مناهج تقويم جودة الترجمة الآلية: مقياس تير TER نموذجا

Methods for assessing machine translation quality : TER scale as a model

فاتح فريخة

جامعة الجزائر 2، معهد الترجمة ، fatah.frikha.univ@gmail.com

تاريخ الاستلام: 2018/05/15 تاريخ القبول: 2018/06/15

ملخّص

نهدف في هذا البحث إلى دراسة مدى صلاحية أحد المقاييس الآلية لتقويم الترجمة الآلية، والمتمثل في مقياس تير، لإنتاج تقويم موضوعي يسمح بالاستغناء عن تقويم الإنسان الذي يعاب عليه استهلاكه للوقت والمال والجهد والذي تحكمه الذاتية في أغلب الأحيان، فتطرقنا أولا إلى التعريف بهذا المقياس، ثمّ إلى المنهج الذي يقوم عليه، بعد ذلك شرحنا طريقة حسابه لعلامة تقويم الترجمة الخاضعة للتقويم، قبل أن نختم دراستنا باختبار للمقياس.

الكلمات المفتاح: مقياس تير، تقويم الترجمة الآلية، الترجمة الخاضعة للتقويم، الترجمة المرجعية.

Abstract

This study aims at finding out to what extent one of the metrics for MT evaluation, namely TER score, is reliable for the production of an objective evaluation of MT, for a possible substitution for costly, exhausting, and subjective human evaluation for MT. Therefore, an introduction to the metric is to start with, followed by a discussion of the method TER. Then, TER method for generating MT evaluation scores is explained. Lastly, an experiment of TER is to end up with.

Keywords: TER score, MT evaluation, hypothesis translation, reference translation.

المؤلف المرسل: *فريخة فاتح*

مقدمة

لقد دفعت الحاجة إلى إنتاج تقويم موضوعي وسريع وسهل وغير مكلف للترجمة، ولاسيما الترجمة الآلية، إلى تصميم عدة مقاييس آلية للتقويم، وذلك لغايات ومآرب عديدة ومختلفة أهمها استخدامها في مخابر تطوير نظم الترجمة الآلية بهدف تقويم مدى تأثير التغييرات والتحديثات التي تطرأ على هاته النظم في كل مرة، إلا أن التحدي في تقويم الترجمة الآلية آليا يكمن في حقيقة أن الترجمة الصحيحة تقبل عدة احتمالات مختلفة في كثير من الأحيان وليس احتمالا واحدا، الأمر الذي جعل من تحقيق تقويم الترجمة آليا مهمة عسيرة ومعقدة.

وعلى الرغم من ذلك إلا أنه قد صممت بالفعل عدة مقاييس آلية لتقويم الترجمة الآلية، ولعل أهمها مما شاع بين أهل الاختصاص ومما تتوفر عليه الساحة الآن مقاييس بلو (BLEU) المصمم سنة 2002 (Papineni, Roukos, Ward, & Zhu, 2002; Papineni, Roukos, Ward, Henderson, & Reeder, 2002) ونيست (NIST) المصمم سنة 2002 (Doddington, 2002) وميتيور (METEOR) المصمم سنة 2004 (Lavie, Sagae, & Jayaraman, 2004)، غير أن هاته المقاييس يشوبها كثير من الذاتية والبعد عن الموضوعية إذ أنها تقوم على مطابقة الترجمة الخاضعة للتقويم (Translation Hypothesis) مع مجموعة من الترجمات المرجعية (Reference Translation) التي ترجمها مترجم بشري، وبالتالي فإن تقويم هاته المقاييس لا يخرج عن حكم تقويم الإنسان الذاتي وغير الموضوعي، لذلك فقد حاول الباحثون مرارا وتكرارا التفكير في القضاء على هذه الذاتية أو إنقاصها على أقل تقدير، ولعل من بين أهم الحلول المقترحة والتي من شأنها التقليص من الذاتية في تقويم الترجمة هو تصميم مقياس جديد ذي منهج مختلف عن منهج المقاييس الثلاث المذكورة سلفا يتمثل في مقياس نسبة مراجعة الترجمة (Translation Edit Rate score) المصمم سنة 2005.

فهل يمكن الاعتماد حقا على مقياس تير في القيام بتقويم موضوعي وموثوق للترجمة الآلية وبالتالي الاستعاضة به عن تقويم الانسان الذاتي والمستهلك للجهد والوقت والمال؟

يكمن الهدف من وراء دراستنا لهذا المقياس في محاولة تعريف المترجمين العرب بمختلف التكنولوجيات المستخدمة في ميدان الترجمة، والمتمثلة في نظم الترجمة الآلية علاوة على المقاييس الآلية لتقويم الترجمة الآلية، وهذا في مسعى لعصرنة مجال الترجمة في الجزائر وكافة البلاد العربية وكذلك تحيين معلومات وقدرات ومؤهلات هؤلاء المترجمين.

1-التعريف بمقياس تير

لقد قام بتصميم مقياس تير (TER Score) سنة 2005 الباحث جوزيف أوليف (Joseph Olive) (Olive, 2005) برعاية وتمويل برنامج غايل[1] (GALE Program) التابع لوكالة مشاريع البحوث المتطورة الدفاعية (Defense Advanced Research Projects Agency) المعروفة اختصارا بوكالة داربا[2] (DARPA agency).

ولقد كان الهدف من وراء تصميم مقياس تير هو التوصل إلى إنتاج مقياس بسيط يغني عن المقاييس المعقدة المرتكزة على المعارف وعن مجهودات المقوم الإنسان المضنية والمستهلكة للوقت، وبحيث يمكنه القيام بتقويم ترجمات نظم الترجمة الآلية وذلك لغاية مقارنة فعالية هاته النظم.

وعلى ذلك فإنه من بين أهم الاستخدامات التي يستهدفها مصمم هذا المقياس هي أن يدمج مع نظم الترجمة المساعدة للإنسان ((Computer-aided Translation Tools (CAT Tools)، أين يقوم بتقويم جودة الترجمة عن طريق معالجة النص المترجم لأجل تحديد كم المجهود الذي يتوجب على المترجم الإنسان أن يبدله في المراجعة البَعدية (Post-editing) للترجمة الآلية (أي؛ المراجعة بعد الانتهاء من الترجمة) حتى تصبح الترجمة الخاضعة للتقويم مطابقة للترجمة المرجعية التي يقارنها بها (Zaret, 2016, p.14).

---

[1] برنامج غايل هو البرنامج العالمي المستقل لاستغلال اللغات (غايل) (Global Autonomous Language Exploitation (GALE) program) وهو تابع لوكالة داربا وحاصل على التمويل منها أنشئ سنة 2005، لغاية استخراج المعلومات آليا من مصادر متنوعة ومتعددة اللغات مثل النشرات الإخبارية ومختلف الوثائق، وقد كان الهدف الأسمى لبرنامج غايل هو تمييز الكلام (Speech Recognition) في اللغتين العربية والصينية تحديدا وترجمته إلى اللغة الإنجليزية.

[2] وكالة مشاريع البحوث المتطورة الدفاعية (داربا) هي إحدى وكالات وزارة الدفاع الأمريكية التي تعنى بتطوير التكنولوجيات الناشئة (Emerging Technologies) لفائدة الجيش الأمريكي، أنشئت في شهر فيفري من سنة 1958 ويقع المقر الرئيس لها في ولاية فيرجينيا، إذ تعتبر المطور الرئيس للعديد من التكنولوجيات ذات الأثر الكبير على العالم بأكمله.

وبناء عليه فإن اسم مقياس تير (مقياس نسبة مراجعة الترجمة Translation Edit Rate score) يدل على طبيعة عملية التقويم التي يقوم بإجرائها على الترجمات.

يدعي مصمم مقياس تير بأنه يحقق تقويما لجودة الترجمة، تقترب جودته كثيرا من مستوى جودة تقويم الإنسان لجودة الترجمة اعتمادا على ترجمة مرجعية واحدة، وهو المستوى الذي لا يبلغه مقياس بلو حسبه إلا من خلال أربع ترجمات مرجعية (Zaret, 2016, p.14)، وهذا اعتمادا على التجارب التي أجراها على المقياس على حد تأكيده.

ولأجل بلوغ هذا المستوى من التقويم ذي الجودة القريبة جدا من جودة تقويم الإنسان كان لابد للباحث جوزيف أوليف (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 223) من اللجوء إلى عدة تغييرات أو بالأحرى تحديثات، أهمها اعتماد منهج جديد ومختلف عن مناهج المقاييس المصممة سلفا لمقياس تير أغلبها.

2-منهج المقياس

كانت الرغبة تحدو مصمم مقياس تير الباحث جوزيف أوليف لتصميم مقياس ذي تكلفة منخفضة وبسيط وغير معقد وعلى أكثر قدر ممكن من الموضوعية والموثوقية، الأمر الذي دفع به إلى الابتعاد عن المناهج القائمة على المعنى (Meaning-based Approaches) وكذلك المناهج القائمة على المعرفة (Knowledge-based Approaches)، وحثّه على اللجوء إلى الاعتماد على منهج مستقل عن اللغة ومختلف تماما عن مناهج المقاييس الأخرى (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 223) السابقة له.

وعلى ذلك فإن منهج مقياس تير يقوم على قياس المجهود الذي يتوجب على مقوّم إنسان بدله وذلك من خلال حساب عدد الأخطاء الواجب تصحيحها لأجل جعل ترجمة النظام (System Output) أو الترجمة الخاضعة للتقويم مطابقة دلاليا (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 223) لترجمة مرجعية واحدة أو مجموعة من الترجمات، في حين أن أغلب مقاييس التقويم الآلي الأخرى تقوم بالتقويم عن طريق حساب مدى مطابقة مفردات وتعابير الترجمة الخاضعة للتقويم مع مفردات وتعابير ترجمة مرجعية

أو مجموعة من الترجمات المرجعية، لذلك فإن منهج مقياس تير يختلف عن كل المقاييس الأخرى السابقة له مثل مقاييس بلو[3] ونيست[4] وميتيور[5].

3-حساب علامة التقويم

يقوم مقياس تير بحساب عدد التصحيحات التي يتوقع من المراجع الإنسان القيام بها حتى تصبح الترجمة سلسة (Fluent) ودقيقة المعنى (Accurate)، وذلك من خلال مطابقتها مع كل ترجمة مرجعية متوفرة، بحيث يخصص لكل خطأ يتوقع تصحيحه نقطة واحدة 1، إذ يشمل ذلك خمسة أنواع من الأخطاء هي الحذف (Deletion) والإضافة (Insertion) والاستبدال (Substitution) وتغيير مواضع الكلمات (Shift) وحتى تغيير مواضع علامات الترقيم (Punctuation) أو حذفها إذ تعتبر بمثابة كلمات (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 225).

[3] صمم مقياس بلو الباحث كيشور بابينيني (Kishore Papineni) رفقة ثلاثة من الباحثين هم سليم روكوس (Salim Roukos) وتود وارد (Todd Ward) وواي جينغ تشو (Wei-Jing Zhu) سنة 2002 (Papineni, Roukos, Ward, & Zhu, 2002; Papineni, Roukos, Ward, Henderson, & Reeder, 2002)، إذ يعتبر هذا المقياس من بين أهم المقاييس الآلية المصممة للقيام بتقويم الترجمة الآلية علاوة على أنه رائد في ذلك بحيث يعد من بين أولى محاولات تقويم الترجمة آليا، وهو يقوم بتقويم الترجمة من خلال مطابقة الترجمة الخاضعة للتقويم مع ترجمة مرجعية واحدة أو عدة ترجمات مرجعية، ومن ثّمة فإنه يقوم بحساب نسبة تماثل الترجمة الخاضعة للتقويم مع الترجمة المرجعية ويقدم هاته النسبة على شكل علامة، وبناء عليه فإنه كلما كانت الترجمة الخاضعة للتقويم مطابقة للترجمة المرجعية كلما كانت نسبة التماثل أكبر وكلما كان ذلك أفضل.

[4] لقد شابت مقياس بلو العديد من النقائص ونقاط الضعف (Doddington, 2002) لذلك قام الباحث جورج دودينغتون (George Doddington) بتصميم مقياس أطلق عليه اسم مقياس نيست سنة 2002 في محاولة لتدارك عيوب مقياس بلو، ولقد ارتكز دودينغتون على مقياس بلو تماما في تصميم مقياس نيست إلى درجة أن هذا الأخير يعتبر مجرد نسخة مطورة من مقياس بلو، وبذلك فهو يتبع كذلك منهج مقياس بلو نفسه في إنجاز عملية التقويم.

[5] صمم مقياس ميتيور الباحث ألون لافي Alon Lavie سنة 2004 وقام بإجراء اختبارات عليه بمساعدة من الباحثين كينجي ساجاي Kenji Sagae وشيامسوندر جيارامان Shyamsundar Jayaraman في السنة نفسها (Lavie, Sagae, & Jayaraman, 2004)، وبمساعدة الباحث سانجيف بانيرجي Sanjeev Banerjee سنة 2005 (Banerjee & Lavie, 2005)، وهذا المقياس كذلك مبني على مقياس بلو والهدف من وراء تصميمه هو تدارك النقائص والأخطاء الشائنة لمقياسي بلو ونيست معا، وبالتالي فهو يعتمد المنهج ذاته الذي يقوم عليه هذين المقياسين.

ومن ثمة يختار المقياسُ من بين مجموعة الترجمات المرجعية الترجمةَ التي حصلت الترجمة الخاضعة للتقويم على أقل عدد أخطاء بالنسبة لها، وذلك بهدف حساب علامة التقويم من خلال قسمة عدد المراجعات الذي يمثل عدد الأخطاء المتوقع تصحيحها على عدد كلمات الترجمة المرجعية[6] بالطريقة التي توضحها الصيغة التالية:

$$\text{علامة التقويم} = \frac{\text{عدد المراجعات}}{\text{عدد كلمات الترجمة المرجعية}}$$

ولأجل فهم واستيعاب أفضل لهاته الصيغة لاحظ الاختلافات الموجودة بين الترجمة الخاضعة للتقويم والترجمة المرجعية في المثال 1 الموالي:

المثال 1

الترجمة الخاضعة للتقويم

فند الروسيون التهم الموجه إليهم بخصوص تسميم جاسوس سابق مزدوج في بريطانيا بمادة كيماوية.

الترجمة المرجعية

فندت روسيا التهم الموجهة إليها والمتعلقة بتسميم جاسوس سابق بمادة كيماوية قاتلة في المملكة المتحدة.

على الرغم من أن الترجمة الخاضعة للتقويم في المثال 1 صحيحة دلاليا وسلسة القراءة إلا أن مقياس تير ووفقا للمعايير الخمس التي يقوم عليها سيعدّ ستة 6 أخطاء يتوقع تصحيحها حتى تتطابق الترجمة الخاضعة للتقويم مع الترجمة المرجعية، إذ تظهر هاته الأخطاء من خلال السطر المرسوم تحتها، وهي منقسمة على ثلاثة 3 استبدالات وحذف واحد 1 وإضافة واحدة 1 وتغيير واحد 1 في مواضع الكلمات، فأما الاستبدالات فهي

[6] يقصد بالترجمة المرجعية ترجمة جاهزة سلفا أنجزها مترجم إنسان محترف، تستخدم في ميدان التقويم الآلي للترجمة الآلية بصفتها مرجع أو ترجمة نموذجية لأجل مقارنة ترجمة خاضعة للتقويم بها ومن ثم تصحيحها، تكون الترجمة المرجعية في أغلب الأحيان عبارة عن جملة منفردة تامة دلاليا كما يظهر في المثال 1 تمثل حين جمع العديد منها مدونة (مدونة تقويم) (Papineni, Roukos, Ward, & Zhu, 2002; Papineni, Roukos, Ward, Henderson, & Reeder, 2002).

(روسيا) بـ (الروسيون) و(بخصوص) بدلا من (والمتعلقة بـ) و(بريطانيا) مكان (المملكة المتحدة)، وأما الحذف فيتمثل في كلمة (قاتلة) الحاضرة في الترجمة المرجعية والمتغيّبة عن الترجمة الخاضعة للتقويم، في حين تظهر الإضافة في كلمة (مزدوج) الغائبة عن الترجمة المرجعية والموجودة في الترجمة الخاضعة للتقويم، بينما يتمثل تغيير موضع الكلمات في العبارة (بمادة كيماوية) التي جاءت في الترجمة المرجعية بعد عبارة (جاسوس سابق) مباشرة وقبل التعبير (في المملكة المتحدة)، غير أنها أُخّرت في الترجمة الخاضعة للتقويم إذ ختمت بها الجملة بعد عبارة (في بريطانيا) البديلة للتعبير (في المملكة المتحدة).

المعايير	عدد المراجعات
الاستبدال	3
الحذف	1
الإضافة	1
تغير مواضع الكلمات	1
المجموع	**6**

الجدول I: عدد المراجعات التي عددها مقياس تير في الترجمة الخاضعة للتقويم التي في المثال 1

وفقا لهاته المعطيات وعلما بأن عدد كلمات الترجمة المرجعية هو 15 فإن علامة تقويم مقياس تير للترجمة الخاضعة للتقويم التي في المثال 1 تحسب كالآتي:

$$\text{علامة تقويم تير} = \frac{6}{15} = 0.4 = 40\ \%$$

وبالتالي فإن نسبة المراجعة تبلغ 40 % وهو معدل عال جدا يشير إلى أن المجهود المتوقع أن يبدل في مراجعة الترجمة الخاضعة للتقويم لأجل أن تصبح سلسة ودقيقة دلاليا يصل إلى نصف المجهود تقريبا الذي كان سيبدله مترجم إنسان في ترجمتها، وهو ما لا يتوافق واقعيا مع الترجمة الخاضعة للتقويم التي تبدو على قدر من السلاسة والدقة الدلالية، وهو الأمر الذي لا يمكن للمقياس تمييزه إذ يتطلب معارف الإنسان بالعالم الحقيقي كي يدركه وهي المعارف التي لا يملكها المقياس الآلي، وهذا لأن الإنسان لم يتمكن بعد من بلوغ تطور تكنولوجي كاف يمكنه من برمجة هاته المقاييس الآلية بأنظمة وبرامج تمكّنها من إدراك معارف العالم الواقعي.

يحسب مقياس تير عدد المراجعات (Edits) أو بالأحرى التصحيحات اللازمة في ترجمة خاضعة للتقويم على مرحلتين، تتمثل أولاهما في حساب عدد الإضافات والمحذوفات والمستبدلات من خلال البرمجة الدينامية[7] (Dynamic Programming)، أي كل واحدة منها على حدة، بينما تكمن ثانيهما في إجراء بحث دقيق عن تغييرات مواضع الكلمات مع التشديد على ضرورة اختيار التغيير الذي يحتوي على أقل عدد من الكلمات، وهي القاعدة التي تظهر في المثال 1 أين اختار المقياس التعبير (بمادة كيماوية) بصفته تغيير في مواضع الكلمات ولم يختر التعبير (في المملكة المتحدة) الذي يعتبر تغيرا في مواضع الكلمات أيضا لأنه ورد في الترجمة المرجعية بعد التعبير (بمادة كيماوية) في حين جاء في الترجمة الخاضعة للتقويم قبله، وذلك لأن التعبير الأول يحتوي على كلمتين بينما يشتمل التعبير الأخير على ثلاث كلمات.

قاعدة أخرى مهمة من قواعد مقياس تير هي أنه عندما يجد العديد من الترجمات المرجعية المقابلة للترجمة الخاضعة للتقويم، فإنه يختار من مجموعة الترجمات المرجعية هاته الأقرب دلاليا إلى الترجمة الخاضعة للتقويم، ويدرك المقياس أيها أقرب من خلال حساب عدد المراجعات إذ تمثل الترجمة المرجعية التي تحتوي الترجمة الخاضعة للتقويم على أقل عدد من التصحيحات بالنسبة لها الترجمة المرجعية الأقرب دلاليا إلى هاته الترجمة الخاضعة للتقويم.

4-اختبار المقياس

[7] البرمجة الدينامية هي طريقة رياضية تستخدم في علوم البرمجة لحل المسائل المعقدة من خلال تقسيم كل مسألة إلى مسائل فرعية بهدف إيجاد حل لكل مسألة فرعية على حدة، ومن ثم القيام بذلك إلى غاية حل كل المسائل الفرعية الذي يمثّل حلا للمسألة الأساس.

أجري اختبار على مقياس تير من خلال مقارنة نتائج عملية تقويم المقياس مع نتائج تقويم مقياسين آخرين هما مقياس بلو ومقياس ميتيور، ومن ثمّ حساب نسبة تماثل هاته المقاييس مع نتائج تقويم الإنسان، وذلك عن طريق حساب علاقة التماثل[8] (Pearson Correlation Coefficient) بين نتائج هاته المقاييس كما يظهر في الجداول II وIII وIV، علما بأن تقويمات مقاييس تير وبلو وميتيور جرت باستخدام أربع 4 ترجمات مرجعية، لذلك فإن الرقم أربعة 4 بين قوسين الذي يظهر في الجدول II على تير (4) وبلو (4) وميتيور (4) إنما يدل على أربع 4 ترجمات مرجعية.

المقاييس	نظام1	نظام2	المعدل
تير (4)	%53.2	%46.0	%49.6
بلو (4)	%73.5	%62.1	%67.8
ميتيور (4)	%46.0	%39.2	%42.6

[8] تعرف علاقة التماثل (Pearson Correlation Coefficient) كذلك بالمصطلحين (Pearson R Correlation) و(Pearson Correlation) بينما التسمية الكاملة لها هي (Pearson Product-moment Correlation Coefficient (PPMCC))، وتقوم وظيفة علاقة التماثل على قياس التماثل الخطي (Linear Correlation) القائم بين متغيرين x وy من خلال صيغة رياضية، إذ ينتج عن قياس هذ التماثل قيم تتراوح بين (+1) و (−1)، أين تعبر القيم الموجبة عن تماثل خطي موجب بينما تعبير القيم السلبية عن تماثل خطي سلبي، وبناء على ذلك فإن القيمة (+1) تمثل تماثلا خطيا موجبا تماما في حين تمثل القيمة (−1) تماثلا خطيا سلبيا تماما، بينما تعبر القيمة (0) على أنه لا يوجد أي تماثل خطي على الإطلاق بين المتغيرين المعنيين، ويشيع استعمال هذه العلاقة في العلوم التقنية بعدما قام بتطويرها الباحث كارل بيرسون (Karl Pearson) بناء على فكرة يعود الفضل في ابتكارها إلى الباحث فرانسيس غالتون (Francis Galton) في ثمانينات القرن التاسع عشر.

الجدول II: **علامات تقويمات مقاييس تير وبلو وميتيور لنظامي الترجمة 1 و2** (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 229)

أجري التقويم على مدونة إلكترونية تتألف من مائة 100 جملة محررة باللغة العربية مع ترجمتين 2 مقابلتين لكل جملة من هاته الجمل في اللغة الإنجليزية باستخدام نظامي ترجمة لم يذكر اسميهما، إنما أشير إليهما فقط بالرمزين نظام1 ونظام2 مع ذكر بأن نظام1 يعد واحدا من بين أقل نظم الترجمة أداءً في حين يعتبر نظام2 من أحسنها، كما تحتوي المدونة كذلك على أربع 4 ترجمات مرجعية محررة باللغة الإنجليزية مقابل كل جملة ترجمة في اللغة الإنجليزية (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 227).

تجدر الإشارة إلى أن مقياسي بلو وميتيور يقومان بتقويم الترجمة من خلال مطابقة الترجمة الخاضعة للتقويم مع الترجمة المرجعية، ومن ثّم فإنهما يقومان بحساب نسبة تماثل الترجمة الخاضعة للتقويم مع الترجمة المرجعية ويقدمان هاته النسبة على شكل علامة، وبناء على ذلك فإنه كلما كانت الترجمة الخاضعة للتقويم مطابقة للترجمة المرجعية كلما كانت نسبة التماثل أكبر وكلما كان ذلك أفضل.

وهذا على عكس مقياس تير الذي يقوم بحساب نسبة مراجعة الترجمة كما رأينا في العنصرين السابقين (2 و3)، وهي النسبة التي تمثلها النتائج في الجدول II، لذلك فإنه كلما كانت النسب أكبر كلما كان ذلك أسوء وكلما كانت أقل كلما كان ذلك أحسن.

وبالتالي فإن نتائج تقويم مقياسي بلو وميتيور التي في الجدول II تمثل نسب تماثل الترجمة الخاضعة للتقويم مع الترجمة المرجعية، أما النسب المتبقية لتبلغ 100% فهي تمثل الجزء من الترجمة الخاضعة للتقويم الذي لا يتماثل مع الترجمة المرجعية والذي يعتبر بذلك خاطئا وتجب مراجعته أو تصحيحه أو تعديله، وبذلك فإن هاته النسبة المتبقية التي تمثل نسبة مراجعة الترجمة والتي نحصل عليها من خلال طرح نتائج مقياسي بلو وميتيور التي في الجدول II من 100 هي التي تتوافق مع علامات مقياس تير وهي التي يمكن مقارنتها مع نتائج هذا المقياس التي تمثل نسب مراجعة الترجمة، وبعد فعل ذلك نحصل على النتائج الممثلة في الجدول III.

المقاييس	نظام1	نظام2	المعدل

تير (4)	%53.2	%46.0	%49.6
بلو (4)	%26.5	%37.9	%32.2
ميتيور (4)	%54.0	%60.8	%57.4

الجدول III: نسب مراجعة الترجمات المحصّلة من مقاييس تير وبلو وميتيور

نلاحظ من خلال الجدول III بأن مقياس ميتيور أنتج نسبة مراجعة أكبر من تلك التي حصّلها مقياسي بلو وتير في حين قدّم مقياس تير نسبة أعلى من نسبة مقياس ميتيور، كما يظهر من الجدول كذلك بأن النسب المحصّلة من قبل مقياسي تير وميتيور أقرب إلى بعضها من النسب المقدّمة من قبل مقياس بلو الذي حقق نسبا أقل من المقياسين الآخرين مع فارق كبير عنهما أيضا، غير أننا لا نعلم أيٌّ من نسب هاته المقاييس أقرب إلى الصواب إلا إذا قارناها مع علامة تقويم الإنسان للترجمة الخاضعة للتقويم نفسها، وذلك من خلال حساب علاقات التماثل بين علامات هذه المقاييس الثلاثة وعلامة تقويم الإنسان، وهو الأمر الذي يوضحه الجدول IV، مع العلم بأن تقويم الإنسان جرى من قبل مراجعين اثنين على معياري الفصاحة (Fluency) والملاءمة[9] (Adequacy)، وذلك من خلال وضع كل واحد من المراجعَين علامة على كل واحد من المعيارين، ومن ثم جمع علامتي معياري الفصاحة والملاءمة الخاصة بكل مراجع وقسمة المجموع على اثنين لنحصل على علامة التقويم النهائية الخاصة بكل مراجع، وبعد ذلك جمع علامتي التقويم النهائيتين للمراجعين وقسمة المجموع على اثنين لنحصل على علامة تقويم الإنسان للترجمة الخاضعة للتقويم.

المقاييس	تقويم الإنسان

[9] تجدر الإشارة إلى أن مصطلح الملاءمة (Adequacy) يستخدم في مجال الترجمة الآلية ويقصد به مفهوم الدقة (Accuracy) الذي يستخدم في مجال ترجمة الإنسان خصوصا، وبذلك فهذين المصطلحين مترادفين ويقصد بهما المفهوم ذاته.

تير	0.539
بلو	0.391
ميتيور	0.550

الجدول IV: علاقات تماثل تقويمات مقاييس تير وبلو وميتيور مع تقويم الإنسان (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 229)

يتضح من الجدول IV بأن مقياس ميتيور هو الذي حصل على أعلى نسبة تماثل مع تقويم الإنسان إذ وصلت إلى (0.550)، ثم يأتي بعده مقياس تير الذي حصل على ثاني أعلى نسبة وقريبة جدا من نسبة مقياس ميتيور إذ بلغت (0.539)، بينما حل مقياس بلو أخيرا إذ حصل على أقل نسبة تماثل مع تقويم الإنسان قدرت بـ (0.391) وهي بعيدة جدا عما حققه مقياسي ميتيور وتير.

نستنتج من هاته النتائج بأن جودة تقويم كل هاته المقاييس الثلاثة للترجمة الآلية بعيد كل البعد عن جودة تقويم الإنسان لها، إذ لا تتجاوز نسب تماثل تقويم مقياسي ميتيور وتير مع تقويم الإنسان النصف إلا بنزر قليل 55% و53.9% على الترتيب، في حين لم تبلغ نسبة مقياس بلو النصف حتى 39.1%، وهذا إن دلّ على شيء فإنما يدل على أن جودة التقويم الذي تنجزه هاته المقاييس الثلاث تبقى متوسطة جدا وبعيدة كثيرا عن جودة تقويم الإنسان للترجمة عموما والترجمة الآلية على الخصوص، الأمر الذي يضع هاته المقاييس موضع الشك والريبة من حقيقة قدرتها على القيام بمختلف مهام تقويم الترجمة وأهدافه، وكذلك قدرتها على استخلاف المراجع الإنسان على الأقل في إنجاز تقويم ذي جودة مقبولة إلى حد ما.

لعل من بين أهم الأسباب التي ساهمت في نتائج مقياس تير المتوسطة والبعيدة عن مستوى جودة تقويم الإنسان للترجمة الآلية، هو معيار تغيير مواضع الكلمات الذي يدرج في عداد المراجعات كل تغيير يطرأ على ترتيب الكلمات أو التعابير في جمل الترجمة، على الرغم من أن هاته التغييرات قد لا تؤثر بتاتا على معاني ودلالات جمل الترجمة في غالب الأحيان، وهو ما من شأنه أن يرفع من نسبة مراجعة الترجمة الخاضعة للتقويم على الرغم من أن الأمر لا يستلزم ذلك في كثير من المرات، في حين أن المقوّم الإنسان لا يعتبر تغيير مواضع الكلمات

والتعابير أو ترتيبها إن صح التعبير خطأ ولا يخصم أي نقاط من الترجمة الخاضعة للتقويم على تلك التغييرات، ولاسيما في حالة ما إذا كانت الترجمة الخاضعة للتقويم تؤدي الدلالة نفسها التي يؤديها النص الأصلي ولا تنقص منها شيئا.

نوع الخطأ	علامة مقياس تير
الإضافة	%4.6
الحذف	%12.0
الاستبدال	%25.8
تغيير مواضع الكلمات	%7.2
المجموع	**%49.6**

الجدول V: **نتائج تقويم مقياس تير للترجمة الخاضعة للتقويم**

(Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006, p. 228)

لذلك فإن حل هذا الإشكال لغاية الحصول على تقويم ذي جودة أحسن من قبل مقياس تير يكمن في حذف معيار تغيير مواضع الكلمات من قائمة المعايير التي يقوم عليها هذا المقياس، خاصة وأن نسبة معتبرة من الأخطاء التي يحصيها مقياس تير (7.2%) إنما تقع في صنف معيار تغيير مواضع الكلمات كما يظهر من الجدول V.

خاتمة

يظهر من خلال دراسة مقياس تير بأنه يمكن استخدامه لأهداف تطويرية في مخابر تصميم وإنتاج وتطوير نظم الترجمة الآلية، إذ كل ما يحتاجه المطورون هو تقويم روتيني وسريع وموضوعي وغير مكلف ومتناسق وهو ما يمكن أن يضمنه مقياس تير، على عكس تقويم الإنسان الذي يتسم بالبطء والذاتية والتنافر، إذ لو أعطيت مراجعا إنسانا النص نفسه ليقوم بتقويمه على مرتين في فترتين متباعدتين، بحيث يكون قد قام بتقويم عدد كبير جدا من نصوص بينهما إلى درجة يكون في المرة الثانية قد نسي كيف قوّم هذا النص في المرة الأولى وما هي العلامة التي منحها إياه، فإنه سوف لن يعطي هذا النص في المرة الثانية العلامة نفسها التي منحها له في المرة الأولى لا محالة، بينما مقياس التقويم سيفعل العكس من ذلك تماما وسيمنح العلامة نفسها ما لم تجري على هذا المقياس أية تحديثات بين عمليتي التقويم الأولى والثانية.

وعلى الرغم من ذلك إلا أن مقياس تير لا يمكن الاعتماد عليه في القيام بتقويم الترجمة الآلية لأهداف أخرى مثل تقويم ترجمات نظام ترجمة معين لمختلف أنماط النصوص لأجل مقارنة مدى جودة ترجمة نظام معين لكل نمط من النصوص، أو تقويم ترجمات نظم مختلفة لنوع النصوص نفسه أو أنواع مختلفة لغاية المقارنة أيضا.

كما أنه لا يمكن في الوقت الحاضر الاستعاضة بأي من مقاييس تقويم الترجمة سواء أكان مقياس تير أو غيره من المقاييس عن تقويم الإنسان للترجمة الآلية ولا للترجمة عموما، خاصة وأن هاته المقاييس لا تخلو من الذاتية التي تعاب على تقويم الإنسان وحتى ترجمته، ولاسيما وأن هاته المقاييس تعتمد كلية على ترجمات مرجعية منجزة من قبل إنسان يطبع ترجمته كثيرا من الذاتية والنزعة الإيديولوجية في غالب الأحيان كما هو معلوم، وبالتالي فإن الاعتماد على ترجمة الإنسان الغارقة في بحر الذاتية لإنتاج تقويم موضوعي للترجمة يبدو متناقضا للغاية بل بالإمكان وسمه حتى بالمفارقة.

لذلك ولأجل محاولة تصميم مقاييس تتسم بالموضوعية لتقويم جودة الترجمة الآلية في المستقبل لابد من التفكير في إيجاد طريقة أو منهج لفعل ذلك بحيث يبتعد كل البعد عن الاعتماد على ترجمات الإنسان سواء لاتخاذها كترجمات مرجعية بهدف مقارنة الترجمات الخاضعة للتقويم بها أو استخدامها بأي شكل من الأشكال الأخرى، وبذلك فلابد أن تكون هاته المقاييس مستقبلا مستقلة استقلالا تاما عن الإنسان وذاتيته، وحبذا لو تكون كذلك مستقلة عن اللغة أيضا كأن تعتمد مثلا على طريقة أو منهج الإحصاء بالطريقة التي تفعل ذلك نظم الترجمة الإحصائية (Statistical Systems) أو نظم الترجمة القائمة على الإحصاء (Statistics-based

(Systems، إذ بالإمكان الاستفادة من النتائج التي توصلت إليها أبحاث مطوري نظم الترجمة القائمة على الإحصاء في هذا المجال.

المصادر والمراجع

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 65-72). Ann Arbor, Michigan: Assossiation for Computational Linguistics.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research* (pp. 138-145). Morgan Kaufmann Publishers Inc.

Lavie, A., Sagae, K., & Jayaraman, S. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In R. E. Frederking, & K. B. Taylor (Ed.), *In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004* (pp. 134-143). Washington DC: Springer-Verlag.

Olive, J. (2005). Global autonomous language exploitation (GALE). *DARPA/IPTO Proposer Information Pamphlet.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Assossiation for Computational Linguistics* (pp. 311-318). Philadelphia, Pennsylvania: Assossiation for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., Henderson, J., & Reeder, F. (2002). Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results. *Proceedings of Human Language Technology*, (pp. 132-137). San Diego.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. 200*, pp. 223-231. Cambridge, MA.

Zaret, A. (2016). A Quality Evaluation Template for Machine Translation. *Translation Journal, 19*(1). Retrieved January 3, 2016, from http://www.translationjournal.net/January-2016/a-quality-evaluation-template-for-machine-translation.html