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

Mahmoud El-Haj[1], Paul Rayson[1], and David Hall[2]

[1] School of Computing and Communications, Lancaster University, UK
{m.el-haj, p.rayson}@lancaster.ac.uk
[2] UK Data Archive, The University of Essex, UK
djhall@essex.ac.uk

Abstract. We present quantitative and qualitative results of automatic and manual comparisons of translations of the originally French novel "The Stranger" (French: L'Étranger). We provide a novel approach to evaluating translation performance across languages without the need for reference translations or comparable corpora. Our approach examines the consistency of the translation of various document levels including chapters, parts and sentences. In our experiments we analyse four expert translations of the French novel. We also used Google's machine translation output as baselines. We analyse the translations by using readability metrics, rank correlation comparisons and Word Error Rate (WER).

1 Introduction

Translation in general is a complex task and it is more challenging when translating novels [7], especially ones written by Nobel Prize winners such as the French author, journalist and philosopher, Albert Camus. In a talk by [7], translator of a volume of Camus' Combat editorials, he called it "nonsense" to believe that "good translation requires some sort of mystical sympathy between author and translator" and instead he compared translating to playing the piano, saying that "The sinews and reflexes that translation requires are capable of development through exercise.". [7] believed that "it is different when translating novels, where translators need to have the eyes, ears, and wits to savour its beauties, and that they are obliged to preserve as many of them as they can.". [4] defined the task of the translator as consisting of "finding that intended effect upon the language into which he is translating which produces in it the echo of the original". He used this feature to differentiate translation from poet's work as the latter is never directed at the language but solely at specific contextual aspects.

These different opinions indicate that translation is a complex task and that there are many factors and features that could play a big role in defining translation quality. In our work we focus on the use of statistical metrics to judge translation consistency, comparing our results to qualitative analysis by expert readers for both the Arabic and English translations of the French novel, "The Stranger" (L'Étranger) by Albert Camus[3].

[3] http://en.wikipedia.org/wiki/Albert_Camus

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 116–124, 2014.

In our study we used four translations, male and female professional Arabic translations and male and female professional English translations. We also used Google's machine translation to automatically translate the French novel into Arabic and English, using these as baseline translations to allow us to judge the efficiency of our approach across both human and machine translation.

2 Related Work and Our Hypothesis

[13] demonstrate that readability features can be used in statistical machine translation to produce simplified text translations that could be useful, for example, for language learners and others who want to have a feel for the major content in highly domain-specific documents written in a foreign language.

The study by [6] measures the relationships between linguistic variation and reader perceptions by analysing 74 linguistic features in a set of 80 introductory academic textbooks. The goal was to study whether it is possible to assess student perception of effectiveness, comprehensibly and organisation in the textbooks. The statistical test showed three ("Elaboration and Involvement", "Colloquial Style", and "Academic Clarity") of the 74 linguistic features of variation were significant predictors of perception.

Previous research on translation quality has not taken into account the variability or consistency of the translation text. In our work we propose a novel approach to evaluate translations without the need for a reference translation (gold standard). The approach uses language-independent readability metrics in combination with statistical rank correlation comparisons. The approach overcomes problems found with BLEU [10] and other n-gram based scores. Those problems include the unreliability of the metrics when it comes to evaluating on an individual sentence level, caused by data sparsity, and the dependency of n-gram metrics on word order [9].

Our approach focuses on checking whether the translation process (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. A poor translation would not preserve the original variation in style and readability. Our approach is therefore product-driven making explicit use of the structure of the original novel in our case study.

3 Data Collection and Pre-Processing

In our work we analyse the translation of four human expert translators. Two translations into Arabic and two translations into English with a male and female translator for each language.

The original French novel is divided into two parts with 6 and 5 chapters respectively. We followed the same order with the Arabic and English translations when running the experiments. The analysis and experiments were carried out at three levels: document, part, and chapter. Our approach does not require alignment of sentence level.

The original novel and the translations differ in size. Table 1 shows the number of words and sentences for each language.

Table 1. Data Collections Statistics

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

4 Readability

In order to detect consistency in the translation style we use readability as a proxy for style and then consider how it varies both within and between translations. The readability metrics used in our experiments are Laesbarheds-Index (LIX) [5] and Automated Readability Index (ARI) [12]. LIX and ARI have been used to measure readability of Arabic and French languages and found to correlate with measuring the readability for the English version [2,3,14,1]. We calculated LIX and ARI readability metrics for each part and chapter in addition to the full text of the original novel and the four translations. Table 2 shows the LIX and ARI readability scores for Part 1 and 2 of the novel in addition to the full text for the three languages. The lower the score the 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

Table 3 shows the LIX readability score for each chapter in Parts 1 and 2 for the three languages. The LIX readability scores show consistency across chapters for each language. Taking the French text readability scores, we can see the writer's style is consistent across chapters. Using this finding to judge the translation quality, chapters with readability scores close to the original text are considered to be high quality

Table 3. Part 1 & 2 Chapters LIX Readability (*1_1 stands for Part 1 Chapter 1 and so on.)*

	Part 1						Part 2				
Language	1_1	1_2	1_3	1_4	1_5	1_6	2_1	2_2	2_3	2_4	2_5
French	31.5	32.3	29.3	30.3	31.6	31.1	34.8	32.5	37.8	36.6	38.2
Arabic Male	25.0	28.4	21.9	24.7	22.9	25.0	30.2	26.9	29.8	28.7	23.7
Arabic Female	18.5	22.0	16.7	17.4	20.8	17.9	20.1	19.1	19.9	19.5	18.1
Arabic Google	25.0	28.1	26.3	27.2	30.0	24.7	27.9	27.6	30.0	33.8	27.7
English Male	27.4	29.3	25.0	26.0	28.5	28.7	31.1	29.8	32.4	33.1	28.1
English Female	25.5	27.9	23.9	24.8	25.4	25.9	28.0	27.8	31.7	32.2	27.9
English Google	33.0	39.2	31.2	36.4	38.7	36.2	36.7	35.7	44.4	41.0	37.2

translations. As an example, for Part 1, Chapter 1, the English male translator produces a higher quality than the other human translators.

5 Rank Correlation and Kendall Tau Comparisons

To identify and test the strength of the readability relationship lists shown in Section 4, we used Spearman's Rank Correlation Coefficient and Kendall Tau statistical methods.

Table 4 shows the Spearman's correlation scores for the LIX readability metric. We only report the Spearman's correlation results as the results we observed using Kendall Tau showed the same trends when compared to Spearman's.

Table 4. 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

The correlation scores shown in Table 4 support our consistency hypothesis in Section 4 where we considered translations with readability scores consistent with the original text to be of higher quality. The scores also show the closeness in translation and style consistency between the translations. Taking the English translation as an example we can see the closeness between the male and female translation style as indicated by consistency. Similarly we found male translations to be more consistent with the original French text when compared to the female translations. The Spearman's scores in the table are consistent with the readability scores shown in tables 2 and 3. As we expected, Google's Arabic and English translations were found to be very close and consistent across chapters.

6 Evaluation

To judge the translation performance, two types of evaluation were carried out: automatic and manual. To evaluate the translation performance for both the human translators and Google machine translations we used the Word Error Rate (WER) metric [11]. WER is derived from Levenshtein distance [8], working at the word level instead of the phoneme level. In addition and to ensure quality, we used domain experts to judge the comprehensibility and readability of the four Arabic and English translation. For the experiments in this paper we used one Arabic and one English native speaker and reader. The Arabic speaker read the Arabic translations in addition to the English translations. The speaker extracted some examples for terms that can be considered dialectical and are not used in Modern Standard Arabic (MSA) (Sample in Table 7). The English speaking participant read the English translations in addition to parts of the original French script and made his judgement based on comparing the translation and by taking some examples where the translators used idioms and metaphors (Sample in Table 8). The speakers were asked to judge the texts' readability and comprehensibility by writing short paragraphs describing how they thought the translations compared. Both speakers have PhD degrees and are experts in Arabic or English literature and are expert readers/writers and critics.

Tables 5 and 6 show the WER scores and the number of substitutions, insertions and deletions needed to transform the hypothesis translation (Example: Google English) into the reference one (Example: Human English). The tables also show the number of words in the reference and the hypothesis translations in addition to the number of correct matches. Table 5 suggests the Arabic male and female translations to be closer to each other than to Google translation. Comparing the male and female scores we can see that each translation contained around 25% correct matches from the other. The percentage is higher, *c*34% between the English male and female translations (Table 6).

Table 5. Arabic Translations WER and Levenshtein Stats

Arabic Full Text	WER	Reference	Hypothesis	Correct	Substitutions	Insertions	Deletions
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	1.03	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 6. English TranslationsWER and Levenshtein Stats

English Full Text	WER	Reference	Hypothesis	Correct	Substitutions	Insertions	Deletions
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

Google Arabic baseline translations were not close to the human translation references, with less than 20% considered correct matches. Google English did considerably better with *c*40% and *c*30% correct matches to the female and male reference translations. This suggests translating French to English using Google to be more accurate than translating into Arabic. The Arab expert reader found the male translation to be easier to comprehend, but the use of long sentences made the readability harder. In contrast the female translation was easier to read with the use of short sentences, but difficult to comprehend with the frequent use of dialect mainly used in Levant countries (i.e. Jordan, Lebanon etc.). Table 7 shows the use of dialect by the Arab female translator, referred to by [DL]. The translator also used transliteration [TL] and foreign words [FN] when translating from French in contrast to the Arab male translator who avoided the use of dialect and transliteration. But the male translator did use some common foreign words, e.g. "Billiard". The male translator used simplification (a tendency to produce simpler and easier-to-follow text) when translating, and this is also noticeable with the amount of skipped words (see example in Table 7).

The right most column in the table showing the Arabic Google translation shows a sample of wrong translations [WT] that have been found by the expert reader. The table also gives an insight into the writing style and use of vocabulary between the female and male translators, both translators using a rich vocabulary as seen in the unique words column in Table 1. The English expert reader found the male translator to be more accurate in places. He also found the male translator to be more literal and much more straightforward. Take for example the first example in Table 8 where the word 'audience' is more literal than the female's 'spectators'. This makes the male translation easier to understand, but the language is more dated due to a gap of over 60 years between translations. This raises questions such as: whether the male translator was more familiar with contemporaneous French idiom, or whether the female translator is interpreting the work for a modern audience? The male translator may indeed have been more in tune with the general philosophic and artistic feelings of the time, when he translated the French novel in 1946, only 4 years after it was first published.

7 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 (see Section 6). The expert reader found the two translations to be using a rich vocabulary (see Table 1). The automatic and manual evaluation results are also consistent with the readability and the rank correlation scores shown in the Tables 2, 3 and 4 in Sections 4 and 5, which show the Arabic female translation to be easier to read with low and consistent readability scores.

The rank correlation comparison scores in Table 4 show that the readability scores are consistent across parts and chapters in addition to the full text. The results strengthen our assumption that translations with readability scores close to those of the original French novel are of better quality considering readability and style consistency. The

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

results show that the Arabic male translation readability scores are closer to the original French than the female translation. This was supported by the rank correlation scores that suggest the Arabic male translation to be closer and more consistent with the French original.

Tables 7 and 8 show the differences between the two Arabic translations in the vocabulary usage and the big difference when translating metaphors and idioms. Which explains the low WER scores (Table 5), which found only 20% similarity between the two translations.

Table 8. Idioms and Metaphors Human Comparisons (English/Arabic) and their meaning
Pt: Part, Ch: Chapter, M: Male, F: Female

Pt	Ch	English/Arabic (F)	English/Arabic (M)	Original Meaning
1	2	let a wave of spectators out	disgorged their audiences	those attending the cinema came out
		موجة من المشاهدين	بحشود من المتفرجين	
1	2	we thrashed them	we licked them	defeated the opposition in a comprehensive way
		لقد انتصرنا عليهم	لقد هزمناهم	

Table 6 shows that the English female translation is closer to the male and Google translations with slightly more towards the latter. But when looking at the rank correlation scores (Table 4) we can see that the scores show high correlation between the

English male and female translations, indicating that the male and female translations are more consistent across chapters than is Google.

Table 7 shows through examples the similarity and closeness between the two translation, which most of the time are also consistent with Google translate. This is consistent with the statistics shown in Table 1, which found these translations to be using nearly the same number of words and sentences. Table 8 shows the differences between both translations when it comes to translating idioms and metaphors in sentences. This suggests the two translations are close on a word level rather than on a sentence level. The readability scores in Tables 2 and 3 are consistent with the expert reader judgements. The reader found the female translation to be more readable when compared to the male translation finding the language of the male translation to be a bit dated with over 60 year gap. The results suggest the male translation to be more consistent with the French novel. The results also show the Arabic and English male translations to be consistent when compared to each other.

8 Conclusion and Future Work

The paper shows language-independent readability metrics and rank correlation comparisons can be used to evaluate the translation quality and style consistency without the need for human translation references. To evaluate the results we used human expert readers and the Word Error Rate (WER) metric. The results show that using readability scores in combination with rank correlation comparison is consistent with human judgements about translation quality. As shown in our experiments, we were able to directly compare the Arabic and English human or machine translations to the French original text. The results also suggest that the Arabic and English male translation were closer, based on manual and automatic evaluation. We have shown that consistent readability scores across parts and chapters between original and translated text are indicators of good quality translations. Since it is usually hard to find enough human gold-standard references for any particular text, especially in under-resourced languages such as Arabic, we believe our method moves towards being an alternative economical solution for machine-translation evaluation. The current evaluation tools usually require several gold-standard references to provide a performance score and they measure the closeness between the hypothesis translation and the references without referring to the original text that is in a different language. Other evaluation metrics rely on the quality of the translation references, while our technique relies completely on the closeness of the translation to the consistency of style of the original text.

References

1. Altamimi, A.K., Jaradat, M., Aljarrah, N., Ghanem, S.: Aari: Automatic Arabic readability index iajit first online publication (2013)
2. Anderson, J.: Analysing the readability of English and non-English texts in the classroom with lix. In: The Australian Reading Association Conference. ERIC Institute of Education Sciences (1981)
3. Anderson, J.: Lix and rix: Variations on a little-known readability index. Journal of Reading 26(6), 490–496 (1983)

4. Benjamin, W.: Illuminations. Houghton Mifflin Harcourt (1968), http://books.google.co.uk/books?id=mV06rdTclagC
5. Björnsson, C.H.: Läsbarhet. In: Stockholm: Liber (1968)
6. Egbert, J.: Student perceptions of stylistic variation in introductory university textbooks. Linguistics and Education 25(0), 64–77 (2014)
7. Goldhammer, A.: Translating subtexts: What the translator must know. In: Talk delivered to Brandeis University Translation seminar. CUNY conference on translation. Association for Computational Linguistics, Waltham, MA, USA (1994), http://www.people.fas.harvard.edu/~agoldham/articles/WhatMust.htm
8. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Tech. Rep. 8 (1966)
9. Padó, S., Galley, M., Jurafsky, D., Manning, C.: Robust machine translation evaluation with entailment features. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1. pp. 297–305. ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
11. Popović, M., Ney, H.: Word error rates: Decomposition over pos classes and applications for error analysis. In: In Proceeding of the Second Workshop on Statistical Machine Translation held within ACL 2007. pp. 48–55 (2007)
12. Smith, E., Senter, R., (U.S.), A.F.A.M.R.L.: Automated Readability Index. AMRL-TR-66-220, Aerospace Medical Research Laboratories (1967)
13. Stymne, S., Tiedemann, J., Hardmeier, C., Nivre, J.: Statistical machine translation with readability constraints. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013). pp. 375–386. NEALT Proceedings Series 16, LiU Electronic Press (2013)
14. Uitdenbogerd, A.L.: Readability of french as a foreign language and its uses. In: Proceedings of the 10th Australasian Document Computing Symposium (2005)