Machine translation in asylum reception centres: system selection and multilingual quality evaluation

Lieve Macken
Margot Fonteyne
Arda Tezcan
Ella van Hest
Katrijn Maryns
July De Wilde

Lieve Macken
Gent University;
lieve.macken@ugent.be;
ORCID: 0000-0001-7516-7487

Margot Fonteyne
Gent University;
margot.fonteyne@ugent.be;
ORCID: 0000-0002-9575-7844

Arda Tezcan
Gent University;
arda.tezcan@ugent.be;
ORCID: 0000-0002-8707-6176

Abstract

This paper discusses the selection of a machine translation system for use in asylum reception centres, to translate messages written in English, French, and Dutch into 14 languages, including low-resource ones like Pashto and Somali. Translation quality is assessed through automatic metrics and expert human evaluations to support effective multilingual communication.

Keywords: asylum seekers, reception centres, machine translation, low-resource languages, translation quality evaluation.

Resumen

Este artículo aborda la selección de un sistema de traducción automática para su uso en centros de acogida, destinado a traducir mensajes redactados en inglés, francés y neerlandés a 14 lenguas, incluidas lenguas con pocos recursos como el pastún y el somalí. La calidad de la traducción se evalúa mediante métricas automáticas y evaluaciones humanas realizadas por expertos, con el fin de apoyar una comunicación multilingüe eficaz.

Palabras clave: solicitantes de asilo, centros de acogida, traducción automática, lenguas con pocos recursos, evaluación de la calidad de la traducción.

Resum

Aquest article aborda la selecció d'un sistema de traducció automàtica per a l'ús en centres d'acollida, destinat a traduir missatges redactats en anglès, francès i neerlandès a 14 llengües, incloses llengües amb pocs recursos com el paixtu i el somali. La qualitat de la traducció s'avalua mitjançant mètriques automàtiques i avaluacions humanes dutes a terme per experts, amb la finalitat de donar suport a una comunicació multilingüe eficaç.

Paraules clau: sol·licitants d'asil, centres d'acollida, traducció automàtica, llengües amb pocs recursos, avaluació de la qualitat de la traducció.

Rebuda: 30 de abril de 2025| Acceptació: 14 de novembre de 2025 | Publicació: 15 de desembre de 2025

Revista Tradumàtica. Número 23 (Desembre 2025)
Cita recomanada: Macken, Lieve; Fonteyne, Margot; Tezcan, Arda; van Hest, Ella; Maryns, Katrijn; De Wilde, July (2025). «Machine translation in asylum reception centres: system selection and multilingual quality evaluation». Revista Tradumàtica. Tecnologies de la Traducció, 23, 326-349.
<https://doi.org/10.5565/rev/tradumatica.522>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559
Universitat Autònoma de Barcelona

Ella van Hest
Gent University;
Ella.vanHest@ugent.be;
ORCID: 0000-0001-6160-0374

Katrijn Maryns
Gent University;
Katrijn.Maryns@ugent.be;
ORCID: 0000-0002-8919-3050

July De Wilde
Gent University;
July.DeWilde@ugent.be;
ORCID: 0000-0002-4936-955X

1. Introduction

Reception centres accommodate a highly diverse population with varied linguistic backgrounds. Due to this diversity and the precariousness of their life situation, residents often struggle to access crucial information and express themselves effectively in institutional and interpersonal interactions (Maryns et al. 2023). Reception staff must communicate with residents in many different ways, ranging from conveying simple, practical information, such as house rules or how to use public transport, to conducting complex conversations about legal procedures or socio-emotional needs. To facilitate effective communication, it is crucial that reception centres provide appropriate language support tailored to the nature of the situation. Complex, individualised interactions typically require the involvement of a professional interpreter, whereas less demanding exchanges can be managed effectively through the use of a lingua franca (such as English) or a multilingual support application (such as fedasilinfo.be), or with the help of multilingual staff or fellow residents. Furthermore, asylum seekers and reception staff extensively use machine translation (MT) tools to communicate and comprehend information (Pięta & Valdez, 2024; Vieira, 2024; Valdez et al., 2023; Kivijärvi & Myllylä, 2022; Ciribuco, 2020; Demmans Epp, 2017).

However, each of these forms of language support presents specific challenges. For example, scheduling a professional interpreter in urgent situations can be difficult, and using a lingua franca can raise issues related to language variation and mutual intelligibility (not all varieties of Arabic, for instance, are mutually intelligible). Moreover, a particular challenge in asylum and refugee contexts, including reception settings, is the prevalence of languages of lesser diffusion (LLDs). These are languages that are rarely spoken in the Global North (such as Pashto, Tigrinya, and Farsi), and little to no support for them is available in host countries. This issue affects various forms of linguistic support. For more complex forms of interaction, there is a significant discrepancy between the supply and demand for interpreting services for LLDs (González Figueroa & Pöllabauer, 2025; Gavioli & Wadensjö, 2023). Current initiatives in several countries aim to address this gap by training interpreters with an LLD profile. For example, the QUADA (Quality Interpreting in Asylum Procedures) project developed a handbook for asylum interpreters that addressed different trainee profiles (Bergunde & Pöllabauer, 2019), and the EU-WEBPSI project developed an EU-based web portal for video-mediated public service interpreting (PSI) to improve access to basic services for migrants and refugees.

For simpler forms of communication, research has shown that while MT has great potential to facilitate migrants' access to important information, low translation quality, a lack of domain-specific vocabulary in MT systems, and privacy concerns remain obstacles to its use in asylum settings (Pięta & Valdez, 2024; Valdez et al., 2023; Liebling et al., 2020).

To address these challenges and enhance communication between staff and residents, Fedasil — the Belgian federal agency responsible for the reception of asylum seekers — launched a call for project proposals (AMIF funding, 2023–2025). The aim was to develop a digital platform that could automatically translate messages into the languages most commonly spoken by residents. In response to this call, an interdisciplinary team of Ghent University researchers specialising in linguistic ethnography and translation technology proposed the MaTIAS project. The project aims to develop a prototype of a web platform that will enable staff to send practical messages via WhatsApp in the languages most commonly spoken by residents.

Figure 1 illustrates the concept of the prototype we are developing.

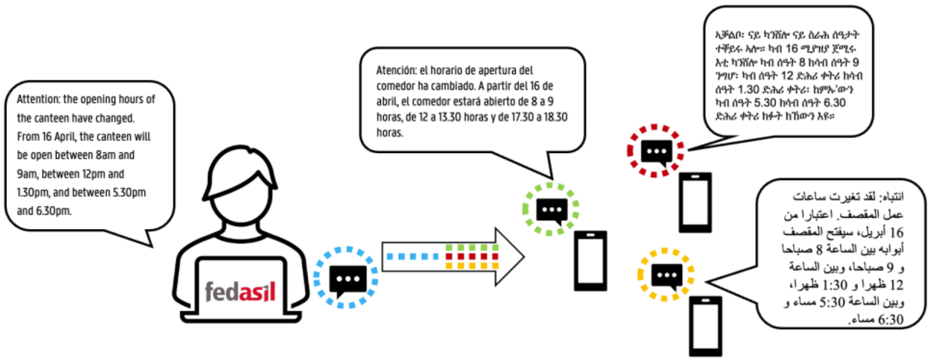

Figure 1: Visualisation of the web platform for sending automatically translated messages via an existing messaging system (e.g. WhatsApp)

This prototype has been specifically designed to enable the rapid and simultaneous delivery of simple messages in multiple languages to one or more residents. Using a web-based platform, staff members can write short messages in Dutch, French or English about upcoming activities or technical issues in the centre, for example. Dutch and French are the two main working languages in the context of asylum procedures in Belgium, while English serves as a lingua franca in many international and multicultural settings. These messages are then automatically translated and delivered to residents in 14 target languages, including LLDs such as Pashto, Somali, and Tigrinya. These 14 languages were selected by Fedasil based on current needs. The system is designed for one-way communication, meaning residents cannot respond to messages. This form of communication complements other, more complex and sensitive interactions, for which alternative forms of language support are employed, such as the use of professional interpreters.

Ethnographic research was conducted in four reception centres to gain insight into current communication practices and identify the necessary preconditions for the development of an effective tool. On-site observations and interviews with staff members (n=30) were conducted to explore how they communicate with residents about practical matters. In addition, document analysis was carried out through the collection of existing written communication materials, such as multilingual posters and flyers. Staff members were also invited to contribute by keeping brief communication diaries or by sharing screenshots and scans of messages they typically send to residents. Furthermore, interviews with residents (n= 45) were conducted to gather their perspectives on the tool, with language support provided by multilingual staff when necessary. In a subsequent phase of the project, the usability of and user satisfaction with the tool will be evaluated. Based on this ethnographic fieldwork, an inventory of around 400 messages was compiled, covering topics including house rules, hygiene and safety, administration and services, opening hours and holidays, appointments, work, and classes (Macken et al. 2025). While the details of the ethnographic research phase that informed the collection of messages and design decisions are discussed extensively elsewhere, this article will focus more closely on the dimension of translation technology. Specifically, it outlines the process of selecting the most suitable MT system for integration into the platform.

The article also reports on a subsequent manual quality evaluation conducted by language experts proficient in the target languages. This evaluation aimed to assess the effectiveness of machine-translated messages in the 14 target languages. Additionally, we examine the predictive value of automatic evaluation metrics, such as ChrF and COMET, for human ratings across languages.

The article is organised as follows. We begin by outlining the challenges faced when using existing MT models to support LLDs. We then report on the selection of the most suitable MT system. Next, we present the human evaluation of the machine-translated messages and explore how these human ratings relate to automatic metrics. The article concludes with a discussion and final remarks.

2. Suitability of current MT models and evaluation metrics for LLDs

Traditional neural MT (NMT) systems typically require substantial amounts of bilingual training data to perform well. This reliance on large-scale, high-quality datasets limits their effectiveness in many low-resource scenarios, where parallel corpora are rare or non-existent, which is the case for most LLDs.

Large language models (LLMs), such as the one powering ChatGPT (OpenAI, 2025), have demonstrated promising translation capabilities, despite not being specifically designed for this task. These models leverage vast amounts of monolingual and multilingual pre-trained data and can perform translation without relying on parallel data, making them appealing for underrepresented languages. However, their performance is highly variable. Their effectiveness tends to be considerably lower for very low-resource languages than for high-resource ones, particularly when those languages are scarcely represented during pre-training (Hendy et al., 2023; Robinson et al., 2023; Xu et al., 2023). In addition, LLMs require substantial computational resources, technical expertise, and maintenance, which can limit their accessibility and practical deployment in many real-world settings.

In addition to the options mentioned above, there are pre-trained or fine-tuned NMT models that are specifically designed for translation tasks and cover many language pairs, including those with very limited resources. Meta's No Language Left Behind initiative represents a notable attempt to overcome these limitations by training multilingual models specifically to improve translation quality for over 200 low-resource languages through a combination of supervised and self-supervised techniques (NLLB Team et al., 2022). Other examples include Facebook's M2M-100 model, which supports direct translation between 100 languages (Fan et al., 2021); Google's multilingual NMT system, designed to leverage shared representations across many language pairs (Aharoni et al., 2019); Microsoft Translator, which incorporates both rule-based and neural techniques to support a wide variety of languages, including low-resource ones (Microsoft, 2025); and ModernMT, a neural adaptive MT system that learns continuously from user feedback and supports customisation for underrepresented languages. While most of these models and services are available off the shelf, many can also be adapted to specific domains to perform optimally on truly low-resource languages. However, achieving consistent and reliable performance for low-resource languages remains difficult and is significantly constrained by the limited availability of relevant data and suitable benchmarks.

MT output quality is typically assessed through automatic and manual evaluation. Automatic metrics are fast, cost-effective, and reproducible, making them ideal for large-scale benchmarking. Manual evaluation, though more resource-intensive, offers greater accuracy by capturing nuances in meaning and fluency that automatic tools may miss.

Automatic metrics can be categorised as word-based, character-based or neural model-based methods. Word-based metrics, such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), rely on overlap at the word level between MT output and a human reference. Although BLEU remains popular due to its simplicity, word-based metrics are less reliable for languages with rich morphology or agglutinative structures, where the same meaning can be conveyed through highly variable word forms. This often results in low BLEU scores despite high semantic adequacy. Character-based metrics, such as ChrF (Popović, 2015), offer a better alternative in such settings, as they evaluate overlap at the character level. This makes them better able to handle inflectional and agglutinative variation.

Neural-based evaluation metrics, such as COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020), represent a newer class of quality evaluation tools. Rather than merely examining surface-level text matching, these metrics calculate semantic similarity between the MT output and a reference translation. Neural-based evaluation metrics tend to show high correlations with human judgements (Freitag et al., 2024). However, these neural metrics have a shared limitation in that they perform poorly in low-resource scenarios or with typologically distant languages. Their effectiveness is significantly tied to how well the underlying pre-trained models generalise to these languages and domains and to how closely the data they were fine-tuned on matches the evaluation context (Zouhar et al., 2024; Hu et al., 2020). As a result, correlations with human judgments weaken in such settings, highlighting the continued importance of robust manual evaluation in low-resource MT settings.

Human evaluation remains the most reliable method of assessing translation quality. Common approaches include direct assessment via rating scales, e.g. the Likert scale (Graham et al., 2013), and pairwise ranking of translations (Bojar et al., 2016). Direct assessment provides absolute scores based on defined quality criteria, making it especially useful for evaluating how well an MT system meets user needs. Ranking performs relative comparisons within a set, which limits cross-language comparability.

To enable meaningful cross-language comparisons, Licht et al. (2022) adopted the cross-lingual semantic text similarity (XSTS) method used by Agirre et al. (2012), which prioritises meaning over fluency. They also introduced a calibration procedure to normalise the assessment scores of different annotators using a separate evaluation set, ensuring comparability across languages and evaluators. This approach was also used in the large-scale evaluation of Meta's No Language Left Behind project (NLLB Team et al., 2022).

3. Selection of the most suitable MT system

To determine the most suitable MT system for the application we have in mind, we applied the following selection criteria:

- Compatibility with the three chosen source languages and the 14 target languages.
- Ease of integration into the web platform, with no need for local deployment or maintenance of models, software or servers.
- Translation quality, as evaluated through automatic quality assessment tools.
- Adaptability to the specific domain, including support for integrating a context-specific translation memory or terminology, with minimal effort required for customisation.

At the outset of the MaTIAS project, four NMT systems were considered: Google Translate, Microsoft Translator, ModernMT, and No Language Left Behind (NLLB). All four systems support the selected target languages. However, the technical implementation of NLLB was ultimately deemed unfeasible, as it required hosting the translation models on local servers, an impractical solution in our context. As a result, the subsequent evaluation focused on Google Translate, Microsoft Translator, and ModernMT.

Each of these three commercial systems offers some degree of domain adaptation. However, Google Translate only supports offline customisation through model retraining, which is not feasible within the context of our project. Microsoft Translator offers customisation capabilities, but not for all 14 target languages. In contrast, ModernMT provides a user-friendly, flexible adaptation process involving the upload of a translation memory (a database containing source sentences paired with their corresponding translations).

We assessed the translation performance of Google Translate, Microsoft Translator, ModernMT, and ModernMT with translation memory (henceforth referred to as ModernMT-Mem) by means of established automatic evaluation metrics that compare MT outputs with professional reference translations. These evaluations aimed to (1) evaluate the effectiveness and suitability of widely used automatic evaluation metrics for low-resource languages, (2) compare the translation quality of the different MT systems across the 14 target languages, and (3) examine the impact of the source language (English, French or Dutch).

A set of 200 Dutch messages was selected from the inventory compiled during the preparatory ethnographic phase of the project. The selection was based on criteria of variation, frequency, and length. These messages were then manually translated into English and French by staff at Ghent University. To ensure consistency and quality across all 14 target languages, we commissioned professional translation agencies (ElaN and Untranslate) to produce the reference translations for the resulting English set of 6,711 words.

This dataset was subsequently split into two subsets. All the even-numbered messages (3,193 English words) and their translations were stored in a context-specific translation memory, which was uploaded to ModernMT to adapt the system to Fedasil's context. The odd-numbered messages (3,518 English words) were reserved for evaluating the quality of the translations produced by the four MT systems: ModernMT, ModernMT-Mem, Google Translate, and Microsoft Translator.

To assess the quality of the automatic translations, we considered all the automatic evaluation metrics available in MATEO (Vanroy et al., 2023). However, neural metrics were excluded early in the process due to limited language coverage and insufficient validation for several target languages. We also encountered issues with tokenisation during the preprocessing stage, particularly with word-based metrics such as BLEU and TER in languages like Tigrinya, where Ge'ez punctuation is not handled properly, as illustrated in Figure 2. Consequently, the character-based metric ChrF proved to be the most robust and reliable option across all target languages.

REF: ኣብቲ ማእከል እንስሳ ዘቤት ኣይፍቀድን እዩ።

MT: ኣብቲ ማእከል እንስሳ ዘቤት ኣይፍቀድን እዩ ።

Figure 2: Example of a tokenisation problem in Tigrinya

As illustrated in Figure 3, the results demonstrate that ModernMT-Mem generally outperformed its baseline version without memory, except in the case of Georgian. Notably, the customised version of ModernMT achieved the highest ChrF scores for nine out of the 14 language pairs. Google Translate excelled in three language pairs, while Microsoft Translator led in one. Additionally, Figure 3 reveals considerable variation in translation quality across the target languages, with Persian (fa) obtaining lower ChrF scores and Tigrinya (ti) obtaining considerably lower ChrF scores.

These findings demonstrate that integrating the translation memory into ModernMT improved translation quality. Due to its adaptability and robust performance, ModernMT was chosen for integration into the MaTIAS project. Its capacity to efficiently incorporate domain-specific translations is well aligned with the project's objectives.

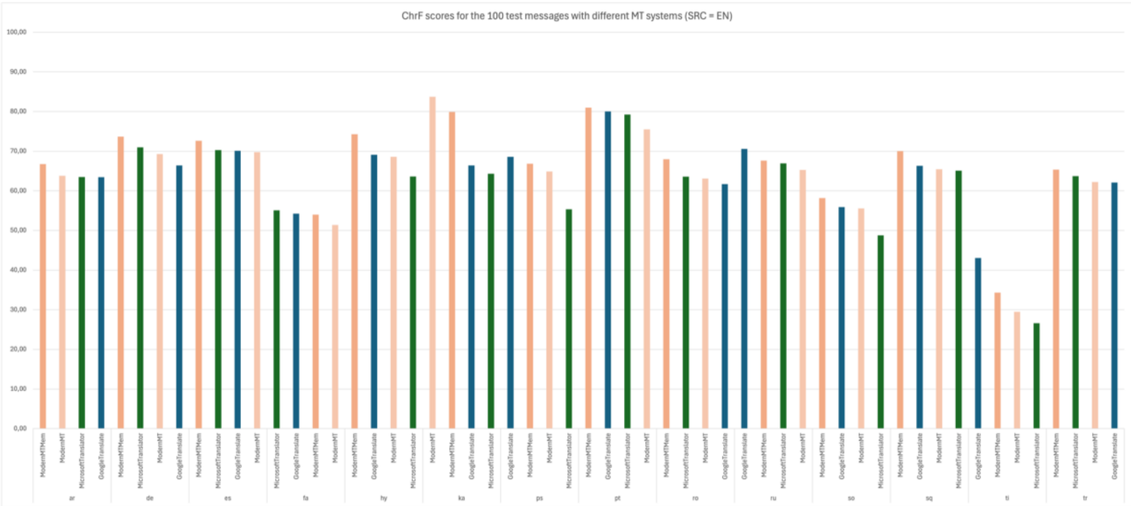

Figure 3: ChrF scores for the 100 messages in the evaluation set translated from English by four MT systems (ModernMT-Mem = orange; ModernMT = light orange; Microsoft Translator = green; Google Translate = blue)

We further investigated the impact of the source language (English versus French and Dutch) in the customised version of ModernMT. Figure 4 displays the ChrF scores for translations generated using ModernMT-Mem. The results clearly illustrate the influence of the source language, with translations from English consistently achieving higher ChrF scores compared to those from French or Dutch.

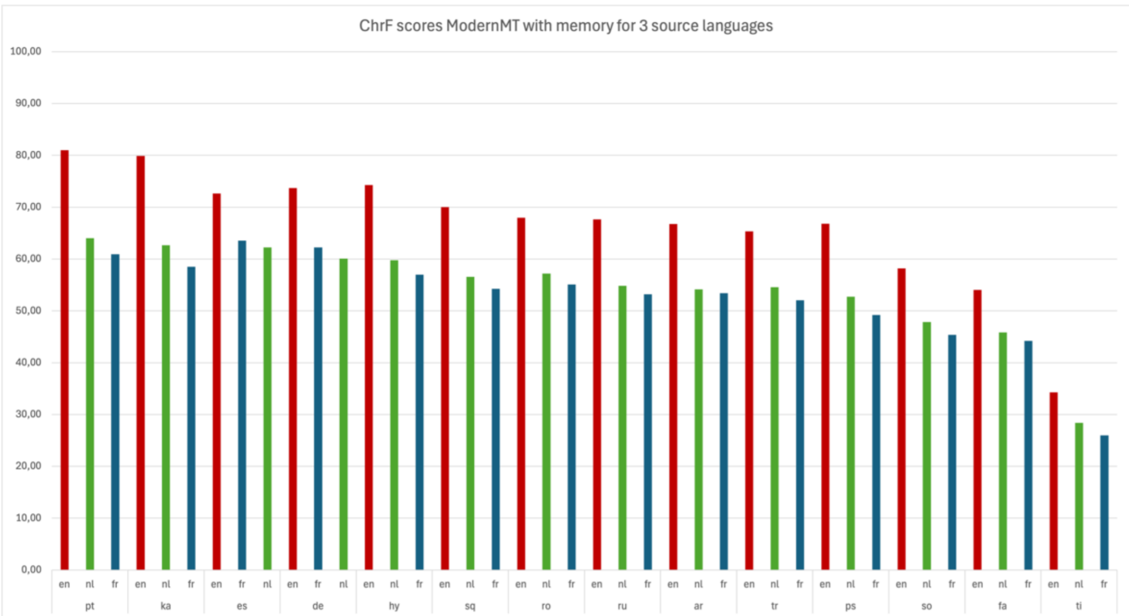

Figure 4: ChrF scores for the 100 messages in the evaluation set translated from English, Dutch, and French into the 14 target languages, sorted by ChrF score

4. Human assessment of the usefulness of the machine-translated messages

While automatic evaluation metrics were initially used to compare the translation quality of three different systems — Google Translate, Microsoft Translator, and ModernMT — a manual evaluation was carried out to reliably assess the usefulness of the automatically translated messages in the 14 target languages, with a customised version of ModernMT with translation memory as the selected MT system. Given that the conveyance of meaning is more critical than fluency in our use case, we adopted the XSTS method described by Licht et al. (2022).

4.1. Assessment method and design

In order to keep human evaluation manageable, we limited manual assessment to the two main source languages: Dutch and French. The actual test set consists of 90 messages extracted from the test set of 100 odd-numbered messages used in the automatic evaluation. Messages that were too long (more than five lines) were shortened to four or five lines, depending on the context. From the original test set, we removed eight messages because they contained Fedasil-specific terms that are difficult for outsiders to understand (e.g. 'orange card'). Two messages were reserved for inclusion as attention checks. We used a custom script to partition the remaining 90 messages equally into two subsets (one with translations from Dutch and one with translations from French), ensuring equal variation in quality (measured by ChrF) in both subsets across all 14 target languages.

As it is difficult to find suitable bilingual evaluators with high proficiency in Dutch, French, and one of the 14 target languages, we decided to present the human-translated English message next to the automatically translated message in the evaluation procedure. To make it easier for the evaluators to compare messages consisting of several sentences, we made sure that the messages were displayed sentence by sentence (see Figure 5 for an example). The evaluators used the following five-point scale to score the preservation of meaning in the messages during MT:

1. The messages are not equivalent, share very few details, and may be about different topics.
2. The messages share some details but are not equivalent. Some important information differs or is missing.
3. The messages are mostly equivalent but some unimportant details differ.
4. The messages are paraphrases of each other. They are equivalent in meaning but not in expression.
5. The messages are exactly and completely equivalent in meaning and expression.

Message 1	Message 2
Dear resident,	ساکنین گرامی،
you will soon have an appointment with 'Kind en Gezin' (Flemish child welfare agency).	به زودی قرار ملاقاتی با Kind en Gezin خواهید داشت.
Kind en Gezin is about half an hour's walk from the centre.	حدود نیم ساعت پیاده روی از مرکز فاصله دارد.
The address is Caverstraat 16, Jabbeke.	آدرس Caverstraat 16, Jabbeke است.
Click here for directions.	برای دیدن مسیرها اینجا کلیک کنید.

Figure 5: Example of an English-Persian message used in the evaluation procedure

In order to obtain meaningful human ratings comparable across language pairs, we adopted the approach of Licht et al. (2022) and constructed a calibration set of 20 English-English messages, pairing the English human reference with an English MT version translated from one of the target language human-produced reference translations. Prior to evaluating 90 messages within their assigned language pair, evaluators first scored the calibration set of 20 messages. This calibration set is used to determine each evaluator's average scoring tendency. If an evaluator's average score for the calibration set is higher than the consensus score, it indicates a lenient scoring pattern; if it is lower, it suggests a stricter approach. To correct for these biases, we apply a linear adjustment. For example, if the consensus score for the calibration set is 3.8 and an evaluator gives an average of 4.1, we subtract 0.3 from all of their scores in the main evaluation task.

To create the calibration set, 45 messages were selected from the context-specific translation memory and translated into English using the standard ModernMT system. To introduce variation in translation quality, different target languages were used as input, including German, Arabic, and Georgian. These 45 English translations were then evaluated by 20 language experts with diverse linguistic backgrounds (e.g. Mandarin, Dutch, French, Greek, Polish, Serbian). From this initial pool, 20 messages were selected based on their quality range and the level of agreement among evaluators. The consensus score for the calibration set was then defined as the average score assigned to these 20 messages by the panel of 20 language experts.

The manual evaluation procedure was set up in Qualtrics and was structured as follows:

- **A welcome page** explaining the evaluation task and allowing evaluators to choose their target language. The evaluators were not explicitly told that they had to evaluate the MT output.
- **Informed consent form** for asking for explicit consent to participate in the study.
- **Commitment request** for asking the evaluators to promise to provide thoughtful responses.
- **Background questionnaire** for collecting information about the evaluators' language skills, education, and experience in a language-related profession.
- **Instructions and guidelines** for explaining the scoring scale using English-English examples. In the subsequent assessment, the guidelines could be easily displayed by clicking on a button.
- **Calibration task** involving scoring 20 English-English messages; messages were randomised in Qualtrics.
- **Evaluation task** involving scoring 90 bilingual messages (English and selected target language; 45 messages were automatically translated from French and 45 from Dutch); messages were randomised in Qualtrics.

We implemented some quality measures in Qualtrics to improve evaluator engagement. These include a five-second delay before the button for proceeding to the next message appeared, to prevent mindless clicking, and the inclusion of two attention-check messages (i.e. messages containing the human translation instead of the automatic translation in the target language).

4.2. Participants

A total of 28 evaluators participated in the actual manual evaluation, with two evaluators assigned to each language pair. Evaluators were recruited through various channels, including:

- The Register of Sworn Interpreters and Translators in Belgium
- Training centres for social interpreters and translators
- A social interpreting and translation agency
- Translation-related social media groups (GentVertaalt and Vertalerskoffiehoek)
- LinkedIn posts
- Contacts within Ghent University

All the evaluators have a high level of English, with four at upper intermediate, 14 at advanced, and 10 at native-like level. The majority (18 out of 28) have a degree in languages. All evaluators were remunerated for their participation in the evaluation task, which was expected to take 90 minutes. External evaluators received €90 (excluding VAT), while university staff received €25.

4.3. Results

All but one of the evaluators gave the attention-check messages a score of 4 or 5. For the 90 messages in the evaluation set, the median number of seconds an evaluator spent on a message was 37.5. If we include the 20 English-English messages that made up the calibration set, the median number of seconds per message rises to 40.5, indicating that it took some time for the evaluators to familiarise themselves with the scoring scale.

To assess the reliability of the human evaluations, we calculated inter-annotator agreement between the two evaluators assigned to each of the 14 language pairs. To do so, we used the intra-class correlation coefficient (ICC), specifically ICC(2,1C), as implemented in the irr R package (version 0.84.1). The ICC was computed on bias-adjusted (shifted) scores, which correct for individual evaluator tendencies identified through a 20-item English-English calibration set. The ICC is appropriate for interval-level data and continuous score comparisons, particularly when evaluating consistency between raters. This makes it especially suitable for our setting, as we applied a linear shift to adjust for evaluator bias based on the calibration set, transforming the original ordinal ratings into continuous ratings. We used the consistency-based version (C), which evaluates whether raters are consistent in their relative rankings, while ignoring systematic differences in absolute scoring levels.

Table 1 presents the ICC scores by target language. The ICC scores range from 0.06 to 0.64. These indicate varying levels of agreement, from slight to moderate, depending on the language pair. ICC values above 0.5 are generally interpreted as indicating moderate reliability, with values below 0.4 indicating poor reliability. Languages such as Persian (fa), Tigrinya (ti), Turkish (tr), Russian (ru), and Armenian (hy) show high inter-annotator consistency (ICC = 0.64, 0.64, 0.58, 0.54, and 0.50 respectively), while Spanish (es) and German (de) exhibit lower agreement (ICC = 0.06 and 0.29 respectively). The particularly low ICC scores for Spanish appear to be influenced by the first kappa paradox, a phenomenon where agreement scores are deflated despite high raw agreement due to skewed score distributions.

The lower ICC scores for several languages indicate that translation evaluation is inherently subjective. Variation in scores across languages may result from differences in evaluator interpretation, translation ambiguity, or linguistic complexity. When reporting results across languages in the following paragraphs, we use the average scores of both evaluators.

Language	ar	de	es	fa	hy	ka	ps	pt	ro	ru	so	sq	ti	tr
ICC	0.31	0.29	0.06	0.64	0.50	0.35	0.51	0.32	0.44	0.54	0.30	0.38	0.64	0.58

Table 1: Inter-annotator agreement by target language, calculated via ICC(2,1C)

Table 2 presents the mean shifted scores for each target language. The first row shows the results for all 90 messages, the second for only the 45 messages that were translated from French, and the third for only the 45 messages translated from Dutch. Notably, Tigrinya (ti) and Somali (so) have the lowest mean scores (2.73 and 2.96 respectively), while all other languages have mean scores above 3.4. Arabic (ar) and Spanish (es) have the highest mean scores (4.14 and 4.00 respectively). For nine out of the 14 languages, the mean values are slightly higher for translations from French than for translations from Dutch.

	ar	de	es	fa	hy	ka	ps	pt	ro	ru	so	sq	ti	tr
Src = All	4.14 (1.15)	3.40 (1.11)	4.00 (0.87)	3.33 (1.35)	3.45 (1.17)	3.47 (1.25)	3.67 (1.09)	3.80 (1.31)	3.96 (1.07)	3.69 (1.08)	2.96 (1.31)	3.98 (1.15)	2.73 (1.18)	3.84 (1.05)
Src = Fr	4.21 (1.18)	3.56 (1.02)	4.05 (0.83)	3.54 (1.24)	3.56 (1.18)	3.45 (1.2)	3.63 (1.11)	3.99 (1.23)	4.16 (0.98)	3.91 (1.06)	2.88 (1.37)	3.96 (1.2)	2.66 (1.09)	3.99 (1.04)
Src= Nl	4.07 (1.13)	3.24 (1.17)	3.95 (0.91)	3.12 (1.42)	3.34 (1.17)	3.49 (1.31)	3.72 (1.08)	3.62 (1.36)	3.75 (1.13)	3.46 (1.05)	3.04 (1.26)	4.01 (1.11)	2.81 (1.25)	3.69 (1.05)

Table 2: Mean shifted scores by target language for all 90 messages, for the 45 translated from French, and for the 45 translated from Dutch; standard deviation in brackets

Figure 6 presents another view of the data and visualises the averaged shifted scores for the 90 messages across all 14 target languages. Please note that the linear shift applied to individual evaluators can be either negative or positive, resulting in shifted scores that may fall below 1 or exceed 5. Scores above 3 (shown in green) indicate that the translated messages are useful for the intended application because they preserve sufficient meaning. In contrast, translations with scores below 2 are considered problematic. As can be seen from the figure, the quality of translations from Tigrinya (ti) and Somali (so) is likely too low to be reliable for our purposes. Only 29% of Tigrinya messages and 40% of Somali messages received an averaged shifted score higher than 3. Translation quality for Persian (fa) also appears to be suboptimal, although 57% of messages received an average shifted score higher than 3.

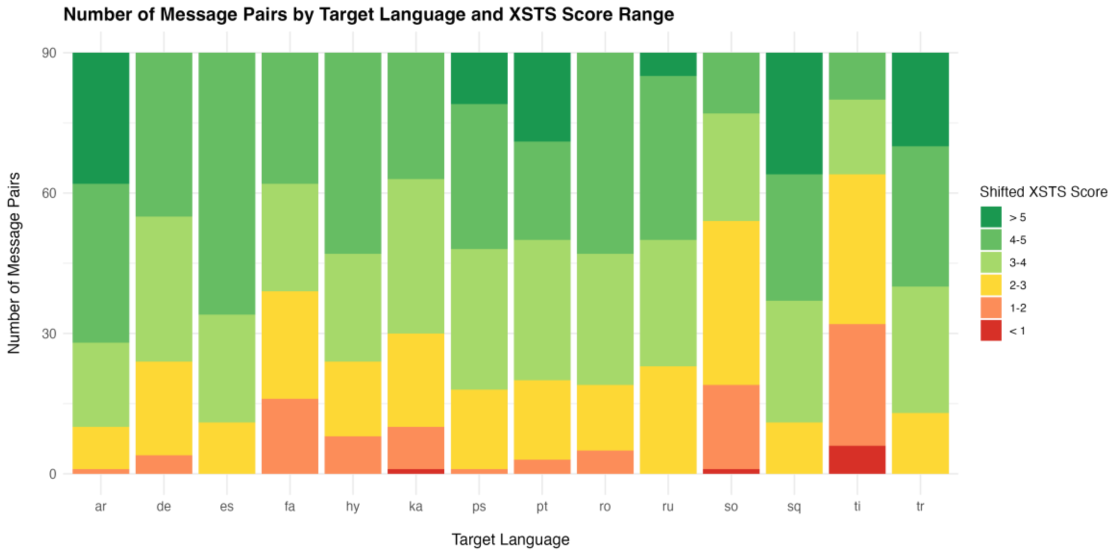

Figure 6: Stacked bar plots with averaged shifted scores by target language

In order to gain insight into the reasons behind the poor quality of the translations, we selected 15 messages that obtained low overall scores and asked five evaluators, who spoke different target languages, to comment on the quality of the translations.

There are recurring issues that suggest pivot translation through English, particularly for certain language pairs. Below are two examples illustrating this point:

1. **Fr**: La douche **est en panne** dans notre immeuble. Vous pouvez donc utiliser les douches du bloc 26 jusqu'à ce que nos douches soient réparées.
En: The shower **is down** in our building. So you can use the showers in block 26 until our showers are repaired.
2. **Nl**: We hebben gemerkt dat sommige bewoners niet in hun eigen kamer slapen. We herinneren u er graag aan dat u moet slapen in het bed en de kamer die u werden toegewezen. **Veranderen van kamer** wordt enkel toegelaten om uitzonderlijke medische redenen, na akkoord van de directie.
En: We've noticed that some residents don't sleep in their own room. We would like to remind you that you must sleep in the bed and room that have been assigned to you. **Changing rooms** is only permitted for exceptional medical reasons, with the approval of the management.

The original messages were written in either French or Dutch. The English version, provided for reference, was translated by the authors using ModernMT. In the first example, the English automatic translation of 'est en panne' is 'is down', which is ambiguous in English as it can mean either 'is not functioning' or 'is located'. According to the Turkish, Persian, and Romanian evaluators, however, the MT reflected the 'location' meaning, which did not correspond with the original French message. The same issue arises in the second example. The English automatic translation of 'Veranderen van kamer' is 'Changing rooms', which has apparently been interpreted as a noun phrase by the Turkish MT model, resulting in an incorrect translation of 'Dressing rooms'.

The following two examples show Dutch compounds that cause translation problems.

1. Opgelet: De **refteruren** zijn gewijzigd. (En: Please note that the canteen opening times have changed.)
2. Je kind draagt een **fluohesje** onderweg naar school. (En: Your child must wear a fluorescent jacket to and from school.)

The compounds 'refter+uren' (En: canteen opening times) and 'fluo+hesje' (En: fluorescent jacket) were incorrectly translated into multiple target languages. In the first example, the second part of the Dutch compound ('uren') was omitted in the translation, altering the meaning to 'the canteen has been changed'. In the second example, the second part of the compound ('hesje', an uncommon word in Dutch) resembles the more common word 'flesje' (En: bottle), leading to the incorrect translation 'fluorescent bottle'. This latter error can be attributed to MT systems often translating at the sub-word level for uncommon words.

Several evaluators provided spontaneous feedback after participating in the task. The evaluators of Romanian and Persian translations highlighted gender bias in the machine-translated texts. Those evaluating Somali, Albanian, and Tigrinya commented on the use of regional language variants, raising concerns about dialectal preferences. For Somali, the translations seemed to favour Northern Somali; standard Somali terminology would have been more appropriate. For Albanian, the translations seemed to favour the Tosk dialect (spoken mainly in southern Albania); this can cause misunderstandings, especially among Albanian speakers from Kosovo or North Macedonia (Gheg dialect). For Tigrinya, the translations seemed to favour Tigrinya from the Tigray region of Ethiopia, which has inherited many words from Amharic, another official language of Ethiopia.

5. Linking human ratings to automatic metrics

For the manual evaluation, we employed 28 language experts to assess the translation quality of 90 messages (totalling 2,316 English words) in 14 target languages. This valuable dataset allows us to explore how well automatic evaluation metrics, such as ChrF and COMET, predict human ratings across languages. We fitted a series of linear mixed-effects (LME) models using the bias-adjusted (shifted) XSTS scores as the dependent variable and tested two widely used automatic evaluation metrics (ChrF and COMET) as predictors, through a stepwise modelling approach that incorporated additional predictors such as message length, target language, and source language.

All models were constructed using the lmer function from the lmerTest package (Kuznetsova et al., 2017) in R. Random intercepts were included for message ID, target message ID, and evaluator ID to account for repeated measures and rater variability. To assess how much of the variance in the models we built is explained by the fixed effects only and the fixed and random effects together, we computed the marginal and conditional R^2 values with the r.squaredGLMM function in the MuMIn package (Bartoń, 2023). To assess the relative importance of the individual fixed effects in these models, we computed semi-partial R^2 values with the r2beta function in the r2glmm package (Jaeger et al., 2017). The alpha level for statistical significance was set at .05 for all models. We used residual plots and Q-Q plots to verify the assumptions of linearity, homoscedasticity, and normality.

For the ChrF-based models (see Table 3), we observed that ChrF was a statistically significant predictor across all models ($p < .001$). However, its unique contribution to explained variance remained modest, with a semi-partial R^2 ranging from 0.036 in the simplest model to 0.022 in the full model that included language variables. The addition of message length improved model fit (R^2_m from 0.036 to 0.087), indicating that longer messages tended to receive lower ratings. Target language contributed further, with a R^2_m of 0.171. Significant negative effects were found for German, Somali, and Tigrinya, indicating that the MT output for these languages was rated significantly lower by human evaluators. No significant effects were found for source language as a predictor.

Model comparison using AICc indicated that the model including ChrF and message length provided the best fit to the data, with the lowest AICc (7219.5). Although the inclusion of target language improved the marginal R^2, this improvement appears to be driven primarily by a subset of languages (German, Somali, and Tigrinya) which showed significantly lower human ratings. This suggests that the target language variable captures language-specific deviations in translation quality rather than contributing uniformly across all languages. Consequently, while the model fit improved modestly, the added complexity may not be justified for languages where MT performance aligns more closely with the average.

Predictors	R^2_m	R^2c	Semi-partial R^2 (ChrF)	Target language significant?	AICc (weights)
ChrF	0.036	0.479	0.036	–	7237.6 (<.001)
ChrF, message length	0.087	0.486	0.034	–	7219.5 (0.857)
ChrF, message length, target language	0.171	0.517	0.022	de, so, ti	7224.1 (0.086)
ChrF, message length, target language, source language	0.172	0.518	0.022	de, so, ti	7224.9 (0.056)

Table 3: Results of the LME model predicting human ratings from ChrF scores and additional predictors (marginal and conditional R^2 values (R^2_m and R^2c), semi-partial R^2 for ChrF, significance of the target language factor, and AIC scores)

We replicated the modelling pipeline using COMET instead of ChrF. However, we excluded the Tigrinya assessments, since the pre-trained language model used to train COMET does not include Tigrinya. The results are presented in Table 4. The model with COMET as sole predictor failed to converge. COMET was a statistically significant predictor in all other models ($p < .001$). The results largely resemble the results obtained for ChrF.

The unique contribution of COMET to explained variance remained modest, with a semi-partial R² ranging from 0.058 in the simplest model to 0.043 in the full model that included language variables. Target language contributed further (R^2_m = 0.171). Significant negative effects for three target languages (German, Somali, and Farsi) were found, indicating that the MT output for these languages was rated significantly lower by human evaluators. No significant effects were found in the model including the source language. Again, model comparison using AICc indicated that the model including COMET and message length provided the best fit to the data, with the lowest AICc (6638.1).

Predictors	R^2_m	R^2c	Semi-partial R² (COMET)	Target language significant?	AICc (weights)
COMET, message length	0.116	0.454	0.058	–	6638.1 (0.989)
COMET, message length, target language	0.169	0.481	0.045	de, so, fa	6647.8 (0.008)
COMET, message length, target language, source language	0.170	0.482	0.043	de, so, fa	6649.4 (0.004)

Table 4: Results of the LME model predicting human ratings from COMET scores and additional predictors (marginal and conditional R² values (R^2_m and R^2c), semi-partial R² for COMET, significance of the target language factor, and AIC scores)

6. Discussion

Our study highlights the potential and limitations of MT systems for multilingual communication in asylum reception centres. Integrating a context-specific translation memory into ModernMT improved translation quality for most language pairs, as evidenced by higher ChrF scores. This suggests that customisation for a specific domain is crucial for achieving accurate, contextually appropriate translations in real-world applications.

However, the results also reveal persistent challenges. One of the primary challenges was the variability in translation quality across different target languages. Translation quality for languages such as Tigrinya, Somali, and Persian remained suboptimal, as demonstrated by automatic and manual evaluation scores. These findings reflect broader issues in MT, where low-resource and typologically distant languages continue to suffer from limited training data and inadequate model generalisation.

A manual evaluation was conducted using the XSTS methodology to gain insight into the perceived usefulness of the translations. While most languages achieved scores indicating acceptable preservation of meaning, variability in inter-annotator agreement (ICC scores) highlights the subjective nature of translation quality assessment. Although the calibration procedure mitigates evaluator bias to some extent, the relatively low level of agreement for certain languages, such as German and Somali, highlights the need for further refinement of evaluation protocols or the involvement of a larger, more diverse group of evaluators.

From a practical point of view, the relatively high scores for most languages indicate that MT can support operational needs in asylum settings for most target languages. However, for critical information delivery in languages with low scores, alternative strategies remain essential. Spontaneous feedback from evaluators also revealed that MT output often reflects regional language variants, which may not align with the linguistic reality of target user groups.

An important finding of our evaluation is that messages translated from English achieved consistently higher quality scores than those translated from Dutch or French. This may be explained by the greater availability of parallel data and better representation of English in the training corpora of most commercial MT systems, including ModernMT. Consequently, translation models are generally more robust when English is the source language, particularly in low-resource settings. This has practical implications for multilingual communication strategies: where feasible, drafting source messages in English can improve the quality and comprehensibility of translations into a broader range of target languages.

LME models confirmed that automatic metrics such as ChrF and COMET are statistically significant predictors of human ratings, but their explanatory power remains modest. This underlines the importance of human evaluation, particularly in low-resource settings where automatic metrics may not fully capture translation quality. Interestingly, message length emerged as a consistent predictor of lower human ratings, suggesting that longer messages may introduce more opportunities for translation errors or ambiguities. Additionally, examples indicate that some translation issues arise from pivoting through English, which may degrade output quality. Finally, while automatic metrics remain essential for MT development and benchmarking, our results reaffirm the indispensable role of human evaluation, particularly for low-resource languages. Future work will focus on integrating user feedback loops into the prototype. This will enable us to improve MT performance further by extending the translation memory with additional manual or post-edited translations.

To conclude, we want to underscore that successful implementation of MT tools in reception centres depends not only on the technical performance of the system but also on users' understanding of how MT works and how it can be applied effectively in different communicative situations. Residents and staff must be aware of both the potential and limitations of MT to use the tool appropriately; for instance, knowing which types of messages, such as simple instructions or routine notices, can be reliably translated, and which, such as legal or healthcare consultations, require human support. Research shows that without such awareness, users may either over-rely on MT, leading to misunderstandings, or underuse it, missing opportunities for effective communication (Martindale & Carpuat, 2018; Deng et al., 2022). Within the MaTIAS project, these insights have been translated into a deliberate investment in training and awareness-raising initiatives for both staff and residents (Macken et al. 2025). These initiatives are designed to embed practical MT skills and knowledge directly into daily routines, enabling staff and residents to recognise the types of messages that MT can reliably support, understand its limitations, and use it in combination with other language support strategies. By integrating these practices into everyday workflows, the project aims to foster greater user trust and engagement, ensuring that the MT tool is not only technically available but also effectively and confidently employed to support communication across languages in the complex, multilingual environment of reception centres.

7. Author contributions

Conceptualization: LM, AT and MF; Data curation: MF; Formal analysis: MF and LM; Funding acquisition: JWD, KM and LM; Investigation: MF; Methodology: LM, AT and MF; Project administration: JDW & EvH; Resources: MF, LM and EvH; Software: MF and LM; Supervision: LM and AT; Validation: MF; Visualization: MF; Writing - original draft: LM, AT, EvH; Writing - review & editing: LM and KM.

8. Funding

This project is co-financed by the European Commission under the Asylum, Migration and Integration Fund (AMIF 093-133).

9. References

Agirre, Eneko; Cer, Daniel; Diab, Mona; Gonzalez-Agirre, Aitor (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: Eneko Agirre; Johan Bos; Mona Diab; Suresh Manandhar; Yuval Marton; Deniz Yuret (eds.). **SEM 2012: The First Joint Conference on Lexical and Computational Semantics Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012).* Association for Computational Linguistics, pp. 385–393. <https://aclanthology.org/S12-1051/>. [Accessed: 20251217].

Aharoni, Roee; Johnson, Melvin; Firat, Orhan (2019). Massively Multilingual Neural Machine Translation. In: Jill Burstein; Christy Doran; Thamar Solorio (eds.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, pp. 3874–3884, <https://aclanthology.org/N19-1388/>. [Accessed: 20251217].

Bartoń, Kamil (2023). *MuMIn: Multi-model inference*. <https://cran.r-project.org/package=MuMIn>. [Accessed: 20251217].

Bergunde, Annika; Pöllabauer, Sonja (2019). Curricular design and implementation of a training course for interpreters in an asylum context. *Translation & Interpreting*, v. 11, n. 1, pp. 1–21. <https://www.trans-int.org/index.php/transint/article/view/560/325>. [Accessed: 20251217].

Bojar, Ondřej; Chatterjee, Rajen; Federmann, Christian; Graham, Yvette; Haddow, Barry; Huck, Matthias; Jimeno Yepes, Antonio; Koehn, Philipp; Logacheva, Varvara; Monz, Christof; Negri, Matteo, Névéol, Aurélie; Neves, Mariana, Popel, Martin; Post, Matt; Rubino, Raphael; Scarton, Carolina; Specia, Lucia; Turchi, Marco; Verspoor, Karin; Zampieri, Marcos (2016). Findings of the 2016 Conference on Machine Translation. In: Ondřej Bojar; Christian Buck; Rajen Chatterjee; Christian Federmann; Liane Guillou; Barry Haddow; Matthias Huck; *et al.* (eds.). *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers.* Association for Computational Linguistics, pp. 131–198 <https://doi.org/10.18653/v1/W16-2301> [Accessed: 20251217].

Ciribuco, Andrea (2020). Translating the village: Translation as part of the everyday lives of asylum seekers in Italy. *Translation Spaces*, v. 9, n. 2, pp. 179–201. <https://doi.org/10.1075/ts.20002.cir>. [Accessed: 20251217].

Demmans Epp, Carrie (2017). Migrants and Mobile Technology Use: Gaps in the Support Provided by Current Tools. *Journal of Interactive Media in Education*, n. 1. <https://jime.open.ac.uk/articles/jime.432>. [Accessed: 20251217].

Deng, Wesley Hanwen; Mehandru, Nikita; Robertson, Samantha; Salehi, Niloufar (2022). *Beyond General Purpose Machine Translation: The Need for Context-specific Empirical Research to Design for Appropriate User Trust*. arXiv:2205.06920v1. <https://doi.org/10.48550/arXiv.2205.06920>. [Accessed: 20251217].

Fan, Angela; Bhosale, Shruti; Schwenk, Holger; Ma, Zhiyi; El-Kishky, Ahmed; Goyal, Siddharth; Baines, Mandeep; Celebi, Onur; Wenzek, Guillaume; Chaudhary, Vishrav; Goyal, Naman; Birch, Tom; Liptchinsky, Vitaliy; Edunov, Sergey; Auli, Michael; Joulin, Armand (2021). Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, v. 22, n. 107, pp. 1-48 <https://jmlr.org/papers/volume22/20-1307/20-1307.pdf>. [Accessed: 20251217].

Freitag, Markus; Mathur, Nitika; Deutsch, Daniel; Lo, Chi-Kiu; Avramidis, Eleftherios; Rei, Ricardo; Thompson, Brian; Blain, Frederic; Kocmi, Tom; Wang, Jiayi; Adelani, David Ifeoluwa; Buchicchio, Marianna; Zerva, Chrysoula; Lavie, Alon (2024). Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In: Barry Haddow, Tom Kocmi, Philipp Koehn, Christof Monz (eds.). *Proceedings of the Ninth Conference on Machine Translation, November 2024*. Association for Computational Linguistics, pp. 47–81. <https://aclanthology.org/2024.wmt-1.2/>. [Accessed: 20251217].

Gavioli, Laura; Wadensjö, Cecilia (eds.). (2023). *The Routledge handbook of public service interpreting*. London: Routledge. <https://doi.org/10.4324/9780429298202>. [Accessed: 20251217].

González Figueroa, Laura Andrea; Pöllabauer, Sonja (2025). Interpreter provision and training of interpreters for languages of limited diffusion: an overview of the situation in Austria. *FITISPos International Journal*, v. 12, n. 1, pp. 122–136. <https://doi.org/10.37536/FITISPos-IJ.2025.12.1.390>. [Accessed: 20251217].

Graham, Yvette; Baldwin, Timothy; Moffat, Alistair (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In: Antonio Pareja-Lora; Maria Liakata; Stefanie Dipper (eds.). *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. *Association for Computational Linguistics*, pp. 33-41. <https://aclanthology.org/W13-2305/> [Accessed: 20251217].

Hendy, Amr; Abdelrehim, Mohamed; Sharaf, Amr; Raunak, Vikas; Gabr, Mohamed; Matsushita, Hitokazu; Kim, Young Jin; Afify, Mohamed; Awadalla, Hany Hassan (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. arXiv:2302.09210v1. <https://doi.org/10.48550/arXiv.2302.09210>. [Accessed: 20250519]. [Accessed: 20251217].

Hu, Junjie; Ruder, Sebastian; Siddhant, Aditya; Neubig, Graham; Firat, Orhan; Johnson, Melvin (2020). XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In: Hal Daumé III; Aarti Singh (eds.). *PMLR, Proceedings of Machine Learning Research*, v. 119, pp. 4411-4421. <https://proceedings.mlr.press/v119/hu20b.html>. [Accessed: 20251217].

Jaeger, Byron C.; Edwards, Lloyd J.; Das, Kalyan; Sen, Pranab K. (2017). An R statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, v. 44, n. 6, pp. 1086–1105. <https://doi.org/10.1080/02664763.2016.1193725>. [Accessed: 20251217].

Kivijärvi, Antti; Myllylä, Martta (2022). Layered Confinement in Reception Centers: A Study of Asylum Seekers' Experiences in Finland. *Journal of Immigrant & Refugee Studies*, v. 22, n. 4, pp. 1–14 <https://doi.org/10.1080/15562948.2022.2132341>. [Accessed: 20251217].

Kuznetsova, Alexandra; Brockhoff, Per B.; Christensen, Rune H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, v. 82, n. 12, pp. 1–26. <https://doi.org/10.18637/jss.v082.i13> [Accessed: 20251217].

Licht, Daniel; Gao, Cynthia; Lam, Janice; Guzman, Francisco; Diab, Mona; Koehn, Philipp (2022). Consistent Human Evaluation of Machine Translation across Language Pairs. In: Kevin Duh; Francisco Guzmán (eds.). *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track).* Association for Machine Translation in the Americas, pp. 309–321. <https://aclanthology.org/2022.amta-research.24/>. [Accessed: 20251217].

Liebling, Daniel J.; Lahav, Michal; Evans, Abigail; Donsbach, Aaron; Holbrook, Jess; Smus, Boris; Boran, Lindsey (2020). Unmet Needs and Opportunities for Mobile Translation AI. In: *Proceedings of the 2020 CHI Conference on Human Factors in*

Computing Systems. Association for Computing Machinery, pp. 1-13. <https://doi.org/10.1145/3313831.3376261>. [Accessed: 20251217].

Macken, Lieve; van Hest, Ella; Tezcan, Arda; Lumingu, Michaël; Maryns, Katrijn; De Wilde, July (2025). Machine Translation to Inform Asylum Seekers: Intermediate Findings from the MaTIAS Project. In: Pierrette Bouillon; Johanna Gerlach; Sabrina Girletti; Lise Volkart; Raphael Rubino; Rico Sennrich; Samuel Läubli; Martin Volk; Miquel Esplà-Gomis; Vincent Vandeghinste; Helena Moniz; Sara Szoc (eds.) *Proceedings of Machine Translation Summit XX : Volume 2.* European Association for Machine Translation, pp. 77–78. <https://aclanthology.org/2025.mtsummit-2.10/>. [Accessed: 20251217].

Martindale, Marianna J.; Carpuat, Marine (2018). Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In: Colin Cherry; Graham Neubig (eds.). *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track).* Association for Machine Translation in the Americas, pp. 13–25. <https://aclanthology.org/W18-1803/>. [Accessed: 20251217].

Maryns, Katrijn; Smith-Khan, Laura Allison; Jacobs, Marie (2023). Multilingualism in Asylum and Migration Procedures. In: Carolyn McKinney; Pinky Makoe; Virginia Zavala (eds.). *The Routledge handbook of multilingualism*. 2nd ed. Abingdon, Oxo; New York: Routledge, pp. 394–414.

Microsoft (2025) *Microsoft Translator.* <https://translator.microsoft.com/>.

NLLB Team; Costa-Jussà, Marta R.; Cross, James; Çelebi, Onur; Elbayad, Maha; Heafield, Kenneth; Heffernan, Kevin; Kalbassi, Elahe; Lam, Janice; Licht, Daniel; Maillard, Jean; Sun, Anna; *et al.* (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672v3 <https://doi.org/10.48550/arXiv.2207.04672>. [Accessed: 20251217].

OpenAI (2025) ChatGPT. <https://chat.openai.com/>.

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, pp. 311-318. <https://doi.org/10.3115/1073083.1073135>. [Accessed: 20251217].

Pięta, Hanna; Valdez, Susana (2024). Migration and translation technologies. In: Brigid Maher; Loredana Polezzi; Rita Wilson (eds.). *The Routledge handbook of translation and migration*. Abingdon, Oxon; New York: Routledge, pp. 434-450. <https://doi.org/10.4324/9781003287797-33>. [Accessed: 20251217].

Popović, Maja (2015). chrF: character n-gram F-score for automatic MT evaluation. In: Ondřej Bojar; Rajan Chatterjee; Christian Federmann; Barry Haddow; Chris Hokamp; Matthias Huck; Varvara Logacheva; Pavel Pecina (eds.). *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Association for Computational Linguistics, pp. 392–395 <https://doi.org/10.18653/v1/W15-3049> [Accessed: 20251217].

Rei, Ricardo; Stewart, Craig; Farinha, Ana C.; Lavie, Alon (2020). COMET: A Neural Framework for MT Evaluation. In: Bonnie Webber; Trevor Cohn; Yulan He; Yang Liu (eds.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2685–2702. <https://aclanthology.org/2020.emnlp-main.213/>. [Accessed: 20251217].

Robinson, Nathaniel; Ogayo, Perez; Mortensen, David R.; Neubig, Graham (2023). ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In: Philipp Koehn, Barry Haddow, Tom Kocmi, Christof Monz (eds.). *Proceedings of the Eighth Conference on Machine Translation.* Association for Computational Linguistics, pp. 392–418. <https://aclanthology.org/2023.wmt-1.40/>. [Accessed: 20251217].

Sellam, Thibault; Das, Dipanjan; Parikh, Ankur (2020). BLEURT: Learning Robust Metrics for Text Generation. In: Dan Jurafsky; Joyce Chai; Natalie Schluter; Joel Tetreault (eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, pp. 7881–7892. <https://aclanthology.org/2020.acl-main.704/>. [Accessed: 20251217].

Snover, Matthew; Dorr, Bonnie; Schwartz, Rich; Micciulla, Linnea; and Makhoul, John (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers.* Association for Machine Translation in the Americas, pp. 223–231. <https://aclanthology.org/2006.amta-papers.25/>. [Accessed: 20251217].

Valdez, Susana; Guerberof Arenas, Ana; Ligtenberg, Kars (2023). Migrant communities living in the Netherlands and their use of MT in healthcare settings. In: Mary Nurminen; Judith Brenner; Maarit Koponen; Sirkku Latomaa; Mikhail Mikhailov; Frederike Schierl; Tharindu Ranasinghe; Eva Vanmassenhove; Sergi Alvarez Vidal; Nora Aranberri; Mara Nunziatini; Carla Parra Escartín; Mikel Forcada; Maja Popovic; Carolina Scarton; Helena Moniz (eds.). *Proceedings of the 24th Annual Conference of the European Association for Machine Translation.* European Association for Machine Translation, pp. 325–334. <https://aclanthology.org/2023.eamt-1.32/>. [Accessed: 20251217].

Vanroy, Bram; Tezcan, Arda; Macken, Lieve (2023). MATEO: MAchine Translation Evaluation Online. In: Mary Nurminen; Judith Brenner; Maarit Koponen; Sirkku Latomaa; Mikhail Mikhailov; Frederike Schierl; Tharindu Ranasinghe; Eva Vanmassenhove; Sergi Alvarez Vidal; Nora Aranberri; Mara Nunziatini; Carla Parra Escartín; Mikel Forcada; Maja Popovic; Carolina Scarton; Helena Moniz (eds.). *Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation*, pp. 499–500. <https://aclanthology.org/2023.eamt-1.52/>. [Accessed: 20251217].

Vieira, Lucas N. (2024). Machine translation and migration. In: Brigid Maher; Loredana Polezzi; Rita Wilson (eds.). *The Routledge handbook of translation and migration.* Abingdon, Oxon; New York: Routledge, pp. 221–234. <https://doi.org/10.4324/9781003287797-33>. [Accessed: 20251217].

Xu, Weijia; Agrawal, Sweta; Briakou, Eleftheria; Martindale, Marianna J.; Carpuat, Marine (2023). Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics,* v. 11, pp. 546–564. <https://doi.org/10.1162/tacl_a_00563>. [Accessed: 20251217].

Zhang, Tianyi; Kishore, Varsha; Wu, Felix; Weinberger, Kilian Q.; Artzi, Yoav (2020). *BERTScore: Evaluating Text Generation with BERT*. arXiv:1904.09675v3. <https://doi.org/10.48550/arXiv.1904.09675>. [Accessed: 20251217].

Zouhar, Vilém; Chen, Pinzhen; Lam, Tsz Kin; Moghe, Nikita; Haddow, Barry (2024). Pitfalls and Outlooks in Using COMET. In: Barry Haddow; Tom Kocmi; Philipp Koehn; Christof Monz (eds.). *Proceedings of the Ninth Conference on Machine Translation.* Association for Computational Linguistics, pp. 1272–1288. <https://aclanthology.org/2024.wmt-1.121/>. [Accessed: 20251217].