopted and the rationale behind these choices of emotion frameworks. Although Ekman's basic emotions (Ekman, 1992) and Plutchik's circumplex model (Plutchik, 1980) are the most popular emotion frameworks used in EDFT research, researchers often fail to justify why these specific emotion frameworks are selected (De Bruyne *et al*., 2019). Moreover, Ekman's and Plutchik's frameworks were based on physiology research, which makes them less suited for direct application in NLP tasks that focus on textual data, because the labels used directly impact the insights derived from the data. As a result, even if the model performs well, the insights it provides may still be suboptimal.

Turkish EDFT research also faces this challenge. The emotion frameworks used in Turkish EDFT research lack a clear rationale for why a specific emotion framework was chosen over others. Consequently, there is a need for a new dataset for Turkish EDFT research that uses an empirically grounded emotion framework specifically designed for text-based applications. Unfortunately, this degrades Turkish EDFT from *lower-resourced* to *no-resourced*. At the same time, no research to date has explored the cross-lingual transfer

capabilities of multilingual transformers for Turkish EDFT. This research, therefore, aims to address these gaps by exploring the cross-lingual transfer performance of multilingual transformers for Turkish EDFT, while also introducing a novel dataset grounded in an empirically validated emotion framework.

Related research

Emotion frameworks are a critical component of EDFT research. Since EDFT models are primarily based on supervised classifiers, they require labelled data for training and emotion frameworks provide the required label set for this task.

In psychology research, emotion frameworks are generally divided into two categories: categorical and dimensional (Nandwani and Verma, 2021). Categorical emotion frameworks consist of distinct emotion categories, such as *love, joy* and *sadness* while dimensional models contain a core set of emotions, accompanied by parameters which allows an emotion to be augmented into related ones, e.g., *joy* could be augmented into *ecstasy*, or *anger* into *rage,* depending on the parameters (Sailunaz *et al.,* 2018).

Ekman's Basic Emotions (Ekman, 1992) and Plutchik's Emotion Wheel (Plutchik, 1980) are the most widely used frameworks in EDFT research for categorical and dimensional approaches, respectively. However, there is no clear rationale provided for choosing these specific frameworks over others (De Bruyne *et al.,* 2019). Furthermore, both frameworks are based on the physical expression of emotions. This grounding in physical cues introduces a gap between the model's performance and real-world insights obtained from textual data.

To address this disparity, an emotion framework that is empirically grounded in textual data is needed. De Bruyne *et al.* (2019) developed such an empirically grounded emotion framework by carrying out cluster analysis on emotionally dense data collected from X using emojis, resulting in 255 posts in total. The data was manually annotated in a multi-label setting, with labels based on a comprehensive emotion framework of 25 emotions (Shaver *et al*., 1987). This initial emotion framework served as a starting point for cluster analysis. Clustering was carried out with the dice similarity metric, and the resulting clusters were named after Ekman's basic emotion keywords in case a cluster contained an emotion from Ekman's framework. The result was an empirically grounded emotion framework that coincided with the goals of text-based emotion detection. The labels in this novel emotion framework consisted of *love*, *joy*, *anger*, *sadness* and *nervousness*. In a subsequent study, *nervousness* was substituted with *fear* (De Bruyne *et al.*, 2024).

One of the earlier works in Turkish EDFT involved the creation of a Turkish Emotion Lexicon (TEL) (Tocoglu and Alpkocak, 2019) and a lexicon-based EDFT model made with TEL. TEL was created using the Turkish emotion dataset TREMO (Tocoglu and Alpkocak, 2018). The rule-based EDFT classifier developed with TEL on the TREMO dataset achieved an accuracy of 92% (Tocoglu and Alpkocak, 2019).

Subsequently, an extensive study on machine learning applications for Turkish EDFT was carried out, including various statistical and deep learning architectures and two datasets (Tocoglu *et al.*, 2019). The statistical machine learning architectures included in the research were support vector machine (SVM), naïve Bayes (NB), k-nearest neighbors (KNN), and decision trees (DT), while the deep learning architectures included artificial neural network (ANN), convolutional neural network (CNN) and long short-term memory (LSTM) networks. The two datasets used to train and test these models were the aforementioned TREMO and TURTED, created within the scope of this research. Based on two experiments, the results revealed that deep learning models outperformed the statistical models, with LSTM achieving the highest accuracy of around 85% on the TREMO dataset and 75% on TURTED. Additionally, a performance gap of roughly 10% was observed between the TREMO and TURTED datasets in deep learning models included in the research.

The introduction of the transformer architecture (Vaswani *et al.*, 2017) caused a paradigm shift in natural language processing due to its impressive performance across various NLP tasks. The initial transformer architecture was built for machine translation and included a natural language understanding (NLU) module, known as the encoder stack, and a text generation module called the decoder stack.

Inspired by the encoder stack design, Devlin *et al.* (2018) trained a bi-directional encoder stack natural language understanding (NLU) model, named BERT, by utilising self-supervised techniques of masked language modelling and next sentence prediction. This training process, referred to as pre-training, allowed BERT to achieve state-of-the-art (SOTA) results in numerous NLU tasks. Additionally, BERT, being an NLU model, mainly serves a feature extraction model. Therefore, the main advantage of BERT is its ability to be further trained end-to-end with task-specific data, a process known as fine-tuning. By adding task-specific components, such as a classification layer for classification tasks, the fine-tuned model can leverage both the vast amount of data used in BERT's pre-training as well as the task-specific data, resulting in exceptional performance.

The Turkish EDFT performance of BERT-based models was explored and contrasted with preliminary neural network models and statistical machine learning models using the TREMO dataset (Ucan *et al.,* 2021). The BERT-based models included in the study were BERT, multilingual BERT (mBERT), trained on 104 languages, including Turkish, and BERTurk, which was pre-trained with Turkish data exclusively.

Measured with the macro-averaged F1 evaluation metric ($F1_{MACRO}$), experiments carried out using the TREMO dataset showed that BERT-based models outperformed both preliminary neural networks models, notably GRU and LSTM, and statistical machine learning models, notably NB and SVM. Specifically, BERTurk achieved a score of 92 while LSTM and GRU achieved scores of 84 and 83, resulting in a performance increase of eight and nine points, respectively. The performance increase was similarly notable when compared to the statistical machine learning models, with the NB model achieving 82 and the SVM model reaching a score of 84, resulting in a performance increase of ten and eight points, respectively. Furthermore, mBERT achieved a score of 89, which is comparable to BERTurk's 92. As BERTurk is a monolingual model, this suggests that cross-lingual transfer has the potential to achieve SOTA performance in Turkish EDFT.

EmotionalTurk

The emotion framework used for the EDFT datasets has a critical effect on the insights obtained from the machine learning models trained with them. EDFT is similar to sentiment analysis, i.e. it is an application of classifying text into discrete classes. However, unlike sentiment analysis, EDFT labels are not arranged in a manner of overt polarity. They consist of predetermined sets of emotions which have complex relationships to each other. The choice of emotion framework for an EDFT application directly influences the textual patterns that the model can recognise. If the framework is not well-suited to the intricacies of textual data, the model's ability to uncover deeper emotional insights will be limited, potentially masking valuable latent information.

As stated already, the emotion framework chosen for an EDFT dataset has a direct effect on the insights gathered from a model trained with it. Thus, an emotion framework not inherently optimal for the area of application medium, i.e. text in this case, results in an EDFT classifier with the potential to create a critical disparity between what is latent in the data to be classified and the classifier output. Consequently, even if an EDFT classifier achieves high scores, these scores do not necessarily mean that the insights gathered

from the model output reflect reality. This was the main driving force behind the creation of EmotionalTurk. Keeping this principle at the core, it was only natural that an empirically grounded emotion framework would be fitting for such a dataset. Therefore, the updated version of De Bruyne *et al.*'s (2024) empirically grounded emotion framework was selected. Furthermore, that initial label set was augmented with the addition of the label *neutral*, in case no emotion could be annotated, the examples of which are abundant in unstructured data, with the rationale behind it being obtaining the goal of mining all latent insights available in the data.

In the same vein as De Bruyne *et al.*'s (2024) previous work, the label set in EmotionalTurk aims to offer accurate insights towards a fairly distributed register of Turkish. Therefore, X has been selected as the data source for this project. X is a social media platform in which users can post short text messages named *posts*, formerly *tweets*, follow other users, repost other users' posts, mention other users in posts or comments and engage in possibly an infinite number of topics. Alongside these properties, X has other features that made it an optimal choice as a data source. X is a multilingual platform. Therefore, it will be easier to create other datasets with the same standards in other languages in future research. It has a large and diverse user base alongside an infinite number of topics, meaning every latent emotion available in text would also be available on X. Lastly, another key advantage of X is in regard to the neutral label included in the emotion framework. X is a platform used by government institutions, government officials, news agencies, and other similar bodies. The content from these accounts is an invaluable resource for the neutral label.

The data source, X, as well as the data collection methodology we applied address the issues found in the TREMO dataset (Tocoglu and Alpkocak, 2018). TREMO is made up of survey responses. The majority of participants, 82% to be exact, belonged to the 14–20-year range. Additionally, 85% of the responses were paper-based. This suggests that most of the data was gathered from survey responses handed to high-school students. Upon inspecting the dataset, it is evident that the TREMO dataset lacks sufficient challenge for SOTA transformers due to limited linguistic diversity and simplicity of syntax. A notable trend is that many examples contain the emotion keyword corresponding to their emotion category. As a result, it is unsurprising that a lexicon-based system can achieve a high accuracy score of 92% on TREMO (Tocoglu and Alpkocak, 2019).

TREMO	Translation	Label
Arkadaş kaybetmek beni üzüyor.	Losing friends makes me sad.	Sadness
Annemin üzüldüğünü görünce üzülüyorum.	I get sad when I see my mother is sad.	Sadness
Okulda ders iptal edildiğinde mutlu olurum.	I am happy when the class is cancelled.	Happy
Mesajlarımın görünüp cevap verilmemesine öfke duyarım.	I get angry when my text messages are seen but not replied.	Anger

Table 1: Randomly selected examples from the TREMO dataset.

Table 1 demonstrates the lack of challenge in TREMO. One entry – 'I am happy when the class is cancelled' – and the fact that an overwhelming ratio of participants are high-school students, suggest that this survey was handed to them during class. Unfortunately, these examples given in **Table 1** are not outliers. The TREMO dataset is not challenging enough.

A key aspect taken into consideration during the data collection process for EmotionalTurk regards bias. The no-bias principle in data collection is important for data integrity, therefore for the quality of the dataset. To further clarify, the no-bias principle entails only that there is no bias in the scope of data collection and does not consider the latent biases available inside the collected textual data. This choice is intentional, because the aim is to gather insights into emotions latently available in the data. Therefore, manipulating the textual data to avoid internal bias is simply a contradiction. In the light of these considerations, data collection was carried out using stopwords only. Using stopwords only prevents collected data to be biased around any kind of topic or trend whatsoever. Furthermore, the text processing in EmotionalTurk is kept to a minimum. This is because, alongside aforementioned rationale, keeping the data raw is beneficial for further research. The minimal text processing carried out on the data collected is limited to privacy issues. The user handles, URLs and hashtag content were substituted with generic tokens, e.g. *#somehashtag* for the hashtag content. Another key consideration for data collection is the timeframe selected for the collection. The data collection was carried out across a single week at a randomly given time. Therefore, the data collection itself does not have any predetermined bias towards any timeline and reflects the natural trends of that timeline.

In order to convey the label integrity of EmotionalTurk, inter-annotator agreement (IAA) studies were carried out. Two experienced annotators contributed to the IAA. Additionally, they were further trained on labelling emotions via an annotation guideline handbook (Maladry *et al.*, 2024) which was introduced within the scope of the Explainability for Cross-Lingual Emotion in Tweets (EXALT) (Singh *et al.*, 2023). EXALT is a shared task which aims to investigate to what extent the emotion information is transferable across languages. EXALT also uses the empirically grounded emotion framework put forward by De Bruyne *et al.* (2024), therefore, the label set is compatible with EmotionalTurk. After translating it into Turkish, this annotation guideline handbook was used in EmotionalTurk's IAA.

A subset of 100 posts was randomly selected from the collected data for the IAA. The collected data was split into 10 equal subsets and 10 posts were randomly selected from each subset to make up the 100 posts for the IAA. Thus, the IAA subset becomes a fair representation of the collected data as a whole. After the annotation of the IAA subset, the data is examined to find out whether annotators are able to fully comprehend and apply the rules put forward in the guidelines, leading to high-quality and reliable annotations. In case the results are not satisfactory in that regard, the respective run of annotation is considered a *pilot*. Then, annotators receive feedback on the pilot and another IAA subset is created with the same methodology. This, in theory, is an iterative process repeated until annotators reach the desired understanding of the rules put forward by the annotation guidelines. For EmotionalTurk, a total of two pilot runs occurred before reaching the final annotations for IAA.

	Anger	Joy	Love	Sadness	Fear	Neutral	Discard
Anger	20	0	0	0	0	0	6
Joy	0	8	0	0	0	0	1
Love	0	0	4	0	0	0	1
Sadness	0	0	0	6	0	0	1
Fear	2	0	0	0	1	0	0
Neutral	5	0	0	0	0	13	2
Discard	2	1	0	0	0	2	25

Table 2: IAA label confusion matrix for two annotators.

In addition to the discussed emotion labels, the label set for the IAA also included the *discard* label. This discard class is used in specific cases (detailed in the annotation guidelines) where there is, for instance, irony in the content, none of the labels properly fit, or multiple labels fit, and there is no clear justification to whether one is preferable, the content is generated with a bot, or other similar cases.

In order to quantify the IAA score for label integrity, Cohen's Kappa coefficient (Cohen, 1960) was used. The resulting kappa score was 0.70, which on the Landis and Koch scale (Landis and Koch, 1977) is equivalent to substantial agreement.

Emotion	**Count**
Anger	474
Fear	57
Sadness	266
Love	153
Joy	188
Neutral	395
Total	1533

Table 3: Entry counts in EmotionalTurk.

After annotation, the resulting dataset, EmotionalTurk, had 1,533 entries in total. The amount of data instances, 1,533, is deemed sufficient for a gold standard evaluation set. The resulting dataset shows a rather unbalanced class distribution, with *anger* being the class with most data, 474, and *fear* being the class with the least data, 57. The uneven class distribution is somewhat expected and par for the course when the goal is to reflect the distribution that is latently available in real data during a random timeframe.

Cross-lingual transfer in Turkish EDFT

Cross-lingual transfer is a novel method of leveraging higher-resource language data for better performance of applications in lower-resource languages. In essence, it refers to the ability of the model that is trained in a language, e.g. English, to carry out tasks such as EDFT in another language,

e.g. Turkish. A popular approach of applying cross-lingual transfer is based on multilingual word embeddings. During pre-training, the large language model is trained with multiple languages, therefore, the resulting base model has semantic understanding of similar concepts across the languages it has been trained on. The cross-lingual transfer happens, conceptually, when the base model is end-to-end fine-tuned on the target task with data in one language and the resulting fine-tuned model is applied to data in another language.

In the context of the research, Turkish was the lower-resource language, therefore, the target language of application and testing, while English was the higher-resource one. The English dataset from the EXALT shared task (Maladry *et al.*, 2024) was selected as the dataset for fine-tuning. This dataset had the same standards set out for EmotionalTurk, i.e. data was collected from X, the keyword criteria for the search were limited to stopwords, and it had the same text processing strategy as EmotionalTurk, namely substituting hashtag, user handle and URL content out with generic tokens. In the same vein as EmotionalTurk, EXALT-English also had satisfactory IAA results and thus reliable label quality. Compared to EmotionalTurk's 100 examples, the IAA for EXALT-English was carried out with 50 examples only. The IAA experiment resulted in a Krippendorf's Alpha (Hayes and Krippendorf, 2007) score of 0.6, and Fleiss' Kappa (Fleiss, 1971) score of 0.62, signifying substantial agreement.

Emotion	**EmotionalTurk**	**EXALT-English**
Anger	474	1028
Fear	57	143
Sadness	266	560
Love	153	579
Joy	188	1293
Neural	395	1397
Total	1533	5000

Table 4: Data counts per emotion class of the EmotionalTurk and EXALT-English datasets.

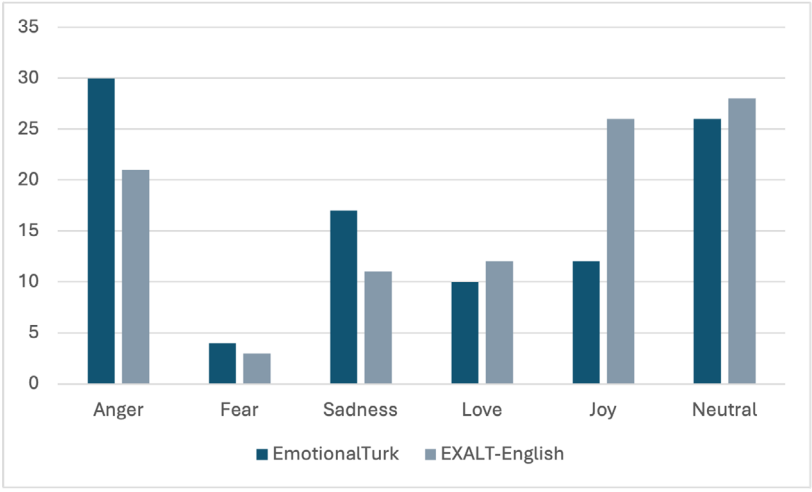


Figure 1: Label distribution by percentage in EmotionalTurk and EXALT-English.

Table 4 contains data counts for each emotion class in EmotionalTurk and EXALT-English datasets and **Figure 1** provides a comparative view of each emotion class representation percentage across these two datasets. As gathered from **Figure 1**, there are both similarities and disparities in their label distributions. The emotions *neutral*, *love* and *fear* occupy a very close percentage in both datasets, while the disparity grows, *sadness* and *anger* classes have acceptable levels of percentage variance. However, the *joy* class representation difference stands out from the rest. The class occupies more than two times the percentage of data in EXALT-English compared to EmotionalTurk. However, the biggest concern regarding the resulting fine-tuned classifier was the lack of data found in the *fear* class. Although the representation of *fear* is similar across two datasets, it is important for EXALT-English to have more examples in *fear*. It is due to the fact that the EXALT-English dataset was the fine-tuning dataset in this study. Therefore, there could be some issues regarding the model's understanding of *fear* due to the lower number of examples in EXALT-English.

XLM-RoBERTa$_{BASE}$ (Conneau *et al.,* 2020) has been selected as the base model to be fine-tuned and tested. XLM-RoBERTa$_{BASE}$ is an improvement on mBERT (Devlin *et al.,* 2018) via the optimisations implemented to the self-supervised training strategies at the pre-training stage. The most recent inclusion of multilingual transformers for the task of Turkish EDFT was mBERT (Ucan *et al.,* 2021) and the cross-lingual transfer capabilities of mBERT were not utilised. The mBERT experiment was a monolingual Turkish EDFT experiment. As reported by Conneau *et al.* (2020), XLM-RoBERTa$_{BASE}$ performs

better in Turkish cross-lingual transfer compared to mBERT with scores of 72.4 and 61.6 respectively. However, these scores might not necessarily reflect the performance gap between the two in the context of Turkish EDFT domain. Therefore, mBERT is also included in the study for comparative analysis between the two.

Evaluation metrics to measure the performance of the resulting models consist of the accuracy score, $F1_{MACRO}$ and $F1_{WEIGHTED}$ scores. The $F1_{MACRO}$ score is calculated by taking the arithmetic means of F1 scores of each class. Therefore, it is an overall indication of the performance of the multi-class classifier. It is important to understand that the $F1_{MACRO}$ score treats each class equally. Thus, in the cases where there is a class imbalance in the dataset such as in EmotionalTurk, a less represented class like *fear* in EmotionalTurk will be treated as equal to a more prevalent class like *anger*. Due to the small amount of training examples available for *fear* in the EXALT-English dataset, the test scores might suffer substantially. Therefore, the $F1_{MACRO}$ score might not tell the whole story of the models' performance. This is where the $F1_{WEIGHTED}$ is useful because it weighs the F1 score of each class in proportion to its representation in the dataset. Including both $F1_{MACRO}$ and $F1_{WEIGHTED}$ is, in this sense, preferable for reporting scores for multi-class classification models.

The experiment setup consists of pre-trained XLM-RoBERTa$_{BASE}$ and mBERT. They are both extended with ANN classifier heads. The fine-tuning setup consists of cross-entropy loss for loss calculations, SoftMax function for probability distribution, Argmax function for selecting the most probable class, and AdamW optimizer with a learning rate of $2e - 5$ for optimising model parameters.

Metric	**mBERT**	**XLM-RoBERTa$_{BASE}$**
Accuracy	33	57
$F1_{MACRO}$	25	43
$F1_{WEIGHTED}$	32	58

Table 5: Performance scores for mBERT and XLM-RoBERTa$_{BASE}$.

Table 5 shows the performance metrics of mBERT and XLM-RoBERTa$_{BASE}$. The performance difference between mBERT and XLM-RoBERTa$_{BASE}$ is an obvious one. All metrics used in the research report that XLM-RoBERTa$_{BASE}$ outperforms mBERT with a significant margin of improvement. Therefore, it is conclusive that XLM-RoBERTa$_{BASE}$ has better cross-lingual

transfer performance in Turkish EDFT compared to mBERT. Furthermore, the XLM-RoBERTa$_{BASE}$ scores, surpassing mBERT, are also the SOTA scores in Turkish EDFT in a cross-lingual transfer setting since this study is the first one to explore cross-lingual transfer performance of multilingual transformers in Turkish EDFT.

Additionally, XLM-RoBERTa$_{BASE}$ reached its peak performance in three epochs while mBERT reached it at the fourth epoch. This could be due to XLM-RoBERTa$_{BASE}$ having a higher number of parameters than mBERT, 279 million to 179 million to be exact. Therefore, XLM-RoBERTa$_{BASE}$ is also more efficient than mBERT.

	Anger	**Fear**	**Joy**	**Love**	**Sadness**	**Neutral**
Anger	374	26	14	22	83	73
Fear	0	0	0	0	0	0
Joy	62	7	127	81	58	57
Love	5	0	37	41	18	14
Sadness	13	9	3	2	95	10
Neutral	20	15	7	7	12	241

Table 6: XLM-RoBERTa$_{BASE}$-based emotion classifier confusion matrix where columns are ground truths and rows are predictions.

As seen from the confusion matrix shown in **Table 6**, the model was not able to correctly classify any examples of the *fear* class. This explains the significant performance disparity between F1$_{MACRO}$ and F1$_{WEIGHTED}$ scores. It is safe to say that having 143 examples of a class is not nearly enough for an XLM-RoBERTa$_{BASE}$-based classifier to learn anything meaningful in a cross-lingual transfer setting.

The model had peak performance on the *anger* class. Correctly classifying 374 entries out of 474, the model achieved 79% accuracy on *anger*. This is the case even though *anger* is not the class with the most instances in the fine-tuning dataset. Therefore, having more data did not proportionally increase the per-class performance of the model. For instance, *joy* has more examples in the fine-tuning dataset than *anger*. Even so, the XLM-RoBERTa$_{BASE}$-based classifier obtained an accuracy score of 67% for *joy* compared to 79% accuracy in *anger*. This phenomenon could be attributed to the cross-lingual transfer setting of the experiment where linguistic features for expressing emotions comes into play.

Conclusion

The article introduced a new dataset, EmotionalTurk, to evaluate Turkish EDFT classifiers. EmotionalTurk debuts a novel and empirically grounded emotion framework to Turkish EDFT research. Furthermore, be it data collection or labelling, EmotionalTurk has been built with strategies which aim for an unbiased dataset. Therefore, this novel dataset aspires to set a new standard for Turkish EDFT datasets. EmotionalTurk has solved the challenge found in TREMO with its data collection strategy and solved the label integrity issues in TURTED due to its automatic labelling scheme. As a result, EmotionalTurk establishes itself as the current standard as far as Turkish EDFT datasets go.

However, EmotionalTurk is by no means a large dataset. Limited to only 1,533 entries, EmotionalTurk is an evaluation benchmark set. This coincides with the goal of the research carried out, i.e. exploring the cross-lingual capabilities of multilingual transformers in Turkish EDFT. The research used the EXALT-English dataset to fine-tune XLM-RoBERTa$_{BASE}$-based and mBERT-based models for cross-lingual Turkish EDFT. The results show that the XLM-RoBERTa$_{BASE}$-based model has significantly outperformed the mBERT-based model both in terms of performance as well as training efficiency. In the same vein, the research also set the SOTA scores for cross-lingual Turkish EDFT.

Future Work

There are two main directions for future research. First, EmotionalTurk should be populated with more data using the same methodology described in this paper. The IAA should also be included in the future additions. The threshold for the IAA in order for the data to be included in EmotionalTurk is 0.6.

The second line of research concerns the next steps in cross-lingual transfer explorations for Turkish EDFT. Although transformers trained with sub-word tokenizers have been achieving SOTA results in many NLP applications, there is high potential for multilingual transformers trained with character-based tokenisers. Investigating the cross-lingual transfer capabilities of character-based multilingual transformers in Turkish EDFT would be an interesting avenue for future research.

References

Cohen, J. (1960). 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement*, 20 (1), 37–46. Available at: https://doi.org/10.1177/001316446002000104.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale [Preprint]. *Arxiv.org*. Cornell University. Available at: https://arxiv.org/abs/1911.02116.

De Bruyne, L., Karimi, A., De Clercq, O., Prati, A., Hoste, V. (2022). 'Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts'. *ACLWeb*. Marseille, France: European Language Resources Association. Available at: https://aclanthology.org/2022.lrec-1.61/.

De Bruyne, L., De Clercq, O., Hoste, V. (2019). 'Towards an Empirically Grounded Framework for Emotion Analysis'. *HUSO*. Available at: https://core.ac.uk/download/pdf/225020418.pdf.

De Bruyne, L., Van der Meer, T., De Clercq, O., Hoste, V. (2024). 'Using State-of-the-art Emotion Detection Models in a Crisis Communication Context'. *Computational Communication Research*, 6 (1). Available at: https://doi.org/10.5117/ccr2024.1.1.debr.

Desmet, B., Hoste, V. (2013). 'Emotion Detection in Suicide Notes'. *Expert Systems with Applications*, 40 (16), 6351–6358. Available at: https://doi.org/10.1016/j.eswa.2013.05.050.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. *Arxiv.org*. Cornell University. Available at: https://arxiv.org/abs/1810.04805.

Ekman, P. (1992). 'An Argument for Basic Emotions'. *Cognition and Emotion*, 6 (3–4), 169–200. Available at: https://doi.org/10.1080/02699939208411068.

Fleiss, J. L. (1971). 'Measuring Nominal Scale Agreement Among Many Raters'. *Psychological Bulletin*, 76 (5), 378–382. Available at: https://doi.org/10.1037/h0031619.

Hayes, A. F., Krippendorff, K. (2007). 'Answering the Call for a Standard Reliability Measure for Coding Data'. *Communication Methods and Measures*, 1 (1), 77–89. Available at: https://doi.org/10.1080/19312450709336664.

Landis, J. R., Koch, G. G. (1977). 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics*, 33 (1), 159–176. https://doi.org/10.2307/2529310.

Maladry, A., Singh, P., Lefever, E. (2024). 'Findings of the WASSA 2024 EXALT shared task on Explainability for Cross-Lingual Emotion in Tweets'. In: *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment Social Media Analysis@ACL 2024*', Bangkok, Thailand: Association for Computational Linguistics. Available at: https://aclanthology.org/2024.wassa-1.43/.

Nandwani, P., Verma, R. (2021). 'A Review on Sentiment Analysis and Emotion Detection from Text'. *Social Network Analysis and Mining*, 11(1), 81. Available at: https://doi.org/10.1007/s13278-021-00776-6.

Plutchik, R. (1980). 'A General Psychoevolutionary Theory of Emotion'. In: *Theories of Emotion*, 3–33. Available at: https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

Sailunaz, K., Dhaliwal, M., Rokne, J., Alhajj, R. (2018). 'Emotion Detection from Text and Speech: a Survey'. *Social Network Analysis and Mining*, 8(1), 28. Available at: https://doi.org/10.1007/s13278-018-0505-2.

Shaver, P., Schwartz, J., Kirson, D., O'Connor, C. (1987). 'Emotion Knowledge: Further Exploration of a Prototype Approach'. *Journal of Personality and Social Psychology*, 52(6), 1061–1086. Available at: https://doi.org/10.1037//0022-3514.52.6.1061.

Singh, P., Maladry, A., Lefever, E. (2023). 'Annotation Guidelines for Labeling Emotion in Multilingual Tweets'. *LT3 Technical Report Series.* Ghent University. Available at: https://biblio.ugent.be/publication/ 01HHSH3CDPHWA217Y73PHAVK5R.

Sosea, T., Pham, C., Tekle, A., Caragea, C., Li, J. J. (2022). 'Emotion Analysis and Detection During COVID-19'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/2107.11020.

Tocoglu, M.A., Alpkocak, A. (2018). 'TREMO: A Dataset for Emotion Analysis in Turkish'. *Journal of Information Science*, 44(6), 848–860. Available at: https://doi.org/10.1177/0165551518761014.

Tocoglu, M. A., Ozturkmenoglu, O., Alpkocak, A. (2019). 'Emotion Analysis From Turkish Tweets Using Deep Neural Networks'. *IEEE Access*, 7, 183061–183069. Available at: https://doi.org/10.1109/ACCESS.2019.2960113.

Tocoglu, M. A., Alpkocak, A. (2019). 'Lexicon-based Emotion Analysis in Turkish'. *TÜBİTAK Academic Journals*. Available at: https://journals.tubitak.gov.tr/elektrik/vol27/iss2/40/.

Ucan, A., Dorterler, M., Akcapinar Sezer, E. (2021). 'A Study of Turkish Emotion Classification with Pretrained Language Models'. *Journal of Information Science*, 48(6), 857–865. Available at: https://doi.org/10.1177/0165551520985507.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017). 'Attention Is All You Need'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/1706.03762.

https://doi.org/10.33919/JCAL.24.2.3

MULTIDIMENSIONAL EVALUATION OF A DOMAIN-SPECIFIC MACHINE TRANSLATION ENGINE

Veridiana Silva

Universidad de Malaga and New Bulgarian University
veridiana.rcs@gmail.com

Abstract

The study presented in this article aims to evaluate the performance of WIPO Translate, a domain-specific machine translation (MT) tool, against three commercially available MT systems: Google Translate, DeepL and OpenAI's ChatGPT 3.5. The evaluation focuses on assessing the translation outputs of patent excerpts across five dimensions, based on the Multidimensional Quality Metrics framework (MQM), namely *Accuracy, Audience Appropriateness, Linguistic Conventions, Style* and *Terminology,* from English to Brazilian Portuguese, deriving an overall quality score from the error annotations. The findings suggest that WIPO Translate, despite being specifically trained on patent corpora and performing reasonably well, does not outperform the commercially available MT systems in terms of translation quality. Generic systems demonstrate higher OQS averages in most cases, with DeepL generally achieving the highest scores among them. Automatic quality metrics were also implemented with BLEU, BERTScore, COM-

ET and COMET-QE. WIPO Translate also ranked lowest in the automatic evaluation metrics. These results indicate that the broad training datasets and advanced algorithms used by general-purpose MT systems provide a competitive edge, even in specialized domains such as patent translation. Future improvements to WIPO Translate could focus on refining its algorithms and expanding its training data to bridge this performance gap. Google and ChatGPT also presented superior performances, the latter being an interesting resource for users to customise the translation through detailed prompts. WIPO Translate still performs well according to the automatic metrics, and can also be very useful as a corpora tool, assisting users in gauging patent style and drafting better patents on their own.

Keywords: *WIPO, patent translation, machine translation, automatic evaluation, quality metrics*

1. Introduction

1.1 Topic relevance and research problems

The present article aims to elaborate on and analyse the quality of WIPO Translate[1], a domain-specific machine translation tool trained exclusively on patent corpora, as the organisation affirms that WIPO Translate outperforms commercially available machine translation engines. We attempt to assess if/to what extent the output is of superior quality through multidimensional analysis of its translation outputs, in the English into Brazilian Portuguese combination, in comparison to three different commercially available MTs, by fulfilling the following goals:

– evaluate the performance of WIPO Translate against three commercially available machine translation systems, namely: Google Translate[2], DeepL[3] and OpenAI's ChatGPT 3.5[4] according to specific dimensions through human evaluations.

– assess the performance of WIPO Translate against commercially available machine translation systems through automatic quality evaluation and estimation metrics, namely BLEU (Bilingual Evaluation Understudy),

[1] Available at: https://patentscope.wipo.int/translate/translate.jsf?interfaceLanguage=es.

[2] Available at: https://translate.google.com/.

[3] Available at: https://www.deepl.com/translator.

[4] Available at: https://chatgpt.com/.

BERTscore (Bidirectional Encoder Representations from Transformers) and COMET/COMET-QE (Crosslingual Optimized Metric for Evaluation of Translation and Crosslingual Optimized Metric for Evaluation of Translation Quality Estimation).

– identify areas of improvement based on the results obtained, if any.

In order to achieve these goals, this study proposes the research questions listed below.

1.2 Research questions

What are WIPO Translate's limitations and advantages compared with other tools based on human evaluations across multiple dimensions (*Accuracy, Audience appropriateness, Linguistic conventions, Style* and *Terminology*)?

What are WIPO Translate's limitations and advantages compared with other tools based on automatic evaluation metrics?

How can the insights gained from the performance comparison be applied to optimise translation workflows in real-world scenarios?

1.3 About the World Intellectual Property Organisation and patent translation

With the aim of regulating intellectual property (IP), the World Intellectual Property Organisation (WIPO)[5] was created in 1967, although its history goes back to 1883 with the Paris Convention.

WIPO is an agency of the United Nations (UN) devoted to the creation of an equitable and inclusive global system of intellectual property; it aims to promote and reward creativity, innovation and economic growth, as well as safeguard public interests in a balanced manner (UN GENEVA, 2024).

WIPO processes and stores millions of documents, such as patent applications, which are written in several languages; as a requirement, they must have their titles and abstracts translated into at least English and French (Pouliquen *et al.*, 2011). As a part of the UN, WIPO currently has 193 member states and receives applications in many different languages, thus creating the need for an automated translation system. To address this need, the World Intellectual Property Organisation developed its own machine translation system, called WIPO Translate. The software is named TAPTA (Translation Assistant for Patent Titles and Abstracts) and was developed to assist translators, either in-house or outsourced by the organization (Pouliquen *et al.*, 2011).

[5] https://www.wipo.int.

1.4 Machine Translation and LLM paradigms

As any specialised service, patent translations are extremely costly. Therefore, inventors and patent professionals may leverage these tools to have an idea about the content, something known as *gisting*. Tinsley (2017) explains that gisting refers to providing an on-demand gist translation of foreign patents for information purposes to determine its relevance. This allows users to quickly understand the essential content of a patent document without needing a full, legally accurate translation, which can be extremely beneficial when dealing with large volumes of documents, where human translation is not feasible due to time and cost constraints, or also to understand more about already patented inventions and see if a novel one is worth pursuing (Rossi and Wiggins, 2013).

In sequence, **Table 1** provides a brief comparison between the current MT approaches in order to facilitate our understanding of what is behind each system's strengths and shortcomings, as well as to present the main ideas about each system.

MT Approach	Main Characteristics	Advantages	Challenges	Examples	References
Rule-based (RBMT)	Based on linguistic rules, including dictionary-based, transfer-based and interlingual methods.	Effective for closely related languages, precise grammatical rules.	Labor-intensive, language pair-specific and struggles with ambiguous sentences.	Dictionary-based, Transfer-based, Interlingual MT	Maučec and Donaj (2020)
Example-based (EBMT)	Relies on translation examples from parallel corpora to generate translations.	Useful for analogous sentence structures, can learn from past translations.	Limited by availability of suitable examples in the corpus, challenges with unseen phrases.	Example-based systems in Japan since the 1980s	Maučec and Donaj (2019)
Statistical (SMT)	Based on statistical models using large bilingual corpora to generate translations.	Data-driven, adaptable with enough training data, handles ambiguity better than RBMT.	Requires large corpora and computational power, struggles with low-resource languages.	IBM's work in 1980s, Moses system, PBSMT	Maučec and Donaj (2019) ; Koehn (2010)

Hybrid	Combines rule-based and statistical methods to improve translation quality.	Leverages the strengths of both RBMT and SMT, efficient use of resources.	Can be complex to implement, computational cost can be high.	Google Translate (initially), Systran, Microsoft	Maučec and Donaj (2019); Rivera-Trigueros (2022)
Neural Machine Translation (NMT)	Uses artificial neural networks to generate translations, often with encoder-decoder models.	Fluency, can account for context, adaptable to large datasets.	Requires vast training data, slow training process, challenges with low-resource languages.	Google Translate (current), OpenNMT	Rivera-Trigueros (2022); Maučec and Donaj (2019)
Large-Language Models (LLM)	Utilizes deep learning on vast text data, excels with few-shot learning and can be flexible with prompts.	Flexible, can handle diverse tasks with minimal training, fluent text generation.	Struggles with low-resource languages and complex semantic linkages, requires large computational power.	GPT-3.5, Gemini, other LLMs used for translation	Zhao *et al.* (2023) ; Brown *et al.* (2020); Briva-Iglesias *et al.* (2024) Vaswani *et al.*, (2017).

Table 1: *Comparative view on MT approaches and LLM for machine translation.*

1.5 Large language models as MT systems

Large language models (LLMs) are advanced artificial intelligence (AI) systems that use deep learning and vast amounts of text data to perform countless language-related tasks by understanding and generating human-like text (Jana *et al.*, 2024).

The foundational algorithm for LLMs in machine translation is the transformer algorithm, introduced by Vaswani *et al.* (2017): it features self-attention, which enables each word to consider every other word in the input sequence for improved context capture compared to earlier models reliant on sequential processing, like recurrent neural networks (RNNs) or long short-term memory networks (LSTMs). By employing specific instructions or commands known as prompts, these AI applications can perform tasks proficiently with a minimal number of training examples. LLMs are more flexible and require less training and fine-tuning on substantial datasets, unlike NMT systems that require extensive training and often show limit-

ed flexibility beyond their trained purpose (Briva-Iglesias *et al.*, 2024). The self-attention mechanism is crucial for weighing the importance of different words in the input while generating each output word, facilitating the model's ability to understand long-range dependencies and subtle nuances necessary for high-quality translation (Vaswani *et al.*, 2017).

LLMs present some advantages in comparison to traditional machine translation paradigms. As LLMs can better tackle long-range dependencies, they are able to generate more contextually appropriate translations due to their self-attention mechanism, producing outputs with coherent sentence structures (Vaswani *et al.*, 2017). Their improved context awareness also enables LLMs to grasp richer contextual information, which is especially useful in complex or ambiguous scenarios and possibly leading to better translations (Briva-Iglesias *et al.*, 2024). Another advantage to the use of LLMs in translation is transfer learning, meaning that these systems can leverage large-scale multilingual corpora pretraining, then generalising it to new languages or domains without extensive task-specific data (Arivazhagan *et al.*, 2019).

Despite their advantages, LLMs present notable disadvantages compared to other MT paradigms; for example, LLMs have poor sample efficiency, requiring massive amounts of data input during pre-training, indicating inefficiencies in their learning processes. Additionally, another concern is that LLM training requires massive processing power, leading to significant energy consumption. These models are also susceptible to inheriting biases from the training data, which can result in prejudiced or stereotypical results (Brown *et al.*, 2020).

1.6 Machine translation systems used in this study

1.6.1 WIPO Translate

The first MT engine we will discuss and the object of this study is WIPO Translate; it supports the ten official languages of the Patent Cooperation Treaty (PCT): Arabic, German, Spanish, French, Korean, Japanese, Portuguese, Russian and Chinese. WIPO developed its own translation engine trained on an extensive corpus of manually translated patents. This system was originally named TAPTA, which stands for Translation Assistant for Patent Titles and Abstracts. The TAPTA model was first based on domain-aware statistical machine translation (SMT) using factors (i.e. an SMT considers the technical domain each text segment belongs to as an additional parameter); and it relied on the open-source toolkit Moses (Pouliquen *et al.*, 2011).

WIPO Translate has continually developed its machine translation system and has shifted from an SMT-based model to neural machine translation technology (NMT) since September 2016. As of September 2017, WIPO provides neural machine translation in all the 10 official languages of the Patent Cooperation Treaty (PCT), powered by the open source Marian NMT 1. According to the organisation, it is especially effective when translating technical sentences and phrases, as it was developed using big data generated from published patent documents and is continuously updated with the texts of the same domain.

1.6.2 Google Translate

Google Translate was launched in April 2006 by Google; initially, it used statistical machine translation (SMT) methods, analysing vast amounts of human-translated texts to generate translations by identifying patterns rather than following strict linguistic rules (Sutrisno, 2020). Early versions of Google Translate heavily relied on SMT technology from Systran, enabling translations between English and other languages (Savoy and Dolamic, 2009). Despite its usefulness, SMT had limitations, such as difficulties in accurately capturing context and idiomatic expressions, prompting Google to seek improvements.

In 2016, Google Translate transitioned from SMT to neural machine translation (NMT) with the launch of the Google Neural Machine Translation system (GNMT). This new system uses deep learning techniques to improve translation accuracy by considering entire sentences as units rather than breaking them into smaller parts. GNMT employs a deep long short-term memory (LSTM) network with attention mechanisms and low-precision arithmetic during inference computations to enhance speed and accuracy (Wu *et al.*, 2016). The model operates within a sequence-to-sequence learning framework, comprising encoder and decoder networks connected through an attention module, which allows the decoder to focus on different regions of the source sentence during decoding (Wu *et al.*, 2016).

1.6.3 DeepL

DeepL was developed by Linguee and introduced the DeepL Translator in 2017. Its success stems from using convolutional neural networks (CNNs) and extensive datasets, including billions of translated sentences, which help the neural networks learn language patterns and nuances more effectively than traditional systems (Alzubaidi *et al.*, 2021). DeepL's architecture leverages CNNs to process entire sentences or paragraphs simultaneously,

enhancing translation speed and context capture, unlike recurrent neural networks (RNNs) that process sequences one word at a time (Morita, 2024).

1.6.4 ChatGPT 3.5

ChatGPT 3.5 was developed by OpenAI and is a language model known for generating human-like text. Therefore, it is not a machine translation engine *per se*, but it has been increasingly used for this purpose, as it allows the user to customise the desired translation product through prompts (Briva-Iglesias *et al.*, 2024).

The version used in this research is GPT 3.5, which is the standard version available to users free of charge. This version was chosen for this study considering it may be a useful resource to inventors and professionals in the patent field, who may look to avoid costs at the first moment of a patenting process.

The mentioned tools were selected due to their widespread use. Additionally, at the time of this experiment, neither WIPO Translate nor Google Translate offered the Brazilian Portuguese variant; while DeepL did and ChatGPT was prompted to translate into Brazilian Portuguese as the target language, and this was done in order to ensure some balance concerning the language variant aspect.

A better understanding of the strengths and shortcomings of WIPO Translate in comparison to a commercially available MT may also assist on efficient translation workflow for other organisations or businesses that are also highly technical and domain-specific (Moslem *et al.,* 2022), possibly assisting on streamlining the translation process for specialised content, reducing the time and effort required for human translators to handle post-editing of domain-specific texts and future translations.

2. Methodology

This research compares WIPO Translate's performance against that of Google Translate, DeepL and OpenAI's ChatGPT 3.5 through both automatic evaluation and human evaluation, since they both present their benefits and drawbacks. Automatic evaluation is generally recognised as being objective and inexpensive, while manual evaluation is often claimed to be subjective and can be slow and expensive to perform (Castilho *et al.*, 2018). However, automatic MTE metrics are useful to compare the output of an MT system to one or several reference translations and manual approaches can be useful to assess complex linguistic phenomena beyond mere adequacy and fluency. In addition, manual approaches have the benefit of requiring bilingual competence, while automatic approaches are faster and more cost-effective (Castilho *et al.*, 2018).

2.1 Human evaluation: experiment design, annotator instructions and evaluation

Three evaluations were conducted, analysing the excerpts from three different patent texts, each belonging to a different domain and patent classification. Therefore, we had a total of nine evaluations, performed according to specific criteria and severity weights explained in detail ahead in the *Human evaluation theoretical survey* section.

Patent texts are usually dense and lengthy; with this in mind, and also due to the fact that participants were working voluntarily, we opted to have one assessment per evaluator, as to not overload them. The study was blind; therefore, the evaluators did not know which system they were assessing, and could only see a number designation. In this exercise, System 1 corresponded to WIPO Translate, System 2 to Google Translate, System 3 to DeepL and System 4 to ChatGPT standard version. The analyses were done per system and on a segment level; segments were not randomised because in a later stage the evaluators were asked to state which system they preferred per dimension. Before starting the proper experiment, we carried out a few initial steps, which are described below.

Data collection and source of data: The experiment had precisely 3496 words collected from 3 different patents that were collected from WIPO's database, PATENTSCOPE. We obtained these excerpts from the Claims and Descriptions sections from patents with application codes: CA3211095, CA3225517 and CA3225541. We collected excerpts of the most recent patents (at the time the experiment was conducted, March 2024) to try and minimise the possibility that these texts were used in WIPO Translate's model retraining. The texts were collected on March 5th, 2024, and were published on the WIPO database on March 3rd and 4th, 2024. However, we cannot guarantee that these texts were not previously used in WIPO Translate's model retraining; we contacted them enquiring about this, but unsuccessfully.

Segmentation: The excerpts were organised using their own divisions, such as claim by claim, or in cases where the sentences were shorter, they were combined into segments of two or three sentences. This was done to ensure similar word counts for evaluators, but more importantly, so that the translations would be done within a context, and not isolated. The excerpts of Patent #1 had a total of 1549 words; Patent #2, 1034 words, and Patent #3, 913 words.

Text normalisation: The patent texts collected were publicly available on PATENTSCOPE, the WIPO patent database. Due to the fact that these patent applications were quite recent at the time of collection, their whole

documentation was not yet available publicly. The texts were originally made available on their site through optical character recognition and therefore contained spacing errors which were corrected, as well as stamps such as *Data submitted on,* which were removed. After that, texts were pasted unaltered to all the machine translation systems (WIPO Translate, Google Translate, DeepL, and ChatGPT 3.5).

Participants: The experiment was done with the assistance of 9 annotators/evaluators who worked as volunteers on this task. All of them were professional translators in the English > Brazilian Portuguese combination, although not necessarily translators of technical texts. All the participants have a degree in Translation or in a translation-related field, with diverse levels of professional experience, although most of them were professionals for more than 3 years. At the end of the experiment, they answered a brief questionnaire about their background/professional experience.

Consent and survey: At the beginning of the experiment, participants were informed about the experiment's purpose, and were asked to sign a data use consent form. At the end of the experiment, they responded to a questionnaire about their professional background, as well as their impressions about the experiment. This questionnaire also asked them about their preferred system per dimension evaluated, which is why the systems' names were anonymised, but their outputs were *not* randomised.

2.2 Human evaluation theoretical survey

The data were organised in an electronic spreadsheet containing the source text in English and the output of the different systems in the following columns. The evaluators were instructed to read the source text and then read the translation outputs provided. Next, they annotated the error(s) found and their level of severity from a drop-down menu. Since more than one error or error dimension could be found, the guidelines were to annotate each error they encountered, being one error per line, while assessing the *Accuracy, Audience appropriateness, Linguistic conventions, Style* and *Terminology* dimensions for the whole *segment*. The evaluation dimensions of this experiment were based off of the Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics (Lommel *et al.*, 2014), which offers the possibility of adapting dimensions the researcher wishes to explore in their evaluation.

In this study, the idea was to have *all* errors found in the segments annotated. So, for example, if a segment presented an error in *Style* and another in *Terminology*, annotators would dedicate one line to *each error* and

then assign their respective severity weight, for each and every error (not by dimension). Additionally, evaluators could come across repeated errors. According to the MQM framework, repeated errors could be accounted for only once with a higher weight, or every time the error was found, but with an 'adequate' weight. In this paper, we chose to account for each and every occurrence given the very precise nature of patent texts.

According to the MQM Framework website[6], there are seven high-level dimensions in the MQM error typology, also known as *MQM-Core*: *Accuracy, Audience Appropriateness, Design and Markup, Linguistic Conventions, Locale, Style, Terminology* and *Custom*. Even though each dimension presents a subset of more specialised error classes (making up the *MQM-Full* framework), these high-level dimensions can in fact be used as error categories on their own, as the researcher determines which dimensions are more relevant for their work, as well as the required granularity level. In this research, we chose to use the *MQM-Core* framework, i.e. considering only the dimensions at a higher level.

The MQM-Full framework would have been interesting considering the texts translated were patents, and therefore highly specialised. However, since we counted on volunteer translators/annotators, the full framework would have been extremely time-consuming and overwhelming to them.

The top-level dimensions considered for the experiment regarded as the most pertinent ones to patent translations were *Accuracy, Audience appropriateness, Linguistic conventions, Style* and *Terminology*; we provide a brief summary of why each dimension was considered relevant below:

Dimensions	**Relevance**
Accuracy	**What it represents:** Assesses whether the translation correctly reflects the source text meaning, without additions, omissions or distortions. **Why it is important:** Accuracy is essential, as even small errors or misinterpretations can change the meaning of claims and descriptions, causing legal or technical issues. Mistakes can invalidate patents, lead to infringement disputes or result in the loss of rights, making precise translations crucial for maintaining the intended scope of the patent.

[6] Available at: https://themqm.org/resources/globalreadiness2024.

Dimensions	Relevance
Audience Appropriateness	**What it represents:** Measures how well the translation aligns with the expectations, knowledge level and cultural context of the target audience. **Why it is important:** The translation must suit the intended audience or region. For example, using European Portuguese instead of Brazilian Portuguese can cause confusion, resulting in delays or rework during the patent filing process. This dimension ensures the content aligns with the audience's expectations and avoids communication breakdowns.
Linguistic Conventions	**What it represents:** Ensures the translation follows grammar rules, spelling, punctuation and syntactic norms of the target language. **Why it is important:** Ensuring adherence to grammar, sentence structure, punctuation and syntax rules prevents ambiguity and confusion. Unclear translations can compromise the patent's readability and legal enforceability, making it difficult for readers to understand the intended meaning.
Style	**What it represents:** Evaluates whether the tone, formality and register are appropriate for the translation task and consistent throughout. **Why it is important:** A consistent, formal style aligns with the technical nature of patent documents, helping to convey information clearly. Inconsistent style can distract readers and reduce the professionalism and clarity of the content.
Terminology	**What it represents:** Focuses on the correct and consistent use of domain-specific terms. Ensures specialised terms match the context and adhere to glossaries. **Why it is important:** Patents often contain technical terms and specific jargon that must be translated consistently and accurately to maintain the legal and technical integrity of the document. Incorrect or inconsistent terminology can lead to misunderstandings, potentially weakening the patent's protection and causing legal disputes.

Table 2: Description of dimensions and error types
(adapted from: https://themqm.org/the-mqm-full-typology/)

The MQM framework also determines the severity of the errors encountered, which are classified as *neutral*, *minor*, *major* and *critical*, with weights of 0, 1, 5 and 25 respectively. The MQM framework defines these weights as such (DFKI, QTLaunchPad, 2015). **Table 3** summarises the main aspects regarding the severity weights.

Severity Level	Description
Neutral (0)	No penalty; the issue is minor or beyond the translator's control (for example, lack of guidelines or style guides)
Minor (1)	Small impact on content, but does not hinder usability or clarity.
Major (5)	Significantly affects understanding or usability.
Critical (25)	Renders the content unfit for use, risking serious physical, financial or reputational harm.

Table 3: Error severity weights
(adapted from: https://themqm.org/the-mqm-full-typology/)

The MQM scoring model allows for adjustments to meet specific stakeholder needs, for example which dimensions are relevant or if the examiners consider the full error typology or the core one. This methodology evaluates translation quality based on error counts, severity levels and error type weights, producing a score that indicates whether the translation meets specified quality standards. In the calibrated approach, the examiners set a passing threshold for the scores, which is useful to streamline quality assurance (QA) procedures in the context of language service providers (LSPs), for example. However, since our goal was more focused on the quality ratings than the usability of translation outputs, we opted for the non-calibrated scoring.

The non-calibrated MQM model follows a series of steps involving key metrics like evaluation word count (EWC), reference word count (RWC), and maximum score value (MSV, 100). Errors are weighted based on type and severity, with penalties assigned according to the error's impact on usability (e.g., minor, major, critical). The quality score (QS) is calculated by subtracting penalty points, scaled to a per-1000-word basis, from the MSV, resulting in a percentage-like value that reflects the overall translation quality (Lommel *et al.*, 2014).

The figure below exemplifies the scoring sheet, accounting for the error annotations. We analysed the outputs of four MT systems, of the 3 source texts; each text was analysed by 3 different annotators, therefore totaling 36 overall quality scores (OQS). Considering that the MQM methodology subtracts penalties from a maximum score value of 100, the OQS would correspond to the percentage of the sample that is error-free. Figure 1 in sequence illustrates a sample of the non-calibrated MQM scoring model, in which 97.55% of the sample is error-free.

MQM Scorecard: Top-Level Error Typology with 4 Severity Levels							
	Error Severity Levels:	Neutral	Minor	Major	Critical	Error Type Penalty Total	
	Severity Penalty Multipliers:	0	1	5	25		
ET Nos	Error Types	Error Counts				Error Type Weights	Error Type Penalty Total ETPTs
1	Terminology	2	7	7	0	1.0	42.0
2	Accuracy	4	14	7	1	1.0	74.0
3	Linguistic conventions	1	23	9	0	1.0	68.0
4	Style	5	7	3	0	1.0	22.0
5	Locale convention	1	12	5	0	1.0	37.0
6	Audience appropriateness	0	2	1	0	1.0	7.0
						Absolute Penalty Total (APT):	250.00
	Evaluation Word Count (EWC):	10184				Per-Word Penalty Total (PWPT):	0.0245
	Reference Word Count (RWC):	1000				Overall Normed Penalty Total (ONPT):	24.55
	Penalty Scaler (PS):	1.00				Overall Quality Score (OQS):	97.55
	Max. Score Value (MSV):	100.00					

Figure 1: MQM scoring model example.

Once all scores were calculated, we proceeded to conduct the inter-annotator agreement for said scores. The inter-annotator agreement (IAA) is an important measurement in machine translation (MT) evaluation, because it ensures consistency and reliability in subjective judgements made by human evaluators. Artstein and Poesio (2008) discuss that since MT quality assessments often involve human evaluation of fluency, adequacy or overall translation quality, differing opinions between annotators can introduce noise and bias into the evaluation process. A high agreement level among annotators minimises subjectivity, enhancing the objectivity of linguistic evaluations and helps in refining MT systems through accurate and consistent feedback. Furthermore, Callison-Burch *et al.* (2006) emphasise that low inter-annotator agreement can undermine the credibility of evaluation outcomes, leading to false conclusions about MT performance.

2.3 Automatic evaluation metrics

We implemented two automatic evaluation metrics, namely BLEU and BERTScore, providing one gold reference translation to this end.

BLEU (Papineni *et al.* 2002) was chosen because it provides a standardised method for evaluating translations, allowing for consistent comparisons across different models and studies (Reiter, 2018), including some of the related works cited in this research. BLEU Counts reflect how many n-grams (sequences of n-words) from the translated text match the reference text; BLEU Totals represent the total number of n-grams in the machine-translated text; BLEU Precisions (%) is the ratios of BLEU counts to BLEU totals, showing the accuracy of translations at different n-gram levels; the BLEU Brevity Penalty (all systems have a brevity penalty of 1.0, indi-

cating no penalties for translation length); all the systems presented system lengths and reference lengths around 1897 words, with slight variations.

BERTScore (Devlin *et al.*, 2019) is an embedding-based metric which uses the BERT model to generate contextual embeddings for each token (word or subword) in the sentences being compared. These embeddings capture semantic meanings by considering the context in which each token appears and utilise contextual embeddings from pre-trained language models like BERT. The metrics compute similarity scores between candidates and reference sentences by comparing the embeddings of their tokens. BERTScore results provide insights into the performance of different machine translation systems by evaluating three key metrics: precision, recall and F1 score.

After analysing the results from human evaluation and automatic scores, we decided to conduct a quality estimation, given the discrepancies found, following the recommendations of Kocmi *et al.* (2024). We implemented quality estimation after the quality evaluation with the reference-based COMET-MQM_2021 model and the reference-free COMET-QE-MQM_2021, using Google Collab Premium for GPU access.

COMET (Rei *et al.*, 2020) is a neural framework developed to assess the quality of machine translations. It takes advantage of recent advances in cross-lingual pre-trained language models to assess translation quality more correctly, considering both the source text and the target reference translation. COMET models are trained with human judgements on a variety of quality measures, including direct assessments, human-mediated translation edit rate (HTER) and multidimensional quality measures (Rei *et al.*, 2020). COMET-QE is a reference-free form of COMET that is used to estimate translation quality when a reference translation is not available. It uses similar neural architectures and training approaches to COMET, but it focuses solely on judging translation quality based on the original text. COMET-QE has shown significant improvements in selecting high-quality translations in low-resource settings, as demonstrated in experiments with Swahili, Kinyarwanda, and Spanish, where it outperformed other methods by up to 5 BLEU points (Chimoto and Bassett, 2022).

3. Results

3.1 Human evaluation

We assessed the four systems' performance with the general error count, according to the MQM Framework, aiming to reach an overall quality score (OQS) of the different patent texts, in each evaluation, then calculating the systems' average with each patent text. **Table 4** presents the overall quality scores and their averages for each system analysed.

	WIPO	OQS	OQS AVG		Google	OQS	OQS AVG
WIPO PATENT 1 (1549 w)	Evaluation 1	**89.22**	**88.92**	**Google PATENT 1** (1549 w)	Evaluation 1	**88.19**	**90.62**
	Evaluation 2	**92.64**			Evaluation 2	**95.48**	
	Evaluation 3	**84.89**			Evaluation 3	**88.19**	
WIPO PATENT 2 (1034 w)	Evaluation 1	**63.44**	**74.71**	**Google PATENT 2** (1034 w)	Evaluation 1	**73.60**	**69.41**
	Evaluation 2	**97.22**			Evaluation 2	**98.65**	
	Evaluation 3	**63.46**			Evaluation 3	**35.98**	
WIPO PATENT 3 (913 w)	Evaluation 1	**61.12**	**77.22**	**Google PATENT 3** (913 w)	Evaluation 1	**75.36**	**84.99**
	Evaluation 2	**91.02**			Evaluation 2	**97.70**	
	Evaluation 3	**79.52**			Evaluation 3	**81.93**	
	DeepL	**OQS**	**OQS AVG**		**ChatGPT**	**OQS**	**OQS AVG**
DeepL PATENT 1 (1549 w)	Evaluation 1	92.45	**92.70**	**ChatGPT PATENT 1** (1549 w)	Evaluation 1	93.29	**91.67**
	Evaluation 2	93.74			Evaluation 2	88.44	
	Evaluation 3	91.93			Evaluation 3	93.29	
DeepL PATENT 2 (1034 w)	Evaluation 1	84.04	**71.31**	**ChatGPT PATENT 2** (1034 w)	Evaluation 1	81.43	**92.08**
	Evaluation 2	97.68			Evaluation 2	99.79	
	Evaluation 3	32.21			Evaluation 3	95.02	
DeepL PATENT 3 (913 w)	Evaluation 1	90.91	**92.55**	**ChatGPT PATENT 3** (913 w)	Evaluation 1	81.60	**87.51**
	Evaluation 2	96.06			Evaluation 2	89.27	
	Evaluation 3	90.69			Evaluation 3	91.68	

Table 4: Overall quality scores (OQS) and their averages for all systems.

Analysing the performance of the four machine translation systems—WIPO, Google, DeepL and ChatGPT—across three patents reveals notable differences in the overall quality scores. Google Translate presented the lowest OQS average with 69.41; in turn, ChatGPT presented the highest average OQS with 97.64. Both DeepL and WIPO Translate which is this study's main focus are in the middle.

At the end of the evaluation, annotators were asked to answer a brief questionnaire with their impressions about which system had the best performance across the dimensions, in addition to any comments they would like to share. **Table 5** illustrates the average of their preferences per dimension, with the number of votes each system got. In the evaluation of machine

translation systems, System 1 was WIPO Translate, System 2 Google Translate, System 3 DeepL, and System 4 was ChatGPT 3.5. System 1 was not voted amongst annotators' preferences.

Dimension	**Preferred System**	**System 3 (DeepL)**	**System 4 (ChatGPT 3.5)**	**System 2 (Google)**
Accuracy	System 3	7	1	1
Audience Appropriateness	System 3	6	2	1
Linguistic Conventions	System 3	7	1	1
Style	System 3	6	2	1
Terminology	System 3	6	2	1

Table 5: Annotators' preferences for the different MT systems evaluated.

Overall, System 3 emerged as the consistently preferred system across all evaluated dimensions, indicating its strong overall performance. System 4 showed some preference in specific areas like terminology and style but was less consistent across other dimensions. In contrast, System 2 was the least preferred in all categories, highlighting areas needing significant improvement.

The participants found the machine translation evaluation for patent texts both interesting and challenging, primarily due to their lack of expertise in patent translation. Many evaluators struggled with assigning error categories accurately because of the technical nature of the text and the absence of visual reference materials. Despite these challenges, the opportunity to reflect on everyday translation errors and compare different machine translation outputs was appreciated.

Several participants noted that the specialised terminology of patent texts required extensive online research to determine correctness, making the evaluation process more difficult. The lack of visual aids referenced in the text further complicated their understanding and evaluation. Deciding which dimension to evaluate was particularly challenging, as critical errors often appeared in multiple categories, such as terminology and style. This complexity was compounded by the highly technical nature of the text, which some felt required more experienced professionals for accurate annotation.

In terms of system performance, System 3 was generally perceived as having the best overall quality and consistency; System 4 was also noted for its acceptable translations, especially in *Terminology* and *Accuracy.* Conversely, Sys-

tem 1, WIPO and Google were less favored. Additionally, some participants highlighted that the source text itself seemed awkward and confusing, likely impacting the machine translation outputs. Despite these difficulties, Systems 3 and 4 stood out as the more reliable options among the systems evaluated. WIPO Translate was not featured in any of the annotators' preferences.

In summary, the outputs of four machine translation systems were evaluated for the performance; the data for annotated errors across the dimensions studied here, namely *Accuracy*, *Audience appropriateness*, *Linguistic conventions*, *Style* and *Terminology*. The annotations were conducted with the assistance of 9 evaluators, all of which were professional translators, although not specialised in technical translations. A significant limitation of the human evaluation was that, due to the length of the samples and their complexity, it was not possible to have the same evaluators analysing each patent. For this reason, we believe there is great variability in the errors annotated, and the severity with which they were perceived by annotators. For this reason, we conducted an inter-annotator agreement according to Krippendorff's Alpha. We believe this is the most appropriate metric for this study, as it can adjust for chance agreement and handles different levels of measurement, including ordinal (i.e. severity levels), and because it supports data from multiple annotators (Krippendorf, 2013).

Krippendorff's alpha ranges from -1 to 1, where 1 indicates perfect agreement among annotators, 0 suggests agreement at the level expected by chance, and negative values imply systematic disagreement that is worse than chance. In our case, Krippendorff's Alpha presented weak agreements between the raters, for all the three patents, as shown by Table 6.

Patent text	**Krippendorff's Alpha Agreement**
Patent 1	0.10271326716169993
Patent 2	0.6366675272710243
Patent 3	0.04489514112883619

Table 6: Inter-annotator agreement calculated through Krippendorff's Alpha.

These low values suggest that the annotators are inconsistent in how they assess the severity of errors in the machine translation outputs, with varying interpretations of error categories and severities across different systems. Several factors could explain this low agreement. Annotators may differ in their judgements, with one seeing an error as 'minor' while another might classify it as 'major.' There could also be ambiguity in the guidelines, leading to inconsistent application of the criteria for categorising errors. Furthermore, the complexity or subjectivity of the task, particularly in areas like

style or appropriateness, might naturally lead to more disagreement among annotators.

This confirms the divergences observed in the annotation spreadsheets, as well as a deficiency in the guidelines for annotators. However, inter-annotator agreements in general are useful when the annotation categories are consistent across annotators—that is, all three annotators should use the same set of categories. In the current scenario, our annotators could identify errors in several dimensions for the same segment of the same text. This inconsistency means that the IAA score will likely provide little meaningful insight, as it will always remain low due to the varying focus of the annotators.

3.2 Automatic evaluations: BLEU, BERTScore and COMET/COMET-QE

The evaluation of machine translation systems—WIPO Translate, Google Translate, DeepL and ChatGPT 3.5—across three patent texts using BLEU, BERTScore and COMET metrics provided a comprehensive comparison of their performance. Table 7 summarises all the automatic evaluation results below.

	Automatic Evaluation – Patent 1				**Automatic Evaluation – Patent 2**				**Automatic Evaluation – Patent 3**			
System	**BLEU Score**	**BERTScore Precision / Recall / F1**	**COMET**	**COMET-QE**	**BLEU Score**	**BERTScore Precision / Recall / F1**	**COMET**	**COMET-QE**	**BLEU Score**	**BERTScore Precision / Recall / F1**	**COMET**	**COMET-QE**
WIPO Translate	48.4	0.8885 / 0.8949 / 0.8915	0.0118	0.0579	53.4	0.911 / 0.901 / 0.906	0.005	0.04	49.6	0.911 / 0.901 / 0.906	0.0131	0.06
Google	54.4	0.9099 / 0.9175 / 0.9136	0.0137	0.0605	57.2	0.925 / 0.922 / 0.923	0.007	0.05	62.6	0.9379 / 0.9334 / 0.9356	0.0166	0.07
DeepL	55.9	0.919 / 0.917 / 0.918	0.014	0.0623	56.7	0.920 / 0.921 / 0.921	0.007	0.05	55.6	0.9332 / 0.9215 / 0.9273	0.0157	0.06
ChatGPT 3.5	55.8	0.9162 / 0.9127 / 0.9144	0.0145	0.0621	58	0.924 / 0.922 / 0.923	0.007	0.05	54.6	0.9174 / 0.9062 / 0.9118	0.0143	0.06

Table 7: Comparison of automatic evaluations using BLEU, BERTScore and COMET.

BLEU scores were low across all systems, however WIPO Translate presented the lowest results for all three patents evaluated. On the other hand, BERTScores (Zhang *et al.*, 2020) were quite high for all the systems, albeit lower for WIPO Translate. We believe this behaviour is due to BLEU being a metric that rewards short and more similar translations to the source text. In contrast, BERTScore is based on contextual embeddings and the model is able to grasp content and different translation choices beyond an established translation.

WIPO Translate, although trained exclusively on patent corpora, presented results indicating that it usually performs well but consistently falls behind the other systems in key evaluation metrics. When examining the BLEU scores, which measure the overlap between the MT output and reference translations, WIPO Translate scores lower than the commercial systems. For instance, in Patent 1, WIPO Translate achieved a BLEU score of 48.4, while Google Translate, DeepL and ChatGPT 3.5 scored 54.4, 55.9, and 55.8, respectively. This trend continues across Patents 2 and 3, where WIPO Translate's BLEU scores remain lower than those of the other systems.

Similarly, BERTScore, which evaluates precision, recall and F1 scores by leveraging pre-trained language models to assess the semantic similarity between MT outputs and references, shows higher scores for the commercial systems. For Patent 1, WIPO Translate had a BERTScore F1 of 0.8915, compared to Google's 0.9136, DeepL's 0.918, and ChatGPT's 0.9144. This pattern persists in the subsequent patents, with the commercial systems consistently achieving higher BERTScore metrics.

The COMET metric, which combines multiple aspects of translation quality, further underscores the superior performance of the commercial systems. For Patent 1, WIPO Translate scored 0.0118, while Google Translate scored 0.0137, DeepL 0.014, and ChatGPT 3.5 0.0145. This trend continues across Patents 2 and 3, where the commercial systems again outperform WIPO Translate.

4. Discussion

In this section, we will elaborate on our findings in the light of this study's research questions. The present work aimed to elaborate on and analyse the quality of the World International Property Organisation's machine translation system, which was developed and trained exclusively on patent corpora, therefore making it a domain-specific machine translation engine. WIPO affirms that WIPO Translate outperforms commercially available machine translation engines.

The human evaluation and automatic evaluation metrics used analysed if/to what extent WIPO Translate's output is of superior quality in comparison

to 3 other generic and more commercial engines, namely Google Translate (also referred to as *System 2*), DeepL (*System 3*), and ChatGPT 3.5 (*System 4*), and these tools were chosen due to their widespread use. ChatGPT 3.5 is a large language model (LLM), therefore, while it is not a machine translation engine *per se*, it has been increasingly used to this end, as it may be catered to the users' specific demands or prompts. In this study we chose the standard version reasoning that users would not use a paid version if their main interest in patents is just for gisting, such as understanding the content of similar patents already submitted to patenting processes.

The first goal of this study was to describe and compare WIPO Translate and the three generic machine translation systems chosen: Google Translate, DeepL and OpenAI's ChatGPT 3.5. The study was primarily focused on these tools' performance across the dimensions of *Accuracy*, *Audience Appropriateness, Linguistic Conventions, Style* and *Terminology*, following the MQM Framework methodology. The following aim was to experiment and analyse the performance of WIPO Translate with various automatic quality evaluation metrics, Bleu and BERTscore.

Research question 1: How does WIPO Translate's performance compare with the proposed generic machine translation systems?

The present study aimed to evaluate the performance of WIPO Translate, a domain-specific machine translation (MT) tool, against three commercially available MT systems: Google Translate, DeepL and OpenAI's ChatGPT 3.5. The evaluation focused on translations of patents from English to Brazilian Portuguese, assessed through overall quality scores (OQS) derived from human evaluations. WIPO performed consistently lower than all other machine translation systems. Evaluators who participated in the annotation did not select it among their preferred choices. In conclusion, Google Translate, DeepL and ChatGPT 3.5 demonstrated superior performance compared to WIPO Translate across all evaluated metrics. Even though these systems are not domain-specific, they benefit from extensive training on diverse datasets and advanced algorithms which impact on higher quality translations. On the other hand, while WIPO Translate performs adequately within its specialised domain, it does not reach the same level of accuracy and fluency as the broader-purpose commercial systems. Our findings suggest that even for patent translation tasks, relying on general-purpose MT systems may yield better results than using a domain-specific tool like WIPO Translate.

Research question 2: What are WIPO Translate's limitations and advantages compared with other tools based on automatic evaluation metrics?

System limitations

Consistency: We found that one of the main limitations of WIPO Translate is its inconsistency in performance across different texts. The BLEU scores, BERTScores, and COMET scores present significant variations, indicating that WIPO Translate can produce high-quality translations, but not always.

Audience appropriateness: WIPO Translate struggles with tailoring translations to the intended audience of this study, which is directed at Brazilian Portuguese. This result was expected, as WIPO does not offer the Brazilian Portuguese specific variant among its language combinations. Yet, Google also did not at the time this experiment was conducted (March, 2024), though presenting better performance in this dimension.

Terminology management: WIPO Translate presented inconsistency in terminology, with errors in this dimension impacting the *Accuracy* as well, as seen by a similar number of errors in these categories in some evaluations. This is a sensitive point as mistranslations or inappropriate use of specialised terms may lead to financial losses and delays in patent processes.

Research question 3: How can the insights gained from the performance comparison be applied to optimise translation workflows in real-world scenarios?

The data showed that commercial systems, while not trained on patent corpora, performed better than WIPO Translate in the dimensions evaluated. In a *real-world* scenario, this could mean that using this tool may not be as useful as Google, DeepL and ChatGPT. ChatGPT, as an LLM, is especially interesting because it allows for the user to customise the translation through detailed prompts, as it was done on this research, albeit being a very simple one ('translate these patent texts into Brazilian Portuguese'). WIPO Translate still performed well, and can also be useful as a corpora tool, possibly assisting users to gauge patent style and better draft patents on their own.

Limitations of this study

This research faced a few limitations that impacted the implementation of the human analysis. While all participants were professional translators, their backgrounds were varied and not always technical. Those professionals who had technical translation experience had not previously worked with patents, which may have hindered their judgement in the sense that the style and structure of these texts was completely new and even odd to evaluators. While looking at the data, the annotated spreadsheets, it is possible to notice annotations can vary greatly from one another, which Krippendorff's Alpha confirmed. This may have been caused both by the professionals' inexperience or lack of familiarity with the style, but also a lack of understanding of

the dimensions evaluated, or poor/divergent understanding of the instructions. Even though all of them received precise instructions and a guide for further reference, they would have benefitted from a live training session demonstrating an evaluation. This study did not provide annotators with a technical glossary; this would have facilitated their task in spotting terminology errors or inconsistencies, as this dimension demands considerable research and may not have been properly annotated in view of this difficulty.

Conclusion and future work

In conclusion, this study evaluated the performance of WIPO Translate, a domain-specific machine translation (MT) tool, against three commercially available MT systems: Google Translate, DeepL and OpenAI's ChatGPT 3.5. The evaluation focused on the translation of patent excerpts from English to Brazilian Portuguese, assessing five dimensions based on the Multidimensional Quality Metrics (MQM) framework: *Accuracy, Audience appropriateness, Linguistic conventions, Style* and *Terminology*, culminating in an overall quality score derived from error annotations. The findings reveal that, while WIPO Translate performs reasonably well due to its specialised training on patent corpora, it does not outperform the commercial MT systems in terms of translation quality. Generic systems, particularly DeepL, demonstrated higher overall quality scores (OQS) on average. WIPO Translate also underperformed in the automatic evaluations in comparison to the other generic systems. The broad training datasets and advanced algorithms employed by general-purpose MT systems provide them with a competitive edge, even in specialized domains like patent translation. Future enhancements to WIPO Translate could include refining its algorithms and expanding its training data to bridge the performance gap. Despite its current limitations, WIPO Translate remains a useful tool for understanding patent style and assisting users in drafting patents. We hope the results will assist future researchers and developers of domain-specific machine translation engines in developing and/or fine-tuning translation models to better understand and generate accurate translations within that specific domain.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., Farhan, L. (2021). 'Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions'. *Journal of Big Data*, 8(1), 1–74. Available at: https://doi.org/10.1186/s40537-021-00444-8.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., Wu, Y. (2019). 'Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/1907.05019.

Artstein, R., Poesio, M. (2008). 'Inter-Coder Agreement for Computational Linguistics'. *Computational Linguistics*, 34(4), 555–596. Available at: https://doi.org/10.1162/coli.07-034-R2.

Bonyadi, A. (2020). 'Google Translate Performance in Translating English Passive Voice into Indonesian'. *PIONEER: Journal of Language and Literature*. Available at: https://unars.ac.id/ojs/index.php/pioneer/article/view/1292.

Briva-Iglesias, V., Cavalheiro Camargo, J. L., Dogru, G. (2024). 'Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain?'. *Arxiv.org.* Cornell University. Available at: https://doi.org/10.48550/arXiv.2402.07681.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C. Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). 'Language Models are Few-Shot Learners'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/2005.14165.

Callison-Burch, C., Osborne, M., Koehn, P. (2006). 'Re-evaluating the Role of BLEU in Machine Translation Research'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 249–256. Available at: https://aclanthology.org/E06-1032.pdf.

Castilho, S., Doherty, S., Gaspari, F., Moorkens, J. (2018). 'Approaches to Human and Machine Translation Quality Assessment'. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) *Translation Quality Assessment. Machine Translation: Technologies and Applications*, vol. 1, 9–38. Springer, Cham. Available at: https://doi.org/10.1007/978-3-319-91241-7_2.

Chimoto, E., Bassett, B. (2022). 'COMET-QE and Active Learning for Low-Resource Machine Translation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. Available at: https://aclanthology.org/2022.findings-emnlp.348/.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/1810.04805.

German Research Center for Artificial Intelligence [DFKI] and QTLaunchPad (2015). 'Multidimensional Quality Metrics (MQM) Definition'. *Web.archive.org.* Available at: https://web.archive.org/web/20160425060914/http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

Jana, S., Biswas, R., Pal, K., Biswas, S., Roy, K. (2024). 'The Evolution and Impact of Large Language Model Systems: A Comprehensive Analysis'. *Alochana*, 10-AJ2169. Available at: https://alochana.org/wp-content/uploads/10-AJ2169.pdf.

Jooste, W., Haque, R., Way, A. (2021). 'Philipp Koehn: Neural Machine Translation – Cambridge University Press'. *Cambridge University Press*. Available at: https://www.researchgate.net/publication/353142102_Philipp_Koehn_Neural_Machine_Translation.

Kocmi, T., Zouhar, V., Federmann, C., Post, M. (2024). 'Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies'. *Arxiv.org*. Cornell University. Available at: https://arxiv.org/abs/2401.06760.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, P. (2017). 'Neural Machine Translation'. *Statistical Machine Translation*. *Arxiv.org*. Cornell University. Available at: https://arxiv.org/abs/1709.07809.

Koehn, P., Knowles, R. (2017). 'Six Challenges for Neural Machine Translation'. In: *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. Vancouver. Association for Computational Linguistics. Available at: https://aclanthology.org/W17-3204/.

Krippendorff, K. (2013). 'Computing Krippendorff's Alpha-Reliability'. University of Pennsylvania Annenberg School for Communication. Available at: https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

Lommel, A., Uszkoreit, H., Burchardt, A. (2014). 'Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics'. *Revista Tradumàtica: Tecnologies de la Traducció*, (12), 455–463. Available at: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n 12p455.pdf.

Maučec, M., Donaj, G. (2019). 'Machine Translation and the Evaluation of its Quality'. *Recent Trends in Computational Intelligence. IntechOpen*. Available at: https://doi.org/10.5772/intechopen.89063.

Morita, T. (2024). 'Positional Encoding Helps Recurrent Neural Networks Handle a Large Vocabulary'. *Arxiv.org*. Cornell University Available at: https://arxiv.org/abs/2402.00236.

Moslem, Y., Way, A., Haque, R., Kelleher, J. D. (2022). 'Domain-Specific Text Generation for Machine Translation'. In: *Proceedings from the 15th Conference of the Association for Machine Translation in the Americas*. Available at: https://aclanthology.org/2022.amta-research.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). 'Bleu: A Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania,

USA. Association for Computational Linguistics. Available at: https://aclanthology.org/P02-1040.pdf.

Pouliquen, B., Mazenc, C., Iorio, A. (2011). 'Tapta: A User-Driven Translation System for Patent Documents Based on Domain-Aware Statistical Machine Translation'. In: *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation. Available at: https://aclanthology.org/2011.eamt-1.2/.

Rei, R., Stewart, C., Farinha, A., Lavie, A. (2020). 'COMET: A Neural Framework for MT Evaluation'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Association for Computational Linguistics. Available at: https://aclanthology.org/2020.emnlp-main.213/.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3), 393-401. Available at: https://doi.org/10.1162/coli_a_00322.

Rivera-Trigueros, I. (2022). 'Machine Translation Systems and Quality Assessment: A Systematic Review'. *Computers and the Humanities*, 56, 593–619. Available at: https://doi.org/10.1007/s10579-021-09537-5.

Rossi, L., Wiggins, D. (2013). 'Applicability and Application of Machine Translation Quality Metrics in the Patent Field'. *World Patent Information*, 35(1), 20–26. Available at: http://dx.doi.org/10.1016/j.wpi.2012.12.001.

Savoy, J., Dolamic, L. (2009). 'How Effective is Google's Translation Service in Search?' *Communications of the ACM*, 52(10), 139–143. Available at: https://doi.org/10.1145/1562764.1562799.

Sutrisno, A. (2020). 'The Accuracy and Shortcomings of Google Translate. Translating English Sentences to Indonesian'. *Education Quarterly Reviews*, Vol. 3(4). Available at: https://ssrn.com/abstract=3747888.

Tinsley, J. (2017). Machine translation and the challenge of patents. In: *Multilingual Information Management*. Springer, 295–316. Available at: https://doi.org/10.1007/978-3-662-53817-3_16.

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, Ł., 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., Macduff, H., Dean, J. (2016). 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'. *Arxiv.org.* Cornell University. Available at: https://arxiv.org/abs/1609.08144.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2020). 'BERTScore: Evaluating Text Generation with BERT'. In: *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*. Available at: https://openreview.net/pdf?id=SkeHuCVFDr.

Zhao, Y., Zhang, J., Zong, C. (2023). 'Transformer: A General Framework from Machine Translation to Others'. *International Journal of Automation and Computing*, 20(4), 514–538. Available at: https://doi.org/10.1007/s11633-022-1393-5.

https://doi.org/10.33919/JCAL.24.2.4

CAN THE USE OF TEXT-TO-SPEECH TECHNOLOGY SUPPORT POST-EDITORS WORKING INTO THEIR L2?

Foteini Kotsi[1], Todor Lazarov[2], Joke Daems[3]

[1]Ghent University, [2]New Bulgarian University, [3]Ghent University

Abstract

Speech technologies have benefited greatly from the rise of artificial intelligence (AI) and they have developed over the past few years. Many industries have harnessed their power with great gains. In our research, we investigated to what extent professionals in the localisation industry accept machine translation post-editing (MTPE) with the help of speech technologies, and more particularly before and after a relevant experiment. We also examined their attitudes towards this approach in MTPE depending on their academic and professional background. Participants who achieved significant gains in terms of error corrections and relative gain in terms of time spent post-editing, no matter how negative or positive they were before the experiment, became open to TTS-enabled MTPE. The academic background of the participants did not seem to influence their attitude heavily. However, their professional background created some patterns not only in their attitude before the experiment, but also after its completion.

Keywords: *machine translation, post-editing, text-to-speech technology*

1. Introduction

In recent years, MTPE has attracted the interest of researchers. With the advancements of different technologies, the face of machine translation post-editing (MTPE) has started to change and so have the research objectives. Despite the fact that speech technologies, namely automatic speech recognition (ASR) and text-to-speech (TTS) technology, have seen immense growth in recent years, research on speech technologies and MTPE has been limited, especially for TTS technology, leaving the vast majority of post-editors unaware of its potential.

There are different scenarios in which the target segment is read aloud in the language service industry (LSI) for the convenience of the linguists themselves. For instance, in the European Space Agency, translators, revisers and reviewers have altered the traditional translation – editing – proofreading (TEP) model and instead they all use track changes and pass documents among themselves for a translation + face-to-face review model (Ciobanu, 2014). In this process, a colleague reads back to the translator their work and in this way the colleague is doing a monolingual target language review at the same time as the translator is doing a bilingual revision (*ibid.*). There are also cases where translators use TTS technology during the revision step of their own translations and this 'could help linguists spot both fluency and accuracy errors more easily' (Ciobanu and Secară, 2019). Similarly, there are translators that have introduced both ASR and TTS into their workflow. More concretely, in this scenario the translator is using ASR to pronounce their translation into the target text and then have TTS technology read the target text to them. This is believed to be an effective method of catching *speakos* (errors that might have occurred as the linguist spoke their translation through ASR) (Ciobanu, 2014).

Given all that, we would like to see how a group of linguists of various professional and academic backgrounds, unaware of this potential, feel towards the integration of TTS technology to the MTPE workflow before and after a relevant experiment and whether their backgrounds affect their attitude in any significant way.

2. Related research

By reviewing the related literature, it becomes evident that ASR has received more attention when it comes to the implementation of speech technologies in LSI workflows (Dragsted *et al.*, 2011; Ciobanu, 2014; Ciobanu, 2016; Ciobanu and Secară, 2019). However, in more recent years, researchers have also explored the potential of TTS technology in the LSI.

More specifically, in the later years, two main studies have been completed regarding localisation services and TTS. Firstly, Ciobanu *et al.* (2019) focused on translation revision, but the TTS technology was only enabled in the source text. It should be mentioned that the participants were a mix of professional translators and trainees and they worked in memoQ.[1] The results mostly focused on quality and they were promising for the use of TTS. Secondly, Brockmann *et al.* (2022) focused on error annotation with the TTS technology enabled for both the source and target segments. The participants were students and the experiments were carried out in Microsoft Word 365[2]. Given the nature of the task, the results were given in error over- and under-annotation terms, providing good indications for quality while working with TTS.

Even though the aforementioned studies indicated a positive attitude of their participants towards TTS-enabled revision and error-annotation, and evidence points to quality and accuracy gains, they did not report on the different backgrounds of the participants and how this can affect the attitude of professionals towards post-editing with TTS.

2.1 Research questions

In order to investigate these research grounds, which, to the best of our knowledge, remain unexplored, we formulated the following research questions (RQs):

1. How does the academic background of post-editors affect their attitude towards working with TTS?

As mentioned above, there are studies that have so far touched upon the attitude of professionals and students towards LSI workflows with TTS integration. However, none of them has taken their academic background into consideration. With our experiment, we investigated whether there is any correlation between an academic background in translation technology and the attitude towards working with TTS in MTPE.

To answer this question, we reported on the attitudes of linguists in the LSI towards the use of TTS in localisation workflows depending on their academic background. By gathering the participants' responses to three questionnaires, we determined whether their academic background affected their attitude towards the use of TTS technology in the machine translation post-editing process.

2. How does the professional background of post-editors affect their attitude towards working with TTS?

[1] Available at: https://www.memoq.com/.

[2] Available at: https://www.microsoft.com/en-us/microsoft- 365/word.

Similar to the academic background, the specific professional background in correlation to the attitude towards TTS-enabled localisation tasks has not been explicitly explored. In our experiment we investigated the attitudes of professionals experienced in multimodal translation and traditional translation and examined whether their professional background affected their overall attitude towards the proposed MTPE method.

To answer this question, we reported on the attitudes of linguists in the LSI towards the use of TTS in MTPE depending on their professional background. By gathering the responses of the participants to three questionnaires, we determined whether their professional background affects their attitude towards the use of TTS technology in the machine translation post-editing process.

3. Methodology

Our experiment consisted of five main steps: pre-experiment questionnaire, pilot study, post-practice questionnaire, main experiment and post-experiment questionnaire. The pre-experiment questionnaire included questions about the academic and professional background of the participants, as well as their initial attitude towards post-editing with TTS technology. During the pilot study, the participants completed two MTPE tasks with the help of TTS technology in order to get as comfortable as possible with this new way of working and the interface. The post-practice questionnaire was intended to collect feedback on the interface as well as report on the attitude of the participants after they had a first impression on TTS-enabled MTPE. The main experiment consisted of two tasks, one carried out with the help of TTS technology and one without it. During the main experiment, the quality and the time spent (temporal effort) were recorded to examine whether the performance at that point affected the attitudes of the participants. In the post-experiment questionnaire, the participants had the opportunity to express their opinion on TTS-enabled MTPE and give their remarks.

During the selection process for participants, there were specific criteria to take into account. First of all, they had to be native Greek speakers and their L2 had to be English. Moreover, they had to hold a bachelor's degree in Translation and have between two and ten years of experience in the LSI. Since speech technology applications are more widespread in captioning and subtitling, we were aiming to include professionals with experience in these areas and explore whether they tend to have a more positive attitude towards TTS technology or not, due to being more accustomed to multimodality in translation. Another factor that was taken into consideration was the familiarity of technology in translation through additional studies other than their

bachelor's degrees. This would also allow us to examine any positive pre-dispositions towards applying new technologies to a localisation service. The total number of participants was eight. The data about the exact synthesis of the group was gathered by the responses to a corresponding questionnaire. Full MTPE was carried out since most of the participants did not have extensive experience in MTPE and could struggle with the notion of focusing only on critical errors and making minimal changes.

To ensure that the level of difficulty of the texts used in the two tasks would not affect the participants' attitude, both texts used during the main experiment referred to pieces of news with the same overarching topic, the Common Agricultural Policy of the European Union, coming from the same news website. As a first step, the articles were edited as to have almost identical length: 298 and 297 words, respectively. In order to ensure that the difficulty levels of the texts were comparable, both source texts were run through the Text Readability software for Greek developed by the Centre for Greek Language[3]. The scores obtained are presented in **Table 1**:

Text	B2	C1	C2
Text 1	31.53%	51.68%	15.65%
Text 2	22.17%	58.44%	19.04%

Table 1: CEFR scores for Text 1 and Text 2.

To determine which MT engine would be used, we compared the raw output of ModernMT[4], DeepL[5] and eTranslation[6] for both of the texts. The outputs were ranked by the researcher segment by segment. More specifically, for each segment the best performing engine was ranked as first, the second-best as second, and the worst as third. The best and worst performing engines, when taking each engine's overall performance in the entirety of the texts, were excluded in order to avoid conditions where too much or too little effort would be required by the participants. The second-best performing engine according to the ranking of all the segments for both texts was eTranslation, with DeepL performing the best and ModernMT performing the worst, by relatively small margins. As a final step, the eTranslation output for both texts was annotated for errors, using the TAUS DQF-MQM error typology[7]. No critical errors were detected and the major and minor errors

[3] Available at: https://www.greek-language.gr/certification/readability/index.html.

[4] Available at: https://www.modernmt.com/.

[5] Available at: https://www.deepl.com/en/translator.

[6] Available at: https://commission.europa.eu/resources-partners/etranslation_en.

[7] Available at: https://www.taus.net/.

were comparable in numbers. The texts were also alternated in order and use to ensure the reliability of the results.

MateCat[8] was selected for this experiment. It is a cloud-based free CAT tool with a straight-forward, clean interface, very popular among professionals. Compared to other similar CAT tools, MateCat has an added advantage. It offers a MTPE effort measuring feature. More specifically, MateCat counts the time spent post-editing in each segment and in the overall document, as well as the percentage of text changed both at segment and document level.

The Google Chrome Read Aloud[9] browser extension was selected among other TTS tools. Read Aloud was installed as an extension to each participant's Chrome browser. It provides a wide range of voices, especially for English. It allows the user to modify its shortcuts, the speed of the speaker, the pitch, and the volume. What set Read Aloud apart from other TTS tools for this particular experiment was its variety of display modes. This allowed the participants to only operate the TTS function by use of the shortcuts without having an extra element in their view of their workbench.

Lastly, the keystroke logging tool Inputlog[10] was employed, in order to collect data regarding the MTPE temporal effort.

4. Results

In this section, the results from all three questionnaires throughout the experiment are analysed and correlated with the actual results regarding temporal effort and quality where appropriate.

4.1 Pre-experiment questionnaire

As mentioned above, the pre-experiment questionnaire was a means to record the academic background, work experience, and the attitude of the participants towards speech technologies in the LSI and TTS in MTPE in particular. In this segment, the collected data is briefly discussed.

75% of the participants had between two and six years of experience. Only one of the participants had between zero and two years of experience, while another one had between eight and ten years of experience, balancing out the average in the years. All participants obtained their bachelor's degrees in translation from the Ionian University between 2014 and 2021. During the experiment, half of the participants had completed or were attending master's degree programmes directly related to translation and technology.

[8] Available at: https://www.matecat.com/.

[9] Available at: https://chromewebstore.google.com/detail/read-aloud-a-text-to-spee/hdhinadidafjejdhmfkjgnolgimiaplp?pli=1.

[10] Available at:https://www.inputlog.net/.

More specifically, three out of these four participants have either completed or are currently attending the master's programme in Translation Technology of the Dublin City University, and the fourth participant obtained their Master's in Science of Translation from the Ionian University, which has as one of its main pillars a group of courses in new technologies in translation. The remaining participants either had received master's degrees in a domain not directly related to translation technology or had responded that they did not possess a master's degree. This dichotomy in the participants allows us to examine if the academic background of post-editors can affect their views and attitudes toward TTS-enabled PE.

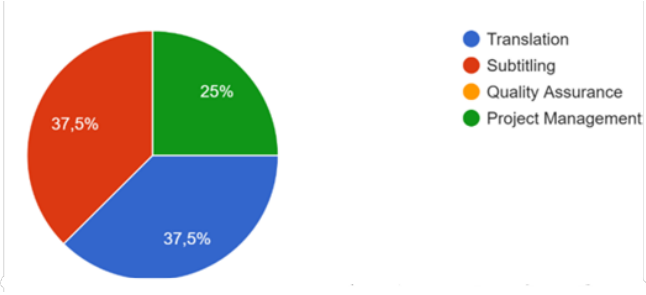


Figure 1: Participant distribution according to professional background.

Similarly, the domain of the participants' work experience divided them in three main groups, as demonstrated in **Figure 1** above. Firstly, 37.5% of the participants responded that the main focus of their work experience had been translation. Secondly, another 37.5% answered that they have mostly worked with subtitling. Lastly, 25% marked project management as their main-experience position in the LSI. In the professional background options, the questionnaire included quality assurance as an alternative, which did not reflect any of the participants' experience. This distribution allows us to compare different views on TTS-enabled PE, depending on whether the post-editor has experience in traditional translation, multimodal translation, or quality assurance of the final product.

Regarding the CAT tools the participants were familiar with, all of them had at some point worked with SDL Trados, and just one of them had never used MateCat. 37.5% of the participants answered that they used SDL Trados the most, while working with MateCat was the second most popular option with 25%. Phrase and memoQ gathered lower percentages.

Regarding the extent to which the participants were of the possibility to use TTS technology in their preferred CAT tools, they seemed to be split into two groups, with 50% of them having verified whether they could make

use of it in the specific tool, while the other half not having investigated the possibility.

On the other hand, the participants familiar with multimodal translation were more aware of voice services in subtitling tools. Only 37.5% had not looked into the possibility of the tool they were using for subtitling offering automatic transcription.

A 50-50 split could be observed in the preferred MT engine used by the participants, between Google Translate and DeepL. This condition created an even playing field for all participants since none of them was more familiar than the rest with the chosen MT engine.

Regarding the overall attitude of the participants towards the use of TTS within their preferred tool, on a scale of one to ten on average, the participants gave the possible use a chance of 7.6/10, as shown in **Figure 2** below.

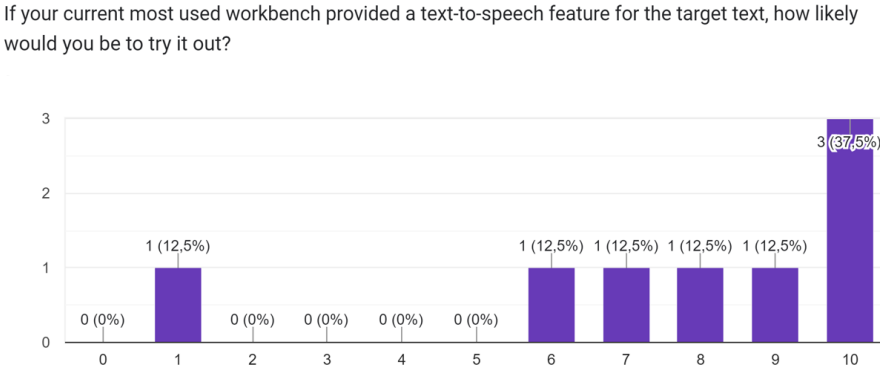


Figure 2: Participant pre-experiment attitude towards working with TTS.

More specifically, the group with academic background in translation technology, regardless of their work experience, gave this alternative a 6.75/10 on average, whereas the other half of the participants gave it an 8.5/10. It should be noted that the first group's average was affected by the score of one given by one of the participants.

The group of three participants experienced in subtitling gave the possibility of trying TTS during post-editing a 10/10. The group of three professionals mostly experienced in translation gave the possibility of experimenting with TTS in their translation workbench a 5.3/10 and the two participants with experience in project management awarded an average score of 7.5.

If we examine the participants' attitudes combined with their years of experience, we can notice a trend of more experienced professionals being more positive towards using TTS while post-editing. More specifically, the

participants with experience of up to four years gave the possibility of using TTS while post-editing an average of 6.8/10. It should be noted that the participant who had marked this likelihood with one was included in this group, and they had the least years of experience in the industry, being the only one having proclaimed up to two years of experience. Within this group, the participants with two to four years of experience had an average of 8.25/10. The participants with over four years of experience, with all three of them representing a different professional background, responded with an average of 9/10 to the same question.

4.2 Post-practice questionnaire and feedback

The Post-practice questionnaire consisted of one open-ended question regarding the experience the participants had during the first practice text, their feedback, concerns or possible remarks. In this section, we discuss briefly all the data collected after the practice text, including the responses from the questionnaire, the relevant Inputlog files from each participant, and the post-edited text delivered on MateCat along with the relevant findings on the CAT tool regarding the TER and the time spent post-editing.

Regarding the findings from the open-ended questionnaire, half of the participants mentioned that it was not clear to them what the advantages of using TTS in the machine translation post-editing process were. One of the participants mentioned the speed of the Read Aloud function as a disadvantage. Three out of the eight participants mentioned a technical issue with the Read Aloud function reading the source text when they forgot to mark the correct segment for the extension to read aloud before pressing the shortcut launching it.

By taking into consideration the comments in the questionnaire and after examining the shortcuts registered in the Inputlog files, we had a clear image of the weak points in the process. More specifically, when most participants wanted to listen to a new segment, they would press the shortcut that launched the Read Aloud extension. The disadvantage of this method is that the window of the app pops up every time and it starts from the beginning of the text. Not only does this create issues with the flow of the TTS, but it also minimises the post-editor's view of their workbench. Having this in mind, we fine-tuned the instructions to avoid these two issues for a better user experience during the main experiment. Similarly, we added the instruction to experiment with the speed of the Read Aloud during the second practice text and before getting started on the main experiment.

Lastly, by checking the TER measured for each task on MateCat, it was made clear that many participants were prone to preferential changes. Given

that, we also stressed in the main experiment instructions that preferential changes should be avoided. MateCat data also revealed some discrepancies in the time recorded on the CAT tool and Inputlog. To avoid similar issues during the main experiment, a special note was included in the instructions that each task had to be completed in one go and, if any of the participants had to step away from their computer, the MateCat tab should be closed.

In order to make the purpose of the study as clear to the participants as the conditions allowed without biasing them, we communicated to them that TTS would be the one difference between the two main tasks of the experiment, meaning that one task would be TTS-enabled and the other one not. They were instructed to view TTS as an extra tool available to them for their convenience during the MTPE process, and that at the end of the main experiment they would be asked to fill in a questionnaire regarding the two different experiences they had post-editing with and without the aid of TTS.

Upon receipt of the updated instructions and targeted feedback, the participants were also assigned the second practice text. They were instructed to post-edit the new practice text with TTS by applying all the tips, feedback and instructions received before moving forward with the two main texts. In this manner, the participants had a second opportunity to get more familiar with the TTS-enabled machine translation post-editing process by also applying best practices that would ameliorate their experience.

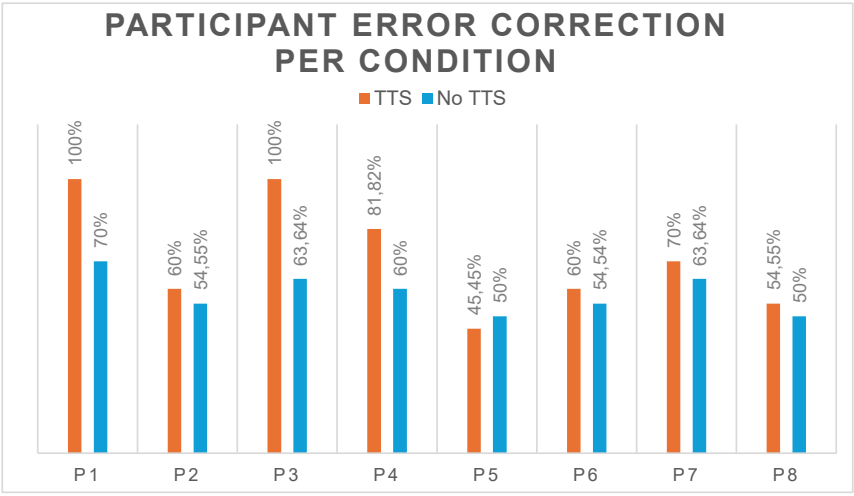


Figure 3: Participant error correction percentage per condition.

4.3 Main experiment results

To measure whether the post-edited texts were of acceptable quality, we used the number of annotated errors corrected to the total number of anno-

tated errors and expressed it as a percentage, as seen in **Figure 3**. As you can see, P1, P3 and P4 were the ones that benefitted the most from the use of TTS in terms of quality. P2, P6, P7 and P8 had lower gains but still performed better than when they were working without TTS. P5 is the only participant that scored lower in terms of quality when working with TTS and it should be mentioned that they were by a great margin the most sceptical towards TTS-enabled MTPE by giving the possibility of working in this way a 1/10.

Enabling TTS was not that beneficial for as many participants in terms of temporal effort. This was to be expected to an extent, since the added actions of pressing the shortcuts and actually listening to the target text require more time than traditional MTPE. However, P2, P3 and P4 worked faster while post-editing with the use of TTS technology, as demonstrated in **Figure 4**. The remaining participants were faster when working without it.

To conclude, the benefits in terms of quality when it comes to the use of TTS in MTPE seem to be more prominent. In terms of time, the results seem to differ depending on the participant.

4.4 Post-experiment attitude

All of the above results are to some extent indicative of the participants' disposition towards TTS-enabled machine translation post-editing. To determine their actual attitude, we asked our group to re-evaluate the possibility of working with TTS while post-editing on a scale of one to 10.

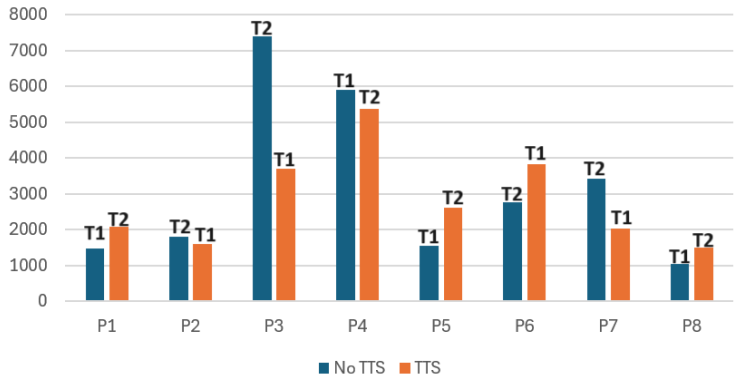


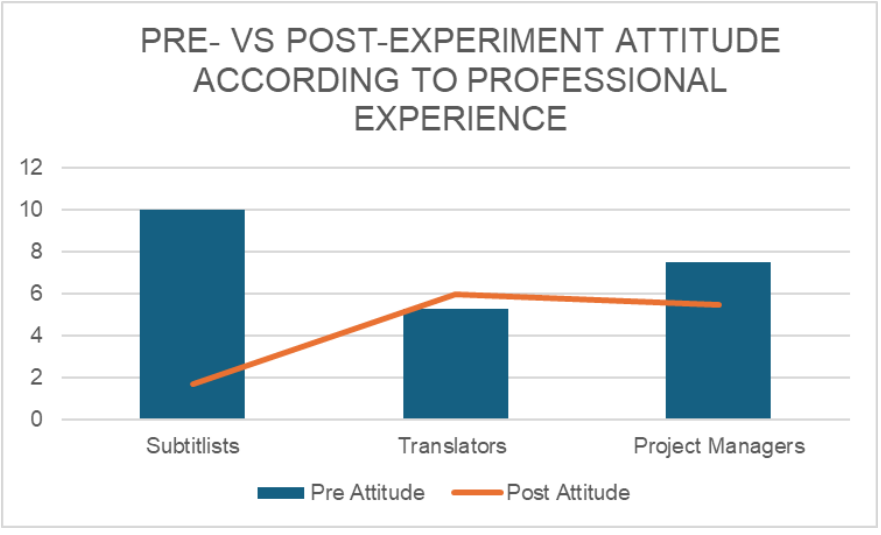
Figure 4: The time each participant spent post-editing under each condition in seconds.

For reference, the initial score during the pre-experiment questionnaire was on average across all participants a 7.6/10. The average after the completion of the experiment had dropped significantly to 4.25/10. Let us have a closer look at the data in order to interpret this significant decrease.

First of all, P1, P3 and P4 were positive throughout the questionnaire and were also the ones with the greatest gains in quality and time savings; they either remained steady at their evaluation or increased it. More specifically, P1 and P3 remained at 6 and 7, respectively, and P4 went from an 8 to a 9/10. P5, after having expressed no satisfaction with TTS performance, about which they were hesitant from the beginning, went from a 1 to a 2. However, it should be mentioned that the negative attitude was also in line with the decrease they had in error correction on the TTS-enabled task that also took the participants significantly longer to complete. Lastly, the participants that had low error correction and time gains were the ones that had the biggest shift in their attitude for the worst. More specifically, P6 and P7 went from 10 to 1 and P7 went from 10 to 3. P2, who started off from 9, dropped to 5.

Focusing on the academic background of the participants, the ones with an academic background related to translation technology gave an average of 6.75 in the same question during the pre-experiment questionnaire and landed on a mere 4/10, with the remaining 4 participants without master's degrees in a related field averaging at 5.5. There were no significant changes in the comparison here, since from the very beginning the first group was more negative by 1.75 compared to the latter.

Figure 5: Comparison of the attitude of each professional group's pre- and post-experiment attitude.



The most interesting data arise when we focus on the work experience of the participants, as featured on **Figure 5**. P6, P7 and P8 were our subtitlist group, which, as mentioned in the Pre-experiment questionnaire, gave the

possibility of using TTS while post-editing a perfect 10. When asked in the post-experiment questionnaire how likely they were to reuse TTS technology while post editing, P6 and P8 responded with a 1 and P7 with a 3. There was a relative drop on the average of our project manager group as well. P1 remained at a 6/10, but P2 went from 9 to 5. Lastly, the only group with a raising average was the one of the translators with P3 remaining steady at 7, P4 went from 8 to 9 and P5 from 1 to 2.

4.5 Participant feedback

At the end of our Post-experiment questionnaire, we included an open-end question, encouraging the participants to voice any comments or concerns regarding the experiment. This was the second time they were asked to do this during the experiment and it was an opportunity for them to express openly any issue or remark. The first was our post-practice questionnaire.

The main positive aspect of all the answers was that, contrary to the Post-practice questionnaire, no technical issues were reported about the main experiment or no concerns regarding the procedure since any issues were resolved between the practice sessions and the main experiment. For example, there were no remarks regarding the speed of the voicing since that was included in our updated instructions.

However, many interesting remarks from the participants helped us interpret their attitudes more efficiently. Both P2 and P6 explained that they personally did not prefer working with TTS since they had been used to reading the source and/or target segments aloud on their own, listening to their own voice while working on a text and, naturally, the use of TTS forced them to change the way they had been used to.

On the other hand, P4 said they perceived the authentic English accent as an added benefit during the machine translation post-editing process to improve quality and mentioned that TTS could be a useful addition to other localisation services as well.

An interesting remark by P5 was that they understood the use of TTS for accessibility purposes to blind and visually impaired professionals, but they did not perceive any benefits while working with it.

All in all, the feedback given at this stage was particularly important since it came by the people who experienced the whole procedure under normal working conditions on CAT tools. By the end of the experiment, they had completed three separate TTS-enabled tasks, they had received and applied targeted feedback on how to use it and they were able to come to their own conclusions regarding its benefits or lack thereof. More particularly, P8, who had mentioned in their Post-practice feedback they were 'uncertain of its

benefits' now characterised it as unhelpful. On the other hand, P3 said it could be time-consuming, but they considered it helpful. There were also participants that highlighted the objectivity element of each individual's way of working and they clarified that even though it had not been helpful to them, it could be beneficial to a colleague.

5. Discussion

We designed and carried out this experiment in order to explore any potential links between the academic and/or professional background of professionals and their attitude towards working with TTS while post-editing. To answer these questions, we employed both quantitative and qualitative methods in order to obtain a clearer picture.

1. How does the academic background of post-editors affect their attitude towards working with TTS?

Starting from the academic background, surprisingly from the very beginning, the participants with master's degrees in a field related to Translation Technology seemed to be less open to the idea of machine translation post-editing with TTS. Their views remained similar throughout the experiment and by its completion they had lowered their probability of machine translation post-editing with TTS again, being still lower than the average of the rest of the participants. Probably, their attitude stems from the familiarity that these individuals have with accessibility services. It is true that with the EU accessibility directives, the face of localisation has changed completely, with new services being added to the industry. However, this rise, as supported by participants' commentary, has probably led to professionals in the LSI believing that the potential of TTS and ASR is mostly being used for accessibility services and the localisation industry does not necessarily need to harness the potential of these new technologies to its advantage.

2. How does the professional background of post-editors affect their attitude towards working with TTS?

On the other hand, professional experience seemed to be a decisive factor in the attitude of the participants since it determined the tendencies of our group. More specifically, the most striking pattern was that of the subtitlists. All three of them started off the experiment by claiming that, if given the opportunity to post-edit with TTS, they were 10/10 likely to do it. All three of them decreased their evaluation almost identically, to 1/10 and 3/10.

Even though the reasoning given by all three of them regarding this tremendous switch in attitude was not exactly the same, we can analyse it by looking at the effect speech technologies have had on subtitling. More spe-

cifically, as we have mentioned already, according to Ciobanu and Secară (2019), audiovisual translation has reaped the most benefits from ASR and TTS. To give a more indicative example, subtitling by ear is the service for which a subtitlist is given no additional resources other than the original video, without script, or subtitle template. The use of ASR, very often coupled with automatic spotting, allows a subtitlist to decrease their workload by a significant amount and, with the accuracy rates of ASR today, to almost skip one or two entire steps in a very challenging service. Viewing TTS-enabled machine translation post-editing as an advancement similar to this could explain the enthusiasm going into the experiment, as well as the disappointment when realising that the benefits of this new proposed way of working were not as evident or indisputable as the group expected.

Our project manager group did not seem to move in unison as the subtitlists did. There were no specific patterns to discuss with this group neither in terms of their starting attitude nor their post-experiment attitude. More specifically, they did not start off particularly positively or negatively and ended the experiment just as neutral towards the idea of post-editing with TTS tools.

The other group of main interest, which was the one of our translators, did present some specific patterns in their behaviour. Two out of three started moderately optimistic, with the third one being as negative as permitted by the scale towards the idea. However, one of them remained at the same position in their evaluation after the experiment and the other two went higher by one. It is particularly important that this group had the more positive trajectory in the experiment since they would be the main target group for this new working method. This could also be due to the familiarity they already had with the service of machine translation post-editing and the overall familiarity they had gained by working with MT output.

It was also very encouraging to see how the participants with significant improvements in their overall performance had either stayed as positive as they were at the beginning of the experiment or became even more positive in their evaluation of TTS-enabled machine translation post-editing, indicating that they may have been aware of this amelioration in their performance.

In conclusion, participants' professional backgrounds influenced their attitudes towards TTS more than their academic background, with subtitlists starting off enthusiastic, which was not the attitude they demonstrated by the end of the experiment. Translators, on the other hand, had a more positive evolution in their attitude. Overall, participants who experienced significant improvements tended to perceive TTS-enabled machine translation post-editing more favourably, suggesting a potential awareness of their en-

hanced performance. These insights could form a basis for further research on TTS-enabled MTPE.

References

Brockmann, J., Wiesinger, C., Ciobanu, D. (2022). 'Error Annotation in Post-Editing Machine Translation: Investigating the Impact of Text-to-Speech Technology'. In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 251–259. Ghent, Belgium: European Association for Machine Translation. Available at: https://aclanthology.org/2022.eamt-1.28.

Ciobanu, D. (2014) 'Of Dragons and Speech Recognition Wizards and Apprentices'. *Revista Tradumàtica : Traducció i Tecnologies de La Informació i La Comunicació* 12, 524–538. Available at: https://doi.org/10.5565/rev/tradumatica.71.

Ciobanu, D. (2016). 'Automatic Speech Recognition in the Professional Translation Process'. *Translation Spaces,* 5(1), 124–144. Available at: https://doi.org/10.1075/ts.5.1.07cio.

Ciobanu, D., Ragni V., Secară A. (2019). 'Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking'. *Informatics,* 6(4), 51. Available at: https://doi.org/10.3390/informatics6040051.

Ciobanu, D., Secară A. (2019). 'Speech Recognition and Synthesis Technologies in the Translation Workflow'. In: *The Routledge Handbook of Translation and Technology*. Routledge.

Da Silva, I., Alves F., Schmaltz M., Pagano A., Wong D., Chao L., Leal A., Quaresma P., Silva, G. (2017). 'Translation, Post-Editing and Directionality: A Study of Effort in the Chinese-Portuguese Language Pair'. In: *Translation in Transition: Between cognition, computing and technology*. Benjamins Translation Library.

Dragsted, B., Hansen, I. (2009). 'Exploring Translation and Interpreting Hybrids. The Case of Sight Translation'. *Meta: Journal Des Traducteurs / Meta: Translators' Journal*, 54(3), 588–604. Available at: https://doi.org/10.7202/038317ar.

Dragsted, B., Mees, I., Hansen; I. (2011). 'Speaking Your Translation: Students' First Encounter with Speech Recognition Technology'. *The International Journal for Translation and Interpreting,* 3(1), 10–43. Available at: https://www.trans-int.org/index.php/transint/article/view/115/87.

Koponen, M. (2012). 'Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations'. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 181–190. Montréal, Canada: Association for Computational Linguistics. https://aclanthology.org/W12-3123.

Koponen, M. (2016). 'Is Machine Translation Post-Editing Worth the Effort? A Survey of Research into Post-Editing and Effort'. *The Journal of Specialised Translation*, 131–148. Available at: https://www.researchgate.net/publication/299345698_

Is_Machine_Translation_Post-editing_Worth_the_Effort_A_Survey_of_Research_into_Post-editing_and_Effort.

O'Brien, S. (2006). 'Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output'. *Across Languages and Cultures,* 7(1), 1–21. Available at: https://doi.org/10.1556/Acr.7.2006.1.1.

O'Brien, S. (2007). 'An Empirical Investigation of Temporal and Technical Post-Editing Effort'. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association,* 2(1), 83–136. Available at: https://www.jbe-platform.com/content/journals/10.1075/tis.2.1.03ob.

O'Brien, S. (2011). 'Towards Predicting Post-Editing Productivity'. *Machine Translation,* 25(3), 197–215. Available at: https://doi.org/10.1007/s10590-011-9096-7.

Stasimioti, M., Sosoni, V., Chatzitheodorou, K. (2021). 'Investigating Post-Editing Effort: Does Directionality Play a Role?' *Cognitive Linguistic Studies,* 8(2), 378–403. Available at: https://doi.org/10.1075/cogls.00083.sta.

https://doi.org/10.33919/JCAL.24.2.5

CAN CAT TOOLS AND LARGE LANGUAGE MODELS TRANSLATE FROM UNRECOGNISED LANGUAGES?

Emanuela Tchitchova

New Bulgarian University,
Computational and Applied Linguistics Research Centre
etchitchova@nbu.bg

Abstract

First designed as a case study for high-school students competing in the field of linguistics, this experiment is based on the hypothesis that CAT tools and large language models have arrived at a stage where they can translate texts from languages that have not been explicitly listed as source languages. We took an excerpt in old French *ancien français* from Chrétien de Troyes's novel *Le Chevalier de la Charrette*, versified into octosyllables. We machine translated it into Bulgarian – first from Latin and French with Phrase, and then without specifying the source language with ChatGPT. Phrase yielded better results with Latin than with modern French. Still, ChatGPT's translation was much better and could be considered as a gisting translation. We had a reference human translation – into modern French, non versified, and into Bulgarian, versified. When comparing the results, we found that Phrase

couldn't provide a gisting translation, while ChatGPT rendered the general events and characters present in the text, even though on a contextual level there were still a lot of discrepancies; detail was missing; the text was rendered in prose. The experiment was conclusive that LLMs manage to render the general meaning of a text from an unknown dead language. It was illustrative for the students who were introduced to the specificities of medieval French and were given some linguistics problems to solve, based on the particularities of this language and its versification. The experiment shows the extent to which cultural heritage such as medieval knights' novels can be rendered accessible to modern-day public, without having to resort to professional translation.

Keywords: *gisting translation, translating ancient languages, computer-aided translation of Latin, machine translation*

Context and methodology

The experiment described in this study was designed for two purposes: first, present some linguistic particularities of the evolution of the French language through the ages to students in Bulgarian high schools, competing in national and international Linguistics Olympiads[1], and second, illustrate some of the modern language technologies and their way of working.

Having in mind that large language models rely on neural networks and machine learning in order to re-create the meaning of a text, this hypothesis had to be further complicated by a new variable, so that the testing can give interesting results. We chose Old French, a language derived from Gallo-Romance (itself derived from Vulgar Latin) – sufficiently far away both from Latin, from which it originated, and from Modern French, so that there would be less interference on a lexical and grammatical level with linguistic systems of both languages.

Historically, the French language has passed through 5 stages, ranking from Latin to modern French. Old French (*ancien français*) corresponds to the middle stage, the history starting with Vulgar Latin (1), Gallo-Romance (2), Old French (3: dating from the end of the 8th century to the 14th century); Middle French (14th to 17th century); Modern French (from 17th century onwards) (Cerquiglini, 1993) (Lodge, 1993).

[1] https://ioling.org/index.html.

Old French had the following particularities: it kept its Latin lexicon, even though a multitude of phonetic changes occurred; the Latin case system greatly eroded (only the nominative and accusative cases remained); it kept its pro-drop feature from Latin, but major phonetic changes had already taken place and consequently many words were widely different from their Latin origin. The incursion of the Frankish language[2], as well as the Viking invasions created another superstrate, and further phonetic changes occurred, such as diphthongisation, palatalisation and accent stress, making Old French incomprehensible for Latin language users of the time (9th-14th century).

Thus, the experiment stayed more on the side of comparative linguistics than on the side of language technologies. In the present paper, we will discuss the technological part of it.

Our research question was whether a CAT-tool such as Phrase[3] can translate from unrecognised languages, e.g. Old French, into modern languages, and we aimed to compare its translation and performance to the one proposed by ChatGPT, a large language model using neural networks. We decided to take Old French for several reasons. First, for the purposes of translation, we had to specify a source language, and Phrase supports both Latin and Modern French. Even though it is a language *per se* and cannot be understood by speakers of both Latin and French, we supposed that it could at least bear some similarities with them. Secondly, we wanted to have a target language as far away as possible from French, and this was Bulgarian (for our high-school public, this was the common language). We chose a text, *Le Chevalier de la Charrette* from Chrétien de Troyes, which has a human official translation both into modern French and into Bulgarian. This text is deemed to have been created between 1177 and 1181, by order of Marie de Champagne (thus using the *champenois* language from the *langues d'oïl*), a great supporter herself of courtly love and knights' novels. It is the first text mentioning Lancelot as a prominent knight of King Arthur's and this feature makes it an addition to the *matière bretonne*, the Arthurian legend (Uitti,

[2] It influenced all of the languages found in the regions above the Loire river, known since as the *langues d'oïl*. (Pellegrini, 1980) (Roegiest, 2006).

[3] Phrase is a CAT tool which currently uses a multitude of engines. As stated, under the Phrase Language AI engine, it lists 8 other engines used in the same profile and the two jobs, from Latin and from modern French (Amazon Translate, DeepL, Google Translate, Microsoft Translator, Phrase Next GenMT (GPT-4o), Phrase NextMT, Rozetta Translate, Tencent). This tool has been used to work with university students preparing for a State exam in translation and thus was suitable for the experiment conducted with secondary students, showing the possibilities of CAT tools in an academic environment.

1995) (Lacy, 1991). Chrétien's writings prepared Lancelot's prominent place in subsequent English literature.

Setup of the experiment

Diving into the experiment itself, we started by comparing the three texts that we already had: the Middle French original, its official translation into Modern French (non-versified) and its Bulgarian official translation (versified). The choice of versification or non-versification stands for the needs of understanding the full meaning of the message (Modern French) or for those of enjoying the full effects of a literary work (Bulgarian). In the French literary field, while it is virtually impossible to fully enjoy the work in the original language (as it has evolved way beyond its medieval state), one can still see the verses and the rhymes, recognise certain terms and thus see at least the structure that was meant to be enjoyed at the time. In Bulgarian, all this would have been lost, if not for the effort of the translators to find a suitable versified version. Of course, this version has the disadvantage of lacking some information which was impossible to fit into the tight octosyllable verse.

We have to address as well some other particularities of translation from / into an ancient language (Verma *et al.*, 2023). First, as with Latin, it has no native speakers, thus no possibility to verify the usage of terms or the *natural* way of expressing in the language. Translating from a dead language means that we decipher rather than enjoy the richness and versatility of expression, having little information on how the natural language sounded and which constructions were considered plain, everyday language, compared to more sophisticated language. Secondly, there is a knowledge of historical and cultural context, which no longer exists. Even with the current progress of AI in this field, there is no certainty about the extent of its knowledge about past times, the customs, activities, and relations between people[4].

We chose an excerpt from the novel based on the presence of events and clearly defined characters, for several reasons. First, this excerpt can be read and understood on its own, even though the characters have already been mentioned before in the novel. Their interactions and the events in which they participate are self-explanatory. This serves us well in translating the

[4] Translation tools for ancient languages have existed since 2004, when the site Sacred-Texts.com was created, offering extensive corpora of ancient texts (the Bible in all languages of the Antiquity; Sumerian texts; eposes…), many of them translated into English. A curious example to be found there are Nostradamus's writings (first half of 16th century); orthography and spelling are still not normalised (they predate the big reform of the French Academy's Dictionary, in the 17th century). This, again, raises the important question of recognising a language as such, even when it does not have a uniform norm.

excerpt later with the help of neural networks. Second, in this excerpt some key features of middle French are present, such as cases (nominative and accusative), feminine and masculine, and they have more than one occurrence.

Here is the excerpt, with the fully recognisable and the fully translated parts from Middle French into Modern French and Bulgarian:

Markings for:

– Exact translations (over the three texts); in Modern French: words recognisable by etymology;

– Paraphrases of terms (non-existent or having changed meaning over time);

– Bulgarian text: interpretations of meaning.

Original text	**Translation into modern French**	**Bulgarian translation**
Atant s'an va chascuns par lui.	Chacun s'en va sur le chemin de son choix;	
Et cil de la charrete panse	Et celui de la charrette reste plongé dans ses pensées	А оня, в колесарка возен
Con cil qui force ne desfanse	Tout comme une personne privée de force et de défense	от мисъл зла не е тормозен -
N'a vers Amors qui le justise;	Contre Amour qui le maintient sous sa juridiction;	от Любовта е завладян.
Et ses pansers est de tel guise	Sa méditation est d'une intensité telle	Такъв е неговият (блян),
Que lui meïsmes en oblie,	Qu'il perd le sens de lui-même;	че себе си дори забрави,
Ne set s'il est ou s'il n'est mie,	Il ne sait pas s'il existe ou s'il n'existe pas,	какво наистина да прави,
Ne ne li manbre de son non,	Il ne se rappelle pas son nom,	а името си пък съвсем,
Ne set s'il est armez ou non,	Il ne sait s'il est armé ou non,	дали е с (копие и шлем),
Ne set ou va, ne set don vient;	Il ne sait pas où il va, ni d'où il vient;	отде дойде, къде отива,
De rien nule ne li sovient	Il ne se souvient de rien,	и всяка мисъл разсъдлива
Fors d'une seule, et por celi	Hormis d'une seule chose, et, à cause d'elle,	освен една – и няма той

Original text	**Translation into modern French**	**Bulgarian translation**
A mis les autres en obli ;	Il a mis les autres choses en oubli ;	от нея нито миг покой.
A cele seule panse tant	Il pense tant à cette seule chose	Не слуша нищо, в тая мисъл
Qu'il n'ot ne voit ne rien n'antant.	Qu'il n'entend, ne voit ni ne comprend rien.	и сляп, и глух се е улисал.
Et ses chevax mout tost l'en porte,	Et son cheval l'emporte à vive allure,	А конят не запени гръд
Ne ne vet mie voie torte,	En n'empruntant jamais de mauvais chemin	по лъкатушещия път –
Mes la meillor et la plus droite ;	Mais toujours le meilleur et le plus direct;	напряко стигна до една
Et tant par avanture esploite	Il s'empresse si habilement que par aventure,	най-ненадейна долина,
Qu'an une lande l'a porté.	Il l'a conduit dans une lande.	където спря забързан ход.
An cele lande avoit .i. gué	Dans cette lande, il y avait un gué	Там имаше река и брод,
Et d'autre part armez estoit	Sur l'autre rive duquel se trouvait, tout armé,	и рицар също от онази
Uns chevaliers qui le gardoit ;	Un chevalier qui en assurait la garde;	страна зад брода, да го пази.
S'ert une dameisele o soi	Et celui-ci avait avec lui une demoiselle.	До него млада хубавица -
[...]		на коня си една девица.
[...]		
Et cil qui fu de l'autre part	Et celui qui fut sur l'autre bord	А онзи там оттатък каза:
S'escrie : « Chevaliers, ge gart	S'écrie: « Chevalier, je garde	"Хей, рицарю, аз брода пазя,
Le gué, si le vos contredi. »	Le gué, et je vous en défends la traversée. »	да минете посмейте само!"
Cil ne l'antant ne ne l'oï,	Ce dernier ne l'entend ni ne l'écoute,	А този тука дързостта му

Original text	Translation into modern French	Bulgarian translation
Car ses pansers ne li leissa,	Car son penser ne le lui permet pas;	не чу, потънал в мисли цял.
Et totes voies s'esleissa	Toutefois, avec ardeur,	Летеше конят зажаднял
Li chevax vers l'eve mout tost.	Le cheval s'élança à toute vitesse vers l'eau	със всички сили към водата.
Cil li escrie que il l'ost :	L'autre lui crie de se détourner	А онзи викаше оттатък:
« Lai le gué, si feras que sages,	Du gué, que ce sera prudent de sa part,	"Ехей, от брода се махни,
Que la n'est mie li passages ».	Car par là on ne trouve point de passage.	минава пътя по-встрани."
Et jure le cuer de son vantre	Et il jure par le coeur qui bat dans sa poitrine	С длан на сърцето се закле:
Qu'il le ferra, se il i antre.	Qu'il le transpercera de sa lance s'il y met le pied.	нагази ли, ще пати зле.
[...]		
Cil panse tant qu'il ne l'ot pas,	Il pense toujours si fort qu'il ne l'entend pas,	Не го и чу замислен той,
Et li chevax en es le pas	Et soudain, le cheval	а конят му за водопой
Saut en l'eve et del chanp se soivre,	Saute dans l'eau, abandonnant le champ,	чак до гърди се потопи
Par grant talant comance a boivre.	Et, en s'y adonnant à coeur joie, il commence à boire.	и жадно от водата пи.
[...]		
Lors vient li chevaliers avant	Le chevalier s'avance alors	Дойде сред брода онзи. Тук
Enmi le gué, et cil le prant	Jusqu'au milieu du gué, et l'autre le saisit	отляво рицарят в юмрук
Par la resne a la main senestre	Par la rêne de la main gauche	юздата стисна му, защото
Et par la cuisse a la main destre ;	Et de la main droite par la cuisse ;	отдясно сграбчи му бедрото,
Sel sache et tire et si l'estraint	Il le tient, le tire et le serre	задърпа с мощ ожесточена,
Si duremant que cil se plaint,	Tellement fort que l'autre se plaint	пазачът-рицар чак застена –

Original text	Translation into modern French	Bulgarian translation
Qu'il li sanble que tote fors	Qu'il lui semble effectivement	изтръгват, стори му се, цяло
Li traie la cuisse del cors.	Qu'on lui arrache du corps la cuisse;	парче от неговото тяло.
Se li prie que il le lest	Et il le supplie de le laisser	Помоли го да го остави:
Et dit : « Chevaliers, se toi plest	En disant: « Chevalier, s'il te plaît	"Я според рицарските нрави -
A moi conbatre par igal,	Que nous nous combattions d'égal à égal	да бъдем като равен с равен –
Pran ton escu et ton cheval	Prends ton bouclier et ton cheval	вземи щита и коня славен,
Et ta lance, si joste a moi. »	Et ta lance, et joute avec moi."	и копието за двубой."
[...]		
Lors li cort sore et si le haste	Alors, il fonce sur l'autre et le presse	Нахвърли се ожесточен.
Tant que cil li ganchist et fuit ;	Si fort que celui-ci lui abandonne la partie et s'enfuit;	Пазача тук на бяг удари
Le gué, mes que bien li enuit,	Le gué – chose qui le contrarie beaucoup –	да мине брода, но не свари,
Et le passage li otroie.	Et le passage, il les lui octroie.	преследван по петите той,
Et cil le chace tote voie	Et ce dernier le pourchasse sans relâche	и свърши целия двубой,
Tant que il chiet a paumetons.	Jusqu'à ce qu'il tombe sur ses mains.	че по очи беглецът падна.
[...]		
Lors li vient sus, l'espee treite ;	Il avance donc jusqu'à lui, l'épée toute prête;	Към него тръгна с меч в ръка.
Et cil dit, qui fu esmaiez :	Et, épouvanté, l'autre dit:	А оня сгърчи се в тревога:
« Por Deu et por moi l'en aiez	" Pour Dieu et pour moi, accordez-lui	"Към мене в името на бога
La merci que je vos demant. »	La grâce qu'elle implore et que je vous demande moi aussi	бъдете милостив, ви моля."

Original text	Translation into modern French	Bulgarian translation
Et cil respont : « Se Dex m'amant,	Et il répond: « Que Dieu en soit témoin,	"Свидетел ми е бог, че воля
Onques nus tant ne me mesfist	Jamais nul ne se comporta à mon égard si méchamment	по-зла от твоята не знам,
Se por Deu merci me requist,	Que, s'il invoquait Dieu en me demandant grâce,	от бога после нямаш срам –
Que por Deu, si com il est droiz,	Pour Dieu, et comme il est juste	да те щадят за твойто зло.
Merci n'an eüsse une foiz.	Je refusasse de la lui accorder une seule fois.	Това и с други е било.
Et ausi avrai ge de toi,	J'aurai donc pitié aussi de toi	Така и с тебе ще направя,
Car refuser ne la te doi	Car je ne dois pas te la dénier	защото милостта се дава,
Des que demandee la m'as ;	Puisque tu me l'as demandée;	след като я измолиш ти.
Mesençois me fianceras	Mais c'est à condition que tu t'engages,	Ала едно запамети:
A tenir, la ou ge voldrai,	Là où je voudrais, à te constituer	Да идеш, та макар без стража,
Prison quant je t'an semondrai.	Prisonnier, quand je t'en donnerai l'ordre.	като затворник, дето кажа.

We can see that the Modern French translation often has full verses corresponding directly to the derived words from Old French, thus recognisable for anyone who is aware of the linguistic changes that took place:

Phonetic changes:
– vocalisation of "l" : *voldrai>voudrai* ;
– reduction of diphthongs : *avoit>avait* ;
– *s* in middle syllables falling out : *joste>joute* ; *desfanse > défense*; *resne>rêne*;
– advancement and rounding of /ɔ/>/u/: *obli>oubli*; *cort>court* ; *tote>toute* ; *por>pour*.

Morphological changes:
– accusative for nominative: *chevaliers>chevalier*; *ses chevax>son cheval*, *ses pensers>son penser, sa pensée* (after taking up the Frankish final *s* for plural, this accusative form is left out);
– introduction of subject pronouns into Modern French (Frankish influence) : *demandee la m'as > tu me l'as demandée* ;

– strict word order in Modern French against free word order in Old French : *et ausi avrai ge de toi > j'aurai donc […] aussi de toi*;
– vocabulary changes: *chiet>tombe* ; *droiz>juste* ; *esmaiez>épouvanté* ; *ferra>transpercera (frappera)* ; *oï>écoute* ;

Orthographic changes :

– lack of spelling rules in Old French (generally tending towards phonetic spelling) : *ge/je >je*, *mes>mais*, *ausi>aussi*, *chace>chasse*;
– lack of diacritic signs (accents): *demandee>demandée*; *la>là*; *ou>où*;
– phonetic spelling (Old French) against etymological spelling (Modern French) : *com>comme*; *demant>demande*; *respont>répond*; *pran>prends*; *conbatre>combattions*; *cors>corps*; *sanble>semble*.

Finally, there are such words that have stayed exactly the same (*car*, *toi*, *une*, *main*, *cuisse*, *lande*, *dit*). This allows CAT tools and AI to derive general meaning – not only because of the inherent semantics of the text, but also in what pertains to cultural context.

The translation into Bulgarian, as we already stated, was done for the purposes of enjoying the text. For that reason, we see much more paraphrased text than translated faithfully from Old French. However, the main story is preserved: there are two main characters, both knights. One of them is guarding a ford (*gué*), forbidding its entry, and the other comes and inadvertently enters the river. A battle ensues, where the imposter is defeated but instead of being killed, he is liberated with the promise to accept to be made prisoner as soon as the victor decides. The logical connections in the text, as well as the main figures of speech, are preserved to the best that the translators could and some cunning interpretations have been made, so that the meaning is conveyed in the space of the octosyllable:

Original text	**Modern French**	**Bulgarian**	**English**
Et ses pansers est de tel guise	Sa méditation est d'une intensité telle	Такъв е неговият блян,	*Such is his dream*
[…]	[…]	[…]	[…]
Ne set s'il est armez ou non,	Il ne sait s'il est armé ou non,	дали е с копие и шлем,	*If he is with a javelin and a helmet*
Ne set ou va, ne set don vient ;	Il ne sait pas où il va, ni d'où il vient ;	отде дойде, къде отива,	*Where he came from, where he is going*
[…]	[…]	[…]	[…]
Et cil qui fu de l'autre part	Et celui qui fut sur l'autre bord	А онзи там оттатък каза:	*And the one on the other side said :*

Original text	Modern French	Bulgarian	English
S'escrie : « Chevaliers, ge gart	S'écrie: « Chevalier, je garde	"Хей, рицарю, аз брода пазя,	*- Hey, knight, I guard the ford,*
[…]	[…]	[…]	[…]
Et totes voies s'esleissa	Toutefois, avec ardeur,	Летеше конят зажаднял	*The thirsty horse was flying in*
Li chevax vers l'eve mout tost.	Le cheval s'élança à toute vitesse vers l'eau	със всички сили към водата.	*With all his might towards the water*

Results

Our experiment concerned the machine translation provided by Phrase, from Latin and modern French as source languages into Bulgarian as a target language. After conducting this experiment, we decided to go further and see how ChatGPT (version 3.5, at the time) addressed the issue, so we translated an even smaller excerpt with it. To confirm our findings, we contrasted it with an existing – by November 2024 – GPT especially trained in translation into Old French, but this time from the 13th-14th centuries (and not 12th, as Chrétien's text).

We did not evaluate the translation quality using automated metrics, as this would have been outside the scope of the experiment, aiming to show how large language models work. This would be a suitable expansion of the current hypothesis for future research. Nonetheless, in the following results, we have marked, in the same way that we did for the original human translations, the discrepancies between the original and its consecutive translations by different mechanisms, in order to visualise the results for the reader.

Markings for:
- transliteration;
- copy of the original segment;
- correct translation;
- incorrect contamination of meaning;
- interference with words from other languages and incorrect translation.

Original text	From Latin into Bulgarian	From Modern French into Bulgarian
Atant s'an va chascuns par lui.	Атанте с'ан ва часкунс пар луй.	И сега ще вземе всички със себе си.
Et cil de la charrete panse	И целият свят на каруцата мисли	И cil de la charrete panse
Con cil qui force ne desfanse	Със сила, която не се отказва	Con cil qui force ne desfanse

Original text	**From Latin into Bulgarian**	**From Modern French into Bulgarian**
N'a vers Amors qui le justise ;	Няма стихове за любовта, която го оправдава.	Не разполага с Amors, които да го оправдаят;
Et ses pansers est de tel guise	И неговите броня е от такъв вид	И неговите пансери са такава маска
Que lui meïsmes en oblie,	Който сам забравя,	Нека ме забрави,
Ne set s'il est ou s'il n'est mie,	Не знам дали е или не е.	Не знам дали е така, или не,
Ne ne li manbre de son non,	Не, не е ли мъка от неговото име.	Ne li manbre de son non,
Ne set s'il est armez ou non,	Не знам дали е въоръжен или не,	Не задавайте дали е въоръжен или не,
Ne set ou va, ne set don vient ;	Не се опитвай, не се опитвай да дойдеш;	Ne set ou va, ne set don vient ;
De rien nule ne li sovient	Няма за какво, никой не помни.	От нищо ли не се възползвате
Fors d'une seule, et por celi	За един, и за този	Fors d'une seule, et por celi
A mis les autres en obli ;	А е поставил другите в задължение;	е поставил другите в obli ;
A cele seule panse tant	Само един ум мисли толкова.	Само за този един корем
Qu'il n'ot ne voit ne rien n'antant.	Да не вижда нищо, нито пък нещо.	Той не може да види нищо.
Et ses chevax mout ost l'en porte,	И косата му много бързо я носи,	А конете му винаги го носят,
Ne ne vet mie voie torte,	Не не вет ми воие торте,	Ne vet mie voie torte,
Mes la meillor et la plus droite ;	Мес ла мейлор ет ла плюс дройт ;	Mes la meillor et la plus droite ;
Et tant par avanture esploite	И толкова напред приключение	И толкова много от advance esploite
Qu'an une lande l'a porté.	Когато една поляна я е носила.	че го носи блато.
An cele lande avoit .i. gué	В тази страна имаше .i. битка	В това тресавище имаше .i. брод
Et d'autre part armez estoit	И от друга страна, това беше въоръжено.	А от другата страна, въоръжени с оръжия.
Uns chevaliers qui le gardoit ;	Някои рицари, които го пазеха;	Рицар, който го пази;
S'ert une dameisele o soi	С'ерт une дамейзеле о сои	S'ert une dameisele o soi
[...]	[...]	[...]

Original text	From Latin into Bulgarian	From Modern French into Bulgarian
Et cil qui fu de l'autre part	И той, който беше от другата страна	И cil qui fu de l'autre part
S'escrie : « Chevaliers, ge gart	Пише се : Рицари, ге гарт	S'escrie : "Knights, ge gart
Le gué, si le vos contredi. »	Блатото, ако ви противоречи.	Le foré, si le vos contredi."
Cil ne l'antant ne ne l'oï,	Цел не л'антант не не л'ои,	Cil ne l'antant ne l'oï,
Car ses pansers ne li leissa,	Като тези панцири не оставят ли	Защото пансерите му никога не го пускат,
Et totes voies s'esleissa	И всички пътища се отделиха	И всички пътища бяха издигнати
Li chevax vers l'eve mout tost.	Ли чевакс верс л'еве моуст тост.	Li chevax vers l'eve mout tost.
Cil li escrie que il l'ost :	Цил ли е да напише, че той е гост?	Cil li escrie que il l'ost :
« Lai le gué, si feras que sages,	« Лай ле гюе, ако направиш, че мъдри,	«Lai le gué, si feras que sages,
Que la n'est mie li passages ».	Нека не е преминаването.	Que la n'est mie li passages".
Et jure le cuer de son vantre	И по правото на сърцето на корема му	И да се закълне в короната на своята гордост
Qu'il le ferra, se il i antre.	Ако той го направи, ако влезе.	Той ще го направи, ако може.
[...]	[...]	[...]
Cil panse tant qu'il ne l'ot pas,	Целта е да се мисли, докато не я има.	Цил превръзки, стига да не се от,
Et li chevax en es le pas	И ли шевакс в е пас	Et li chevax en es le pas
Saut en l'eve et del chanp se soivre,	Скок в ева и от полето се събужда,	Saut en l'eve et del chanp se soivre,
Par grant talant comance a boivre.	Пара грант талант комансе а боивре.	Par grant talant comance a boivre.
[...]	[...]	[...]
Lors vient li chevaliers avant	Скоро идва рицарят	Когато идва li рицари преди
Enmi le gué, et cil le prant	Вземи го, и той го взема	Enmi le gué, et cil le prant
Par la resne a la main senestre	Чрез лентата на лявата ръка	С помощта на смолата в зловещата ръка
Et par la cuisse a la main destre ;	И от бедрото до дясната ръка;	И за бедрото до ръката на Дестре;

Original text	**From Latin into Bulgarian**	**From Modern French into Bulgarian**
Sel sache et tire et si l'estraint	Вземи, дръпни и ако го притиснеш	Sel sache et tire et si l'estraint
Si duremant que cil se plaint,	Ако той се оплаква, докато спи,	Толкова твърдо, че се оплаква,
Qu'il li sanble que tote fors	Да му се струва, че всичко е наред.	Qu'il li sanble que tote fors
Li traie la cuisse del cors.	Ли трае на бедрото на корса.	Li milie la cuisse del cors.
Se li prie que il le lest	Се ли приеме, че той е прочетен	Se li prie que il le lest
Et dit : « Chevaliers, se toi plest	И каза: Рицари, да ти е приятно	И каза: "Knights, se toi plest
A moi conbatre par igal,	Аз трябва да се боря за равенство,	A moi conbatre par igal,
Pran ton escu et ton cheval	Пран тон еску и тон шевал	Pran вашия escu и вашия кон
Et ta lance, si joste a moi. »	И твоята копие, ако е при мен.	И твоето копие, което ми е толкова скъпо."
[...]	[...]	[...]
Lors li cort sore et si le haste	Когато сърцето е горе и ако бързината	Lors li cort sore et si le haste
Tant que cil li ganchist et fuit ;	Докато този, който е ганчист, беше.	Tant que cil li ganchist et fuit ;
Le gué, mes que bien li enuit,	Ла гюе, моите, които добре му е вечер.	Le gué, mes que bien li enuit,
Et le passage li otroie.	И преминаването му е друго.	И пасажът ли otroie.
Et cil le chace tote voie	И той преследва всяка пътека	И тя chace tote voie
Tant que il chiet a paumetons.	Докато той мълчи за пауметоните.	Стига да има цигулка.
[...]	[...]	[...]
Lors li vient sus, l'espee treite ;	Когато той идва, мечът блести;	След това се появява отгоре с кръстосана сабя;
Et cil dit, qui fu esmaiez :	И той каза, който беше изтласкан:	Et cil dit, qui fu esmaiez :
« Por Deu et por moi l'en aiez	« За Бога и за мен, нека го имате.	"Por Deu et por moi l'en aiez
La merci que je vos demant. »	Благодаря, което ви искам.	Благодарностите, които ти дължа.
Et cil respont : « Se Dex m'amant,	И той отговори: « Ако Бог ме обича,	И диша: "Se Dex m'amant,

Original text	From Latin into Bulgarian	From Modern French into Bulgarian
Onques nus tant ne me mesfist	Наистина, никой не ме е наранил.	Onques nus tant ne me mesfist
Se por Deu merci me requist,	Се поради Божия милост ме поиска,	Se por Deu merci me requist,
Que por Deu, si com il est droiz,	Какво за Бога, ако е правилно,	Que por Deu, si com il est droiz,
Merci n'an eüsse une foiz.	Благодаря, че не беше веднъж.	Merci n'an eüsse une foiz.
Et ausi avrai ge de toi,	И ще имаш смелостта да бъдеш с теб.	А също и за вас,
Car refuser ne la te doi	Отказвам колата, не ти дължа.	Защото отказвайки да ви го даде.
Des que demandee la m'as ;	Веднага щом поисках мъжа;	Des que demandee la m'as ;
Mesençois me fianceras	Мес енсуа ме финсира	Mesençois me fianceras
A tenir, la ou ge voldrai,	Да държиш, където искаш,	A tenir, la ou ge voldrai,
Prison quant je t'an semondrai.	Затвор колко пъти ще те призова.	Затвор, когато ви видя.

We can see that the results in Latin are much better than those in Modern French. The CAT tool was able to extrapolate certain Old French words from the rules and statistics for Latin, but not from Modern French. These are interesting results, knowing that both languages – Latin and Modern French – have strict spelling and we would have expected similar results, if not tending towards Modern French. At the same time, even with Vulgar Latin there has never been a strict norm (Gaspar, 2020); interpretation depends on time and place and these particularities have been introduced in the translation, bearing in mind that it was made by an AI-powered engine, under the name of Phrase Language AI.

Apart from the unrecognised segments and the correctly translated words and segments, there were still particular expressions from Latin which had never been literal and which gave again non-literal derivatives in French. Such was the expression found in old French *tote voie / totes voies*, which in modern French gave *toutefois* (however). It derives from *tota via erras* (Bougy, 2007), meaning in Latin 'you are wrong all the way' (as we may use it in modern English as well). There are a few instances of it in the text, some of which were literally translated and others were either transliterated or the source segment was copied in the target text. The CAT tool has to have some previous input for this kind of expressions; a corpus for statistic-based machine translation may be considered, if not a rule-based system for machine translation from ancient languages (if the goal is exact translation, of course).

Translation with ChatGPT

Here are the results gained in translating with ChatGPT in November 2023. To measure the precision of the translation, we did not feed the entire excerpt used for Phrase to the LLM: we wanted to see how it fared without having too much context. Here are the results:

Original text	**GhatGPT translation (november 2023)**	**Remarks**
Atant s'an va chascuns par lui.	Всеки се отдалечава от себе си.	Incorrect translation
Et cil de la charrete panse	И този в колесницата си мисли,	OK
Con cil qui force ne desfanse	както този, който не е премахнал силата	Incorrect translation : enjambement over to the next verse
N'a vers Amors qui le justise ;	на Любовта;	
Et ses pansers est de tel guise	и мислите му са такива,	OK
Que lui meïsmes en oblie,	че сам се забравя,	OK
Ne set s'il est ou s'il n'est mie,	не знае дали съществува или не,	OK
Ne ne li manbre de son non,	не си спомня за името си,	OK
Ne set s'il est armez ou non,	не знае дали е въоръжен или не,	OK
Ne set ou va, ne set don vient ;	не знае къде отива, не знае откъде идва;	OK
De rien nule ne li sovient	не помни нищо,	OK
Fors d'une seule, et por celi	освен едно, и заради тази една	Incorrect grammatical agreement of terms because of the feminine "rien" in old French
A mis les autres en obli ;	забравя всички други;	
A cele seule panse tant	на тази една мисли толкова,	The same incorrect agreement
Qu'il n'ot ne voit ne rien n'antant.	че не забелязва и не стойността на нищо.	Incorrect translation
Et ses chevax mout tost l'en porte,	И конете го носят бързо,	OK
Ne ne vet mie voie torte,	не се свърта по крив път,	Incorrect translation only of the term "свърта"
Mes la meillor et la plus droite ;	а най-добрият и най-прав път;	OK
Et tant par avanture esploite	и така по съдба	Incorrect translation of logical links
Qu'an une lande l'a porté.	е изпрат на една поляна.	"изпрат" doesn't exist in Bulgarian

In this excerpt, the syntaxic coherence is much more pronounced. ChatGPT has learned how Bulgarian syntax works and tries to make connections and to arrive to a certain coherent sentence. The second sentence of the excerpt, being quite long and having a lot of subordinate clauses with repetitive structure, could not be translated correctly by Phrase. We have to concede that Phrase segmented it into much shorter chunks and had to interpret meaning only inside these chunks, not considering the whole passage. Certain cases are over-interpreted as plurals and this creates a meaning shift; however, here with ChatGPT, there is no discrepancy as to the meaning of the story, which is the big difference with Phrase.

The comparison between ChatGPT's translation and Phrase's translation has to do with the engines used by Phrase. As we have seen above, some of the engines listed indeed use neural networks extensively, which should have allowed us to have similar results to those of ChatGPT. The only difference that we can make is the fact that under Phrase the text is already segmented and the interpretation of terms is done inside the same segment.

ChatGPT proves useful for the use of gisting translations of texts in Old French. Gisting translations are proposed as a service by different language service providers and range from machine translation gisting (without any post-editing)[5] to summarising a text[6] that the client needs for translation. While almost all of the discovered ancient texts in French are accessible through a modernised version, this ability of AI can prove useful when constituting and analysing corpora of ancient texts for linguistic, cultural, educative or other purposes. The original texts are often found online without their respective adaptation or translation into Modern French due to author's rights (these cover the modern edition and the modern text, but not the ancient text, which has been created centuries ago). So for the purposes of using cultural heritage in writing form (novels, inscriptions, apocryphal biblic texts and other writings) from ancient times, it will be useful to have some possibility to gist these texts.

Contrasting with an existing Old French generative pre-trained transformer (GPT)

By the present moment, November 2024, there is still an existing GPT model for translating text into Old French, specified for 14th century French

[5] Here are some definitions given on LSPs' pages: United Language Group, https://www.unitedlanguagegroup.com/blog/what-is-machine-translation-gisting; Pairaphrase, https://www.pairaphrase.com/how-it-works/translation-wizard/;

[6] CETRA: https://www.cetra.com/blog/gisting-cost-effective-translation-services/, an experiment conducted by Web Corpora, http://web-corpora.net/acad/gisting/.

(by a 2-centuries difference with Chrétien's text)[7]. Having this in mind, as well as the fact that any recreation of dead languages means that there will be less richness and more of a fixed norm of the language (and the problems of drawing a single system for a non-normativised language, such as French in the middle ages), we tried to 'verify' ChatGPT's performance by translating the Bulgarian version we see above with this GPT model. Here are the results:

Markings for:
– *incorrectly translated segments*;
– synonyms / lexical variants;
– unrecognised case in Bulgarian, rendered as a plural, affecting translation in old French.

ChatGPT-translated Bulgarian text	**ChatGPT-generated Old French text**	**Original text**
Всеки се отдалечава от себе си.	*Chascuns s'eschappe de soi meisme.*	Atant s'an va chascuns par lui.
И този в колесницата си мисли,	Et cil en sa charete cuide,	Et cil de la charrete panse
както този, *който не е премахнал*	come cil *qui n'a pas osté la*	Con cil qui force ne desfanse
силата на Любовта;	*force d'*Amours;	N'a vers Amors qui le justise ;
и мислите му са такива,	et ses pensées sont teles	Et ses pansers est de tel guise
че сам се забравя,	qu'il se oublie,	Que lui meïsmes en oblie,
не знае дали съществува или не,	ne set se il est ou n'est,	Ne set s'il est ou s'il n'est mie,
не си спомня за името си,	ne se sovient de son nom,	Ne ne li manbre de son non,
не знае дали е въоръжен или не,	ne set se il est armé ou non,	Ne set s'il est armez ou non,
не знае къде отива, не знае откъде идва;	ne set où il va, ne set d'où il vient;	Ne set ou va, ne set don vient ;
не помни нищо,	ne retient riens	De rien nule ne li sovient
освен едно, и заради тази една	fors une chose, et pour ceste une	Fors d'une seule, et por celi
забравя всички други;	oublie toutes les autres;	A mis les autres en obli ;
на тази една мисли толкова,	à celle seule il pense tant	A cele seule panse tant
че не забелязва и не стойността на нищо.	*qu'il ne voit ne pris ne valeur de nulle autre chose.*	Qu'il n'ot ne voit ne rien n'antant.
И конете го носят бързо,	Et les chevaus le portent hastivement,	Et ses chevax mout tost l'en porte,

[7] GPT Traduction en ancien français: https://chatgpt.com/g/g-1sZgBwKz0-traduction-en-ancien-francais.

ChatGPT-translated Bulgarian text	**ChatGPT-generated Old French text**	**Original text**
не се свърта по крив път,	ne s'arreste en nul sentier tort,	Ne ne vet mie voie torte,
а най-добрият и най-прав път;	mais le meilleur et plus droit chemin;	Mes la meillor et la plus droite ;
и така по съдба	*et si par destinee*	Et tant par avanture esploite
е *изпрат* на една *поляна*.	est mené en une *praerie.*	Qu'an une lande l'a porté.

This translation clearly shows that recognised particles were already introduced into the text and that recognisable terms in Bulgarian (*крив – torte*; *толкова – tant*; *освен – fors*; *Любовта – Amours*; *колесница – charrette*; including whole segments – *не знае дали съществува или не – ne set se il est ou n'est*, the verb *exister*, to exist, corresponding to the Bulgarian *съществувам* having been introduced much later), gave correct translation into the reconstructed Old French GPT. The Old French GPT has clearly some synonyms introduced (*pense/cuide; si/tant*), only one of which can be found in the original text. Again, versification was left out, aiming strictly at the meaning of the text. An interesting result here is the fact that the GPT's translation has conserved many of the original structures and thus the rhythm, albeit not always the rhyme.

We have rich vocabulary in this excerpt concerning thinking, remembering and forgetting. The cunning language that Chrétien used can be seen in contrasting the two verses *oublie toutes les autres* (Old French GPT) – *A mis les autres en obli* (original text). Another interesting detail is that the Old French GPT repaired a grammatical error in the Bulgarian text, namely an agreement: *освен **едно**, и заради **тази една** – fors une chose, et pour ceste une.*

Conclusion

We deemed the experiment conclusive on the previously set hypotheses and research questions. Throughout the experiment, the main features of the two technologies – CAT tools and LLMs – were successfully put forward. The aim was to show the discrepancies between stages of development in the French language, using in practice a modern translation technology. All of the particularities that we needed to show: phonetic, morphological and syntactic changes from Latin to Modern French were illustrated. Furthermore, a successful model for gist-translating ancient languages was put into practice: ChatGPT.

For a larger and replicable study of ChatGPT's resourcefulness in translating Old French texts, it would be suitable to have a much larger context and introduce standardised metrics (like BLEU or TER) to assess translation quality across tools. For the needs of the present experiment, ChatGPT was provided with very limited context, which skewed its performance and

did not allow for leveraging its contextual understanding. A larger corpus would lead us to a clearer comparative analysis. There is such a possibility, as the entire cultural heritage in medieval French language is available online, through the Gallica site of the French National library. For an adequate corpus, Chrétien de Troyes' novels can be compiled; another corpus may contain the French texts of the Arthurian legends (*la matière bretonne*), or all the writings made at Marie de France's command.

Such a study can provide practical uses for gist translating of ancient texts: browsing through cultural heritage texts, creating corpora for analysing language development through the ages and making these sources accessible to history and linguistics' students. A specific use of translation in the case of the Arthurian legends can be the creation of a trans-linguistic corpus of texts putting together the whole of the existing saga. ChatGPT proves useful for the reverse translation as well, when a supplementary verification is needed, in the case of Old French.

References

Bougy, C. (2007). 'Les « pseudo-adverbes » concessifs neporuec, neporquant, nequedent et l'adverbe toute(s)voie(s) en ancien français'. *Syntaxe sémantique*, Issue 8, 107–124.

Cerquiglini, B. (1993). *La naissance du français.* 2nd edition. Paris: Presses Universitaires de France.

Gaspar, C. (2020). 'Orthography as Described in Latin Grammars and Spelling in Latin Epigraphic Texts'. *Acta Classica Universitatis Scientiarum Debreceniensis*, Issue 56, 61–72.

Lacy, N. J. (1991). 'Chrétien de Troyes'. In: N. J. Lacy. *The New Arthurian Encyclopedia.* New York: Garland, 88–91.

Lodge, R. A. (1993). *French: From Dialect to Standard.* London: Routledge.

Pellegrini, G. B. (1980). 'Substrata'. In: R. Posner, J. N. Green, eds. *Trends in Romance Linguistics and Philology. Volume 1: Romance Comparative and Historical Linguistics.*The Hague: Mouton, 43–72.

Roegiest, E. (2006). *Vers les sources des langues romanes: Un itinéraire linguistique à travers la Romania.* Leuven: Acco.

Uitti, K. D. (1995). *Chrétien de Troyes Revisited.* New York: Twayne Publishers.

Verma, S., Gupta, N., B C, A., Chauhan, R. (2023). 'A Novel Framework for Ancient Text Translation Using Artificial Intelligence'. *Advances in Distributed Computing and Artificial Intelligence Journal*, 11(4), 411–425.

ISSN (online): 2815-4967