Developing a 'Neo-Metric' for the Evaluation of Translation Memory

Gareth Watkins

Cardiff University, UNITED KINGDOM

ABSTRACT

A novel system by which Translation Memory (TM) tools can be evaluated would be desirable for translation industry stakeholders. Such evaluation systems are commonly known as metrics. While metrics devised to evaluate Machine Translation (MT) are well developed, there has been limited work on the creation of a TM tool metric. This paper discusses the creation of a 'Neo-Metric' designed to help translators, translation professionals and translation industry stakeholders as they evaluate TM software. Having identified a need for such a metric, a brief literature review is conducted. Variables to be included in the Neo-Metric are discussed, and an initial metric is introduced. This first version of the metric is briefly tested, weaknesses are identified, and based on these initial tests, a refined metric is developed for more in-depth testing along with testing in the field, both to be discussed in future papers.

KEYWORDS: CAT tools, evaluation, metric, translation memory

1. Introduction

A translation memory (TM) system[1] is a software program that aids human translators by, among other things, saving source text (ST) segments together with target text (TT) segments as a Translation Unit (TU) in a TM database as the translator translates. In this way the TUs can be reused for the translation of similar or updated documents or within the same document if any segments are repeated.

Furthermore, if no exact match[2] exists, the TM system will return any similar segments – more commonly known as fuzzy matches. If the sentence or similar sentences do not exist in

[1] Commonly used examples include SDL Trados Studio 2015, Déjà Vu X3, Wordfast Classic and Wordfast Pro.
[2] Also known as a 100% match.

the database the translator will translate as normal and the TM system will save the new sentence and its translation as a TU in the database for future use. Many TM systems also offer a type of match which can be labelled as a Context Match (CM). In fact, Context Match is the term used by SDL Trados. Other tools use other terms, for instance memoQ uses the term *101% match*, while Déjà Vu X3 uses the term *Guaranteed Match*. In essence, what is described by these tool-specific terms is a match which is likely to be extremely accurate because it has appeared in exactly the same context in the new ST as it has in past documents.

TM software has also been referred to as a 'TM tool', a 'TM system' and 'TM technology' in the literature (see, for example, Austermühl 2001; Bowker 2002; Kenny 2011). In this paper the term TM system is used. Furthermore, a clear distinction is made between the TM system and the TM database which the system builds. Moreover, when both system and database are under discussion, the term TM is used without reference to tool or database.

Metrics, generally speaking, are systems which aid evaluators as they go about the business of evaluating a tool. While metrics devised to evaluate Machine Translation (MT) are well developed, there has been limited work on the creation of a metric designed to assist translators, Language Service Providers (LSPs) and clients alike during evaluation of TM tools. Reinke (2000) makes a distinction between MT and TM systems, in that while MT creates translations, a TM system simply stores TUs and retrieves them for re-use when needed. He adds that while data processing in TM equates to data retrieval, data processing in MT equates to data production. Further, Reinke asserts that as the technologies are so different, the need for different evaluation criteria is implied.

A 'Neo-Metric' which would attempt to quantify the usefulness of a TM system, or which would quantify the gain in efficiency experienced when using TM, is desirable. This research was conducted in the context of the Welsh language translation industry.[3] There is a lack of awareness of TM in Wales, and the potential benefits of TM require underlining to translators and clients alike. A method of evaluating TM systems may assist in this respect. The creation of a methodology that would allow industry stakeholders to evaluate TM would be of benefit to everyone involved in the Welsh translation industry.

[3]There is no reason why the methodology could not be used for language pairs other than Welsh/English however.

The main purpose of developing this 'Neo-Metric', then, is to evaluate a TM system's ability to retrieve and suggest useful matches, and, by extension, the system's matching algorithm. This with a view to assisting potential purchasers in choosing a system, or to helping stakeholders ascertain whether TM is appropriate to their needs.

Before designing a metric it was necessary to discern whether or not differences between different TM tools are noticeable and whether or not the tools are different enough to justify the effort made in testing individual tools. Had no noteworthy differences been discovered, then a simple comparison of translation using a single TM system with translation performed manually, similar to work conducted by Brkić et al. (2009), could have been conducted. There would be no need to test individual tools if there were no discernible differences between tools. However, the tools were found to be significantly different in respect of matches returned and fuzzy scores assigned to those matches.

In order to devise this metric, in this paper a short literature review is conducted, the main aims of the metric are set out, and a metric is proposed, tested and developed in readiness for testing in the 'real world'.

2. Existing Metric Research

While TM evaluation literature is comparatively rare, it certainly does exist. However, as noted by Whyman and Somers (1999:1,269), "much TM evaluation […] centers on comparing features of different TM systems, or concentrates more particularly on the overall user-friendliness of the software product". The evaluations conducted by Benis (1999; 2007), Holloway (1996—referenced in Whyman and Somers, 1999), Shadbolt (2002) and Mikuličková (2010) certainly come under these headings. Zerfass (2002a) lists some basic functions that can be compared and compares some features of different TM tools in a later article (Zerfass 2002b). Höge (2002:31-36), too, provides a brief comparison of some of the key TM technologies of the time, although this is not the main focus of her research. Www.proz.com provides a software comparison tool.[4] In addition, "individual tools are often

[4] See http://www.proz.com/software-comparison-tool (accessed 1 February 2015)

reviewed in professional journals" (Kenny 2006:305). Moreover, "there are only isolated instances in the literature of experimental research or attempts to provide general frameworks for those wishing to pursue research in the area (see for example, Bowker 2003, 2005)" (Kenny 2006:305).

Much research has been conducted into the *potential* gain in efficiency when using a TM system, if not a methodology for measuring that gain. Somers (2003b:42), for instance, argues that while a TM may increase productivity by up to 60% on occasion, it is unlikely that such a high gain will be seen every time a TM is used. Somers suggests that it would be more realistic to expect an average of a 30% gain in productivity, and that, as this figure is an average, on occasion the gain in productivity will be much lower. While reviewing and comparing different TM products, Benis (1999) wrote "[a]ll the programs tested will significantly increase your productivity on the sort of texts that are good TM fodder". TM is most useful when a text is either very repetitive internally, or when a text has a high degree of similarity with related or similar texts (Hartley 2009:117). Gow (2003:14) advises that a TM system is most useful when used with repetitive texts such as business/commercial texts, legal texts, scientific texts or technical texts. The efficiency gained is highly dependent on the type of texts translated.

Due to the lack of standards in Language Technology (LT) evaluation, the European Commission funded the Expert Advisory group for Language Engineering Standards (EAGLES) project 1993-1996 (Quah 2006:143). The EAGLES group set about creating a flexible, modifiable and user-centric framework for evaluation (Quah 2006:143). Several benchmark tests were specified for TM. Of most interest here is the suggested evaluation process for exact and fuzzy matches.

When considering exact matches, EAGLES advise that a corpus of related bi-texts[5] should be created. A subset of the corpus should then be used to populate a TM database. A different text or texts from the corpus should then be translated with the TM system using the newly populated TM database. The percentage of segments translated and the percentage of correct translations should be noted and then scores should be calculated "based on notions like recall

[5] A bi-text is a 'merged document composed of both source- and target-language versions of a given text' (WordSense.eu Online Dictionary 2016).

and precision" (EAGLES 1995). EAGLES correctly note that the results will depend largely on the materials collected for the corpus (EAGLES 1995). EAGLES do not state what size the corpus should be, however in discussing a TM database's size more generally, state that "[t]he optimum size is likely to depend on text and translation type" (EAGLES 1995). In the EAGLES spirit of flexibility, presumably both corpus and text for testing should be as large as possible, within the constraints[6] of the evaluation project.

In any case, Reinke (2000) argues that there are two major problems with EAGLES' test scenario: firstly, that the corpus is created using related texts (that is, relating to the same subject field and text type) is not sufficient. Reinke insists that the corpus should be created using a text and an updated or modified version of that text. This may be a mistake, because, if the evaluation is to be user-centric, the texts used to evaluate the system should be representative. If the user does not normally translate texts that are related in the sense that Reinke uses, then the results of the evaluation may not be appropriate to the end user. What Reinke suggests will, however, give a general measure of the tool's performance.

Secondly, Reinke argues that, as TM systems do not translate as such, the correctness of the translation retrieved should not be measured. Rather, the relevance of the suggestion should be measured.

When considering fuzzy matches, EAGLES advise that, once again a TM database should be created. A text should be run through the TM system then systematically modified or changed before being run through the TM system again. The recall of the TM system should be measured following the changes. Reinke (2000) criticizes this approach as being too vague, in that it "does not specify more complex syntactic and semantic variations". The EAGLES framework has influenced or has been used by many researchers as a point of departure for their work (see, for instance, Reinke, 2000; Rico, 2001; Palacz, 2003 and Gow, 2003).

Gow believes the time wasted in reading through suggested fuzzy matches to be an important consideration. As such, in designing her metric, Gow weights against fuzzy and multiple fuzzy matches. Clearly, a translator will need to put extra effort into reading the fuzzy matches; however Gow's weighting is questionable. Gow initially assigns a time loss penalty

[6]That is, constraints in respect of time, manpower and money dedicated to evaluation.

of -1 for a 100% match, -2 for one fuzzy match and -3 for multiple fuzzy matches. Later on in her thesis, following initial testing of the metric, she modifies the penalties, but as the modifications are built on the original penalties they still seem arbitrary. While Gow's metric provided useable results, the logic behind the values assigned to the penalties is not instantly understandable. As such, her metric may not closely relate to translators in the real world.

Both Hodász (2006) and Whyman and Somers (1999) ignore 100% matches, valid choices in the context of their own work; however, when considering productivity, exact matches are hard to ignore. Translators should not be interested in translating un-translated, unconnected patches of text,[7] but documents as a whole. The 100% matches should not be filtered out because translators will need to see the un-translated segments in context. In addition, while translators may save some time as they are not reading the TT or ST, any time gained in ignoring 100% matches while translating could potentially be lost during additional editing while proofreading.

Whyman and Somers' metric aims to evaluate the relationship between the level at which the fuzzy threshold is set and how useful a match is. They hope to "provide a method for determining the cut-off point, the 'usefulness threshold"' (Whyman and Somers 1999:1269). The user interface and general user friendliness of the tool are ignored, as the function of the matching algorithm is what is considered important here. This is highly appropriate. A human translator will be able to adapt to a program, given enough time. However, the algorithm that calculates the fuzzy matches to be suggested is outside of the control of all but the developer.

Whyman and Somers liken TM system functionality to Information Retrieval (IR) software, and use IR performance measures of recall and precision in their metric. In order to weigh the precision and recall results, Whyman and Somers attempt to create a way of deciding how good a suggestion is in a way that is not subjective. They decide that measuring the number of keystrokes required to adapt a suggestion in order that it becomes an appropriate translation (also known as edit distance) would be a good measure of how much effort a translator would need to expend. They base their calculations on the Levenshtein distance algorithm.[8]

[7]Although, according to García (2009:201), this is what translators are sometimes or indeed often asked to do.
[8]Levenshtein distance is calculated by defining insertions, deletions and substitutions as edit operations and assigning a cost of 1 to each of these operations (Leusch et al. 2003:240).

Whyman and Somers (1999:1274) create a modified version of the Levenshtein distance algorithm, one that takes into account the modern word processor's ability to simplify some tasks with the use of a mouse controlled 'drag and drop'. They calculate keystrokes as follows:

Operation	Key-strokes	Explanation
Insert	1 + *c*	click to locate + number of characters to be typed in
Delete	2	highlight + cut
Replace	1 + *c*	highlight + number of characters to be typed in
Move	2	highlight + drag
Swop	4	= 2 moves
Adjacent swop	2	where strings are adjacent

(Whyman and Somers 1999:1,275)

Whyman and Somers are not alone in using a form of edit distance for evaluation (see, for instance, Akiba et al., 2001; Akiba et al., 2003 and Civera et al., 2005). Despite this, Hodász (2006:2,046) maintains that counting keystrokes "could be subject to criticism regarding subjectivity and small numbers of judges". Whyman and Somers recognize that, as what constitutes a good translation is subjective (and so edit distance may be different depending on who performs the editing), the best way in which to calculate the appropriateness of a match is by counting the number of steps it would take to change the ST of the match into the original ST. It is the current author's belief that, if evaluation is to be as scientific as possible, then any element of subjectivity should be minimised. However, if an evaluator is attempting to evaluate anything for his/her own use, then subjectivity is less of an issue. Translators will have their own translation style, so translation decisions and decisions as to how to modify a match will be similar or relatively consistent when similar or identical translation problems or TM suggestions are presented.

Giving suggested matches a score based on any form of edit distance, or in fact any score whatsoever, while giving the translator an idea of how efficient the system under test is, does not show the translator what that efficiency means in real terms. Tate sums this issue up best when discussing MT evaluation:

> How far off is a metric score of .35 from a score of .50 when you are dealing with translated outputs? Is the .15 score difference really that significant? […] [T]here

> have been no empirical results indicating how useful a translation is based on these scores (Tate 2008:1).

Pym (2011:6) insists that, to those providing translation or translation services, the most important consideration is the amount of time invested in a task. While other considerations are no doubt important, the importance of time cannot be denied. However, time has been measured by only a handful of researchers during experiments involving TMs, for example, Guerberof (2008; 2009), Yamada (2011) and Wallis (2006).

This lack of interest in time as a variable (much less an output score) is hard to resolve or understand, considering that, as Gow would have it, "[a]nother element in determining match quality is time. If two matches are equally valid, the better of the two is the one that saves the user the most time" (2003:48).

The short review of literature identifies several themes, most important of which is the relative lack of attention given to time. Despite possibly being the most relevant variable to translators, time is largely ignored as a variable, and metrics do not advise of or output a figure of potential time savings when a particular system is used.

Other than Gow (2003), who levies additional penalties for multiple matches, it appears that there is a general lack of interest in the effect of multiple matches on productivity. Neither Hodász (2006) nor Whyman and Somers (1999) concern themselves with multiple matches nor 100% matches when designing their respective metrics. Guerberof's (2008; 2009) setup does not allow her to consider the impact of multiple matches on translation speed. Wallis (2006:69) notes that, when discussing corpus-based resources, Bowker advises that less experienced users are more likely to read many suggestions where possibly only one or two need consideration. Wallis argues that, as a TM system is in some respects a kind of corpus-based tool, the same can be said to be true of how people treat matches suggested by the TM system. That is to say that a less experienced of TM will more likely read multiple suggestions even if reading only one or two is necessary. The current author suggests, therefore, that ignoring time lost when multiple matches are suggested is a mistake. Multiple matches are a reality of TM that will affect the efficiency of TM, although this is dependent on the translator. If the retrieval function of a TM system and its effect on productivity is to be evaluated, then a more holistic approach, which considers all types of suggestion or match,

needs to be devised. Moreover, despite being criticized by Höge (2002:171), edit distance should be utilised in this holistic approach as it is an effective and relatively simple method of evaluation.

The themes identified in this short review of available metric literature provide a road map for the creation of the Neo-Metric.

3. Developing the Neo-Metric

The first aim is to create a metric to evaluate match retrieval in TM systems in a way that provides useful results, thus enabling comparisons between translation without the use of TM systems and translation using different TM systems.

It could be proposed that the simplest way to test would be to time human translators with and without TM. Human translators could be asked to participate in two separate timed experiments. A text would be translated without the aid of a TM system, then re-translated with a TM system.

However, Höge advises that such methodology would result in the test subject having to complete the same task twice. Höge (2002:133) notes that "[c]onsequently, in the first test round a translator would be [presented] with more problems, for which he/she has to develop strategies than in the second case". She further notes that asking a translator to complete one task without software support and a different task with support would also affect the test results, as would using one translator to translate a text without support and a different translator to translate with support.

One way of solving this issue is to ensure that a suitable period of time will have elapsed between the two experiments in order that the translator will have had time to forget the original text and translation. However, this method would also be problematic as will be illustrated in the following paragraphs.

Gow (2003) does not include human translators in her metric. That the skills of translators are not considered is a mistake. One of the oft-mentioned advantages of TM is that the translator

is left in control. The translator, then, is part of the process of translation when TM is used. As the translator is part of the process during real life use, the translator's skills and abilities, or more accurately the translator's level of skill and level of ability, should be part of the process when evaluating a tool for said translator.

That is not to say that factors influencing the performance of a human translator should be incorporated into the metric. Indeed, doing so should be avoided as it would create too many variables. Such factors include, but are not limited to, sleep deprivation (Dinges 1992:177; Tilley and Brown 1992:242), consumption of caffeine (Lieberman et al. 2002:260), health (Smith 1992:215), motivation (Fitts and Posner 1967:26), temperature (Brooke and Ellis, 1992:126; Hygge 1992:95), light levels (Megaw 1992:261) and possibly ionization levels (Farmer 1992:257).

Due to the numerous factors that could potentially affect a translator's performance, a test schedule whereby a translator is timed while translating a text with a TM system one week and timed again with the same text but without the use of TM the next, or vice versa, would not necessarily provide accurate, useable results. In short, then, in order to achieve the first aim, the metric should take into account the translator's skills but separate these skills from the translator. The metric should somehow sidestep 'human frailties'. This should be achievable. Whyman and Somers use the skill of a translator to judge whether a suggested match should be considered—they make "a subjective evaluation of its usefulness for translation, based on previous translation experience" (Whyman and Somers 1999:1,277). This use of experience does not affect the metric results in the long run, it only identifies whether a suggested translation should be deemed valid or not.

That the metric could be used to assist translators in deciding whether or not to invest in a TM system would be a desirable contribution of this paper. A translator who is not experienced in using such software will not be in a position to adequately evaluate the different systems if his/her evaluation is based on timing himself/herself using the different packages.

The second aim is to create a metric to evaluate TM systems that returns results that are relevant to the translator.

Hodász (2006:2,045) differentiates between automatic and manual evaluation. Automatic evaluation, Hodász says, involves comparing TM results to a 'gold standard', that is the text fed into the TM system will have already been translated separately by human translators in order that the TM output can be evaluated. Thus, the skills and opinions of a translator or group of translators are not required after the 'gold standard' text is translated. Hodász argues that the manual method of evaluation, which involves either a translator giving a suggested translation a score "usually on a 1-4 scale from 'absolutely useless' to 'no changes needed'" or the counting of post-edit steps following modification of a match by a translator, mirrors real life usage of the system.

As discussed above, Whyman and Somers (1999:1,274-76) designed a weighting system based on edit distance.

As also discussed above, Gow (2003:80) weighs against fuzzy matches in her metric. If the metric being designed and considered here is to make use of edit distance, the question arises, how can the effort taken to read fuzzy matches be incorporated into a metric that uses keystrokes as a measure of how good a fuzzy match is? The measurements appear to be incompatible. However, instead of counting the number of changes made to the fuzzy match, the time taken to make those changes can be counted. The time taken to make one modification or change could equate to the time taken to translate one word.

This method in itself may be considered arbitrary. However, as it would be based on an individual translator's recorded variables, in respect of their general translation speed, it fits in with the goal of creating a customizable metric, which in turn makes the method slightly less arbitrary. In addition, it could be argued that translating, say, a 6-word sentence from scratch using a TM system is the same as making 6 changes to the contents of the cell where a match would appear had one existed.[9] Assigning the same time penalty to a change as would be assigned to translating one word is therefore appropriate. In any case, it is certainly convenient to define the relationship as such, pending further testing.

Similarly, instead of assigning an arbitrary weighing to a fuzzy match as Gow (ibid.) does, the time taken to read the fuzzy match can be noted and added to the time taken to change the

[9]Providing a one to one relationship exists between ST and TT.

fuzzy match. This time can then be compared to the time taken to translate the sentence manually.

The present author proposes that in order for a metric to return results which are not only meaningful but meaningful on a level which can be understood by, and are appropriate to, translators, the results must be in a format which the translators can relate to. By using a method that takes into account the time taken to implement edits, a method that does not simply count the edits themselves, a method that also considers the time taken to read the suggested matches, fuzzy or otherwise, such results should be obtainable.

In order to achieve the second aim, any metric would need to take the texts that are normally translated by an individual translator into account. But how would one go about customizing the metric to suit individual translators? A TM database would need to be created in order to test the different systems using the metric. The simplest answer, then, would be to tailor the TM database to an individual translator, to create a TM database using materials he/she previously translated.

With this in mind, one of the first steps in designing a test schedule would be to obtain a corpus of bilingual materials representative of a translator's work. Once obtained, these materials would be aligned and imported into the TM system(s) under evaluation. In order to test the TM system a text would need to be translated using that system, or at least 'run' through the system so that suggested matches can be noted and evaluated. This text would also need to reflect the type of text that the translator would normally translate. In line with the suggestion of EAGLES when evaluating exact matches, such a text could be obtained by keeping one of the bilingual texts obtained from the translator aside, that is, one of the texts obtained would not be included in the alignment process.

As mentioned in Section 2, some texts are better suited to translation with TM than others. When implementing the proposed metric, the quantity and quality of suggested matches will vary depending on the text being 'translated' and the content of the TM database therefore. This is a strong feature of the metric. It contributes to the customisability of the metric and results will be more tailored to an individual translator. If, having followed this process, a translator is offered few usable matches, then it is likely that their work is not suited to TM. In

that case, the evaluation would be successful in that the translator would know that TM is not suitable for their needs.

In order to make comparisons of the nature described in the proposed metric, the speed at which a translator would normally translate a given sentence and the speed at which a translator is able to read would first need to be ascertained. Both these measurements should be measured using the same units (that is, seconds). The translation speed could either be provided by the translator or, preferably, the translator could be asked to translate a short text (of the type that he or she would normally translate) while being timed. The speed at which the translator reads could be measured in a similar way to that proposed by Ziefle (1998: 555-568), whereby the translator would read a text of a specific length and be timed.

Having created a TM database and imported it into a TM system, located a suitable text for translation, established the reading speed and established the translation speed, and having arranged for the translator to translate the text while using the TM system or obtained a previously translated bilingual text, the text can be analyzed one segment at a time. The time saved or lost over manual translation can then be calculated.

The complexity of the text, and, therefore, the effect which text complexity has on the average speed figures of translators, has been omitted from the metric calculations. Should the metric be adopted by the translation community, it will eventually be tailored to an individual translator, using materials that they are used to translating. Stylistic features of a text within the domains which they translate should therefore be familiar to them, as should specialist, complex terms.

The process of measuring translation and reading speeds fits in perfectly with the aims that any metric should take into consideration the skills of the translator and that any metric should be customizable to the individual translator. As the process of translation while using a TM system is not itself timed, the issue that the translator may be unfamiliar with different systems and their inner workings can be sidestepped. The average number of words translated per minute and the average number of words read per minute would be measured and applied to a metric. The metric itself would then be applied to the finished translation produced by the translator. Even if a human translator who is inexperienced in using the TM system under test

is used in order to evaluate and modify the suggested matches, the metric could be applied to the modifications and results could be obtained without having to consider the time taken for the translator to get used to the software. Moreover, once the translator's base variables have been calculated and a representative TM database created from representative texts, the translator would not actually be required to translate a text with the TM system at all. The metric would not only be independent of the system, but would also be independent of the translator's IT skills or any other external factors that could affect performance.

4. Prototype Neo-Metric

If the metric under development here is to take into consideration everything discussed in the previous Sections, and if results produced by the metric are to be meaningful and are to be of use to evaluators, the metric will become necessarily complex. The metric, then, will be introduced in two stages. A prototype metric will introduce the main functions of the metric to the reader in a less complex, more simplified form and will allow the reader to better understand and follow the thought processes of the current author in designing the metric. Following a pattern set by Gow (2003), the prototype metric will be introduced, tested and modified in light of test results. Following testing and discussion of results in Sections 5 to 7, a more sophisticated metric will be introduced in Section 8.

In implementing this prototype, the following steps were adhered to:

Pre-testing steps

1. Ascertain how many words are translated per hour by the translator (**WTPH**).

2. In order to calculate the average time taken to translate one word in seconds (**TTrans**), divide the number of seconds in an hour (3,600) by the number of words translated per hour.

3. Ascertain how many words the translator is able to read per minute (**WRPM**).

4. In order to calculate the average time taken to read one word in seconds (**TRead**), divide 60 seconds by the number of words read per minute.

5. Obtain a cross-section of bitexts previously translated by the translator or the translation company/department.

6. Align all but one of the cross section of texts to create a TM database in a TMX format. The ST of the one text not aligned will act as the text to be 'translated'; the TT as the 'gold standard' translation.

7. Import TMX into TM system.

8. Feed the text to be 'translated' into the TM system.

Analysis using metric

1. Compare the matches suggested by the TM system with the 'gold standard' translation segment by segment.

2. For each new segment, count the number of words in the source sentence (**SW**).

3. In order to calculate how much time it would take to translate the sentence manually (**T1**), multiply SW by TTrans.

4. For each 100% or fuzzy match, count the number of words in the match (**MW**).

5. For each fuzzy match, compare suggested matches to the 'gold standard' translated text in order to ascertain the number of changes required.

6. If a 'context match' was suggested, the time saved over translation without TM will be equal to T1.

7. If a 100% match was suggested and no modification was required, multiply MW by TRead in order to ascertain the time taken to read the suggested match (**TRM**). The time saved over translation without TM will be equal to T1 minus TRM.

8. If a 100% match is suggested but modification is required, multiply number of words in the match by TRead in order to ascertain the time taken to read the suggested match (**TRM**). In order to ascertain the time taken to modify the suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod.

9. If one fuzzy match is suggested, multiply number of words in the match by TRead in order to ascertain the time taken to read the suggested match (**TRM**). In order to ascertain the time taken to modify the suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod.

10. If multiple fuzzy matches are suggested, multiply the number of words in the matches by TRead in order to ascertain the time taken to read the suggested matches (**TRM**). In order to ascertain the time taken to modify the best-suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod.

5. Testing the Prototype

Prior to commencement of full testing with a more sizeable text a 'mini test schedule' using only one simple sentence as an ST was designed in order to ascertain how the metric performed when the figures 'fed in' were of extremely low or high values. A small TM database, consisting of four similar TUs was created and imported into a TM system. Of course, in practice, a much larger TM database would be used, along with many test sentences. The following is merely an example, one used in order to test and prove the concept. In any case, the following test sentence was fed into the TM system:

Will the First Minister make a statement on action being taken by WAG[10] to reduce the prevalence of tipping?

[10] Welsh Assembly Government

The following sentence was used as a gold standard:

A wnaiff y Prif Weinidog ddatganiad am y camau sy'n cael eu cymryd gan LlCC i leihau nifer y bobl sy'n **tipio**?

Current author's back translation:

Will the First Minister make a statement on action being taken by WAG to reduce the number of people who **tip**?

The TM system offered the following suggestion:

A wnaiff y Prif Weinidog ddatganiad am y camau sy'n cael eu cymryd gan Lywodraeth Cynulliad Cymru i leihau nifer y bobl sy'n **ysmygu**?[11]

Current author's back translation:

Will the First Minister make a statement on action being taken by WAG to reduce the number of people who **smoke**?

The metric was then applied with:

'average' values, i.e. WRPM= 180, WTPH= 300
'extreme low' values, i.e. WRPM= 40, WTPH= 100
'extreme high' values, i.e. WRPM= 350, WTPH= 1,000
'mixed extremes', i.e. WRPM= 40, WTPH= 1,000
'mixed extremes', i.e. WRPM= 350, WTPH= 100
'extreme low' and 'average' values, i.e. WRPM 40, WTPH 300
'extreme high' and 'average' values, i.e. WRPM 350, WTPH 300
'average' and 'extreme low' values, i.e. WRPM 180, WTPH 100
'average' and 'extreme high' i.e. WRPM = 180, WTPH = 1,000

[11]In the gold standard translation, the suggestion offered by the TM and the current author's back translations the final word has been bolded in order to illustrate the difference between the two sentences.

6. Results of Prototype Testing

Following the methodology suggested in Section 5 the following results were obtained. N.B. the calculation of the first test scenario is illustrated in full, the remaining results are available in Table 1.

With 'average' values, where WRPM is 180 and WTPH is 300:

Average time taken to translate one word in seconds (TTrans) is equal to 3,600 divided by the number of words translated per hour:

$$TTrans = \frac{3600}{300}$$

TTrans=12s

Average time taken to read one word in seconds (TRead) is equal to 60 divided by the number of words read per minute:

$$TRead = \frac{60}{180}$$

TRead = 0.33s

The number of words in the source sentence (SW) is 19

Time it would take to translate the sentence manually (T1) is equal to SW multiplied by TTrans:

T1 = 19 X 12

T1 = 228s

The number of words in the suggested match (MW) is 24

Time it would take to read the suggested match (TRM) is equal to MW multiplied by TRead:

TRM = 24 X 0.33

TRM = 7.92s

Time it would take to modify the suggested match (TMod) is equal to Number of Changes multiplied by TTrans:

TMod = 4 X 12

TMod = 48s

Time Saved over translation without a TM is equal to T1–TRM – TMod

Time Saved = 228–7.92 – 48

Time Saved[12] = 172.08s

Time saved as a percentage: 75.47%

Table 1 — Time saved as per initial metric testing

WRPM	WTPH	Time Saved (s)	Time Saved (%)
40	100	504	73.68
350	1,000	49.92	72.98
40	1,000	18	26.32
350	100	535.92	78.35
40	300	144	63.16
350	300	175.92	77.16
180	100	532.08	77.79
180	1,000	46.08	67.37

7. Discussion of Results and Evolution of the Prototype Neo-Metric

[12] This is the time saved using the one 19-word example sentence only as per Section 5.

The metric appears to have produced useful results. According to the metric, *for the single sentence under test at least*, the use of TM will equate to a very substantial time saving in all but one of the test scenarios. It should be noted, however, that the test match was chosen due to its high degree of similarity to the test sentence, this to simplify the initial testing and to prove the concept. It should be further noted that the metric, if adopted by the translation community, should be used on a more substantial text containing multiple sentences.

It appears that translators who read slowly but translate quickly will find TM less useful; although this has been exaggerated by the extremes of slowness of reading and speed of translation. Moreover, the results obtained may exaggerate the usefulness of TM systems. In the proposed metric, the time taken to read the ST during translation is not considered. This raises the question of whether T1 needs to be modified to consider the time taken to read the ST. To ascertain the 'actual' translation time, the time taken to read the ST may have to be subtracted from the translation time suggested by the metric as will be demonstrated.

For example, if a TM returns a 100% match, according to the current metric, the only element to be considered is the time taken to read that match. In reality the new ST (as opposed to the match ST) would need to be read in order for the 100% match to be validated. Reading the new ST will take time and this time should be considered when calculating time saved, else the metric will return results that are biased towards TM use. Simply subtracting the time taken to read the ST from T1 would not solve the issue. When calculating the amount of time saved, TRM and TMod are subtracted from T1. TMod is calculated by multiplying the number of changes by TTrans. This has the effect of including the time taken to read the word that is to be modified twice. This being the case, TTrans itself needs to be modified. The metric needs to be modified so that TRead is calculated first. The calculation for TTrans will then be modified so that TRead is subtracted. Thus an 'actual' or 'pure' translation speed can be calculated. Applying the changed metric with WRPM of 40 and WTPH of 1,000 returns a result of -4.5s, which means that in the case where a translator reads very slowly and translates quickly, given the test sentence and the test TM database, the use of a TM system will be 4 ½ seconds slower than manual translation. In order that time saved as a percentage of the time taken to translate the text without TM can be calculated, the time taken to read the

new ST will need to be added back to T1.[13] The results when applying the modified metric to the test values, along with the original results can be seen in Table 2.

Table 2 — Comparison of original and modified metric results.

WRPM	WTPH	Time Saved (Original Metric)	Time Saved (Modified Metric)
180	300	75.47%	73.30%
40	100	73.68%	70.39%
350	1,000	72.98%	69.25%
40	1,000	26.32%	-6.58%
350	100	78.35%	77.97%
40	300	63.16%	53.29%
350	300	77.16%	76.04%
180	100	77.79%	77.07%
180	1,000	67.37%	59.99%

It is possible that Welsh language reading speeds and English language reading speeds may differ. Brief testing suggests that this is the case. As both languages will be read during the assessment of a match returned by a TM system, the English and Welsh reading speed will need to be calculated and included separately in the calculation of the metric.

During testing, the process of deciding what exactly constitutes a change proved to be problematic. For instance, should deleting a word from a suggested match be considered a change? If so, should a deletion be given the same weighting and time penalty as an addition of a word to a suggested match or the act of replacing one word with another? As noted above, Whyman and Somers experienced similar problems when devising their metric and they suggest a modification of the Levenshtein distance algorithm in order to solve the issue (Whyman and Somers 1999:1,275). Whyman and Somers's suggestions can be adapted to suit the metric being developed here.

Any action that requires two or more keystrokes can be considered a change. Following this logic, *two* changes are required to modify the match during the above testing, that is, one for the click, drag and deletion of *Lywodraeth Cynulliad Cymru* [Welsh Assembly Government], followed by the typing of *LlCC* [WAG], one for the click, drag and deletion of *ysmygu* [to smoke], followed by the typing of *tipio* [to tip]. However, in modifying the match, the

[13]As T1 is calculated using TTrans.

translator may decide to delete all irrelevant text before entering the new text. If the translator decides to work this way, then *four* changes are required, the click, drag and deletion of *Lywodraeth Cynulliad Cymru* being the first change, the click, drag and deletion of *ysmygu* being the second change, the typing of *LlCC* being the third change, and the typing of *tipio* being the fourth and final change. As each modification is assigned a time penalty, this is important.

The time taken by a translator to decide which suggested match is best, or to decide whether a match is worth modifying or not has not yet been considered in the design of the metric. Much translation literature appears to be concerned with the decision-making process (see, for example, Chen 2008; Darwish 1999). As decision making is such an important factor during translation in general and as, when using TM systems, translators will potentially need to decide on the worth of multiple suggested matches, the time taken to make those decisions will need to be considered. There is, however, no 'easy fix' for the inclusion of decision-making time in the metric. A general formula for time taken for a person to decide between two or more different options does not appear to exist. As such, the time taken to choose between multiple suggestions, or indeed whether or not to use a suggestion at all, will need to be calculated.

8. The Refined Neo-Metric

The results of testing and proceeding evaluation point towards the need to modify the initial suggested metric in respect of the addition of decision time, the addition of language specific reading time, and the modification of the calculation of T1. The modified metric steps should be as follows:

Pre-testing Steps

1. Ascertain how many words the translator is able to read per minute in English (**WRPM_en**).

2. In order to calculate the average time taken to read one English word in seconds (**TRead_en**), divide 60 seconds by the number of English words read per minute.

3. Ascertain how many words the translator is able to read per minute in Welsh (**WRPM_cy**).

4. In order to calculate the average time taken to read one Welsh word in seconds (**TRead_cy**), divide 60 seconds by the number of Welsh words read per minute.

5. Ascertain how many words are translated per hour by the translator (**WTPH**).

6. In order to calculate the average time taken to translate one word in seconds (**TTrans**), divide the number of seconds in an hour (3,600) by the number of words translated per hour and subtract TRead_en.

7. Ascertain how fast the translator is able to make a decision (**DT**).[14]

8. Obtain a cross section of bitexts previously translated by the translator or the translation company/department.

9. Align all but one of the cross section of texts to create a TM database in a TMX format. The ST of the one text not aligned will act as the text to be 'translated'; the TT as the 'gold standard' translation.

10. Import TMX into TM system.

11. Feed the text to be 'translated' into the TM system.

Analysis using metric

1. Compare the matches suggested by the TM system with the 'gold standard' segment by segment.

2. For each new segment, count the number of words in the source sentence (**SW**).

[14] That is, the time taken by the translator to decide which of the matches suggested by the TM system should be used, if any.

3. In order to calculate how much time it would take to translate the sentence manually (**T1**), multiply SW by TTrans.

4. For each 100% or fuzzy match, count the number of words in the match (**MW**).

5. For each fuzzy match, compare suggested matches to the 'gold standard' translated text in order to ascertain the number of changes required.

6. If a 'context match' was suggested, the time saved over translation without TM will be equal to T1.

7. If a 100% match was suggested and no modification was required, multiply MW by TRead_cy in order to ascertain the time taken to read the suggested match (**TRM**). The time saved over translation without TM will be equal to T1 minus TRM minus DT.

8. If a 100% match is suggested but modification is required, multiply number of words in the match by TRead_cy in order to ascertain the time taken to read the suggested match (**TRM**). In order to ascertain the time taken to modify the suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod minus DT.

9. If one fuzzy match is suggested, multiply the number of words in the match by TRead_cy in order to ascertain the time taken to read the suggested match (**TRM**). In order to ascertain the time taken to modify the suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod minus DT.

10. If multiple fuzzy matches are suggested, multiply the number of words in the matches by TRead_cy in order to ascertain the time taken to read the suggested matches (**TRM**). In order to ascertain the time taken to modify the best-suggested match (**TMod**), multiply the number of changes needed by TTrans. The time saved over translation without TM will be equal to T1 minus TRM minus TMod minus DT.

9. Conclusion

Having identified the need for a TM metric, a prototype 'Neo-Metric' has been developed, tested, and refined. Many issues with the metric have been ironed out during this initial phase of testing. The need to modify the T1 variable and to distinguish between Welsh and English reading speeds has been identified. The need for care when calculating the number of changes required to a suggestion and the importance of including decision time has been underlined.

The use of a 'gold standard' could be challenged. As EAGLES (1995) note, "there is no one correct translation of a given segment". However, an individual translator is unlikely to disagree with himself/herself regarding an appropriate translation in any significant way. Performance aside, it can be assumed that an individual translator will approach a translation problem in a similar way and make similar decisions time after time. It is worth re-iterating here, as mentioned in Section 2, that translators will have their own translation style, so translation decisions and decisions as to how to modify a match will be similar or relatively consistent when similar or identical translation problems are presented. Providing the 'gold standard' is representative of how the translator would normally translate the text in respect of style and vocabulary, using the 'gold standard' as reference is acceptable, and certainly less time consuming than the translator having to translate from scratch during evaluation.

It should be noted that the refined metric is not a finished product. Indeed, this paper aims to introduce a methodology, or to test a theoretical concept, not to produce a polished end product. More in-depth testing has yet to be discussed. Moreover, the process of testing the metric in more realistic scenarios has yet to be recounted, that is to say the process of testing the metric 'in the field' with real life translators has yet to be described. Nonetheless, many interesting ideas which will inform future papers to be written by the current author have been uncovered.

Further research aside, the metric produces *results* which are easy for a user to understand. The metric itself is, however, far from simple. Some may question the metric's practicality. This complexity is necessary in order that an accurate result is obtained and in order that the metric remains highly customisable. As has been mentioned in Section 4, the metric has been

introduced in two stages partly to facilitate ease of understanding of the complexity. Nonetheless to expect a translator to dedicate time to calculating each variable for use with the metric may be considered optimistic. A tool designed to facilitate variable collection or calculation (to include calculation of decision time) is under development by the current author. This tool will ease the would be tester's burden. Despite this, collecting and processing variables and materials and then applying the metric to several TM systems remain time consuming and complex procedures. Future papers and articles will inform solutions to these issues.

One thing yet to be considered is the potential effect of the metric on the translation industry and its stakeholders. The methodology discussed in this paper may affect tool and industry development. Benito makes an interesting point regarding that which drives tool development. Benito notes that while TM tool research and development was, at first, aimed at making a translator more efficient, the emphasis changed to one of reducing costs for those commissioning translators and/or LSPs. Benito advises:

> This change is perhaps most clearly demonstrated by the emergence of de facto standard discount schemes based on the results of TM pre-processing: words that appear in segments which have a corresponding exact or fuzzy match in the TM are charged at lower rates (Benito, 2009:2).

He implies that the main obstacle to developing a system that would make a translator more efficient is that such a system may not result in a reduced cost for LSPs or translation commissioners, and argues that '[t]he development of new metrics to cover consistency and subsegment-level translation reuse will be the critical first step towards developing new approaches to marketing the next generation of TM technology' (Benito, 2009:7). Presumably, paying a translator on an hourly basis would also suffice. However, if Benito is correct, the methodology used in designing the metric under development here could be used as a starting point in creating Benito's suggested metrics. This in turn will signal a change in the way that translators, LSPs and clients charge and pay for their translations. It is interesting that, more recently, Zetzsche (2016:27), has suggested that the industry could use edit distance instead of word counts as a method by which payment or pricing can be calculated. However, simple edit distance may not be sufficient. The more robust and customisable form of edit distance introduced in this paper may be an answer to the issue of payment calculation

which will be fair to translator and client alike and enable the translator to be more competitive.

Issues of complexity and potential effect on the translation industry aside, the main purpose of this paper, that is to introduce a novel method of quantifying the efficiency gained due to a TM system's ability to recall TUs has been realised.

E-mail watkinsg13@cardiff.ac.uk

References

Akiba, Yasuhiro, Kenji Imamura and Eiichiro Sumita (2001) 'Using Multiple Edit Distances to Automatically Rank Machine Translation Output', in *Machine Translation Summit VIII*, Santiago de Compostela.

Akiba, Yasuhiro, Eiichiro Sumita, Hiromi Nakaiwa, Seiichi Yamamoto and Hiroshi Okuno (2003) 'Experimental Comparison of MT Evaluation Methods: Red Vs. Bleu', in *MT Summit IX*, New Orleans.

Austermühl, Frank (2001), *Electronic Tools for Translators*, Manchester/Northampton, Mass.: St. Jerome.

Benis, Michael (1999) 'Translation Memory from O to R', originally published in ITI Bulletin. Available online at [http://www.textum.pl/tlumaczenia/portal_tlumaczy/informacje/technologie/artykuly/angielski/translation_memory_from_O_to_R.html] (accessed 1 February 2015).

Benis, Michael (2007) 'Déjà Vu: Taking a Second Look at CAT in a Mature Market', *ITI Bulletin*, November-December 2007: 28-32.

Benito, Daniel (2009) 'Future Trends in Translation Memory', *Revista Tradumàtica*, 7:1-8.

Bowker, Lynne (2002) Computer-Aided Translation Technology: A Practical Introduction, Ottawa: University of Ottawa Press.

Brooke, S. and H. Ellis (1992) 'Hyperbaric Environments', in Andrew P. Smith and Dylan Jones (eds) *Handbook of Human Performance Vol.1, Physical Environment*, London/New York: Academic Press, 177-210.

Chen, Shuming (2008) 'Cultural Presupposition and Decision-Making in the Functional Approach to Translation', *Journal of Humanities and Social Sciences*, 4(1): 83-89.

Civera, Jorge, Juan M. Vilar, Elsa Cubel, Antonio L. Lagarda, Sergio Barrachina, Francisco Casacuberta and Enrique Vidal (2005) 'A Novel Approach to Computer-Assisted Translation Based on Finite-State Transducers', in Anssi Yli-Jyra, Lauri Karttunen and Juhani Karhumäki (eds) *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP)*, Helsinki: Springer, 32–42.

Darwish, Ali (1999) 'Towards a Theory of Constraints in Translation — Draft Version 0.2'.Available online at [http://www.translocutions.com/translation/constraints_0.1.pdf] (accessed 1 February 2015).

Dinges, David F. (1992) 'Probing the Limits of Functional Capability: The Effects of Sleep Loss on Short-Duration Tasks', in Wilkinson, Robert T., Roger J. Broughton and Robert D. Ogilvie (eds) *Sleep, Arousal, and Performance: A Tribute to Bob Wilkinson*, Boston: Birkhäuser, 176-188.

EAGLES (1995) 'Benchmarking translation memories'. Available online at [http://www.issco.unige.ch/en/research/projects/ewg95/node157.html] (accessed 1 February 2015).

Farmer, E.W. (1992) 'Ionization', in Andrew P. Smith and Dylan Jones (eds) *Handbook of Human Performance Vol.3, State and trait*, London/New York: Academic Press, 237-260.

Fitts, Paul Morris and Michael I. Posner (1967) *Human Performance*, Belmont/California: Brooks-Cole.

García, Ignacio. (2009) 'Beyond Translation Memory: Computers and the Professional Translator', *The Journal of Specialised Translation*, 12: 199-214. Available online at [http://jostrans.org/issue12/art_garcia.pdf] (accessed 1 February 2015).

Gow, Francie (2003) 'Metrics for Evaluating Translation Memory Software', MA thesis, University of Ottawa.

Guerberof, Ana (2008) 'Productivity and Quality in the Post-Editing of Outputs from Translation Memories and Machine Translation: A Pilot Study', Minor Dissertation thesis, Universitat Rovira i Virgili.

Guerberof, Ana (2009) 'Productivity and Quality in the Post-Editing of Outputs from Translation Memories and Machine Translation', *Localisation Focus* 7(1): 11-21.

Hartley, Tony (2009) 'Translation and Technology', in Jeremy Munday (ed.) *The Routledge Companion to Translation Studies*, London: Routledge, 106-127.

Hodász, Gábor (2006) 'Evaluation Methods of a Linguistically Enriched Translation Memory System,' in *LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

Höge, Monika (2002) 'Towards a Framework for the Evaluation of Translators' Aids' Systems', Ph.D. thesis, University of Helsinki.

Hygge, S. (1992) 'Heat and Performance', in Andrew P. Smith and Dylan Jones (eds) *Handbook of Human Performance Vol.1, Physical Environment*, London/New York: Academic Press, 79-104.

IEEE Global History Network (2015) 'Vladimir I. Levenshtein - GHN: IEEE Global History Network'. Available online at [http://www.ieeeghn.org/wiki/index.php/Vladimir_I._Levenshtein] (accessed 1 February 2015).

Kenny, Dorothy (2006) 'Book Review Uwe Reinke, Translation Memories. Systeme – Konzepte – Linguistische Optimierung', *Machine Translation*, 20: 305-309.

Kenny, Dorothy (2011), 'Electronic Tools and Resources for Translators', in Kirsten Malmkjær and Kevin Windle (eds), *The Oxford Handbook of Translation Studies*, Oxford Handbooks in Linguistics, Oxford: Oxford University Press, 455-474.

King, Margaret (1999) 'Setting Standards for Evaluation', In *MT Summit VII*, Singapore.

Leusch, Gregor, Nicola Ueffing and Hermann Ney (2003) 'A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation', in *MT Summit IX*, New Orleans.

Lieberman, Harris R., William J. Tharion, Barbara Shukitt-Hale, Karen L. Speckman and Richard Tulley (2002) 'Effects of Caffeine, Sleep Loss, and Stress on Cognitive Performance and Mood During U.S. Navy Seal Training', *Psychopharmacology*, 164: 250-256.

Megaw, E. (1992)'The Visual Environment', in Andrew P. Smith and Dylan Jones(eds)*Handbook of Human Performance Vol.1, Physical Environment*, London; New York: Academic Press, 261-296.

Mikuličková, Eliška (2010) 'Computer Assisted Technology - Comparison of Programs', Bachelor thesis, Tomas Bata University.

Palacz, Błażej (2003) 'A Comparative Study of CAT Tools (MAHT Workbenches) with Translation Memory Components', MA thesis, Adam Mickiewicz University.

Pym, Anthony (2011) 'Democratizing Translation Technologies – the Role of Humanistic Research', in *Luspio Translation Automation Conference*, Rome.

Quah, Chiew Kin (2006) *Translation and Technology*, Houndmills/New York: Palgrave Macmillan.

Reinke, Uwe (2000) 'Reinke: Evaluating the Performance of Translation Memories'. Available online at [http://mt-archive.info/IAI-2000-Reinke-2.pdf] (accessed 1 February 2015)

Rico, Celia (2001) 'Reproducible Models for CAT Tools Evaluation: A User-Oriented Perspective', in *Translating and the Computer 23*, London.

Shadbolt, David (2002) 'Finding the Right Language Technology Tools', *Language Technology Supplement, MultiLingual Computing and Technology*13 (51): 18-23.

Smith, Andrew P. (1992)'Colds, Influenza and Performance', in Andrew P. Smith and Dylan Jones (eds) *Handbook of Human Performance Vol.2, Health and Performance*, London/New York: Academic Press, 197-218.

Somers, Harold (2003) 'Translation Memory Systems', in Harold Somers, (ed.) *Computers and Translation [Print and Electronic Book]: A Translator's Guide*, Amsterdam/Philadelphia: J. Benjamins, 31-47.

Tate, Calandra (2008) 'A Statistical Analysis of Automated MT Evaluation Metrics for Assessments in Task-Based MT Evaluation', in *Association for Machine Translation in the Americas (AMTA)* Waikiki.

Tilley, A. and S. Brown (1992) 'Sleep Deprivation', in Andrew P. Smith and Dylan Jones (eds) *Handbook of Human Performance Vol.3, State and trait*, London/New York: Academic Press, 237-260.

Wallis, Julian (2006) 'Interactive Translation Vs Pre-Translation in the Context of Translation Memory Systems: Investigating the Effects of Translation Method on Productivity, Quality and Translator Satisfaction', MA thesis, University of Ottawa.

Whyman, Edward and Harold Somers (1999) 'Evaluation Metrics for a Translation Memory System', *Software - Practice and Experience*, 29(14): 1265-1283.

WordSense.eu Online Dictionary (2016) 'bitext (English)'. Available online at [http://www.wordsense.eu/bitext] (accessed 1 March 2016).

Yamada, Masaru (2011) 'The Effect of Translation Memory Databases on Productivity'. Available online at [http://isg.urv.es/publicity/isg/publications/trp_3_2011/yamada.pdf] (viewed 16 February 2015).

Zerfass, Angelika (2002a) 'Evaluating Translation Memory Systems', in *LREC 2002: Language Resources in Translation Work and Research*, Las Palmas de Gran Canaria.

Zerfass, Angelika (2002b) 'Comparing Basic Features of TM Tools', *Language Technology Supplement, MultiLingual Computing and Technology*, 13(51): 11-14.

Zetzsche, Jost (2016) 'Rethinking Payments', *ITI Bulletin*, January-February 2016: 26-27.

Ziefle, Martina (1998) 'Effects of Display Resolution on Visual Performance', *Human Factors* 40(4): 555-567.

Appendix A – List of Acronyms Used

CAT	Computer Aided Translation
cy	Welsh language
DT	Decision Time
EAGLES	Expert Advisory group for Language Engineering Standards
en	English language
IR	Information Retrieval
IT	Information Technology
ITI	The Institute of Translation & Interpreting
LSP	Language Service Provider
LT	Language Technology
MT	Machine Translation
MW	the number of words in the match
ST	Source Text
SW	The number of words in the source sentence
T1	The time it would take to translate a sentence manually
TM	Translation Memory
TMod	The time taken to modify a suggested match
TMX	Translation Memory eXchange
TRead	The average time taken to read one word in seconds
TRM	The time taken to read a suggested match
TT	Target Text
TTrans	The average time taken to translate one word in seconds
TU	Translation Unit
WAG	Welsh Assembly Government
WPH	Words Per Hour
WRPM	Words Read Per Minute
WTPH	Words Translated Per Hour