Inter-Annotator Agreement Analysis for MQM-based Machine Translation Quality Evaluation: A Case Study on English-Italian Translation

Abstract

We present an inter-annotator agreement (IAA) study for Multidimensional Quality Metrics (MQM) annotation of machine translation output. Two professional linguists independently annotated English-to-Italian translations from two neural MT systems (EuroLLM-22B and Qwen3-235B) using the MQM error typology. The source documents were drawn from the WMT 2025 Human Evaluation dataset, specifically selecting segments without prior MQM or other quality annotations. Our analysis reveals a Kendall's tau correlation of 0.317 for segment-level MQM scores, substantially higher than the typical 0.12 reported in WMT shared tasks. While annotators achieved 100% agreement on identifying segments containing errors, significant differences emerged in error density (42 vs. 134 total errors) and category preferences. Span-level analysis shows 50% overlap on error locations, with 48% category agreement on matched spans. These findings contribute to understanding annotator variation in MQM-based evaluation and highlight the importance of multi-annotator setups for reliable MT quality assessment.

1. Introduction

Human evaluation remains the gold standard for assessing machine translation (MT) quality, despite significant advances in automatic metrics. The Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) has emerged as the preferred methodology for fine-grained error annotation, adopted by the Workshop on Machine Translation (WMT) since 2021 (Freitag et al., 2021).

MQM annotation requires identifying error spans in translated text and categorizing them by type (e.g., Accuracy, Fluency, Terminology) and severity (Minor, Major, Critical). This granular approach provides richer feedback than scalar ratings but introduces challenges for inter-annotator agreement. Previous studies report low IAA for MQM, with Kendall's tau correlations around 0.12 (Freitag et al., 2021).

This paper presents a case study examining IAA between two professional linguists annotating English-Italian MT output. We investigate:

1. **Segment-level agreement**: Do annotators assign similar quality scores to segments?
2. **Span-level agreement**: Do annotators identify the same error locations?
3. **Category agreement**: When annotators mark the same span, do they assign the same error type?

Our findings provide insights into the sources of annotator disagreement and practical implications for MQM annotation campaigns.

2. Data and Methods

2.1 Dataset Description

The evaluation corpus consists of 10 text segments in the social media domain, translated from English to Italian. Source documents were obtained from the WMT 2025 Human Evaluation dataset, selecting documents that had not received prior MQM or other quality annotations.

Each segment was translated by two neural MT systems: - **EuroLLM-22B**: A multilingual large language model - **Qwen3-235B**: A large-scale multilingual model

This yielded 20 translation instances (10 segments × 2 systems) for annotation.

2.2 Annotation Setup

Two professional linguists, native Italian speakers with translation expertise, independently annotated all translations using the MQM framework. Annotators are identified by anonymized hashes:

Annotator ID	Experience Level
A-5BFF0F0F	Professional
A-7A8BCDCD	Professional

Table 1: Annotator profiles.

Annotators used the Alconost MQM annotation tool with the following error categories: - Accuracy (Mistranslation, Omission, Addition, Untranslated) - Fluency (Grammar, Spelling, Punctuation, Inconsistency) - Terminology - Style

Severity levels followed MQM conventions: Minor, Major, and Critical.

2.3 Agreement Metrics

We employ multiple metrics following WMT evaluation practices:

Segment-level MQM Score:

$$\text{MQM}_{score} = -\sum(w_i \times e_i)$$

Where weights w_i are: Minor = 1, Major = 5, Critical = 25, Minor/Punctuation = 0.1.

Correlation Metrics: - Kendall's Tau-c: Rank correlation robust to ties - Pearson's r: Linear correlation of segment scores - Spearman's rho: Monotonic relationship

Span-Level Metrics: - Jaccard Index: $J = \frac{|A \cap B|}{|A \cup B|}$ - Precision/Recall: Treating each annotator as reference

Spans were considered matching if they overlapped by ≥30% (following standard practice for span-based evaluation).

---

3. Results

3.1 Annotation Statistics

Table 2 presents summary statistics for both annotators.

Metric	A-5BFF0F0F	A-7A8BCDCD
Total errors annotated	42	134
Annotation time (hours)	1.5	3.5
Time span (first-last)	1.39h	3.11h

Metric	A-5BFF0F0F	A-7A8BCDCD
Errors per hour	28	43
Segments with errors	10/10	10/10
Mean errors per segment	4.2	13.4

Table 2: Annotation statistics by annotator.

A-7A8BCDCD identified 3.2× more errors while working 2.3× longer, resulting in a higher error detection rate (43 vs. 28 errors/hour).

Category Distribution:

Category	A-5BFF0F0F	A-7A8BCDCD
Fluency/Grammar	7 (17%)	54 (40%)
Accuracy/Mistranslation	11 (26%)	28 (21%)
Style	15 (36%)	22 (16%)
Accuracy/Untranslated	1 (2%)	20 (15%)
Terminology	5 (12%)	0 (0%)
Other	3 (7%)	10 (8%)

Table 3: Error category distribution.

Notable differences include A-5BFF0F0F's focus on Style and Terminology errors, while A-7A8BCDCD emphasized Fluency/Grammar and Untranslated content.

Severity Distribution:

Severity	A-5BFF0F0F	A-7A8BCDCD
Minor	33 (79%)	132 (98%)
Major	8 (19%)	2 (2%)
Critical	1 (2%)	0 (0%)

Table 4: Severity distribution.

A-5BFF0F0F assigned substantially more Major/Critical ratings (21%) compared to A-7A8BCDCD (2%).

3.2 Segment-Level Agreement

Table 5 shows MQM scores per segment.

Segment	A-5BFF0F0F	A-7A8BCDCD	Δ
auto_0	-8.0	-11.0	3.0
auto_1	-16.0	-15.0	1.0
auto_2	-9.0	-19.0	10.0
auto_3	-3.0	-13.0	10.0
auto_4	-8.0	-15.0	7.0

Segment	A-5BFF0F0F	A-7A8BCDCD	Δ
auto_5	-2.0	-14.0	12.0
auto_6	-8.0	-14.0	6.0
auto_7	-9.0	-9.0	0.0
auto_8	-7.0	-13.0	6.0
auto_9	-28.0	-19.0	9.0
Mean	**-9.8**	**-14.2**	**6.4**

Table 5: Segment-level MQM scores.

Correlation Results:

Metric	Value	p-value
Kendall's Tau	0.317	0.229
Pearson r	0.530	0.115
Spearman rho	0.458	0.183

Table 6: Segment-level correlations.

The Kendall's tau of 0.317 exceeds the typical WMT MQM agreement of ~0.12 by a factor of 2.6×.

3.3 Span-Level Agreement

Metric	Value
A-5BFF0F0F errors matched by A-7A8BCDCD	21/42 (50.0%)
A-7A8BCDCD errors matched by A-5BFF0F0F	21/134 (15.7%)
Total unique error spans	155
Jaccard Index	13.5%

Table 7: Span-level agreement metrics.

Of the 155 unique error spans identified, only 21 (13.5%) were marked by both annotators.

3.4 Category and Severity Agreement on Matched Spans

For the 21 spans where both annotators identified an error:

Agreement Type	Count	Percentage
Same category	10/21	47.6%
Same severity	15/21	71.4%
Both match	8/21	38.1%

Table 8: Agreement on matched spans.

When annotators agree on error location, they agree on severity more often (71%) than category (48%).

4. Discussion

4.1 Comparison to WMT Benchmarks

Our segment-level Kendall's tau (0.317) substantially exceeds the 0.12 typically reported for MQM in WMT evaluations (Freitag et al., 2021). Several factors may explain this:

1. **Domain consistency**: All segments came from a single domain (social media)
2. **Language pair**: EN→IT may have clearer error patterns than other pairs
3. **Document context**: Both annotators had access to full document context
4. **Annotator expertise**: Both are professional linguists with target language nativity

4.2 Sources of Disagreement

Despite reasonable segment-level agreement, substantial differences emerged:

Error Density: A-7A8BCDCD identified 3.2× more errors. This may reflect: - Different thresholds for what constitutes an "error" - More thorough reading by A-7A8BCDCD (higher time investment) - Different interpretation of annotation guidelines

Category Preferences: - A-5BFF0F0F favored Terminology (12% vs 0%) and Style (36% vs 16%) - A-7A8BCDCD emphasized Grammar (40% vs 17%) and Untranslated (15% vs 2%)

This suggests annotators applied different mental models of translation quality.

Severity Calibration: A-5BFF0F0F marked 21% of errors as Major/Critical versus only 2% for A-7A8BCDCD. This represents a fundamental difference in severity threshold interpretation.

4.3 Implications for Annotation Practice

Our findings suggest:

1. **Multi-annotator setups are essential**: Single-annotator MQM provides unreliable estimates
2. **Calibration sessions needed**: Annotators should align on severity thresholds
3. **Category guidelines**: Clear definitions needed for overlapping categories (Style vs Grammar)
4. **Cost-quality tradeoff**: Higher-paid annotator found more errors but at 1.3× cost per error ($0.76 vs $0.59)

---

5. Conclusion

We presented an inter-annotator agreement study for MQM annotation of English-Italian machine translation. While segment-level correlation (Kendall's tau = 0.317) exceeded typical WMT benchmarks, span-level analysis revealed only 13.5% Jaccard agreement on error locations.

Key findings: - Annotators agree that segments contain errors but disagree on quantity - Category agreement on matched spans is near-chance (48%) - Severity calibration differs substantially between annotators

Future work should investigate calibration techniques to improve span-level agreement while maintaining the granular feedback that makes MQM valuable for MT development.

References

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9, 1460–1474.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12, 455–463.

Appendix A: Per-Segment Error Counts

Segment	A-5BFF0F0F	A-7A8BCDCD	Matched	A-5BFF0F0F only	A-7A8BCDCD only
auto_0	4	11	3	1	8
auto_1	8	15	4	4	11
auto_2	5	15	2	3	13
auto_3	3	13	1	2	12
auto_4	4	15	2	2	13
auto_5	2	14	1	1	13
auto_6	4	14	3	1	11
auto_7	5	9	2	3	7
auto_8	3	13	1	2	12
auto_9	4	15	2	2	13
Total	**42**	**134**	**21**	**21**	**113**

Appendix B: Annotation Timeline

Annotator	Session 1	Gap	Session 2	Total Span
A-5BFF0F0F	07:27–07:54 (Qwen3)	44 min	08:38–08:51 (EuroLLM)	1.39h
A-7A8BCDCD	07:52–08:59 (Qwen3)	67 min	10:06–10:59 (EuroLLM)	3.11h

All timestamps UTC, 2025-12-30.

Appendix C: Data Availability

All annotation data is available as part of the Alconost MQM Translation Gold Dataset on HuggingFace.