The Relationship between Syntactic Complexity and Quality of Nmt Outputs: an Exploratory Study

Huang Yueyue, Li Keru

School of Foreign Languages, Xinhua College of Sun Yat-Sen University, Guangzhou, China

Keywords: Syntactic complexity, neural machine translation, Quality evaluation

Abstract: Propelled by automated translation technology, translators in the current times are in urgent need of more precise guidelines on the ramification of post-editing tasks. Set in English-Chinese pairs, this paper attempts to explore the relationship between syntactic complexity of source text (ST) and quality of Neutral Machine Translation (NMT) output. 40 sentences were extracted from two pieces of legal documents. Three groups were formed based on sentence length: the first group includes 20 sentences with 7-to-36-word length range, the second includes another 20 sentences with longer length ranging from 31 to 68 words, and the third comprises two previous groups as combined to test overall correlation. Syntactic Complexity Analyzer developed by Lu (2010) was adopted to measure the 40 sentences, which were then processed by two versions of free online NMT systems-Google Neural Machine Translation (GNMT) and Systran online translation tools. MT quality evaluation was carried out manually by counting errors at lexical and syntactic level. The overall results suggest a small-to-medium effect size from ST syntactic complexity for NMT quality regardless of different NMT systems, and T-unit-related complexity measurements, mean length of T-unit (MLT) in particular, account for most such correlation. Also, whereas GNMT output quality at lexical level scores significantly higher than that of Systran, error scoring for both systems at syntactic level does not vary significantly.

1. Introduction

Machine Translation (MT), which computational linguists employ to seek technological means in order for automatically transferring text from one natural language to another, has gained ever increasing interest in both academic and commercial fields. It is a challenging task, as words bear with multiple meanings; sentences may be interpreted differently in varied contexts, and grammatical structures in one natural language may not register correspondence in another.

Experts have dived into the field of MT to achieve a better result for translation output. It has now gone through various phases of development from Rule-based Machine Translation (RBMT), Statistical Machine Translation (SMT), Example-Based Machine Translation (EBMT), Hybrid Machine Translation (HMT), to the cutting-edge Neural Machine Translation (NMT).

The latest NMT model, in preference to previous statistical methods, integrates a deep learning module and is trained to maximise the probability of matching a source text to its target one without additional external linguistic information[1]. It has been reported to be able to narrow “the gap between human and machine translators” [2], but studies have yet to reach a full consensus on whether and on what degree NMT could surpass previous MT models. Furthermore, a lack of clear understanding of NMT output quality has led to a multitude of media hyperbole about the potential usefulness of NMT and corresponding displacement of human translators[1].

In language service industry, service providers are now exploring and adopting the approach of incorporating MT and Post-editing (PE) to improve the overall translation quality[3]. However, as translators still confront volatile MT quality, their predicted PE effort is not entirely reliable, and presumed effort is often inconsistent with actual efforts measured in PE process[4]. One way of dealing with such inconsistency is to at least present translators with a rough indicator for raw MT output to deliver a guided judgement of their PE efforts.

This study thus intends to explore one possible indicator-syntactic complexity of ST and its correlation with quality of NMT output. The paradigm is set in legal domain (English-Chinese) to

  DOI: 10.25236/icbsis.2020.027

avoid unduly cultural ambiguity as widely concerned in literary genres. 40 sentences were extracted from two pieces of English legal documents. A Syntactic Complexity Analyser compiled by Lu[5] was adopted to measure the syntactic complexity of the ST, which were processed and transferred into Chinese by two free online NMT systems-Google Neural Machine Translation and Systran online translation system. The former adopts a neural network to deliver MT result, whilst the latter traditionally follows a phrase-based statistical approach, but currently is also undergoing a significant shift to the neural network as claimed[6]. MT quality evaluation was carried out manually by counting errors at both lexical and syntactical level.

The following sections first summarise results from previously published related studies, and explain the rationale to conduct this study. Section 3 and 4 will further provide a detailed explanation on research questions and research methods. Finally, results, discussion and final conclusion are followed in section 5, 6 and 7.

2. Research Background

2.1 Measurements of Syntactic Complexity

Syntactic complexity can be defined as "the ability to produce writing that shows how ideas and large chunks of information are represented with the use of subordination and embedded subordinate clauses," or more broadly as "the range of forms that surface in language production and the degree of sophistication of such forms." Research into this field mostly concerns a variety of topics ranging from L1 development to L2 acquisition, where foci draw on the trajectory of syntactic development in the writings of EFL learners.

Different sets of measures for quantifying syntactic complexity have been proposed so far, mostly to explore one of the following metrics: length of production units (i.e. clauses, sentences, and T-units), amount of subordination, amount of coordination, range of surface syntactic structures, and degree of sophistication of particular syntactic structures.

One key element of particular academic inquiry is T-unit. Since it was introduced in 1965, T-unit has been examined in a variety of research across first language development and second language acquisition. It is defined as "each unit contains one independent clause and its dependent clauses "and believed to be "the shortest units into which a piece of discourse can be cut without leaving any sentence fragments as residue." It is different from a sentence but sometimes they may overlap in certain occasions.

Although syntactic complexity is viewed as part of a multi-dimensional construct, some measures overlap with one another and redundantly measure the same scope. Clarifying the purposes and functions of each measure can help to construct a more systematic construct of syntactic complexity.

Based on Norris and Ortega (2009), a diagram depicting the conceptualisation of a multi-dimensional and hierarchical construction of syntactic complexity was later published in 2015, as shown in Figure 1.

However, as recent studies on syntactic complexity are experiencing a shift from large-grained measures to fine-grained ones, the above diagram mostly incorporates large-grained measures of syntactic complexity. Large-grained measures, by classification, are normally examined under four dimensions (length of production unit, amount of subordination, amount of coordination, and degree of phrasal complexity), whilst fine-grained measures under two dimensions (different types of subordinate clauses and noun modifiers).

Nonetheless, with a focus on tentative exploration in mind, this study will mostly employ the proposed structure of syntactic complexity shown in the above diagram.

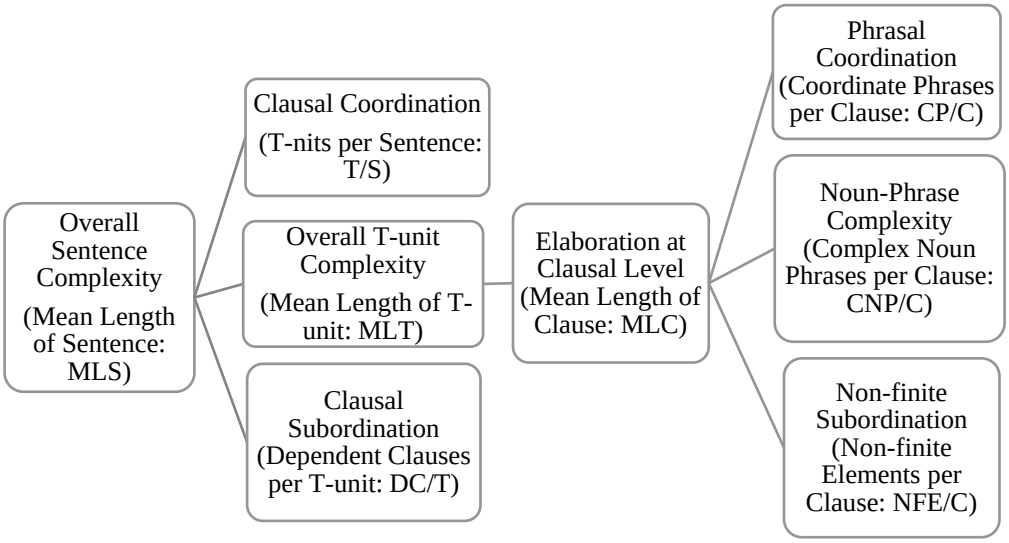

Fig.1 A Multi-Dimensional Construction of Syntactic Complexity

2.2 Mt Quality Evaluation

At present, evaluation for MT quality commonly adopts two paths: manual human evaluation and automatic machine evaluation. Automatic machine evaluation is often conducted using metrics like BLEU or METEOR to compare MT output with "gold standard" human translation. Besides, some choose to evaluate MT output by calculating post-editing rates such as TER (Translation Edit Rate) and HTER (human-mediated TER), etc.

Research into MT quality evaluation conducted through automatic machine evaluation metrics have yielded promising results on NMT quality. One such study using BLEU metrics shows that NMT indeed provides a more reliable version compared with previously used phrase-based MT[2]. Another joint report on NMT quality assessment shows that, in 6 out of 12 language pairs, NMT produces a more desirable level of quality compared with that of PBMT.

However, different evaluation methods may produce mixed results. One study shows that two sets of evaluation measurements conducted by automatic metrics (HTER, METEO and BLEU) and human metrics (fluency and adequacy) produce inconsistent results on the quality of SMT and NMT, suggesting that it might be related to factors such as language pairs and domain context [1].

It is noteworthy that the abovementioned automatic metrics and evaluation are often conducted via European language pairs, which excludes language pairs with wider lexico-syntactic distance such as Chinese-English pairs. Thus, the assessment of MT quality in the current study is presumably undertaken by human translators.

For translation quality assessment performed by humans, a range of measures are available in research and in industry, most of which are commonly operationalized on the sentence or text level of adequacy and fluency. In English-Chinese pairs, calculating MT errors is often included in the assessment of MT output.

2.3 Performance of Nmt Regarding Sentences with Complex Syntax

Current neural machine translation approach typically adopts an attention-based mechanism with two components, the first of which encodes a source sentence *x* and the second decodes *x* to generate a target sentence *y*. Instead of a linear statistical machine translation model, NMT incorporates a non-linear activation function to align and translate the *x* to *y*. Such mechanism reportedly presents an ideal and impressive leap forward in MT models, and its performance has been examined by numerous studies since.

Regarding its performance on sentences with complex syntax, several studies have demonstrated that NMT outperforms previous PBMT models in lab environments. One such study highlights that, in English-French language pairs, NMT exhibits higher reliability in terms of many of the more complex cases of subject-verb agreement, lexico-syntactic divergences, as well as the handling of

purely syntactic deviations. Nonetheless, it is still unclear on which extent NMT can provide a better result regarding sentences with complex syntax in English-Chinese language pairs beyond lab environments. And to clarify the relationship between syntactic complexity and MT quality may equip translators with a clearer orientation on how to regulate post-MT related tasks.

3. Current Study

In light of the above review, it requires further investigation in the nexus between ST syntactic complexity and NMT output quality, especially beyond Indo-European language pairs. With a view to offering practical guidance for translators on their *post hoc* handling of NMT output, the NMT quality evaluation for the extracted translation samples in this study is emphatically predicated on the translation errors each NMT system produces. Focused examination looks into error scoring's correlation with the multi-dimensional metrics of syntactic complexity of the ST. The research questions are raised for current research as follows:

1) What is the relationship between syntactic complexity and quality of NMT output for English-Chinese pairs?

2) How does such a relationship differ in terms of different NMT systems?

4. Research Method

4.1 Syntactic Complexity Scoring

In this study, the syntactic complexity of ST was measured by a Syntactic Complexity Analyser software package compiled by Lu (2010)[5], designed initially with advanced second language proficiency research in vision. The analyser toolkit can process parsed sentence samples and provide fourteen syntactic complexity indices, as shown in Table 1.

Table 1 Fourteen Syntactic Complexity Measures by the Syntactic Complexity Analyzer, Five of Which Are Selected for This Study

Measure	Definition	Label	Measure in this study
Type 1: Length of production unit	Mean length of clause	MLC	√
	Mean length of sentence	MLS	
	Mean length of T-unit	MLT	√
Type 2: sentence complexity	Clauses per sentence	C/S	√
Type 3: subordination	T-unit complexity ratio	C/T	
	Complex T-units per T-unit	CT/T	
	Dependent clauses per clause	DC/C	
	Dependent clauses per T-unit	DC/T	√
Type 4: Coordination	Coordinate phrases per clause	CP/C	
	Coordinate phrases per T-unit	CP/T	
	T-units per sentence	T/S	√
Type 5: Particular structure	Complex nominals per clause	CN/C	
	Complex nominals per T-unit	CN/T	
	Verb phrases per T-unit	VP/T	

As previous studies show that some measures are redundantly overlapping with another, the construction of syntactic complexity in Figure 1 (Section 2.1) thus is conceptualised as a three-layer multi-dimensional structure. Given that professional translators are presumably more sensitive towards explicit syntactic features in ST, the first- and second-layers sub-constructs (i.e. MLT, DC/T, T/S, MLC) thus were taken into consideration in the current research (see Table 1). In addition, since most measures are categorically related to T-unit or clauses, clauses per sentence (C/S) was also included in the study. MLS was excluded as it is equivalent to sentence length in this study, but considered as a baseline for categorising groups for differentiation.

Based on the above considerations, a pilot study was first conducted to explore the relationship between syntactic complexity and MT quality. It is shown in Table 3 as Group 1 (G1), which contains 20 randomly extracted sentences with MLS ranging from 7 to 36 from one legal document.

The initial test detected two measures (DCT, MLT) as prominent indicators for the relationship between syntactic complexity and NMT quality (SYST.LEX&DC/T: *R2= .23, Adjusted R2= .18, p< .05*, SYST.SYN&MLT: *R2= .57, Adjusted R2= .54, p< .001*, GNMT.SYN&MLT: *R2= .45, Adjusted R2= .42, p< .001*).

To further examine the results noted in the pilot study, 20 more sentences with higher MLS value (31-68) were extracted from both legal documents to form a Group 2 (G2). Both Group 1 and Group 2 were later combined as Group 3 (G3). Table 2 outlines the descriptive statistics for the three groups.

Table 2 Description of Mls for G1, G2, G3

	Total	MLS			
		Mean	*Median*	*Std. Deviation*	*Range*
G1	20	20.6	19.5	8.54955	7-36
G2	20	42.6	38	10.33848	31-68
G3	40	31.6	31.5	14.55282	7-68

And Table 3 presents brief descriptive statistics on all the involved syntactic complexity indicators measured in this study.

Table 3 Descriptive Statistics of All the Measures Tested in This Study

		G1	G2	G3
MLT	*Mean*	19.95	38.88	29.41
	Median	18.50	36.50	31.00
	Std. Deviation	8.61	10.62	13.52
TS	*Mean*	1.05	1.15	1.10
	Median	1.00	1.00	1.00
	Std. Deviation	0.22	0.37	0.30

To be continued next page

MLC	*Mean*	11.01	13.83	12.42
	Median	10.17	11.83	11.55
	Std. Deviation	4.94	7.05	6.17
DCT	*Mean*	0.83	1.95	1.39
	Median	1.00	2.00	1.00
	Std. Deviation	0.88	1.42	1.30
CS	*Mean*	1.90	3.55	2.73
	Median	2.00	4.00	3.00
	Std. Deviation	1.02	1.36	1.45

4.2 Quality Assessment of Mt Output

As the study intends to offer translators with a more precise indication for their post-editing tasks, the quality assessment is emphatically judged by counting the errors identified in the MT output. The errors are measured into two types - lexical and syntactic errors, referring to the classification proposed by Luo and Li. Their rating includes two levels, in which the first consists of three types (lexical, syntactic and symbolic errors). Symbol errors mainly concern mismatch on physical units, special symbols and punctuation marks, which therein do not fall into the range of legal source texts for the current study. Eventually, the error counts in this study are classified as shown in Table 4. Errors on level 1 are further elaborated on level 2.

Table 4 Classification of Translation Errors of Mt Output

Level 1	Lexical errors	Syntactic errors
Level 2	On term	On conjunction
	On parts of speech	Wrong word of order
	On noun phrases	On nominal construction
	On abbreviation	On verbal construction
	Missed translation	On prepositional construction
	On reference	On passive voice
		On infinitive structure
		On particles

The evaluation was conducted manually by two experienced translators, who both have 5 years of experience in the industry and currently work in a college-level university, teaching translation courses for BA students. The results in this study were meticulously checked and discussed until both raters reach a consensus.

Two examples are listed here to illustrate the process of error assessment:

Example 1:

ST: Liability for negligent hiring *extends to* any situation *where a third party is injured* by an employer's own negligence in failing to select an employee fit or competent to perform the services of employment.

TT by Systran: 失职雇用 的责任*延伸到*雇主 本身 因未能选择适合或有能力履行 雇佣服务 的雇员而*损害第三方的任何情况.*

TT by humans: 疏忽雇佣 *可包括*由于雇主 本人的过失 ,未能选择合适或能 胜任的 人员担任有关职位而*造成第三方伤害的情况.*

The underlined parts in the ST indicate lexical errors, while the italic and bold parts indicate syntactic errors. Thus, the total number of errors for this MT output by Systran is 3 for lexical errors and 2 for syntactic errors.

Example Two:

ST: Before he *was hired by the defendants,* Deojay had been convicted of burglary *in the third degree* and larceny in the third degree.

TT by GNMT: Deojay *在被告雇用之前,*曾被判犯有*三年级* 盗窃罪 和第三级盗窃罪.

TT by humans: *被告雇佣 Deojay 之前,*Deojay 曾被裁定犯有*三级* 入屋盗窃罪 和三级盗窃罪.

Based on the classification illustrated in Table 3, this example contains one lexical error and two syntactic errors.

Table 5 Tabulates the Descriptive Statistics of Error Count in Three Groups for Systran Output (Syst) and Gnmt Output At Lexical (Lex) and Syntactic (Syn) Level.

Table 5 Statistics on Lexical and Syntactic Error Counts from Systran and Google Machine Translation

	SYST.LEX				SYST.SYN			
	Mean	*Median*	*Std.*	*Range*	*Mean*	*Median*	*Std.*	*Range*
G1	2.1	2	1.12	0-5	1.25	1	1.02	0-3
G2	3.9	3	1.77	1-8	2.55	2.5	1.19	0-5
G3	3	3	1.72	0-8	1.9	2	1.28	0-5
	GNMT.LEX				GNMT.SYN			
	Mean	*Median*	*Std.*	*Range*	*Mean*	*Median*	*Std.*	*Range*
G1	1.5	1.5	1	0-4	1.3	1	1.38	0-4
G2	3.3	3	1.49	2-7	2.7	3	1.38	0-6
G3	2.4	2	1.55	0-7	2	2	1.54	0-6

4.3 Measurements of Correlation between Syntactic Complexity and Mt Output

Statistical analyses were run on IBM SPSS Statistics Version 23 software package. Normality of dependant variance (error counts) and homogeneity of variance assumptions were assessed by examining histograms and normality tests (Shapiro-Wilks), which indicated that half sets of variances were normally distributed, except for three sets in G1 (SYST.SYN: *SW= .43, std.= .51, p = .011*; GNMT.LEX: *SW= .526, std.= .512, p= .011*; GNMT.SYN: *SW= .731, std= .512, p= .004*), one set in G2 (GNMT.LEX: *SW=1.228, std.= .512, p= .001*) and two sets in G3 (SYST.LEX: *SW= .979, std.= .374, p= .003*; GNMT.LEX: *SW=1.024, std.= .374, p= .001*). After log transformation, the non-normality of these six variances was visibly optimised and thus was considered to be desirable enough to account for the slight deviations from normality for parametric tests.

Step-wise multiple linear regression was therefore applied to explore the correlation between syntactic complexity of ST and NMT quality, and, to be more exact, to pinpoint which complexity measure shows significant effect size in predicting the NMT quality. Additionally, since not all sets

of dependant variances demonstrate normality, Wilcoxon signed-ranks test with related samples was adopted to see whether there is a significant difference between the error counts of two NMT systems.

5. Results

5.1 Relationship between Syntactic Complexity and Nmt Quality

The central question of the study was illustrated by the multiple regression analysis to explore the relative strength of each measure in predicting NMT quality across all three groups. Table 6 then reports the most significant test results.

Table 6 Multiple Regression Analysis (Step-Wise Mode) for Measures with Significant Effect Size for Nmt Quality (#: Data Was Log-Transformed to Improve Normality)

G1 (n=20)	*predictor*	*R2*	*Adjusted R2*	*ΔR2*	*F(df)*	*b*	*p*
SYST.LEX	*DC/T*	0.23	0.18	0.23	5.22(1,18)	0.47	0.035
SYST.SYN#	*MLT*	0.57	0.54	0.57	23.34(1,18)	0.75	0.000
GNMT.LEX#	*- NONE -*						
GNMT.SYN#	*MLT*	0.45	0.42	0.45	14.80(1,18)	0.67	0.000
G2 (n=20)	*predictor*	*R2*	*Adjusted R2*	*ΔR2*	*F(df)*	*b*	*p*
SYST.LEX	*MLT*	0.29	0.25	0.29	7.20(1,18)	0.53	0.015
SYST.SYN	*- NONE -*						
GNMT.LEX#	*DC/T*	0.25	0.21	0.25	5.98(1,18)	0.50	0.025
GNMT.SYN	*C/S*	0.21	0.17	0.21	4.79(1,18)	0.46	0.042
G3 (n=40)	*predictor*	*R2*	*Adjusted R2*	*ΔR2*	*F(df)*	*b*	*p*
SYST.LEX#	*MLT*	0.44	0.43	0.44	29.91(1,38)	0.66	0.000
SYST.SYN	*MLT*	0.46	0.45	0.46	32.73(1,38)	0.74	0.000
	T/S	0.55	0.52	0.09	22.42(2,37)	0.30	0.012
GNMT.LEX#	*MLT*	0.30	0.28	0.30	16.03(1,38)	0.55	0.000
GNMT.SYN	*MLT*	0.46	0.45	0.46	32.28(1,38)	0.68	0.000

Initial observation demonstrates that MLT presents a good across-group predictive power for NMT quality of two systems at both lexical and syntactic level, with only one in Group 2 indicating a relatively weak correlation (SYST.LEX&MLT: *R2= .29; Adjusted R2= .25; p< .05*). At the syntactic level of Systran output in G3, two measures (MLT and T/S) enter the regression model as prominent predictors (MLT: *R2=0.46; Adjusted R2=0.45; p< .001*; T/S: *R2= .55; Adjusted R2= .52; p< .05*). Additionally, DC/T and C/S also show significant correlation with NMT quality in three specific occasions, though only accounting for a variance of lower degree (G1.SYST.LEX &DCT: *R2= .23; Adjusted R2= .18, p< .05*; G2.GNMT.LEX &DCT*: R2= .25; Adjusted R2= .21; p< .000;* G2.GNMT.SYN &CS: *R2= .21; Adjusted R2= .17; p< .05*)

5.2 Difference on Performance between Two Nmt Systems

The prerequisite to the second research question is to confirm whether there is significant difference in performance between two NMT systems. Results from Wilcoxon signed-ranks test are displayed in Table 7.

Table 7 Wilcoxon Signed-Rank Test Result

		G1		G2		G3	
		N	*p*	N	*p*	N	*p*
GNMT.LEX cf. SYST.LEX	Negative Ranks	10a	0.002	12a	0.016	22a	0.000
	Positive Ranks	0b		3b		3b	
	Ties	10c		5c		15c	
	Total	20	0.796	20	0.448	40	0.457
GNMT.SYN cf. SYST.SYN	Negative Ranks	6d		5d		11d	
	Positive Ranks	6e		6e		12e	
	Ties	8f		9f		17f	
	Total	20		20		40	

(a. GNMT.LEX < SYST.LEX; b. GNMT. LEX> SYST.LEX; c. GNMT.LEX = SYST.LEX; d.

GNMT.SYN < SYST.SYN; e. GNMT.SYN > SYST.SYN; f. GNMT.SYN = SYST.SYN)

The test results show that, across all three groups, there is significant difference only at the level of lexical performance between GNMT and Systran, in which the former largely reduces the number of lexical errors compared with what Systran does (G1: $p < .01$, G2: $p < .05$, G3: $p < .001$). However, such difference does not appear on the syntactic level (G1, G2, G3: $p > .05$).

6. Discussion

As there is significant deviation of performance between Systran and GNMT at the lexical level of quality, it is propelled to examine the relevant results separately in regression model. With regard to Systran, both models for G2 and G3 identifies MLT as a main predictor variable, but it gives way to DC/T in G1. Despite both measures being possible candidates, a closer examination of the data shows that MLT yields higher value of Adjusted $R2$ and lower p value (G1.SYST.LEX&DC/T: *Adjusted R2= .18, p=.035*; G2.SYST.LEX&MLT: *Adjusted R2= .25, p= .015*; G3.SYST.LEX& MLT: *Adjusted R2= .43, p= .000*). The revelation suggests that MLT appears to demonstrate a stronger predictive power compared with DC/T. As for GNMT, DC/T and MLT are also considered to be promising predictors, even though G1 does not produce statistically significant results, which might be a small probability event considered small sample size of this study. Moreover, across-group comparison illustrates that DC/T yields lower Adjusted *R2* value and higher *p* value, suggesting a weaker predictability compared with MLT (G2.GNMT.LEX&DC/T: *Adjusted R2= .21, p< .05*; G3.GNMT.LEX&MLT: *Adjusted R2= .28, p< .001*).

At the syntactic lexical, where no significant difference is found between GNMT and Systran output quality, MLT is also reflected to present a strong predictive power for the raw output quality in G1 and G3 (G1.SYST.SYN&MLT, G1.GNMT.SYN& MLT, G3.SYST.SYN&MLT, G3.GNMT.SYN& MLT: *Adjusted R2 > 0.4*; $p < .001$). It is also noteworthy that, in Group 3, two measures (MLT and T/S) enter the model as prominent predictors (MLT: *R2=0.46; Adjusted R2=0.45; p< .001*; T/S: *R2= .55; Adjusted R2= .52; p< .05*). Yet as the standardized beta coefficients value of T/S records distinctively lower than that of MLT, the latter is considered to be a more stable predictor (T/S: $b = .30$ < MLT: $b = .74$).

However, also at the syntactic level, results are rather confusing in Group 2-only C/S was found to be weakly correlated to the syntactic quality of GNMT output, whereas no significant correlation was detected in Systran data. Given that this study only examines large-grained measures of syntactic complexity and G2 contains sentences with higher MLS value, it might suggest other hierarchically refined variables, such as fined-grained measures of syntactic complexity which address phrase-level syntactic constituents, potentially correlate with complex sentences above certain threshold. Since most measures (MLT, DCT, T/S) highlighted in this study are conceptually related to T-unit, further study is required to focus more on refined measures on T-unit to uncover more prominent predictors for quality of ST involving complex syntactic structures.

7. Conclusion

Regardless of all, the overall findings of this study indicate a good correlation between ST syntactic complexity and NMT quality regardless of different NMT systems, and such correlation mostly centres on T-unit-related indicators, particularly on MLT. Additionally, the quality of GNMT at lexical level marks a significant advantage over that of Systran, but the performance of both systems at syntactic level does not vary significantly. With these findings in mind, it is thus hoped that translators could gain calculated insight from this study with a targeted guideline on how to pre-empt their post-MT tasks.

Nonetheless, results from this exploratory inquiry can only be interpreted tentatively as the sample size is relatively small and only legal textual domain is concerned. As the study does not involve other genres, it is uncertain that such correlation can be detected in other registers until further examination is conducted. Additionally, as the study only took large-grained measures of syntactic complexity into consideration, further research may need to explore the exact correlation

in terms of fined-grained measures to pinpoint more elaborate indicators for translators.

Acknowledgment

The study is funded by Translation Teaching Project of School of Foreign Languages in Xinhua College of Sun Yat-sen University (2019)

References

[1] Castilho, S., et al. Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics, 108(1), pp. 109-120, 2017.

[2] Wu Y., et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Computing Research Repository arXiv: 1609.08144, 2016.

[3] O'Brien, S., Balling, L. W., Carl, M. (Eds.). Post-editing of machine translation: Process and Application. Newcastle: Cambridge Scholars Publishing, 2014.

[4] Moorkens, J., et al. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation, 29(3-4), pp. 267-284, 2015.

[5] Lu, X. Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), pp. 474-496, 2010.

[6] Crego, J., et al. SYSTRAN's Pure Neural Machine Translation Systems. arXiv e-prints: 1610. 05540, 2016.