An experiment in error analysis of real-time speech machine translation using the example of the European Parliament's Innovation Partnership*

Elisa Di Nuovo[1,*]

[1]University of Turin, via Giuseppe Verdi, 8, 10124 Torino (TO) - Italy

Abstract

In recent years, technological progress has made Machine Translation (MT) a reality. Significant improvements have been obtained using deep learning models as opposed to rule-based and statistical MT models. Human evaluation still remains under-explored. In 2019 the European Parliament (EP) started an innovation partnership with commercial operators, with the purpose of developing a tool exploiting state-of-the-art, real-time Automatic Speech Recognition (ASR) and MT technologies to make parliamentary plenary sessions accessible to D/deaf and hard of hearing. In this paper, we present a quantitative and qualitative error analysis carried out on a test set consisting of 78 short speeches delivered by Members of the EP in 19 languages deployed in the EP prototype by November 2022. The taxonomy used for ASR and MT is adapted from the Multidimensional Quality Metrics framework. Results show that sentence segmentation is the biggest issue in the ASR output—not considered using automatic metrics—which often affects the MT output.

Keywords

Real-time speech machine translation, Error analysis, Human evaluation, Cascade system

1. Introduction

In recent years, the landscape of language translation has been fundamentally transformed by remarkable technological advancements. Machine Translation (MT), once an ambitious aspiration, has now become a tangible reality. This transformation has been primarily fueled by the advent of deep learning models, specifically neural machine translation and transformers. These cutting-edge models have ushered in a new era of translation, eclipsing the limitations of traditional rule-based and statistical MT methods. Deep learning models use the mechanism called attention to improve the performance [1] and have been usually evaluated on offline written translation tasks involving a few language pairs [2].

Very recently, research expanded its focus also on speech machine translation, tackled as a concatenation of Automatic Speech Recognition (ASR) and MT, or as an end-to-end task (i.e. direct translation of speech in language A into text in language B).[1] In the last evaluation campaign of the 19th International Conference on Spoken Language Translation (IWSLT 2022) [5], one of the eight shared tasks focused on real-time speech translation, addressed as translation of ASR output or directly from the audio source and involving English to German, English to Japanese and English to Mandarin Chinese. A novelty of this year campaign is the addition of manual evaluation of real-time outputs.

Like many natural language processing tasks, MT is difficult to evaluate. One of the reasons for this is the non-deterministic nature of translation, i.e. there is more than one correct way to translate from one language into another. Evaluation in shared tasks is usually carried out by means of automatic metrics, BLEU (Bilingual Evaluation Understudy) [6] being the standard for MT evaluation. This metric tries to overcome the nondeterministic nature of translation using multiple references. However, automatic metrics have several limitations [7].[2] On the other hand, human evaluation, if carried out using fined-grained guidelines to limit subjectivity, can give a clearer indication of the MT output quality. However, being resource expensive (i.e. it is hard to find skilled evaluators; skilled evaluators have a high cost), it has been used limitedly and in small studies. To avoid these limitations, crowdsourced annotators have been used. Unfortunately, crowdsourced annotators are frequently inexperienced. As [10] affirm, crowdsourced human evaluation can be used when MT quality is poor, because it can still provide

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*This study and paper was written while the author was working for the European Parliament, in the Unit in charge of the administration and evaluation of the prototype. This is not the official evaluation methodology employed by the European Parliament for evaluation.

*The author, as of 1st October 2023 is employed by the Joint Research Centre, European Commission, Ispra (VA), Italy.

elisa.dinuovo@unito.it (E. D. Nuovo)

0000-0002-4814-982X (E. D. Nuovo)

[1]See for example the studies in [3, 4] for a comparison between cascade and end-to-end systems.

[2]See Moorkens et al. [8] on translation quality assessment and Chatzikoumi [9] for a comprehensive review of automatic and human MT evaluation.

a useful indication; but, as quality improves, it becomes unfit and leads to erroneous claims.[3]

In 2019 the European Parliament (EP) started an Innovation Partnership with commercial operators, with the purpose of developing a tool that can perform real-time ASR and MT from and into all the 24 official languages of the European Union (EU).[4] This partnership has the aim of making parliamentary plenary sessions accessible in near-real time to D/deaf and hard of hearing persons.[5] The challenges faced by this project are manifold: the high degree of multilingualism which is highly ambitious considering the technical limits of current MT particularly in a number of low-resource languages, the presence of non-native accents, the large variety of vocabulary/EU jargon required by the numerous specific domains tackled in the plenaries, the low latency constraints to have transcriptions and translations in near-real time, and the required high quality of the output. By November 2022, 19 language models have been developed and made available via a demo interface. These languages are English (EN), French (FR), German (DE), Spanish (ES), Italian (IT), Polish (PL), Greek (EL), Romanian (RO), Dutch (NL), Portuguese (PT), Bulgarian (BG), Czech (CS), Slovak (SK), Croatian (HR), Lithuanian (LT), Finnish (FI), Hungarian (HU), Swedish (SV) and Slovenian (SL).

In this paper, we present a quantitative and qualitative study—using Word Error Rate (WER) metric [14] and manual human evaluation—on a test set consisting of short speeches delivered by members of the EP in the 19 languages already deployed in the prototype.[6] The aim of this study is to evaluate the quality of both ASR and MT output and to reflect on the different insights of the same text given by different annotators. The manual human evaluation is based on an error taxonomy adapted from the Multidimensional Quality Metrics (MQM) framework[7] and applied to part of the test set covering 6 languages (EN, FR, ES, IT, RO, DE). The MQM framework, developed in the EU QTLaunchPad and QT21 projects, provides a hierarchy of translation errors that can be adapted according to the application. We devised our taxonomy consisting of different error categories for ASR and MT and 3 severity levels (i.e. neutral, minor and major). We decided to exclude critical errors as in [10]. The remainder of the paper is organised as follows: in Section 2 we describe the methodology applied for automatic and human evaluation; in Section 3 we report the results quantitatively and qualitatively analysed per language and per annotator; Section 4 concludes the paper.

2. Methodology description

We evaluated ASR using both automatic metrics, in particular WER and human evaluation, and MT only relying on human evaluation. Human evaluation for both ASR and MT is carried out under the MQM framework. We describe the procedures and the experimental setup in the subsequent sections.

2.1. Automatic evaluation

Automatic evaluation was used only for ASR. The metric used is WER.[8] The test set consists of 92 short speeches (minimum = 01:01; maximum = 05:10; average = 01:39; standard deviation = 00:39) delivered in March and May 2022 plenaries by members of the EP. The speeches are in the 19 languages deployed in the tool by November 2022. See Table 1 for more details. Languages are ordered according to deployment in the tool.[9]

The gold standard of this test set is made of the verbatim transcription of the speeches (often referred to by its French abbreviation, CRE, *Compte Rendu d'Evènement*), manually corrected from the published report available on the website of the EP.[10] The corrections are performed by two native speakers per language and a third annotator is involved to solve disagreement.

2.2. Human evaluation

The error taxonomy is germane to the MQM framework and includes different categories for ASR and MT, sharing the same severity scale—i.e. neutral, minor, major (see Figure 1 in Appendix A for the decision tree). The

[3] One (in)famous claim is that MT has achieved human parity [11, 12, 13].

[4] Specifications of the Innovation Partnership are available here: https://etendering.ted.europa.eu/cft/cft-document.html?docId=58722. All links were last access on 13/05/2023.

[5] Deaf with a capital D denotes individuals who are culturally and linguistically Deaf, often due to congenital deafness or early-life hearing loss. They identify with the Deaf community, characterized by its unique culture, sign languages, and traditions. In contrast, deaf (with a lowercase d) is a general term referring to individuals with a hearing impairment, irrespective of their cultural identification or community affiliation. It describes the audiological condition of partial or complete hearing loss, without specifying sign language usage or cultural ties.

[6] This study and paper was written while the author was working for the European Parliament Unit in charge of the prototype management and evaluation. This is not the official evaluation methodology employed by the Parliament to evaluate the prototype.

[7] MQM website available here: https://themqm.org/.

[8] Script written by Dr. Claudio Fantinuoli available here: https://github.com/fantinuoli/WERvisual/blob/main/wer.py.

[9] During Stage 1 of the project 10 language models were deployed; during Stage 2, 9 other language models were added. This order is maintained in Table 1. Stage 1 models were trained during 2020-2021, Stage 2 languages during 2021-2022. Stage 1 ASR models have been updated in August 2022. Both ASR and MT models are developed by Cedat85 consortium.

[10] Each parliamentary sitting is publicly available and the CRE and videos in the original language are available in the EP website: https://www.europarl.europa.eu/plenary/en/debates-video.html.

Language	# speeches	Time (hh:mm:ss)
EN	9	00:14:56
FR	3	00:04:52
DE	4	00:06:28
ES	4	00:05:03
IT	7	00:15:31
RO	4	00:04:37
PL	10	00:19:10
EL	1	00:01:05
NL	1	00:01:21
PT	1	00:01:10
BG	4	00:07:19
CS	4	00:07:58
SK	4	00:05:42
SL	4	00:06:10
HR	4	00:05:59
LT	4	00:06:13
FI	8	00:13:20
HU	9	00:13:03
SV	7	00:12:00
Total	92	02:31:57

Table 1
ASR automatic evaluation test set.

error categories used for ASR error annotation are *over-segmentation*, *under-segmentation*, *lexical substitution*, *lexical deletion*, *lexical addition*, *morpho-syntactic errors* (e.g. number agreement, part of speech substitution), *terminology* (e.g. named entities and terms). The categories used for MT error annotation are *accuracy* (e.g. meaning is not rendered in its entirety), *punctuation*, *grammar*, *register* (formality, gender-marked pronouns), *terminology* (including the presence of non-words, spelling errors or incorrect terms), *other* and *unintelligible*. *Unintelligible* is used to mark segments containing more than 5 major errors [10]. *Other* should be used in rare cases in which none of the existent error categories apply. Neutral errors weight 0 points, minor errors 1, major 5. Except for unintelligible which weights 5, if minor, and 25, if major. These weights are similar to those used in [10].

We involved four annotators. All received the annotation guidelines and a training. After a few annotations a further meeting was scheduled to clear doubts. We involved four annotators with different backgrounds and knowledge of the languages. For reference, we call them annotator A, B, C and D (henceforth, Ann for annotator). Ann A has a background in Translation studies and is an experienced translator at the EP. Ann B was a trainee at the EP with a master's degree in Translation and previous experience on the MQM framework. Ann C was a trainee at the EP with a master's degree in Translation and no previous experience on ASR and MT evaluation. Ann D is a communications assistant at the EP with a background in interpretation, with no experience on the MQM framework, but with experience on ASR and MT evaluation. In Table 2 we report their self-reported knowledge of the languages according to the CEFR levels.[11]

Annotator	Language	CEFR level
Ann A	RO	Native language
	EN	C2
	IT	C2
Ann B	IT	Native language
	RO	Native language
	EN	C2
	ES	C2
	DE	C2
Ann C	IT	Native language
	EN	C1
	FR	C1
	ES	B2
Ann D	IT	Native language
	EN	C2
	FR	C1

Table 2
Annotators and language knowledge.

The annotated test set consists of 48 documents in 6 languages (EN, IT, FR, ES, RO, and DE): 18 automatic transcriptions (3 speeches per language, with an identification number from 1 to 6) and 30 translations (from and into the 6 above mentioned languages). In Table 3 we report the evaluated task and the involved languages, the number of speeches (with an identification number between brackets to be able to identify them when used as source and target and also in the automatic evaluation results reported in Table 4), and the annotators providing the annotations.

3. Results

3.1. Automatic evaluation

ASR was evaluated in two different scenarios: first, in sessions with more than one speech but all in the same language; second, in sessions with more than one speech, each in a different language. This is possible because the tool has a feature called Language Identification (LID), which is used to identify the language spoken and subsequently transcribe the audio in the identified language. WER results (computed per session) are reported in Table 4. In the table body, from row 2 to 7, we report the WER obtained in the speeches also undergoing human evaluation. This is the reason why we have multiple rows for the same language (e.g. EN LID on with id code

[11] For more details about the CEFR levels see the website: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

Evaluated task	# speeches (source id)	Annotations
ASR RO	3 speeches (1)	Ann A – B
ASR IT	3 speeches (2)	Ann B – C
ASR EN	3 speeches (3)	Ann B – C
ASR ES	3 speeches (4)	Ann B – C
ASR FR	3 speeches (5)	Ann C – D
ASR DE	3 speeches (6)	Ann B
MT EN-IT	3 speeches (3)	Ann B – C
MT EN-RO	3 speeches (3)	Ann A – B
MT EN-FR	3 speeches (3)	Ann C – D
MT EN-ES	3 speeches (3)	Ann B
MT EN-DE	3 speeches (3)	Ann B
MT RO-IT	3 speeches (1)	Ann A – B
MT IT-EN	3 speeches (2)	Ann B – C
MT ES-IT	3 speeches (4)	Ann B – C
MT FR-IT	3 speeches (5)	Ann C – D
MT DE-IT	3 speeches (6)	Ann B

Table 3
Human evaluation test set.

3, indicating the speeches subjected to human evaluation, and then again EN LID on, non subjected to human evaluation).

Language (source id)	LID	WER
All 19, 1 speech each	On	6.45
RO (1)	On	2.77
IT (2)	On	3.22
EN (3)	On	8.98
ES (4)	On	4.94
FR (5)	On	8.91
DE (6)	On	7.81
EN	Off	5.25
EN	On	5.48
IT	Off	5.58
BG	Off	5.83
PL	Off	5.05
PL	On	7.80
HU	Off	9.18
HU	On	9.57
CS	Off	4.03
SK	Off	2.52
SL	Off	5.02
HR	Off	5.63
LT	Off	11.14
FI	Off	5.48
SV	Off	10.78

Table 4
ASR: Averaged WER results. Source id in brackets links the speeches with those in Table 3.

The results show that LID does not have a big impact on WER (e.g. EN, HU), except for PL (almost 3% WER difference), but the main difference in WER is due to different speeches (e.g. IT, in which the 3 speeches with LID on have a lower WER than the 3 with LID off, or EN with LID on in different interventions with more than 4% WER difference).[12]

3.2. Human evaluation

We investigated manual annotation quantitatively and qualitatively. Quantitative evaluation is based on an average score per document and annotator. Qualitative evaluation takes error categories and severities into account.

3.2.1. Quantitative evaluation

Same speeches, different annotators. For each annotator, we calculated a score per document by averaging the segment-level scores. Results are shown in Figures 2–6 in Appendix A. In general, ASR output received higher scores than expected, especially in languages in which WER is lower than 5% (i.e. RO, IT, ES). This can be due to the fact that WER metric does not take punctuation into account, thus over- and under-segmentation issues are not counted. Also, in WER calculation all errors have the same weight (e.g. a missing negation changing completely the meaning of the sentence has the same weight of any other missing token). MT output received lower scores if compared to ASR output. This might mean that some errors in ASR are well handled in translation. When both annotators are native speakers of the target language, their scores are more similar. This is the case of Ann A and B in EN-RO (Figure 2), Ann B and C in EN-IT (Figure 4), Ann B and C in ES-IT (Figure 5). The same applies to Ann C and D (Figure 6) in the annotation of FR-IT MT, although Ann D displays a different annotation behaviour than Ann B and C. In fact, Ann D tends to annotate fewer errors. This could be influenced by their different backgrounds (interpreter vs. translator).

The annotation scores are the most similar when the annotators are native speaker of the target language, as in the annotation of IT-EN MT (Figure 3) and EN-FR MT (Figure 6). However, monolingual annotators (Ann A and C) show more severity in MT judgement into their native language (Figure 2 RO-IT MT and Figure 4 EN-IT MT) when compared with our IT-RO bilingual annotator (Ann B). This seems in line with research on bilingualism and acceptability, where results show that "bilinguals do not reject ungrammatical items with the same certainty as monolinguals" [15].

Averaging the scores attributed by the two annotators (ASR and translation into IT using the ASR output as source, except for IT, that is translated into EN), we obtain the following order (from the presumably best output to the worse): ES (average = 17.3), IT (average = 24.5), FR

[12] Please note that this could be due to pure chance and since the test set is small, we do not report statistical tests.

(average = 24.7), EN (average = 25.3), RO (average = 40.8). The order considering WER would be RO, IT, ES, FR, EN.

Same annotator, different speeches. Here we compare the annotations carried out by Ann B and C. We selected these two annotators because they performed the majority of the annotation task, so it is possible to compare their results in different languages. We report their scores in Figure 7–8, respectively (Appendix A).

According to Ann B (Figure 7), we can order the languages from the best output to the worse: ES (average = 14.33), IT (average = 27.00), DE (average = 29.44), EN (average = 36.67) and RO (average = 46.33). According to Ann C (Figure 8), the order would be: ES (average = 13.67), IT (average = 18.50), EN (average = 26.33) and FR (average = 27.50).

3.2.2. Qualitative evaluation

Each annotator draws a different picture of each text, being that the product of ASR or MT. As far as ASR output is concerned, we report the results in Figures 9–13 in Appendix A.

Despite we did not put a major emphasis on over- and under-segmentation errors during training, as they were considered to be straightforward (at least in their identification), the disagreement in annotations suggests the contrary. In fact, different annotators draw opposite pictures of their presence and importance. For example, in Figure 9, we can notice that Ann A weighs more over-segmentation than under-segmentation errors in RO transcription, while Ann B does the opposite. The system results on punctuation marking, and full stop identification, in particular, seems to be below state-of-the-art performance [16]. Ann C (Figures 10–13) seems to be more severe about morpho-syntactic errors in ASR. The same errors are annotated as lexical substitutions by the other annotators, as in Example 1.

(1) **REF:** Protéger les citoyens de la haine en ligne, voici un bel usage **des** technologies les plus avancées. **Et** voici aussi un usage très approprié de l'Union européenne.
"Protecting citizens from online hate, here is a good use **of the** most advanced technologies. **And** here a very appropriate use by the European Union."
ASR: Protéger les citoyens de la haine en ligne. Voici un bel usage**. Les** technologies les plus avancées. **En** voici aussi un usage très approprié de l'Union européenne.
"Protecting citizens from online hate. Here is a good use. The most advanced technologies. Here also a very appropriate use by the European Union."
FR-IT: Proteggere i cittadini dall'odio online. Qui è un buon uso. Le tecnologie più avanzate. Anche questo è un uso molto appropriato da parte dell'Unione europea.
"Protecting citizens from online hate. Here is a good use. The most advanced technologies. This is also a very appropriate use by the European Union."

Ann C marked the errors in bold in Example 1 as morpho-syntactic errors of a major nature, Ann D as lexical substitution of a minor nature. This is a blurry area, if you consider that both are functional words and in other languages could be rendered morphologically. We think that in a multilingual perspective, these should be treated as morphological being functional words. However, probably they are not major errors, as they do not affect a main idea of the speech (decision tree in Figure 1).

As far as MT output is concerned, we report the results in Figures 14– 20 in Appendix A. We notice the inappropriate use of the unintelligible category. Unintelligible should mark segments in which it is impossible to understand the message and to identify all the errors that led to the incomprehensible segment. The fact that on the same set, different annotators used it or not, it is a clear sign of misunderstanding (Figures 15, 17, 18 and 19). In fact, in Example 2, unintelligible is used in a segment in which it is possible to understand the meaning, although there is a minor grammatical error (*attaccano* 'they attack') and a minor accuracy error (relative clause instead of adverbial clause, *che* 'that' substituting *per* 'to').

(2) **REF:** [...] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**, acest** lucru necesită un răspuns **rapid** și unit.
"[...] countries like the Russian Federation use modern tools to attack states, to attack entities, to endanger European democracy, this requires a rapid and united response."
ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**. Acest** lucru necesită un răspuns**. Rapid** și unit, [...]
FR-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**. Ciò** richiede una risposta**. Veloce** e unito,

In Example 2 we also notice over-segmentation errors in the ASR transcription cascading in MT (*Acest lucru necesită un răspuns. Rapid și unit*). In addition, it seems that Ann B, and in other examples Ann C, annotated the output as if it was a written text and not an oral text transposed in written. Thus, the reference text is only one of the possible transpositions. This is evident looking at punctuation. In Example 2, in fact, Ann B not only marked the over-segmentation error dividing

the noun *răspuns* from its modifiers (*rapid și unit*), but also another over-segmentation error (despite marked as minor) because in the reference this sentence is joint to the preceding one with a comma and not divided by a full stop. However, it must be noted that a full stop there is perfectly acceptable.

Unintelligible errors were also marked when the other annotator only noticed punctuation issues, as shown in Example 3.

(3) **REF:** Putin has thrown the world and Europe back to a time we had hoped never to experience again**. A crisis** of such dignity shows our true **colours** – if we are on the right side of history or choose the [path] path of destruction.
ASR: Putin has thrown the world and Europe back to a time, we would hope to never experienced again **crisis** of such dignity shows our true **colours** if we are on the right side of history or choose the path path of destruction.
EN-FR: Poutine a renvoyé le monde et l'Europe à une époque, nous espérons ne plus avoir connu **de crise** de cette dignité montre nos vraies **couleurs** si nous sommes du bon côté de l'histoire ou si nous choisissons le chemin de la destruction.

(4) **REF:** [...] Ceux qui ont harcelé et appelé au meurtre **sur Internet Samuel Paty**, sont-ils, étaient-ils, des vecteurs de liberté d'expression? Poser la question, c'est déjà y apporter une réponse.
"Were those who harassed and called for the murder of Samuel Paty on the Internet vectors of freedom of expression? To ask the question is to answer it."
ASR: Ceux qui ont harcelé. **Su Internet. Internet. Jsem jej petic.** Et appelé au meurtre sur Internet, Samuel Paty. Sont-ils, étaient-ils des vecteurs de liberté d'expression. Poser la question c'est déjà y apporter une réponse.
EN-IT: Coloro che hanno molestato. **Su internet. Internet. Sono una petizione.** E ha chiesto omicidio su Internet, Samuel Paty. Sono loro, erano vettori della libertà di espressione. Fare la domanda è già fornire una risposta.
"Those who harassed. **On the Internet. Internet. They are a petition.** And called for murder on the internet, Samuel Paty. It's them, they were vectors of freedom of expression. To ask the question is already to provide an answer."

An actual unintelligible error is instead reported in Example 4. LID errors in this case caused unintelligibility in the translation because the same portions of audio were transcribed in different languages (transcribed as IT, PL, and CS). Perhaps including the information about the source language in the translated output could be useful to reduce the impact that LID errors like these have in the MT understanding. Morpho-syntactic errors annotated in the ASR are frequently correct in the MT output. Over-segmentation, instead, in particular when involves a full stop, remains unchanged in the MT output, as MT models usually mirror the punctuation of the source text.

4. Conclusion

We presented a quantitative and qualitative evaluation of the tool that has been developed in the context of a EP's Innovation Partnership. We used WER score and human manual evaluation to evaluate the quality of ASR, and only human evaluation for MT quality. The average WER is 6.43% in the multilingual test set made of 19 languages deployed by November 2022, which is very low but it does not take into account segmentation issues. Human evaluation highlighted the need for refining sentence segmentation, especially in languages in which the WER was very low (e.g. RO and IT). This could indicate that WER by itself is not enough to have a clear picture of the quality of the transcription. However, human evaluation remains a highly subjective task which attains all categories, also those considered clear-cut categories (e.g. sentence segmentation). The annotators' background has an influence on error severity perception and error identification, and should be investigated in detail. In line with what found in [17], we also found that annotators' sensitivity in deepening the error annotation is a main cause of disagreement, in this case due to the attempt to annotate also the consequences of the error. Quantitative results of human evaluation considering the ASR output and its translation into IT (except for IT translated into EN) indicate ES as qualitatively better output, followed by IT, FR and EN, and RO as the worse output. In general, annotators rated ASR output worse than the MT output. However, this might be a consequence of the attitude of annotators putting too much emphasis on the provided reference transcription of the speech, not considering that, especially if punctuation is concerned, it is only one of the possible accepted transpositions. Qualitative results highlighted that different annotators draw different pictures of the same speeches and that a second round of annotations would be necessary to reduce disagreement and to clarify the use of error categories, like unintelligible, frequently improperly applied.

Acknowledgments

I want to thank the European Parliament for giving me the opportunity to conduct this study and the annotators for participating and giving me the authorisation to use their annotations for research purposes. I thank the anonymous reviewers for their precious comments, and I apologies if not all of them have been addressed.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), California, USA, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, A. Tamchyna, Findings of the 2014 workshop on statistical machine translation, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 12–58. URL: https://aclanthology.org/W14-3302. doi:10.3115/v1/W14-3302.

[3] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, M. Turchi, Cascade versus direct speech translation: Do the differences still make a difference?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2873–2887. URL: https://aclanthology.org/2021.acl-long.224. doi:10.18653/v1/2021.acl-long.224.

[4] M. Sperber, M. Paulik, Speech translation and the end-to-end promise: Taking stock of where we are, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7409–7421. URL: https://aclanthology.org/2020.acl-main.661. doi:10.18653/v1/2020.acl-main.661.

[5] A. Anastasopoulos, L. Barrault, L. Bentivogli, M. Zanon Boito, O. Bojar, R. Cattoni, A. Currey, G. Dinu, K. Duh, M. Elbayad, C. Emmanuel, Y. Estève, M. Federico, C. Federmann, S. Gahbiche, H. Gong, R. Grundkiewicz, B. Haddow, B. Hsu, D. Javorský, V. Kloudová, S. Lakew, X. Ma, P. Mathur, P. McNamee, K. Murray, M. Nădejde, S. Nakamura, M. Negri, J. Niehues, X. Niu, J. Ortega, J. Pino, E. Salesky, J. Shi, M. Sperber, S. Stüker, K. Sudoh, M. Turchi, Y. Virkar, A. Waibel, C. Wang, S. Watanabe, Findings of the IWSLT 2022 evaluation campaign, in: Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), Association for Computational Linguistics, Dublin, Ireland (in-person and online), 2022, pp. 98–157. URL: https://aclanthology.org/2022.iwslt-1.10. doi:10.18653/v1/2022.iwslt-1.10.

[6] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[7] B. Dorr, J. Olive, J. McCary, C. Christianson, Machine translation evaluation and optimization, in: J. Olive, C. Christianson, J. McCary (Eds.), Handbook of Natural Language Processing and Machine Translation, Springer, 2011, pp. 745–843.

[8] J. Moorkens, S. Castilho, F. Gaspari, S. Doherty, Translation quality assessment, Machine translation: Technologies and applications, Springer, 2018.

[9] E. Chatzikoumi, How to evaluate machine translation: A review of automated and human metrics, Natural Language Engineering 26 (2020) 137–161.

[10] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, W. Macherey, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, in: Transactions of the Association for Computational Linguistics, volume 9, 2021, pp. 1460–1474. URL: https://doi.org/10.1162/tacl_a_00437. doi:10.1162/tacl_a_00437.

[11] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, M. Zhou, Achieving Human Parity on Automatic Chinese to English News Translation, 2018. arXiv:1803.05567.

[12] A. Toral, S. Castilho, K. Hu, A. Way, Attaining the unattainable? reassessing claims of human parity in neural machine translation, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 113–123. URL: https://aclanthology.org/W18-6312. doi:10.18653/v1/W18-6312.

[13] S. Läubli, R. Sennrich, M. Volk, Has machine translation achieved human parity? a case for document-level evaluation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4791–4796. URL: https://aclanthology.org/D18-1512. doi:10.18653/v1/D18-1512.

[14] S. Nießen, F. J. Och, G. Leusch, H. Ney, An evaluation tool for machine translation: Fast evaluation for MT research, in: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), European Language

Resources Association (ELRA), Athens, Greece, 2000. URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf.

[15] J. C. López Otero, On the acceptability of the spanish dom among romanian-spanish bilinguals, in: A. Mardale, S. Montrul (Eds.), The Acquisition of Differential Object Marking Trends in Language Acquisition Research, John Benjamins Publishing Company, 2020, pp. 161–181.

[16] O. Guhr, A.-K. Schumann, F. Bahrmann, H.-J. Böhme, FullStop: Multilingual Deep Models for Punctuation Prediction, in: Swiss Text Analytics Conference, 2021.

[17] E. Di Nuovo, Introducing VALICO-UD: a parallel, learner Italian treebank for language learning research , Pàtron Editore, 2023.

A. Figures

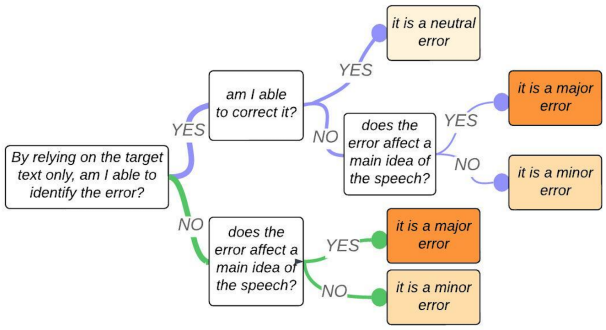

Figure 1: Decision tree used to annotate error severity.

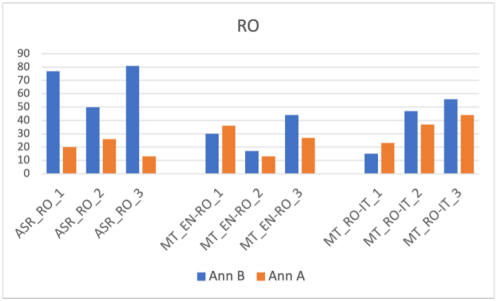

Figure 2: Ann A and B annotations of RO ASR, EN-RO MT and RO-IT MT.

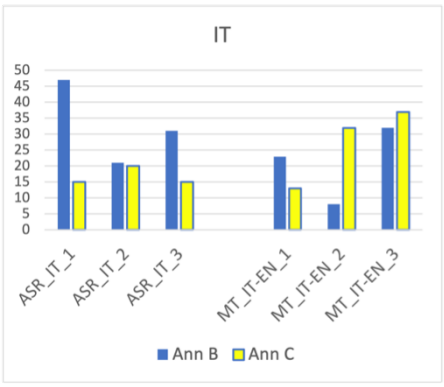

Figure 3: Ann B and C annotations of IT ASR and IT-EN MT.

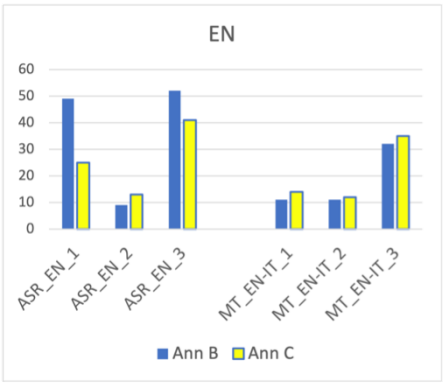

Figure 4: Ann B and C annotations of EN ASR and EN-IT MT.

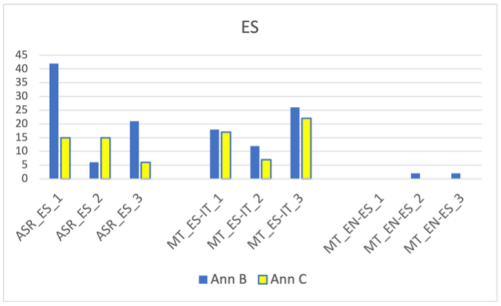

Figure 5: Ann B and C annotations of ES ASR and ES-IT MT; Ann B annotations of EN-ES MT.

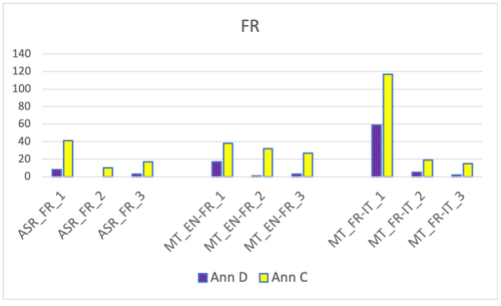

Figure 6: Ann C and D annotations of FR ASR, EN-FR MT and FR-IT MT.

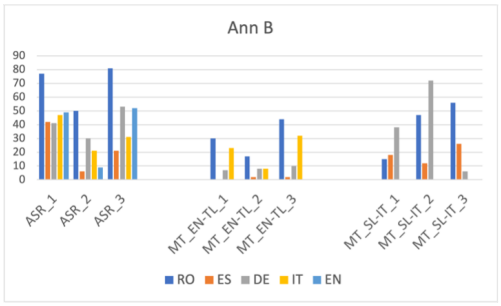

Figure 7: RO, IT, EN, ES and DE annotated by Ann B. In the figure, SL stands for Source Language, TL for Target Language.

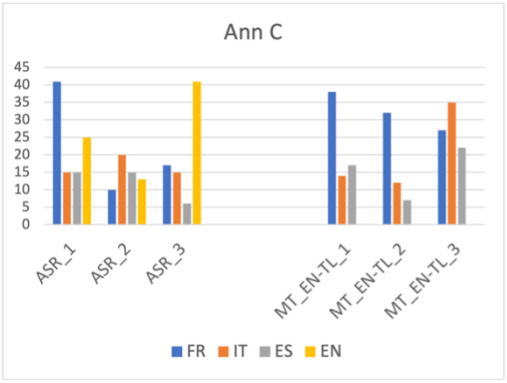

Figure 8: IT, EN, ES, and FR annotated by Ann C. In the figure, SL stands for Source Language, TL for Target Language.

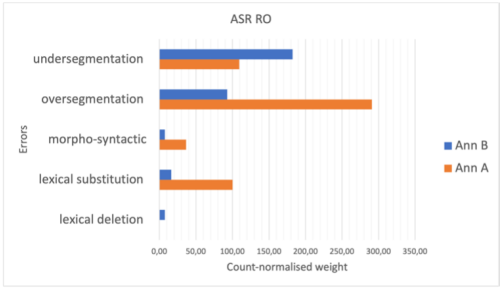

Figure 9: Error categories as annotated by Ann A and B in ASR RO.

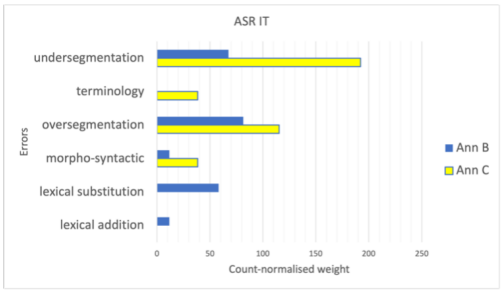

Figure 10: Error categories as annotated by Ann B and C in ASR IT.

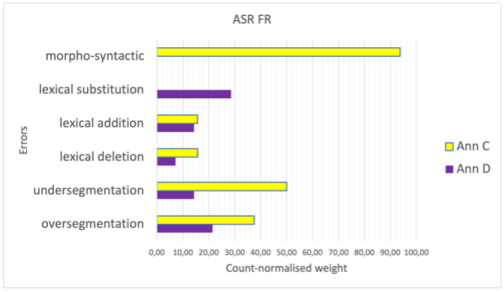

Figure 11: Error categories as annotated by Ann C and D in ASR FR.

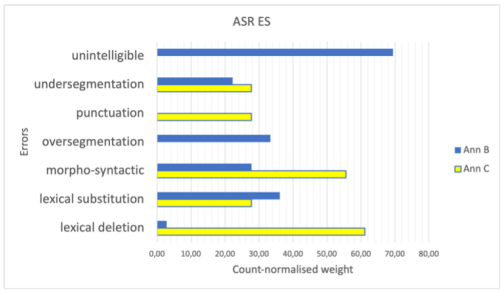

Figure 12: Error categories as annotated by Ann B and C in ASR ES.

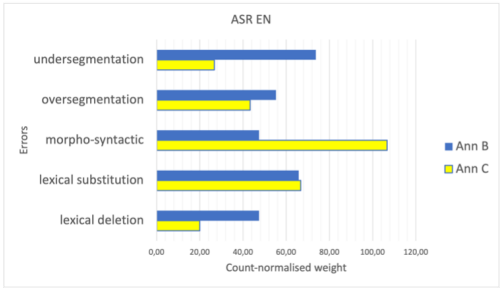

Figure 13: Error categories as annotated by Ann B and C in ASR EN.

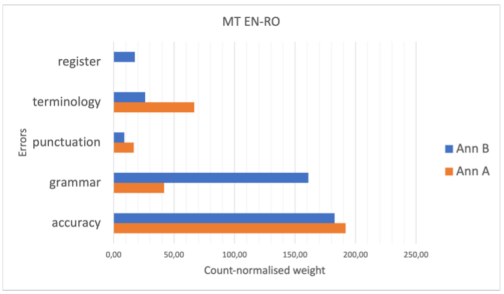

Figure 14: Error categories as annotated by Ann A and B in MT EN-RO.

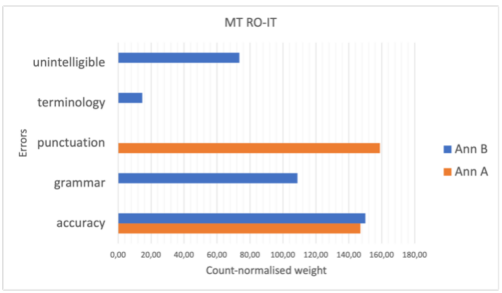

Figure 15: Error categories as annotated by Ann A and B in MT RO-IT.

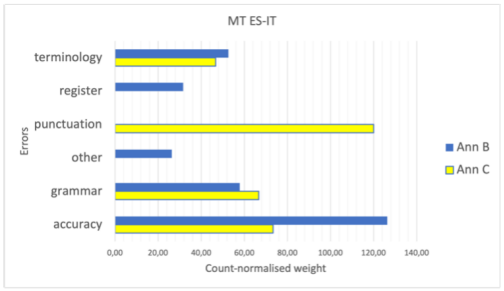

Figure 16: Error categories as annotated by Ann B and C in MT ES-IT.

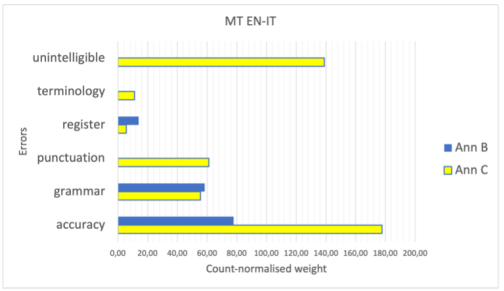

Figure 17: Error categories as annotated by Ann B and C in MT EN-IT.

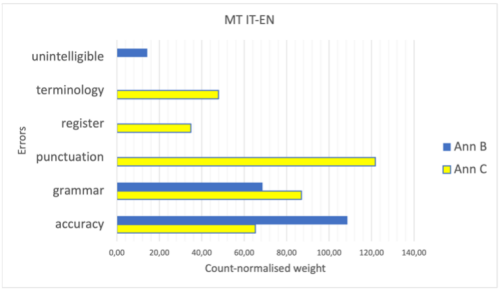

Figure 18: Error categories as annotated by Ann B and C in MT IT-EN.

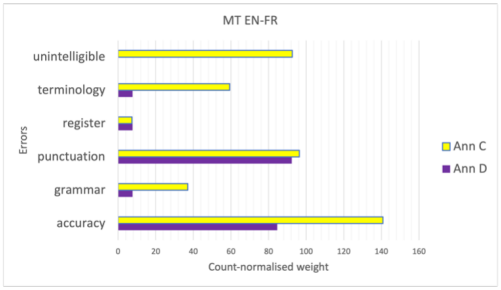

Figure 19: Error categories as annotated by Ann C and D in MT EN-FR.

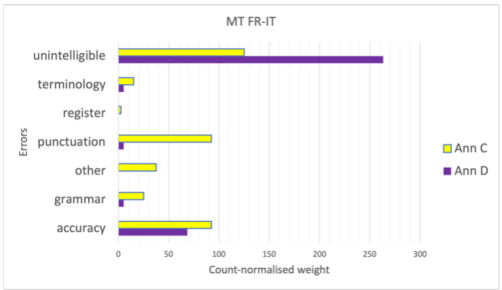

Figure 20: Error categories as annotated by Ann C and D in MT FR-IT.