# Chapter 8. Translation technology and learner performance: Professionally-oriented translation quality assessment with three translation technologies

Katell Hernandez Morin, Franck Barbin, Fabienne Moreau, Daniel Toudic, Gaëlle Phuez-Favris

# Translation technology and learner performance: professionally-oriented translation quality assessment with three translation technologies

**Katell Hernandez Morin, Franck Barbin, Fabienne Moreau, Daniel Toudic, Gaëlle Phuez-Favris**

**Rennes 2 University – LIDILE RESEARCH UNIT (EA3874) TRASILT**

## Abstract

This chapter examines how a three-dimensional translation quality assessment grid (based on error type, effect and criticality) can be used to assess student translation performance with three different tools (standalone TM system, speech recognition and post-edited machine translation).

The study was professionally-oriented, using a technical English-language source text, short deadlines for completion of each translation, and professional quality criteria.

Group and individual performance in the translation of five 500-word extracts were assessed for quality and efficiency, with and without translation tools, using our assessment grid. The factors affecting group and individual performance and possible correlations between tool and performance were studied. The potential usefulness of the grid as a fine-grained training and professional assessment tool is discussed.

**Keywords**: translation tools, translation quality assessment, quality assessment grid, technical translation.

## Introduction

In recent decades, as translation has grown from being not much more than a "cottage industry" to a strategic and highly competitive multi-billion dollar business sector, the issue of translation quality has become central, both to the credibility of the profession, and to the survival of many industry players. The issue of translation quality has been approached from many different angles, but three of them in particular have focused the attention of the research community: how to assess the quality of the end-product (i.e. the translation); how to achieve best quality by optimising the tools used and the processes applied by the translator; and how to ensure that translators are trained to produce optimal quality through their translation skills and best use of the tools and processes available. In this chapter, we bring together these different strands by focusing on the design and use of a three-dimensional, multi-functional assessment grid, which can be used to measure and compare translation quality in both professional and academic contexts. We applied our grid in an experiment involving three different translation tools — traditional translation using a standalone TM system, dictated translation using a speech recognition tool, and post-edited machine translation — and analysed its implications for translator training.

## Related research

The assessment of translation quality has evolved considerably over the past three decades, in line with the evolution of translation markets, tools and practices themselves. Assessment methods have shifted between the holistic approaches of translation as an art form, favoured by literary translation theorists (Berman, 1995) and the quantitative, error-based approaches favoured both in traditional academic contexts and in the many professional translation quality assessment (TQA) methods developed by the language services and the language industry (the Canadian Translation Bureau's Sical model (1986-94), the American Automobile Industry's SAE J2450, the localisation industry's LISA model, the various translation quality index models developed by major language service providers, etc.). In the 1980s, the focus moved from purely text-based approaches to context-based, functionalist approaches (Vermeer, 1979; Reiss, 1981; Nord, 1991) and to combinations of both, integrating both source-text difficulty and end-user effect (Gouadec's "SEPT" TQA system, 1981, 1989). Toury (1995) carried this forward by distinguishing between the "adequacy" of the target text in relation to the source text and the "acceptability" of a translation viewed according to target language standards and end-user expectations. These two concepts, redefined as "adequacy" and "fluency" by the machine translation community (Calisson-Burch et al., 2007) have been widely used in comparative studies of machine-translation (MT) output quality. The widespread use of free online statistical MT has in turn led in recent years to an increasing number of studies focusing on end-user tolerance of varying degrees of MT output quality. Bowker et al. (2007) have explored the relativity of end-user translation quality perception, while Doherty and O'Brien (2014) have studied "usability" (i.e. the ability of someone to perform a task on the basis of translated instructions) as a quality criterion. The idea that translation quality is a relative concept, largely determined by the end-user's own

standards has in fact been given a new lease of life by the mass dissemination of statistical machine translation. Paradoxically, therefore, MT has both reinstated the use of "absolute" quality standards on the one hand, represented by the "reference" translation used in metrics-based TQA systems such as the "Meteor" or "BLEU" scores, and the relativity of end-user assessments on the other hand, based on simple yes/no acceptability and usability tests.

In the field of human translation and translator training, Daems, Macken and Vandepitte (2013) suggest a more subtle two-step approach to the translation quality assessment process, whereby evaluators are asked to assess translations of general newspaper articles produced by trainee translators in two successive stages: first by assessing the target text alone for "acceptability"; then by assessing the "adequacy" of the translation in relation to its source text. In each of the stages, a number of fine-grained error categories were applied, and scores calculated for each type of functional quality.

The assessment grid described in this chapter attempts to go a step further, by combining error type analysis with *four* functional quality criteria, to produce a flexible TQA tool which can be adapted to different professional as well as didactic uses. Our approach posits that in "instrumental translation" (Williams, 2004: xiii), the quality of a translation can be measured a) by the extent to which all the information contained in the source text (ST) deemed relevant for the purpose for which it is intended, is transferred to the target text (TT); b) by the TT's success in fulfilling the function for which it is intended; c) by the degree of clarity and fluency of the TT expected by the end-user in a given context; and d) by the TT's compliance with any standard or specification that may be explicitly set by the employer or contractor or implicitly expected by the end-user of the translation. It also posits that deficiencies (or errors) in the translation's constituent units (generally at sentence level) may affect several criteria and therefore affect the quality of the translation as a whole. Thirdly, it posits with Gouadec

(1981, 1989) that a translation error may have different effects and different degrees of criticality according to the type and aim of the translation.

## Experimental design

The quality assessment grid was applied to the results of an experiment conducted in February-March 2013, in the third year of a three-year study of student translation performance using different translation tools and methods.

The original participants were a group of 19 second year Master's degree students in the translation department at Rennes 2 University (France). All the students were holders of B.A. degrees in applied modern languages, modern languages or translation studies, had already studied for one year in the Master's degree in specialised translation and localisation prior to the experiment, and had been selected for admission to second year on the basis of their first year results, particularly as regards their translation and IT skills. The students were all proficient in the use of office software, had all received basic training in the use of translation memory tools, and had already undergone a three to five month work placement in a translation company, working as translators/proofreaders and/or project managers. However, as none had received any formal training in the use of speech recognition technology during the first year of the degree, a training session was organised prior to the experiment and students were then asked to download and train the system on their own laptops for a week, prior to the experiment. As regards post-editing, students had had an introductory course in the first year of their master's degree.

In order to reduce the risk of bias, only 12 of the original 19 sets of results were retained for analysis. We eliminated from the study those students who were not French native speakers (2 participants), those who had not completed one of the tasks for whatever reason, technical or otherwise (3 participants), and those who demonstrated a particularly negative attitude to one or other of the technologies in the feedback we required students to provide on each tool (2 participants).

The task assigned to the students was to translate five 500-600 word extracts from an American solar pool heating system installation manual. The manual was chosen on the basis of volume, structure, technicality, and accessibility. The overall volume (7,000 words) and structure (24 sections) allowed us to select five separate yet coherent extracts. It was technical in its use of domain-specific terminology, yet sufficiently concrete and descriptive to be understandable by non professional readers. Finally, similar French source material was easily accessible online, although no French translation of the source document was available.

During the first four weeks, all the students translated four successive extracts, using a different translation method each week. In week five, the students were divided into four groups, and a fifth extract was translated, using a different method in each group. The students worked individually on the university premises, during a 2-hour period set aside for this assignment each week. The work schedule was as follows:

- Week 1: Word processor with an Excel glossary (hereafter referred to as “WP”),
- Week 2: Trados Studio 2009 with a prepared (unshared) translation memory (referred to as “TM”),
- Week 3: A speech recognition system (Dragon Naturally Speaking, 2011) coupled with Trados Studio 2009 and a translation memory (referred to as “SR”),
- Week 4: Google Translate, post-edited by the students (referred to as “MT”).

- Week 5: A fifth extract (Trad5) was translated by the students, who were randomly and equally distributed between the three technologies (same number of students assigned for each technology).

The methods were scheduled in decreasing order of familiarity (for the students): from simple word-processing to post-editing. Week 1 was designed as a benchmark phase, enabling us to measure the subsequent impact (if any) of translation tools. Post-editing was placed last because it was the method the students were the least familiar with and because it implied a change of paradigm. Week 5 (same extract translated with different technologies) was designed to check for possible bias due to the nature of the different extracts chosen in weeks 1-4, and compare the impact of technologies on the quality of translation with an identical extract. WP was excluded in the phase, since it was not a tool per se and served as the benchmark method.

The five, non sequential extracts were selected on the basis of length and coherence, from the beginning, middle and end of the manual and dealt with general technical information on solar pool heating installations, solar collector panel connecting instructions, roof mounting instructions, feed line connections, and troubleshooting advice respectively. They were dealt with in chronological order according to the above-mentioned schedule (i.e., extract n°1 in week 1, through to extract n°5 in week 5).

We subjected the extracts to various readability tests in order to identify potentially significant differences which could affect the students' comprehension and translation of the texts. Traditional readability indices, such as Flesch-Kincaid (1975), Dale-Chall (2000) or AMesure were calculated, but did not yield coherent results. Table 1 below gives an overview of these readability indices.

**Table 1 – ST extract readability according to established indices**

| Readability index / Method | Flesch readability | D-Chall score | Amesure |
|---|---|---|---|
| **WP** | 71.4 | 7.2 | 3.0 |
| **TM** | 84.4 | 6.9 | 4.0 |
| **SR** | 81.5 | 6.6 | 4.0 |
| **MT** | 88.8 | 7.5 | 3.0 |

A lower score (71.4 for the WP extract, for instance) with the Flesch readability index (meaning the text is easier to read) should also result in a lower D-Chall score (6 rather than 7); when a higher score (such as 88.8 for the MT extract, meaning the extract is rather complex) should also result in a higher score with the Amesure index (4.0 instead of 3.0, which tends to indicate the text is more fluid - lower density of difficult words, shorter sentences, less syntactic difficulties, etc.). The results did not therefore yield conclusive evidence of differences in readability between the extracts, and highlighted the difficulty of applying such tools to specialised texts as opposed to general ones.

In order to compensate for this deficiency, we then subjected the extracts to a terminology analysis, in order to determine whether differences in terminological "density" could affect the time students would need to spend on term searches for each extract. An initial term extraction conducted with Sketch engine was refined manually with regard to context, to produce a list of term units for each extract. These lists were then compared with each other and with the 460-term reference glossary of solar heating terms, previously prepared using a bilingual English and French installation manual for an almost identical system and made

available to the students during each experimental phase. The results are summarized in the table below:

**Table 2 – Term units in each of the five extracts**

| Method / Extract | WP | TM | SR | MT | Trad5 |
|---|---|---|---|---|---|
| **Nb of words in extract** | **510** | **498** | **524** | **488** | **523** |
| **Nb of term units** | **53** | **48** | **38** | **34** | **39** |
| **Nb of new terms*** | **53** | **44** | **30** | **21** | **21** |
| **% of new terms** | **100** | **91.67** | **78.95** | **61.76** | **53.85** |
| **Nb of terms not in glossary** | **9** | **15** | **9** | **4** | **6** |
| **% of term units not in glossary/Total term units** | **16.98** | **31.25** | **23.68** | **11.76** | **15.38** |

* not encountered in previous extracts

This analysis shows a diminishing number and/or percentage of "new" (not previously encountered) terms in each successive extract, which is only to be expected in an instruction handbook. This should normally have resulted in fewer terminological searches and difficulties for the students as the experiment progressed. Table 2 also shows a relatively small number of terms not available in the bilingual glossary: although the percentage of terms requiring additional research varies between extracts, the actual numbers (4 to 15 items) remain limited. We did not therefore anticipate any major bias due to differences in terminological "density" between the extracts chosen for the study. What we did not (and could not) measure, however, was the semantic complexity of each extract, and the "cognitive load" generated in the translation process.

Resources available to the students also included a translation memory produced by aligning the English and French sections of the manual used to produce the glossary. The alignment only produced a small number (4%) of 100% matches overall, with less than 9% including all

fuzzy matches, but provided scope for a wide range of contextual searches, using key terminology and phraseology. The students also had permanent Internet access and were able to search for additional terminology resources if required, although time constraints limited that possibility.

As regards the quality assessment method, each of the translations produced by the students was assessed by three evaluators, who all had experience in the teaching and practice of professional translation. Due to organisational constraints, the evaluators were also the experimenters, and were therefore aware of the translation method and tool used for each extract, and the student participants were identified by their initials. The translations were assessed using the TRASILT quality assessment grid described below and individual results were compared by the three evaluators, who then agreed on the final quality "score" allocated to each translation.

## The TRASILT three-dimensional, functional assessment grid

Our proprietary quality assessment grid (the TRASILT grid) is based on a three-dimensional translation deficiency analysis: (1) error type, (2) effect on quality, and (3) degree of criticality. The type of error can be defined as the type of linguistic, semantic or formal discrepancy identified in a given target unit between the translation produced and the translation expected by the evaluator. The effect on quality can be defined as the impact of the error type on the various qualitative criteria of the translation according to the nature and aim of the translation. The degree of criticality can be defined as the level of functional impact of the error type.

*Development*

Our quality assessment approach originated from a previous experiment aiming to compare the productivity achieved by Master's degree students using three types of translation technologies (translation memory, machine translation and speech recognition). Our goal was to study the link between the translation times observed using the different tools and the quality of the translations produced. This was then used to determine the tool offering the best compromise to achieve a given level of quality. Our initial qualitative assessment grid was an adapted version of an existing standard professional assessment grid (the LISA model mentioned in Part 2), based on error types (omission, meaning, style, terminology, spelling, grammar, and punctuation) and degrees of criticality (minor, major, and critical), which generate a translation quality index (TQI) by comparing the number of words assessed and the number of errors.

As our first experiment did not yield the expected results, partly because of the heterogeneous level of the students tested and because of the static, non scalable nature of the grid, we decided to (1) review the objectives of our study, and (2) devise our own grid to better take into account the intrinsic variability of translation situations.

As the translators in our sample were all students, we decided to add another dimension to our study, namely the impact of technologies and tools on student performance, while maintaining the objective of a qualitative assessment based on professional criteria.

This led us to devise a multi-criteria assessment grid, which had to meet the following requirements:

- Be applicable to professional translators as well as translators in training,
- Be applicable to any type of human translation process, whatever its nature (sight translation, typed translation, machine translation post-editing, etc.),
- Weigh quality deficiencies differently according to the type of source text and the aim of the target text,
- Reduce subjective biases as much as possible,
- Be easily used and allow automatic calculation of scores.

As none of the existing assessment grids could meet all these requirements (O'Brien, 2012), we decided to design our own grid, with the aim to combine several quality criteria and analyse quality more precisely and dynamically than other grids.

## *Principles*

To avoid starting from scratch, we took as a starting point the error typology used in various existing professional assessment grids.

We then looked for ways to combine error identification with two other dimensions: (a) the type of effect the error has on the quality of the translation, according to the nature of the translation, its goals, target readers and use; and (b) the degree of impact on quality induced by the deficiency, on a scale ranging from 0 to 3. The number of points assessing the criticality of each effect is then calculated, producing the aggregate score for the sample assessed. The higher the score, the lower the quality of the sample.

Considering the three-dimensional nature of our grid, each error type can have multiple effects, which can vary according to the nature of the translation assessed and the goal of the assessment.

## *Grid dimensions*

### *Error typology*

The TRASILT grid consists of nine error types: seven are based on conventional categories (Meaning, Omission/addition, Terminology, Phraseology, Grammar/syntax, Spelling/typography, and Style), and the two other categories are based on professional assessment criteria: Localisation errors (i.e. failure to adapt to target audience or culture) and Desktop publishing or DTP errors (i.e. page layout and formatting problems).

The explanations on the types of errors available in Worksheet 3 of our grid, which are available for the evaluator to check if required, are reproduced in Table 3:

**Table 3 – Error type descriptions**

| Meaning | Omission/addition | Terminology | Phraseology |
|---|---|---|---|
| Ambiguity | Non translation of a meaningful item of the source document | Inappropriate variant (language variety/ professional usage/ In-house usage) | Inappropriate variant (language variety/ professional usage/ in-house usage) |
| Partial mistranslation | Unjustified addition of information with a minor impact on the target text | Inappropriate term (belonging to another domain) | Inappropriate phraseology (belonging to another domain) |
| Complete mistranslation | Unjustified addition of information with a major impact on the target text | Terminological inconsistency (in the document/ with reference material) | Phraseological inconsistency (in the document/ with reference material) |
| Failure to correct source text deficiency | | | |

| Grammar/syntax | Spelling/typography | Style | Localisation | DTP |
|---|---|---|---|---|
| Morpho-syntactical errors | Misspelling | Literal translation | Failure to adapt to target culture | Page layout |
| Word order | Typos | Sentence length | Failure to adapt to target audience | Formatting |
| Sentence structure | Punctuation error | Lack of fluency | Failure to localise facts and figures | Graphics |
| | Typography error | Inappropriate register (formal/informal language) | | Tags |
| | | Inappropriate variety (country-specific spelling or word choice) | | Cross-references |

*Effect typology*

This second dimension is based on four quality criteria: Accuracy, Usability, Readability and Compliance). Accuracy and Readability are akin to the concepts of *Accuracy* and *Fluency* common in machine translation assessment grids, even if our definitions are slightly more restrictive than Koehn's (2007). The other two derive from professional translation: Usability is defined as the ability of a translation to fulfil the function it is given (inform, give directions, warn, etc.), and Compliance is defined as conformity to an explicit or implicit standard (style guides, client-specific standards, imposed terminology, language and/or cultural conventions, etc.).

Effect definitions are given in Table 4:

**Table 4 – Effect typology**

| Accuracy | Usability | Readability | Compliance |
|---|---|---|---|
| error prevents the correct conveyance of information in the source document | error prevents correct use of the product, process or document | error has an impact on the fluency and clarity of the target document | target document does not comply with language-, country-, culture- or client-specific standards, conventions or recommendations |

*Degree of criticality*

The third dimension of our grid refers to the level of functional impact of the error identified. Four levels of criticality impacting the quality of the target document can be distinguished: 0 = no effect/not counted effect, 1 = minor, 2 = major, 3 = critical. These four levels of criticality are applied to the four types of end-user effects on quality and not to the nine types of errors, as is the case in most professional assessment models. Our goal was to assess the consequence of the error and not its cause.

## *Adjustments*

Based on the requirements defined above, the TRASILT assessment grid was tested in two experiments in 2012 and 2013. In our last experiment (described in this chapter), as mentioned before, students translated using different technologies, and were assessed by researchers applying the TRASILT grid.

To reduce the level of subjectivity of human assessment, four evaluators were assigned for each translation during the first testing phase of the TRASILT grid, and only three evaluators in the following phases, as they were more accustomed to using the grid. Comparing our assessments offered two advantages:

On the one hand, the first batch, used as a pre-test of the grid, enabled us to refine the grid and the definitions of each error type and effect induced. On the other hand, the coordination

enabled us to assess translations more objectively, e.g. when examining frequent errors (how each evaluator assessed the error, its effect on quality and criticality), and to reduce variations between evaluators. This step was essential to obtain comparable data for the analysis of results.

The comparison of scores given for each sample when analysing the first batches led us to question the differences between evaluators. We quickly realized that one of the main sources of differences could be found in the way evaluators allocated points differently between the various end-user effects and their degrees of criticality. We thus decided that:

1. The maximum score for each error should be 5,
2. The maximum number of effects on quality for each error should be 2, one major and one minor effect, in order to avoid the dispersion of the types of effects on quality (e.g. allocating 1 for each type of effect). For instance, an error can be awarded a 2-point penalty for Usability and a 1-point penalty for Accuracy, but cannot be penalised on the two other types of effects (Readability and Compliance) nor awarded a 2-point penalty for Accuracy (already allocated to Usability).

After this test period, we considered that the TRASILT grid was operational and could be tested on a larger scale (Toudic et al. 2014).

In sum, our assessment grid is designed to measure both quantitative and qualitative aspects of translation quality through commonly accepted professional error-based quality criteria, while also taking into account the effect of the error on target text functionality and its degree of criticality. The TRASILT grid is designed to be dynamic by allowing error type, effect on translation quality and criticality to be assessed independently or in correlation. It aims to go beyond a simple calculation of points and reveal each translator's weaknesses or strengths according to the technology used, through a detailed study of the end product rendered.

## Results and Discussion

This section will first of all present the overall quality recorded for the translations produced with each of the four translation methods. It will then examine the individual scores recorded for each of the 12 students in the study. These results will then be discussed in relation to possible biases, i.e. technical and organisational issues, textual differences in the extracts translated, individual student attitudes and the specific translation approaches induced by each translation tool and/or method.

### *Overall and individual results in relation to benchmark performance*

The results provided below are those measured by our assessment grid. The scores reflect the aggregate effect of all the translation errors identified in a given translation, weighted according to their degree of criticality for each effect factor. The lower the score, the higher the quality recorded. In this experiment, the four effect factors (Accuracy, Usability, Readability and Compliance) were equally weighted.

*Overall results per method*

**Table 5 – Overall results per method**

| Method used | Average Score Total | Standard Deviation | Median Time (hour) | T-value | P-value |
|---|---|---|---|---|---|
| WP (*baseline*) | 29.33 | 13.18 | 1:32:00 | | |
| TM | 32.58 | 16.84 | 1:15:00 | -0.53 | 0.60 (> 0.05) |
| SR | 37.50 | 12.41 | 1:44:00 | -1.57 | 0.13 (> 0.05) |
| MT | 30.33 | 13.01 | 1:11:00 | -0.19 | 0.85 (> 0.05) |

Table 5 first of all shows that the lowest score (i.e. the highest overall quality) was achieved by students using simple Word Processing (WP), with no specific translation tool, followed by post-edited machine translation (MT), and translation with a translation memory system (TM). Use of a speech recognition system (SR) produced a noticeably higher score (therefore the lowest overall quality), with the lowest standard deviation (12.41), showing that most of the students tended to perform less well when using this method. Conversely, the TM method produced the highest standard deviation (16.84), which seems to signal that some students were more familiar with the method or found it much more effective than others, which is confirmed by student feedback (see the "Attitudes towards the translation technologies used" section).

Nevertheless, it is important to note that the average score differences between the WP method and other translation methods are not statistically significant (according to the Student t-Test with a significance level (p-value) of 0.05, as shown in columns 5 and 6 in Table 5).

These initial results, which tend to show that students actually performed better without the use of translation tools, are all the more surprising as WP was used on the first extract, i.e. before students had time to absorb the subject matter and key concepts. However, they must be qualified by taking into consideration the average time spent on each task. Although students were given a two-hour slot in which to complete each translation, they were also asked to keep track of their actual translation or post-editing times (including time spent proofreading their own work, but excluding additional terminology or knowledge searches and time spent familiarising themselves with the particular tool being used). These measured times show that MT and TM were more effective in terms of productivity, despite producing slightly lower overall quality than simple WP. Again, SR not only produced lower quality, but was significantly slower.

*Individual student performance*

Let us now turn to the individual performance of each of the 12 students in our study, as shown in Table 6 below. The table shows both the students' ranking with each translation method and the quality score obtained (in brackets).

**Table 6 – Student rank and quality score per method**

| STUDENTS | WP *rank (+score)* | TM *rank (+score)* | MT *rank (+score)* | SR *rank (+score)* | 4 Methods (Score Sum) *rank (+score)* |
|---|---|---|---|---|---|
| CC | 1 (15) | 3 (17) | 2 (13) | 2 (24) | 1 (69) |
| BP | 2 (16) | 2 (16) | 12 (56) | 6 (35) | 6 (123) |
| RD | 3 (17) | 9 (42) | 5 (27) | 4 (30) | 5 (116) |
| LTT | 4 (23) | 1 (13) | 6 (31) | 2 (24) | 2 (91) |
| SM | 5 (25) | 8 (34) | 10 (41) | 12 (61) | 10 (161) |
| HG | 6 (28) | 5 (25) | 4 (23) | 1 (22) | 4 (98) |
| CA | 7 (30) | 11 (51) | 11 (43) | 5 (34) | 9 (158) |
| HI | 8 (31) | 6 (32) | 3 (18) | 11 (55) | 7 (136) |
| TL | 9 (31) | 3 (17) | 1 (11) | 8 (37) | 3 (96) |
| BJ | 10 (33) | 6 (32) | 7 (33) | 9 (44) | 8 (142) |
| LM | 11 (39) | 12 (70) | 9 (35) | 7 (36) | 11 (180) |
| CM | 12 (64) | 9 (42) | 7 (33) | 10 (48) | 12 (187) |

In this table, WP was used as the benchmark score, supposedly measuring basic translation competence without the use of translation technology. It shows that the 12 students can be divided into three unequal bands: an upper band of 3 students (CC, BP and RD), a middle band of 7 students with quality scores ranging from 23 to 33, and a lower band of 2, with scores well above the average of 29.33 achieved for this method.

If we examine how quality scores vary over the four methods, we can see that the student who ranked 1st (CC) in the benchmark (WP) translation also ranked very high overall, obtaining low scores and high rankings in each of the other three methods. Similarly, the two students

who ranked bottom of the benchmark scale (LM and CM) also performed worst overall (11th and 12th), although doing marginally better with some of the translation tools.

In most cases, however, the pattern is much more difficult to make out. While some subjects (e.g. LTT, HG, BJ) are relatively consistent, either at the top end or in the middle of the ranking scale, others show quite obvious discrepancies. BP, for instance, who ranks 2nd in the benchmark translation, performs well with TM, but fares very badly with MT. SM, in the top half (5th) of Table 6 in the benchmark translation, comes 10th overall, scoring very badly with SR and MT. Conversely, TL, who comes 8th in the benchmark ranking, achieves 3rd rank overall, thanks to a good performance using TM and post-edited MT.

These results therefore show no absolutely clear-cut pattern emerging from our study of individual performance. There is, however, an indication that a highly competent trainee translator will perform well, irrespective of the tool or method used, and conversely, that tools cannot compensate for inadequate initial translation competence. This is consistent with our earlier findings (Toudic et al., 2013).

A simple correlation analysis of students' ranking with different translation methods yields the following results:

**Table 7 – Translation method Correlation (Spearman ranking correlation coefficient)**

| WP-TM | WP-MT | WP-SR | TM-MT | TM-SR | MT-SR |
|---|---|---|---|---|---|
| **0.524** | **0.042** | **0.543** | **0.349** | **0.333** | **0,214** |

The results show that there is no clear ranking correlation between most of the method pairs (a strong correlation would show results close to 1). The only significant result is the absence of correlation between the WP-MT methods (with a coefficient close to 0), which could

reflect the fact that translation without tools and post-editing calls on very different cognitive processes.

We now need to examine the various factors that may explain the variations in performance observed above, in order to assess which of these factors, if any, has the greatest influence on translation quality as measured by our index.

## *Factors potentially affecting student performance*

### *Technical and organisational factors*

It has to be remembered, first of all, that the WP and MT methods were the simplest in terms of technical process for the students: the WP process consisted in translating a Word document with the help of a glossary and any Web searches necessary to complete the terminology required. The MT process required post-editing a Word translation document generated by Google Translate (again, using the same glossary and any additional Web research required). These two methods ranked first and second in terms of quality (Table 6), with very similar average scores and standard deviations (29.33/30.33 and 13.18/13.01 respectively), but with a clear advantage for post-edited MT in terms of productivity.

The other two methods, TM and SR, which achieved the third (32.58) and "worst" (37.50) average scores respectively, involved the use of a translation memory system (Trados 2009) in addition to the glossary and Web searches. In addition, the SR process included the correction of the speech recognition software errors. As mentioned in the experimental design section (Part 3), only a small number of 100% or fuzzy matches were to be found in the source text extracts used for these two methods. This was mentioned by several of the students, who questioned the usefulness of TM in this case and saw it as a limiting and time-

wasting factor. However, the translation memory did contain substantial partial entries (terms, phrases and parts of speech) that could be extracted through a simple concordance search in the memory. Moreover, our terminology analysis showed that only a small number of terms in each extract were not included in the reference glossary (although slightly more in the TM extract). In spite of this, it seems that not all students made the effort or felt they had sufficient time to search for those missing terms. In using SR technology, students were faced with several additional technical difficulties. The additional software application, combined with the TM system, often resulted in slower computer response times, which may explain the overall poorer productivity apparent in the results. Finally, the students felt they had had insufficient training in translating orally, while at the same time managing the translation memory, glossary and Web searches. This result is not totally surprising as it takes a substantial amount of time to train an SR system for optimum accuracy. Although students were given a full week to train the SR system for their own voices, a substantial part of the cognitive effort of student translators would go into correcting SR errors and not necessarily translation errors, as observed in other comparable studies (Dragsted et al., 2011).

The MT and SR processes achieved the third (32.58) and "worst" (37.50) average scores, respectively.

Another bias that must be considered is the scheduling of the various translation phases in the project. It must be remembered that the experiment was conducted over a period of several weeks, using 5 successive extracts from the same technical document. This might be expected to have induced a gradual acclimatisation to the source text (ST) concepts and terminology, and exponential progress in the quality achieved as the experiment progressed. However, Table 6 shows that students performed best overall in weeks 1 and 4 of the project (WP and MT), while the "worst" results in terms of overall quality were found in week 3 (SR). This

appears to show that most participants did not leverage the knowledge gained from their previous translations to any significant degree.

The average score achieved by students using the various technologies on the final identical sample (Trad5) and presented in Table 8, submitted for translation tends to confirm that the translation phase scheduling had no significant influence on the result: the quality produced in the last project sample (when students were supposedly familiar with the ST) was in fact generally worse than in the previous phases. What seemed to have a stronger effect on quality, again, was the technology used (SR produced the worst quality-highest score, as in the first four weeks of the experiment).

**Table 8 – Average total scores for each method on Trad5 (identical text), week 5**

| Technology used | Average Score Total |
|---|---|
| TM | 39 |
| SR | 42.50 |
| MT | 33.75 |

*Attitudes towards the translation technologies used*

If no clear relationship can be established between the translation quality achieved with the three different translation technologies studied, initial student translation competence and terminological density, to what extent can these variations be put down to collective or individual attitudes towards the tools used in the experiment? To answer this question, we analysed the detailed student feedback elicited through a questionnaire at the end of each phase of the experiment. As pointed out in the "experimental design" section, the most overtly hostile attitudes were filtered out of the study, by not including the students concerned in the

results subjected to analysis. It is, however, interesting to examine whether anomalies in the results can to some extent be explained by student perceptions of the technologies they used.

According to this feedback, of the three translation tools used in the experiment, TM was by far the students' preferred technology, possibly because it is also the most familiar. Among the positives mentioned was the fact that TM helped ensure greater target text consistency and the fact that this technology was more familiar and easier to implement, a feeling seemingly borne out by the average time spent on translating with this method (1:15, see Table 5), which is faster than WP and SR, and similar to MT. However, a comparison of individual student scores and attitudes shows a number of surprising discrepancies. The student with the lowest error score for this method (LTT) for instance, expressed dissatisfaction with the usefulness of the translation memory used for this task, but not with the tool and method per se. At the other end of the scale, LM, the student with the worst performance using this translation tool, found the method "adequate for the task, as it helps with the repetitions."

As regards machine translation, most students acknowledged that it was the fastest method and most were surprised by the relative quality of the raw MT output. However, they remained critical of the method because of the intellectual effort required to identify and correct the errors generated by the system, and to produce a readable and idiomatic target text. In this respect, they far preferred the use of a TM system, which they felt gave them more control over the translation process. Criticism was also levelled at the inconsistency of MT with regard to terminology, which required more effort to harmonize (we must remember that "untrained" MT was used in this experiment). If we examine the feedback from individual students, we find that three of those who performed best in this exercise (TL, CC, and HG) were favourably impressed by the MT output, but that this was also the case for CA, who ranked 11th out of 12. BP, whose post-editing performance was well "below par" (12th compared to 2nd in the benchmark translation) was very critical of the raw MT quality, while

being aware of her limitations in this particular exercise ("Google Translate may have inserted errors that I would not have committed in my own translation, and that I may not have detected through proofreading"). Other students similarly reported feeling unduly constrained by the unedited text and being unable to find a more fluent translation. Some reported that they felt that translating the text from scratch would have been more effective, but that this would have defeated the purpose of the exercise. However, the overall level of quality achieved with this method (higher than TM and SR) shows that most participants were able to successfully overcome the limitations of the method.

According to student feedback, speech recognition was the technology that generated the most questions and mixed feelings: while one student wrote that she/he "particularly enjoyed working with DNS [Dragon Naturally Speaking], which I found very reactive", another stated: "I hated working with this particular tool, because it was very slow and tiring to have to repeat the same sentences, and I would often forget my original sentence." Students who performed well with this tool generally appreciated the freedom afforded by SR technology, both in terms of ergonomics (less keyboard interaction) and in terms of the translation process (greater spontaneity, ability to detect problems by having to think through the sentence before voicing a translation, etc.). Others, however — particularly those who performed least effectively — stressed the technical constraints and the time required to sufficiently master the tool. Some pointed out that the technology is not suited for technical translation, where figures, abbreviations, the frequent use of upper-case letters, proper nouns and specific technical terms required additional effort and time in the proofreading phase (which is borne out by the longer average time spent on this method). However, even the most critical generally recognised the potential of the tool with other types of source texts and with more extensive training, and stated that they would be prepared to experiment further with the technology.

*Impact of translation tool or method on student performance, according to error typology and effect*

Our analysis of results on the quality produced using these four methods of translation would not be complete without an examination of the influence that the translation tools have on student performance in terms of translation error typology and error effect. Beyond the overall quality effect already discussed above, we examine whether different tools lead to different types of errors and different end-user effects, and how this reflects the translation process implemented by the participants. Our detailed grid makes this type of fine-grained analysis possible.

Results in Table 9 below show the breakdown of error-scores, weighted for effect and criticality according to our assessment grid, for each translation method. The percentages are calculated on the basis of the total quality score for each method (e.g. meaning-related errors represent 44.3% of the total of weighted error-scores awarded to all 12 students when using the WP method). They reveal clear differences in the types of errors identified in translations performed with the different translation technologies. Furthermore, for each error type, a variance analysis (mean comparison test) was performed to determine if the mean difference found between the four translation methods is statistically significant. Results shown in Table 9 (last row) represent the F-test score and its associated probability (p-value with a significance level of 5%). Only a p-value close to 0.05 (values represented in bold in the table) allows us to reject the null hypothesis (mean equality) and therefore to conclude that the mean difference between the different methods may be significant. Bold percentages for the methods signal the highest values per error type.

**Table 9 – Error typology score percentages by method (weighted for effect and criticality)**

| *Methods* | Meaning | Style | Terminology | Localisation | Addition/ omission | Grammar /syntax | Spelling | Phraseology | Dtp |
|---|---|---|---|---|---|---|---|---|---|
| **WP** | 44,3% | **17,3%** | 13,1% | **8,5%** | 8,2% | 5,7% | 2,8% | 0,0% | 0,0% |
| **TM** | **55,0%** | 10,5% | 8,4% | 2,3% | 12,0% | 3,6% | 6,6% | 0,8% | 0,8% |
| **SR** | 31,8% | 10,7% | **20,7%** | 4,4% | **17,3%** | **7,6%** | 5,3% | **1,3%** | **0,9%** |
| **MT** | 54,9% | 10,7% | 14,6% | 1,4% | 6,0% | 4,4% | **7,4%** | 0,5% | 0,0% |
| Variance analysis (DLL=3) | F=0.988, p=0.407 | F=1.233, p=0.309 | **F=3.231, p=0.031** | F=2.611, p=0.063 | **F=3.489, p=0.023** | F=0.846, p=0.476 | F=1.502, p=0.227 | F=1.536, p=0.218 | F=1.748, p=0.171 |

Looking first of all at the statistical significance of the above results, we find that the differences between the four methods for a given error type are not consistently significant (as shown by the statistical test results in the last line of Table 9). Only two error categories, "Terminology" and "Addition/omission" show significant differences (with p value <0.05) between the translation tools used.

Although these two categories do not correspond to the most common errors made by students (the most frequent errors are meaning errors), results tend to differentiate the SR method from other methods.

The "terminology" error rate found with the SR method (20.7% of the total error-score for that method) is particularly high compared to other methods, especially with respect to TM (8.4% of the total error-score).

These differences have to be seen in the light of our terminology analysis (cf. "Experimental design" above). This showed that the TM extract contained the highest percentage of terms not found in the glossary, and a high percentage of terms not previously encountered, while the SR extract not only contained fewer term units, but had a smaller proportion of "new" terms and terms not in the glossary. This would seem to indicate that while TM induced a greater awareness of terminological inaccuracy or inconsistency, SR led students to focus on sentence-level translation, and to neglect terminology checks and searches which would interrupt or slow down the process. They may therefore have relied more on intuition or

memory, while a number of other terminology errors could simply be due to mispronunciations or misinterpretation by the system.

For the "Addition/omission" category, Table 9 shows again a significant difference between the SR sample and the other samples. In the SR sample, this category represents 17.3% of the total error-score for that method, against 12% for TM, 8.2% for WP and 6% only for MT. Again, the very nature of oral translation with SR technology may explain this relatively higher percentage of omissions of words or parts of speech, due to the extra effort needed to memorise full sentences before dictating them to the system, or to the constant shifting between the reading process and the dictation process. On the other hand, as already observed by Daems, Macken and Vandepitte (2013, 70), the MT process naturally tends to lead to a smaller number of omissions than a "human" translation, since MT systems translate every source text item more literally than a human translator would generally do. Again, it is important to remember that students could correct the specific errors induced by one process or another during the revision or post-editing phase, but the lack of extra time or sufficient practice may explain the fact that they did not always do so (producing better overall quality with the WP process).

If we now turn to the error categories where the score differences were not found to be statistically significant in Table 9, general observations and tendencies may still be determined. We can first note that distortions of meaning represent by far the largest proportion of error-scores found with each of the four methods. This is particularly true (with almost 55% of the total) in the TM and MT samples. In translations produced with WP and SR, on the other hand, this type of error accounts for less than half the total of error-points, with the lowest percentage found with speech recognition (31.8%). The MT results can be explained by the fact that untrained raw statistical machine translation output is, by its very nature, uncontextualised, and requires careful checking against source text context, which our

trainee translators were unable — or unwilling — to do, as shown in the "Attitudes towards the translation technologies used" section. The lower percentage of meaning-related error-scores in the SR results may be explained by the mental process of sight translation associated with speech recognition technology, which forces the translator to analyse the source sentence meaning and formulate the target utterance mentally before committing the target sentence to the system. SR technology forces the trainee translator to consider a wider context and to take on board implicit meaning before voicing the translation, as described by Zapata Rojas (2012) and found in our earlier study (Toudic et al., 2013).

One could argue that these "better" results obtained with the SR method in meaning transfer could be due to the nature of the sample used in week 3, but the additional analysis conducted in the fifth phase (Trad5) of the project (students translating another identical sample with different methods) tends to confirm this tendency, as shown by the results presented in Table 10.

**Table 10 – Breakdown of error types percentages (weighted for effect and criticality) from an identical text with three translation tools (Trad5)**

| *Methods* | Meaning | Style | Terminology | Localisation | Addition/ omission | Grammar /syntax | Spelling | Phraseology | Dtp |
|---|---|---|---|---|---|---|---|---|---|
| **TM** | 63,6% | **13,3%** | 13,3% | 0,0% | 4,9% | 2,1% | 0,7% | 2,1% | 0,0% |
| **SR** | 41,8% | 10,0% | **22,4%** | 0,0% | **11,8%** | **8,2%** | 2,9% | **2,4%** | **0,6%** |
| **MT** | **64,4%** | **13,3%** | 7,4% | 0,0% | 6,7% | 3,7% | **3,7%** | 0,7% | 0,0% |
| Variance analysis (DLL=2) | F=0.252, p=0.782 | F=0.062, p=0.941 | **F=4.852, p=0.037** | - | F=0.607, p=0.566 | F=0.558, p=0.591 | F=2.824, p=0.112 | F=0.553, p=0.594 | F=1, p=0.405 |

In the text sample (Trad5) studied in Table 10, meaning errors again represented the highest percentage of error-scores in all three methods (MT, TM and SR), but, while they made up almost two-thirds of the total of weighted error scores for translations using TM and MT (63 and 64% respectively), the percentage fell to less than half (41.8%) with the speech recognition tool.

As with the previous results, significant differences between translation methods could be observed for the "Terminology" category errors, with a marked difference between the MT method (7.4% of all error-scores) and SR method (22.4% of all error-scores). Contrary to what was observed in the previous results, however, differences in the "Addition/omission" were not statistically significant for this text sample.

One last dimension in our analysis can be provided by the specific "error-effect" feature of our grid (Table 11).

**Table 11 – Breakdown of error effects in the first four weeks (total score for all 12 students)**

| Method | Accuracy | Usability | Readability | Compliance |
|---|---|---|---|---|
| WP | 110 | 74 | 72 | 50 |
| TM | 125 | 146 | 59 | 74 |
| SR | 128 | 157 | 62 | 103 |
| MT | 110 | 110 | 63 | 81 |
| Variance analysis (DLL=3) | F=0.299, p=0.826 | **F=2.424, p=0.078** | F=0.145, p=0.932 | **F=2.524, p=0.07** |

Firstly, out of four different types of functional effects, Usability and Compliance are the two effects which showed significant differences among methods (p values very close to 0.05: 0.078 and 0.07 respectively). For both of these effects, the WP method produced the best results, with the lowest deficiency scores (74 and 50, respectively). The SR method is the one which produced the highest deficiency score in that respect, with 157 and 103 in the Usability and Compliance categories. This could be linked to the high percentage of terminology error scores produced by the SR method (Table 10). Overall, the SR method showed the highest scores ("worst" quality) in three indices out of four. Although not statistically significant, the Readability category displayed the second best score (62) with this method. The actual process of dictating the translation, which encourages the students to produce fluent,

understandable parts of speech without writing interferences (Roux et al., 2013), could explain this score and the lowest meaning error scores observed in Table 10 (41.8%). The WP method surprisingly produced the "worst" quality in Readability (highest score, 72), even if this method showed the "best" quality (lowest scores) in all other effects. This result might be explained by the fact that the students were discovering the document when applying this method.

**Table 12 – Breakdown of error effects in week 5 (total score for 4 students per method)**

| Method | Accuracy | Usability | Readability | Compliance |
|---|---|---|---|---|
| TM | 58 | 48 | 24 | 26 |
| SR | 55 | 57 | 23 | 35 |
| MT | 48 | 50 | 14 | 23 |
| Variance Analysis (DDL=2) | F=0.330, p=0.728 | F=0.184, p=0.835 | F=0.728, p=0.509 | F=0.634, p=0.553 |

None of the scores in error effects for Trad5 was statistically significant, which can be put down to the small number of students (4) using each method. As in the first four weeks of the experiment, the SR method proved to be the worst method for Usability and Compliance, and the second "best" for Readability. Further studies should be conducted to confirm these tendencies on a larger sample.

## Conclusion

The study described in this chapter set out to examine the possible impact of different translation tools and methods on the translation quality produced by a group of Master's level translation students, using the TRASILT quality assessment grid as a measurement tool.

The aggregate results for the group did not reveal any statistically significant impact of a particular translation tool on the overall quality produced, but surprisingly, showed that the students produced marginally better quality when using only word processing and the glossary of key terms provided (Table 5). However, if both time and quality are taken into account, post-edited machine translation (using Google Translate) and the use of a translation memory tool appeared to be more efficient.

Using individual student performance without a specific translation tool as the benchmark score, individual scores were then compared. Although the use of dedicated translation tools did not appear to affect the quality produced by the students achieving the highest and lowest benchmark scores, performance was far less consistent for those between the two extremes. Some students performed particularly well or badly with one of the three translation tools while achieving good or average quality scores with the others.

The factors which could affect group and individual performance with different translation tools were then examined. Differences in readability and terminological density between the different extracts taken from the same instruction manual were not proven to be a contributing factor, as shown in the "Experimental design" section. This was borne out by the results in the fifth week of the study, where a same extract was translated using the three translation tools, and where the overall quality ranking (MT / TM / SR) was identical to that achieved in the previous phases. Regarding individual performance, the technical and organisational factors

related to the implementation of the different translation tools could explain a certain degree of quality variability, even though the results of those students who encountered particular technical difficulties in using the speech recognition system, for instance, were excluded from the study. This factor could only be eliminated completely by a detailed examination of individual student profiles (for instance: prior experience in the use of particular translation tools during internships), which was not possible in this study. Student attitudes towards particular translation tools were also examined, and there again, the most openly hostile to MT or speech recognition technology were excluded from the results. However, even among those who did not openly reject one or more of the tools used, personality and behavioural differences (e.g., cognitive processes, acceptance of technology, perception of one's own voice, etc.) are likely to explain certain inconsistencies in individual quality performance.

Using a three-dimensional assessment grid based on error type, error-effect and degree of criticality enabled us to take the quality analysis a step further. We first of all examined the effect of particular translation tools on the types of translation errors found in the aggregate results. The main observation was that, while only the "Terminology" and "Addition/omission" categories showed statistically significant differences according to the tool used, meaning errors accounted for the highest proportion of quality deficiencies with all the translation tools. However, the SR method produced fewer meaning errors overall, compared to translations produced with a translation memory system or post-edited machine translation. The oral translation process imposed by speech recognition was found to have a positive influence on meaning transfer, while omissions were more frequent when using this method, probably due to the memorisation process involved in sight translation. Finally, the four "error-effects" measured by the grid (Accuracy, Usability, Readability and Compliance) were examined for each translation tool or method. Again, WP (with no specific translation tool), produced the highest quality in terms of Accuracy, Usability and Compliance, while SR

(speech recognition) only performed better in terms of Readability. This type of analysis would now need to be extended to individual results, and confirmed by more extensive experimentation on a greater number of participants.

This study has also enabled us to demonstrate the usefulness of the TRASILT assessment grid as a translation quality measurement tool. As a research tool, it can be used in similar studies of tool- or context-related quality assessment. In a pedagogical context, it can be used, as in this study, to highlight a student's particular strengths and weaknesses, both in terms of error types, and in terms of error effect and criticality in a functionalist perspective. In a professional context, its flexibility allows evaluators to vary the weighting accorded to Accuracy, Usability, Readability or Compliance, according to the type of source text and the type of end-result that needs to be achieved.

Finally, the study has highlighted gaps in the grid's functionalities, in particular with regard to source text characterization and how different levels of semantic complexity and cognitive load may impact on quality assessment scores when using different translation tools and methods. Further research, involving wider and more diverse corpora, is required to refine and develop the assessment tool still further.

## References

Berman, Antoine. 1995. *Pour une critique des traductions : John Donne*. Paris : Gallimard.

Bowker, Lynne, and Melissa Ehgoetz. 2007. "Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation." In *Across Boundaries: International*

*Perspectives on Translation Studies*, ed. by Dorothy Kenny, and Kyongjoo Ryou, 209–224. Cambridge: Cambridge Scholars Publishing.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. "(Meta-) evaluation of machine translation." In *Second Workshop on Statistical Machine Translation, Proceedings*, ed. by the Association for Computational Linguistics, 136–158.
http://www.statmt.org/wmt07/pdf/WMT18.pdf

François Thomas, Brouwers Laëtitia, Naets Hubert, Fairon Cédrick. 2014. "AMESURE : une plateforme de lisibilité pour les textes administratifs". In *21ème Traitement Automatique des Langues Naturelles.*
http://cental.uclouvain.be/amesure/index.php

Chall, Jeanne S., and Edgar Dale. 2000. *Readability revisited: The new Dale-Chall Readability Formula*. Brookline Books, US.

Daems, Joke, Lieve Macken, and Sonia Vandepitte. 2013. "Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE." In *MT Summit XIV Workshop on Post-editing Technology and Practice, Proceedings*, ed. by Sharon O'Brien, Michel Simard, and Lucia Specia, 63–71. European Association for Machine Translation.
http://hdl.handle.net/1854/LU-4127483

Doherty, Stephen, and Sharon O'Brien. 2014. "Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking." *International Journal of Human-Computer Interaction* 30 (1): 40–51.
http://www.tandfonline.com/doi/abs/10.1080/10447318.2013.802199#.U6yVG_l_vjU

Dragsted, Barbara, Mees, Inger M., Gorm Hansen Inge. 2011. "Speaking your translation: students' first encounter with speech recognition technology". *The International Journal for Translation & Interpreting Research* 3 (1): 1-43.

http://trans-int.org/index.php/transint/article/viewFile/115/87

Gouadec, Daniel. 1981. "Paramètres de l'évaluation des traductions". *Meta* 26 (2): 99–116.
http://www.erudit.org/revue/meta/1981/v26/n2/002949ar.pdf

Gouadec, Daniel. 1989. “Comprendre, évaluer, prévenir”. *TTR* 2 (2): 35–54. http://www.erudit.org/revue/ttr/1989/v2/n2/037045ar.pdf

Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. 1975. “Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel] ”. *Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.*

Koehn Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. “Moses: Open Source Toolkit for Statistical Machine Translation“, In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ed. by the Association for Computational Linguistics, 177–180. http://anthology.aclweb.org//P/P07/P07-2045.pdf

Nord, Christiane. 1997. *Translating As A Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

O'Brien, Sharon. 2012. “Towards a Dynamic Quality Evaluation Model for Translation”, *Journal Of Specialised Translation* (17): 55–77. http://www.jostrans.org/issue17/art_obrien.pdf

Reiss, Katharina. 1981. “Type, Kind and Individuality of Text: Decision Making in Translation”, trans. Susan Kitron. *Poetics Today* 2 (4): 121–131.

Roux, Franck-Emmanuel, Jean-Baptiste Durand, Emilie Réhault, Samuel Planton, Louisa Draper, and Jean-François Démonet. 2013. “The neural basis for writing from dictation in the temporoparietal cortex.” *Cortex* 50: 64–75.

Toudic Daniel, Hernandez Morin Katell, Moreau Fabienne, Barbin Franck, and Phuez Gaëlle. 2014. “Du contexte didactique aux pratiques professionnelles : proposition d’une grille multicritères pour l’évaluation de la qualité en traduction spécialisée.” *ILCEA* 19. http://ilcea.revues.org/2517

Toudic, Daniel, Hernandez Morin Katell, Moreau Fabienne, and Phuez Gaëlle (forthcoming). “Impact de deux approches technologiques sur un panel d’apprentis traducteurs : aide ou obstacle sur le chemin du sens ?”. In *Actes du colloque international Tralogy II*.

Toury, Gideon. 1995. *Translation Studies and Beyond*. Amsterdam: Benjamins.

Vermeer, Hans J. 1979. “Vom ‘richtigen’ Übersetzen.” *Mitteilungsblatt für Dolmetscher und Ubersetzer* 25 (4): 2–8.

Williams, Malcolm. 2004. *Translation Quality Assessment: An Argumentation-Centred Approach (Perspectives on Translation series)*. Ottawa: University of Ottawa Press.

Zapata Rojas, Julian. 2012. *Traduction dictée interactive : intégrer la reconnaissance vocale à l’enseignement et à la pratique de la traduction professionnelle.* Thèse de doctorat non publiée. Ottawa : Université d’Ottawa. http://www.ruor.uottawa.ca/fr/bitstream/handle/10393/23227/Zapata%20Rojas_Julian_2012_these.pdf?sequence=1