Language Independent Evaluation of Translation Style and Consistency: Comparing Human and Machine Translations of Camus' Novel "The Stranger"

*Mahmoud El-Haj, *Paul Rayson, **David Hall
*Lancaster University, **UK Data Archive

Abstract

We provide a novel approach to evaluating translation performance across languages without the need for reference translations or comparable corpora. We present quantitative and qualitative results of automatic and manual comparisons of translations of the originally French novel "The Stranger" (French: L'Étranger).

Aim and Hypothesis

- Our aim is to check whether the human or machine has correctly preserved the variation in style and complexity in the original language at various document levels including chapters and parts.
- We hypothesise that the readability scores for each block of text in the original and translated versions should be similarly ranked if the translation quality is good.

Data Collection

- We are using human (manual) translation as well as machine (automatic) translation data.
- Two Arabic and two English translations.
- Male and female translator for each language
- English and Arabic machine translations using Google Translate.
- The French novel is divided into: 2 parts with 6 and 5 chapters each.
- Experiments were carried out at document, part and chapter levels.

Table 1: Data Collection Stats

Language	Sentences	Words	Unique Words
French	2,204	30,867	4,928
Arabic Male	951	24,129	6,808
Arabic Female	1,945	24,608	7,363
English Male	2,110	33,583	4,420
English Female	2,131	31,293	3,651

Readability

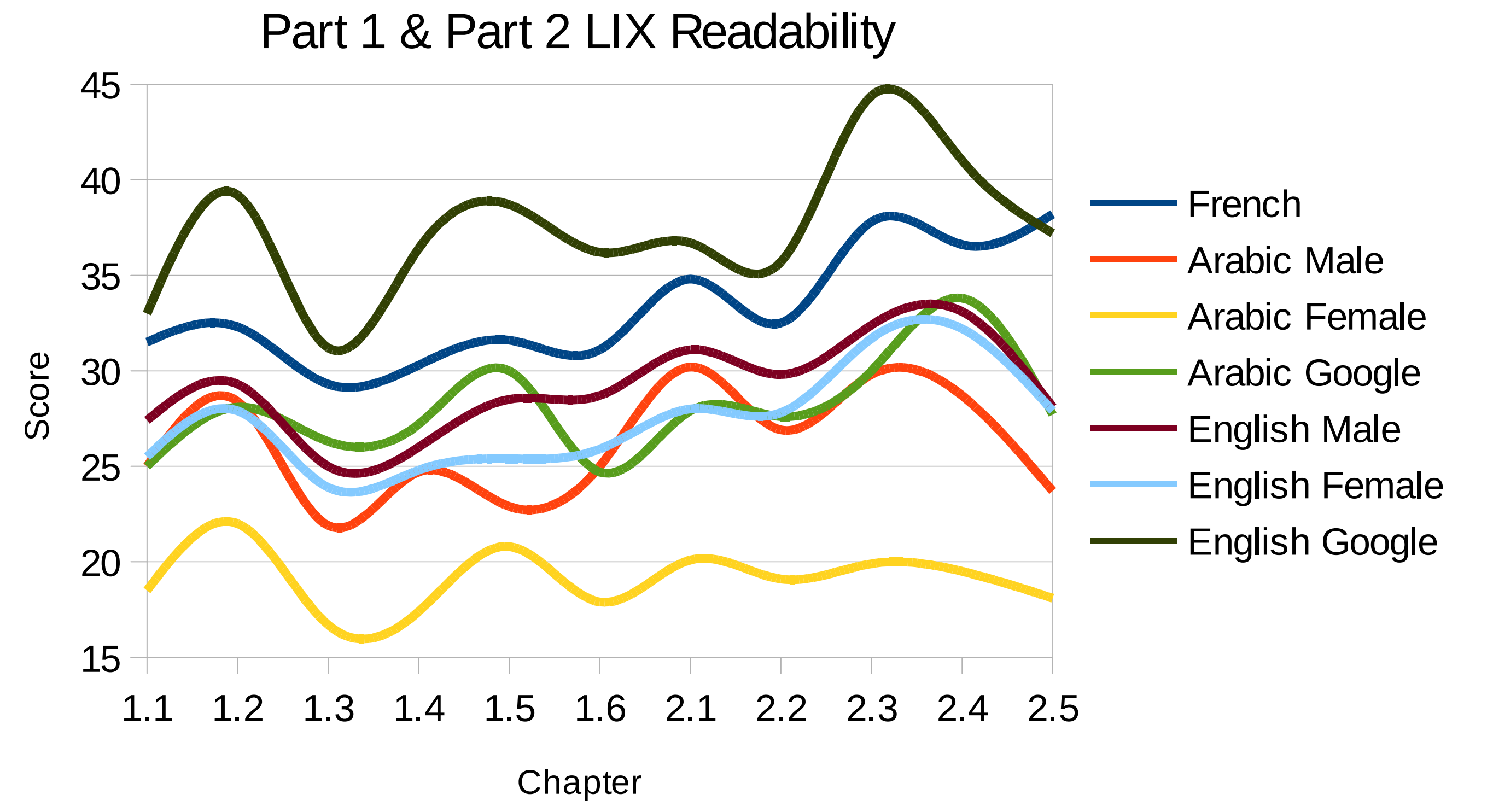
- Detect consistency in translation style using readability.
- We used Laesbarheds-Index (LIX) and Automated Readability Index (ARI).
- Both measures used for Arabic and French and found to correlate with English in previous research.
- For each language we calculated LIX and ARI for each part and chapter in addition to the full text. *Lower scores are easier to read.*

Table 2: LIX and ARI Readability Scores

	LIX			ARI		
Language	Part 1	Part 2	Full Text	Part 1	Part 2	Full Text
French	**33.91**	**38.85**	**36.33**	**6.15**	**7.79**	**6.94**
Arabic Male	24.48	27.39	25.91	6.21	7.36	6.76
Arabic Female	18.54	19.24	18.88	3.85	3.57	3.69
Arabic Google	26.11	28.11	26.97	6.98	7.51	7.20
English Male	27.39	31.07	29.22	4.70	5.90	5.29
English Female	25.43	29.59	27.51	3.78	5.12	4.44
English Google	33.06	37.44	35.31	7.16	9.05	8.12

** French writer's style is consistent across chapters.*

** Using this finding: chapters with readability scores close to the original text are considered to be high quality translation.*



Rank Correlation and Kendall Tau Comparisons

- Correlation scores support our hypothesis: translations with readability scores consistent with the original text to be of higher quality.
- English Male & Female close in translation style as indicated by consistency.
- Male translations more consistent with the original French text when compared to the female ones.
- The Spearman's scores in the table are consistent with the readability scores.
- Google's Arabic and English translations were found to be very close and consistent across chapters.

Table 3: Spearman's for LIX scores

	Arabic Male	Arabic Female	Arabic Google	English Male	**English Female**	English Google
French	0.49	0.29	0.58	0.66	0.47	0.53
Arabic Male	-	0.61	0.49	0.89	0.81	0.54
Arabic Female	-	-	0.74	0.71	0.58	0.70
Arabic Google	-	-	-	0.70	0.70	0.89
English Male	-	-	-	-	**0.92**	0.69
English Female	-	-	-	-	-	0.72

Evaluation

- We used Word Error Rate (WER) metric, derived from Levenshtein distance, working at the word level instead of the phoneme level.
- We used domain experts to judge the comprehensibility and readability of the four Arabic and English translations, one Arabic and one English native speaker and reader.

Table 4: Arabic Translations WER and Levenshtein Stats

Arabic Full Text	WER	Reference	Hypothesis	Correct	Sub	Ins	Del
Female vs. Male	**0.85**	24,867	23,969	6,131	15,424	2,414	3,312
Male vs. Female	**0.88**	23,969	24,867	6,131	15,433	3,306	2,408
Male vs. Google	1.00	23,969	27,028	3,839	18,820	4,369	1,310
Female vs. Google	0.96	24,867	27,028	4,852	18,438	3,738	1,577

Table 5: English Translations WER and Levenshtein Stats

English Full Text	WER	Reference	Hypothesis	Correct	Sub	Ins	Del
Female vs. Male	0.86	31,460	35,012	10,801	17,732	6,479	2,927
Male vs. Female	0.77	35,012	31,460	10,801	17,726	2,931	6,483
Male vs. Google	0.81	35,012	31,379	9,599	18,912	2,868	6,501
Female vs. Google	0.73	31,460	31,379	12,544	14,856	3,979	4,060

Results

- The results in Table 5 show that the Arabic human translators (male and female) are closer to each other than to Google.
- But the high WER scores between the male and female translations (and vice versa) suggest a big difference between the translations vocabulary, which is consistent with what has been reported by the human expert reader.
- The automatic and manual evaluation results are also consistent with the readability and the rank correlation scores.
- readability scores are consistent across parts and chapters in addition to the full text.
- The results support our hypothesis that translations with readability scores close to those of the original French novel are of better quality considering readability and style consistency.

Table 7. Keywords Human Comparisons (*M: Male, F: Female*)

Arabic (M)	**Arabic (F)**	**English (M)**	**English (F)**	**French**	**English Google**	**Arabic Google**
يغلق فمه	[DL] يسد بوزه	shut his trap	Shut your trap	fermer sa gueule	keep his mouth shut	يبقي فمه مغلقا
[FN] البلياردو	[FN] البليار	billiards	billiards	billard	billiards	[FN] البلياردو
حقل الملاهي	[TL] الشان دو مانوفر	Parade Ground	Parade Ground	Champ de Manoeuvres	Field Labourers	[WT] عمال الحقل
skipped	ملك الفرار	Handcuff King	King of the Escape Artists	le Roi de l'évasion	King of Escape	ملك الهروب