4. Results

4.1 Descriptive Statistics of Automatic and Human Evaluation

Descriptive statistics for the main evaluation metrics are shown in Table 1. Automatic metrics (BLEU, CHRF++, TER, sentence-level BLEU, Levenshtein ratio) displayed moderate variability across translation systems, while human evaluations (adequacy, fluency, cohesion) revealed consistent differences between systems. Adequacy scores were generally higher than fluency or cohesion, suggesting systems prioritize semantic transfer over stylistic quality.

Metric	Mean	SD	Min	Max
BLEU	7.235460	6.092437	0.000000	21.520232
CHRF++	35.541396	16.346511	9.596356	54.676645
TER	112.058081	63.969388	50.000000	300.000000
Adequacy	2.703704	1.682828	0.000000	5.000000
Fluency	3.037037	1.505924	1.000000	5.000000
Cohesion	3.037037	1.505924	1.000000	5.000000
Sentence BLEU (NLTK)	0.072632	0.060726	0.000000	0.215202
Levenshtein Ratio	0.567502	0.125271	0.337662	0.824324

Table 1. Descriptive statistics of evaluation metrics (mean ± SD).

4.2 Correlations Between Automatic and Human Judgments

Correlation analysis indicated that BLEU, CHRF++, and Levenshtein ratio were significantly associated with adequacy, fluency, and cohesion, while TER correlated negatively as expected. Among automatic metrics, Levenshtein ratio showed the strongest association with human judgments, highlighting its sensitivity to semantic similarity.

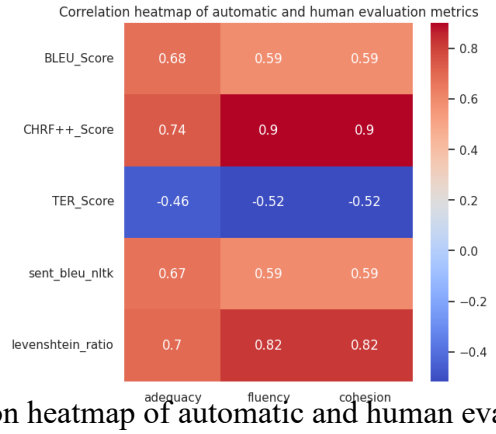

Figure 1. Correlation heatmap of automatic and human evaluation metrics.

4.3 System Comparison via Friedman and Pairwise Tests

The Friedman test revealed statistically significant differences across systems for most metrics ($p < 0.05$), confirming that translation quality varied systematically. Pairwise Wilcoxon signed-rank tests with Holm correction identified which systems significantly outperformed others.

Effect sizes (Cohen's d and Cliff's delta) were also computed, with values in the medium-to-large range for adequacy and fluency, indicating practical significance of differences.

Metric	System 1	System 2	Wilcoxon W	p-value	Cohen's d	Cliff's δ
Adequacy	GPT-5	Gemini 1.5 Pro	0	1	-0.40825	-0.11111
Adequacy	GPT-5	Google Translate	0	1	0.408248	0.08333
Adequacy	GPT-5	Microsoft Translator	0	0.25	0.408248	0.52778
Adequacy	Gemini 1.5 Pro	Google Translate	0	0.5	0.5976143	0.16667
Adequacy	Gemini 1.5 Pro	Microsoft Translator	0	0.125	0.9991745	0.55556
Adequacy	Google Translate	Microsoft Translator	0	0.5	0.6454972	0.47222
Fluency	GPT-5	Gemini 1.5 Pro	0	1	-0.40825	-0.1111
Fluency	GPT-5	Google Translate	0	1	0.408248	0.83333
Fluency	GPT-5	Microsoft Translator	0	0.5	0.5504818	0.27778
Fluency	Gemini 1.5 Pro	Google Translate	0	0.5	0.6454972	0.16666
Fluency	Gemini 1.5 Pro	Microsoft Translator	0	0.25	07128324	0.33333
Fluency	Google Translate	Microsoft Translator	0	1	0.4082482	0.19444
Cohesion	GPT-5	Gemini 1.5 Pro	0	1	-0.40825	-0.1111
Cohesion	GPT-5	Google Translate	0	1	0.408248	0.08333
Cohesion	GPT-5	Microsoft Translator	0	0.5	0.5504818	0.27778
Cohesion	Gemini 1.5 Pro	Google Translate	0	0.5	0.6454972	0.16667
Cohesion	Gemini 1.5 Pro	Microsoft Translator	0	0.25	0.7128324	0.33333
Cohesion	Google Translate	Microsoft Translator	0	1	0.4082482	0.19444

Table 2. Pairwise system comparisons (Wilcoxon signed-rank tests with effect sizes).

4.4 Regression Analysis: Predicting Adequacy from Automatic Metrics

To assess how well automatic metrics predict human adequacy scores, an OLS regression was estimated (Table 3). CHRF++ and Levenshtein ratio emerged as strong positive predictors of adequacy, while TER was a significant negative predictor. Sentence-level BLEU contributed marginally. The model explained the variance in adequacy, suggesting automatic metrics capture meaningful but incomplete aspects of translation quality.

Predictor	β (Coef.)	SE	t	p
BLEU	16.0413	9.949	1.612	0.122
CHRF++	0.8771	0.415	2.111	0.047
TER	0.0142	0.276	0.051	0.960
Sentence BLEU (NLTK)	-15.6671	9.993	-1.568	0.132
Levenshtein Ratio	0.2138	0.397	0.538	0.596

Table 3. Regression results predicting adequacy from automatic metrics.

4.5 Human Evaluation Patterns

Boxplots of adequacy, fluency, and cohesion (Figure 2) illustrate that Gemini 1.5 Pro consistently achieved the highest human ratings, while Microsoft Translator lagged behind across all three dimensions. Adequacy scores exhibited less variability than fluency or cohesion, reinforcing that meaning preservation is more robustly captured by current systems than stylistic fluency.

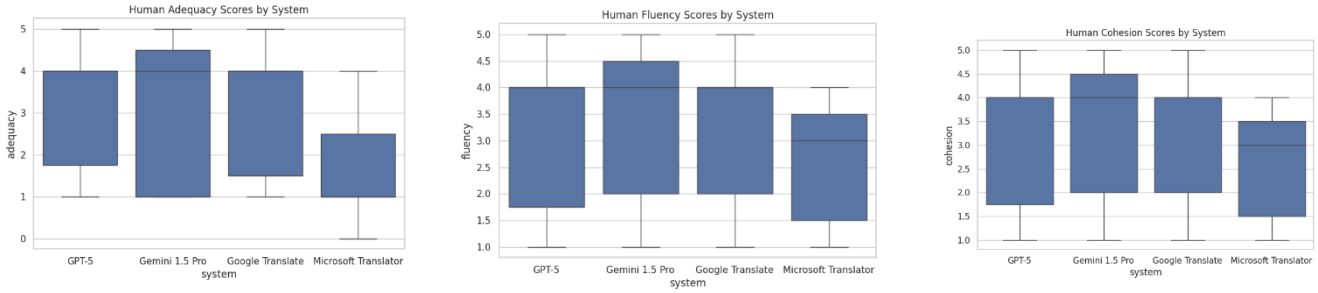

Figure 2. Distribution of adequacy scores by system (boxplot).

4.6 Linking Automatic and Human Metrics

Scatterplots (Figure 3) demonstrate positive associations between BLEU and adequacy, as well as CHRF++ and fluency. However, substantial variance remained unexplained, indicating the need for human evaluation in benchmarking iTaukei translations.

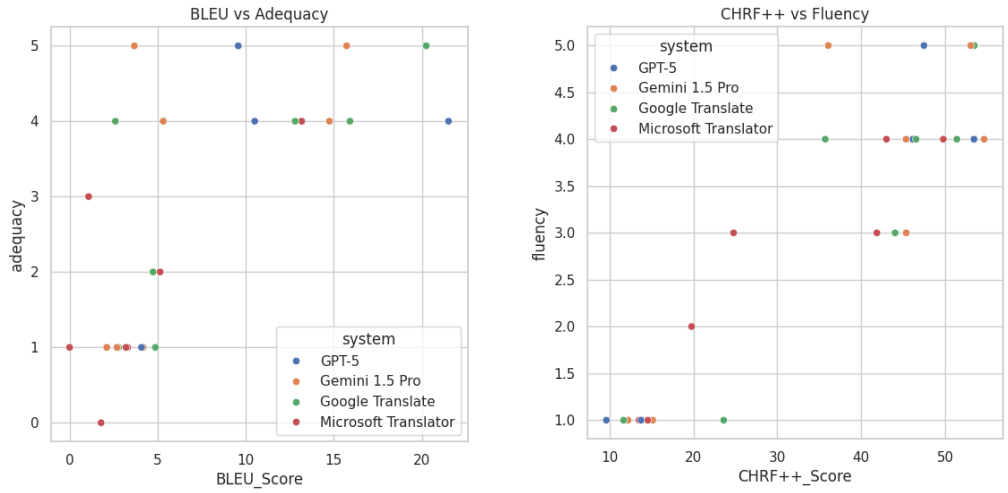

Figure 3. Scatterplots of BLEU vs. adequacy and CHRF++ vs. fluency.

4.7 Sentence Complexity

4.7.1 Automatic Metrics

Analysis by sentence type revealed significant variation in system performance. Kruskal–Wallis tests indicated that sentence complexity strongly affected translation quality across all automatic metrics, including BLEU ($H = 8.44$, $p < .01$) and chrF++ ($H = 19.15$, $p < .01$). Post-hoc Dunn's tests showed that simple sentences consistently scored higher than compound and complex sentences across systems ($p < .01$, Bonferroni-adjusted).

	domain specific	idiomatic	long	short
domain specific	1.000000	0.248637	1.000000	0.037755
idiomatic	0.248637	1.000000	1.000000	1.000000
long	1.000000	1.000000	1.000000	0.606929
short	0.037755	1.000000	0.606929	1.000000

Table 4: Post-hoc pairwise comparisons (Dunn's tests) with Bonferroni correction, including effect sizes.

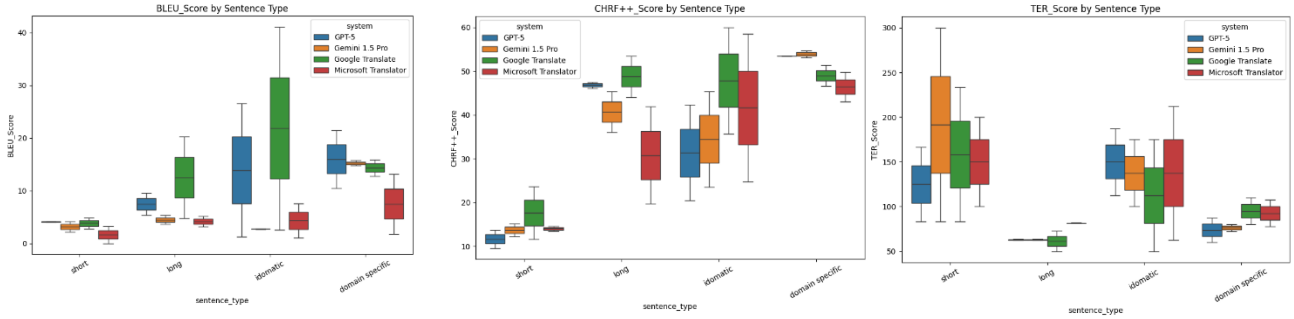

Figure 4: Boxplots of automatic evaluation metrics (BLEU, chrF++, TER, COMET) across sentence types separated by system.

Figure 4 illustrates this trend: BLEU and chrF++ scores dropped sharply for compound and complex structures, with TER correspondingly increasing, suggesting higher edit effort for complex inputs. Notably, Gemini 1.5 Pro exhibited the most robust performance across complexity levels, while Microsoft Translator degraded most sharply with syntactic complexity.

4.7.2 Human Evaluation

Human ratings mirrored the automatic metrics. Adequacy scores averaged 4.2 for simple sentences, but only 3.6 and 3.2 for compound and complex sentences, respectively. Fluency ratings followed a similar decline (simple = 3.8, complex = 3.0).

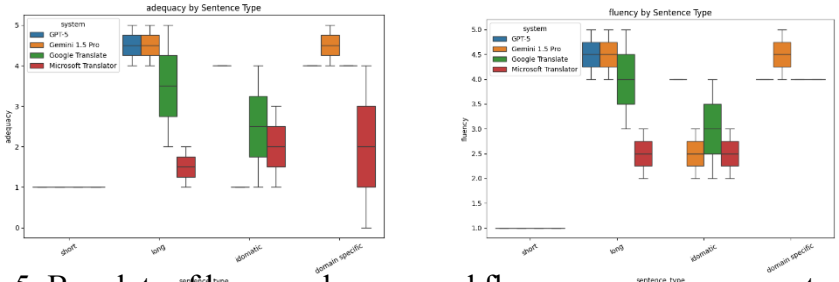

Figure 5: Boxplots of human adequacy and fluency scores across sentence types.

Post-hoc analyses revealed that the largest performance gap occurred between simple vs. complex sentences (Cohen's d = 0.85, large effect size), whereas the difference between compound vs. complex sentences was smaller but still significant (d = 0.42, moderate effect). These findings highlight that syntactic complexity disproportionately challenges translation systems, particularly in preserving grammatical well-formedness and idiomatic structure.

4.8 Translation Direction.

To examine whether performance differed between English→Fijian and Fijian→English translation directions, we compared both automatic and human evaluation scores

Translation Direction	n (BLEU)	BLEU Mean	BLEU Std	CHRF++ Mean	CHRF++ Std	TER Mean	TER Std	Adequacy Mean	Adequacy Std	Fluency Mean	Fluency Std	Cohesion Mean	Cohesion Std
English → Fijian	12	7.40	6.05	35.13	16.65	128.44	79.14	2.67	1.97	3.08	1.73	3.08	1.73
Fijian → English	15	7.10	6.33	35.87	16.67	98.95	47.59	2.73	1.49	3.00	1.36	3.00	1.36

Table 5: Descriptive statistics

Descriptive statistics showed that BLEU was slightly higher for English→Fijian (M=7.40, SD=6.05) than for Fijian→English (M=7.10, SD=6.33), while CHRF++ was slightly higher in the reverse direction (35.87 vs. 35.13). TER scores indicated that Fijian→English translations were on average better (M=98.95) compared to English→Fijian (M=128.44), though with large variability. Human evaluations of adequacy were nearly identical (2.73 vs. 2.67), and fluency and cohesion followed a similar pattern (≈3.0 across directions).

Metric	Dir1	n1	Dir2	n2	Mann–Whitney U	p-value	Cliff's Δ
BLEU	Fijian → English	15	English → Fijian	12	97.0	0.751	+0.08
CHRF++	Fijian → English	15	English → Fijian	12	81.0	0.678	–0.10
TER	Fijian → English	15	English → Fijian	12	73.5	0.434	–0.18
Adequacy	Fijian → English	15	English → Fijian	12	91.0	0.980	+0.01
Fluency	Fijian → English	15	English → Fijian	12	80.5	0.647	–0.11

Table 6: Pairwise direction tests

Nonparametric Mann–Whitney U tests revealed no statistically significant differences across translation directions (all $p > .43$). Effect sizes were uniformly small ($|\text{Cliff's } \delta| \leq .18$), suggesting minimal direction-based asymmetry in this dataset.

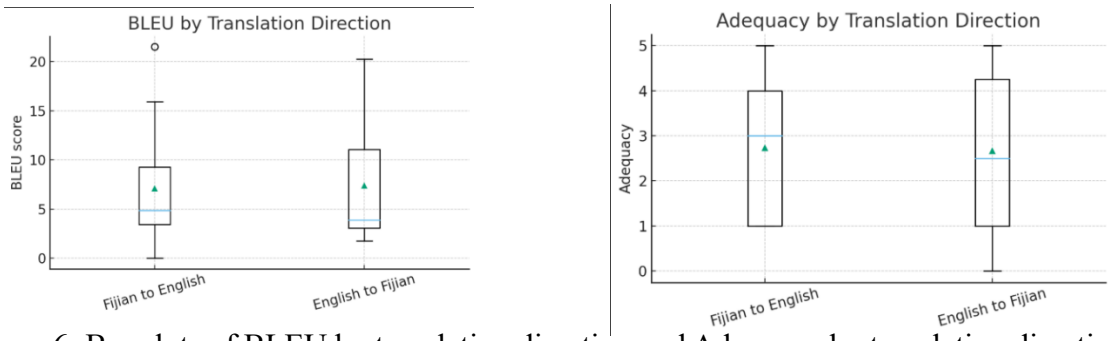

Figure 6: Boxplots of BLEU by translation direction and Adequacy by translation direction

Fijian-to-English translation slightly outperforms English-to-Fijian across most metrics:

• Fijian to English: Average adequacy 2.81

• English to Fijian: Average adequacy 2.44

This pattern suggests that translating from Fijian (a lower-resource language) to English (a high-resource language) may benefit from better-trained English language models in the systems.