

Translation Quality Assessment*

MALCOLM WILLIAMS

University of Ottawa

malcolmwilliams@rogers.com

Abstract:

The relevance of, and justification for, translation quality assessment (TQA) is stronger than ever: professional translators, their clients, translational researchers and trainee translators all rely on TQA for different reasons. Yet whereas there is general agreement about the need for a translation to be "good," "satisfactory" or "acceptable," the definition of acceptability and of the means of determining it are matters of ongoing debate. National and international translation standards now exist, but there are no generally accepted objective criteria for evaluating the quality of translations. What are the problems and issues that stand in the way of consensus and coherence in TQA? This article presents an updated argumentation-centred model to solve some of those problems

Key-words: acceptability, translation standards, evaluation criterion, argumentation-centred model.

Resumen:

La relevancia y la justificación para reflexionar sobre la evaluación de la calidad de las traducciones (TQA) es más fuerte que nunca. Los traductores profesionales sus clientes, los investigadores en traductología y los formadores de traductores, confían en la evaluación de la calidad de las traducciones por razones diferentes. Sin embargo, a pesar de que existe un acuerdo general sobre la necesidad de que las traducciones sean "buenas", "satisfactorias" o "aceptables", la definición de aceptabilidad y de la manera de determinarla es un asunto en continuo debate. Existen ahora estándares nacionales e internacionales de traducción, pero no existen criterios objetivos de aceptación general para evaluar la calidad de las traducciones. ¿Cuáles son los problemas y los elementos que se interponen para que haya consenso y coherencia en la evaluación de la calidad de las traducciones? Este artículo presenta un modelo centrado en la argumentación actualizado como propuesta de soluciones a estos problemas.

Palabras clave: aceptabilidad, estándares de traducción, criterios de evaluación, modelo centrado en la argumentación.

1. INTRODUCTION

The relevance of, and justification for, translation quality assessment (TQA) is stronger than ever: professional translators, their clients, translational researchers and trainee translators all rely on TQA for different reasons. Yet whereas there is general agreement about the need for a translation to be "good," "satisfactory" or "acceptable," the definition of acceptability and of the means of determining it are matters of ongoing debate. National and international translation standards now exist, but there are no generally accepted objective criteria for evaluating the quality of translations. What are the problems and issues that stand in the way of consensus and coherence in TQA? In this article I propose an updated argumentation-centred model to solve some of those problems.

*Article based on a lecture given at Universidad de Antioquia in Medellín, Colombia, as part of the research project "La Construcción del saber del traductor," for which Professor Williams is acting as advisor within the framework of the Graduate programme on Education, Major Didactics of Translation, Universidad de Antioquia.

2. WHAT IS TQA?

First, we have to situate TQA in a broader academic and research context and determine what TQA is about. In fact, people with an interest in translation studies are always evaluating — evaluating sources (their usefulness and authenticity), evaluating authors and their translators (their aesthetic, their influences and how this informs their work), evaluating source texts and evaluating target texts.

TQA is a type of evaluation, but what is evaluation? Michael Scriven, a leading evaluation researcher, defines it as follows: “‘Evaluation’ is taken to mean the determination of merit, worth, or significance” (2007: 1). This definition itself presents a difficulty: How do we define value or worth, be it moral, aesthetic or utilitarian? By extension, evaluation involves asking a question that has challenged thinkers from time immemorial: Is a particular thing *good*?

Just like evaluation in the broad sense, TQA can be quantitative or qualitative: it can be based on mathematical/statistical measurement (as in the case of most academic instruments) or on reader response, interviews and questionnaires (e.g. Nida). TQA can be diagnostic (determining areas for improvement at the outset of a course of study), formative (measuring progress and giving feedback during a course of study) or summative (measuring the results of learning).

Whether our focus is on products, performance or competence, we are essentially trying to determine degrees of *goodness* when we perform TQA. The approach can be prescriptive, assessing translation against criteria of aesthetic effect, usability, and intrinsic compliance with standards of target language correctness and fidelity. At the other extreme, recent research in descriptive translation studies (DTS), and in particular the work of Toury (1981; 1994) and Chesterman (1993), has shifted away from TQA based on absolute criteria and has reflected a relativistic, intersubjective stance on judgments of translation worth. Of course, this dichotomy raises the question as whether TQA is values-free. One of the motivating factors in DTS is precisely, the desire to make translation studies a values-free discipline. Indeed, the tendency in evaluation through the second half of the twentieth century was to avoid making value judgments about target programs. But Sergio Viaggio (1993: 125) makes a very good point about DTS:

Unless we agree with scholars such as Gideon Toury, for whom a translation is whatever a given culture considers it to be, prescriptivism is inescapable. If there is no right, or at least better way of translating, then we are all wasting time, breath and money: there is nothing but language to teach.

Michael Scriven (1993: 13) makes the same point in criticizing what he calls the values-free doctrine:

It makes the blunder of supposing that evaluative statements are all essentially arbitrary, like expressions of taste and preference, when in fact many of them have the solid status of derivations from facts and definitions.

Thus, in my view, TQA cannot and should not be values-free: to be useful, it must be based on criteria of *goodness*. Otherwise, all we do is describe defects and strong points in translations.

Naturally, we strive to be as objective as possible in designing and applying TQA models, and to be successful, we must ensure that our TQA models and procedures pass the test of **validity** and **reliability**.

Validity is the extent to which an evaluation measures what it is designed to measure, such as translation skills (construct validity). Content validity is the extent to which an evaluation covers the skills necessary for performance. For example, is the content of a translation examination an appropriate sample of the content of the course? Does the result of the evaluation accurately predict future performance (predictive validity)?

Reliability is the extent to which an evaluation produces the same results when administered repeatedly to the same population under the same conditions. Thus a TQA system is reliable if evaluators' decisions are consistent and criteria are stable. Are there biases or undue variations in results over time? Is there a mechanism for ensuring that evaluators do not fluctuate between excessive rigour (purism) and extreme flexibility (laxism)? Is the evaluator always objective? Are quality requirements clearly enough defined for decisions on borderline cases to be made with consistency and ample justification?

3. PROBLEMS AND ISSUES IN DESIGNING AND APPLYING TQA

Why is it so difficult to establish and apply a TQA model? There are many reasons. I consider the following ones to be the most important, and they inevitably entail problems of validity or reliability.

- a) *The evaluator*: Does the evaluator have the linguistic or subject-field knowledge required? The client, whose knowledge may be limited, inevitably evaluates the finished product too. Indeed, the client's assessment may be the only one. Further, a number of translation researchers, including Hönig and other functionalists, Dyson (1994) and Kingscott (1996) have implicitly or explicitly given precedence to the reader's response or requirements, not the translator's definition of an adequate translation, as the yardstick for gauging quality.
- b) *Level of target language rigour*: Elegant style is considered essential by some evaluators, but not by others. Some evaluators consider typos and spelling and punctuation errors to be peccadillos and ignore them in their overall assessment, while

others will regard them as serious because they are precisely the errors that the client/end user will detect.

- c) *Seriousness of errors of transfer*: The same inconsistency is apparent in the assessment of level of accuracy. Some evaluators will ignore minor shifts in meaning if the core message is preserved in the translation, while others will insist on total "fidelity," even if an omission of a concept at one point is offset by its inclusion elsewhere in the text. Reasons (b) and (c) underlie the frequent complaints about evaluator "subjectivity."
- d) *Sampling versus full-text analysis*: TQA has traditionally been based on intensive error detection and analysis and has therefore required a considerable investment in human resources. It takes time. One means of obviating the problem has been sampling — the analysis of samples of translations instead of whole texts. Yet this approach has shortcomings. First, the evaluator may not take into account any "compensatory" efforts that the translator has made in unsampled parts of the text. Second, the evaluator may not have taken into account the co-text in order to grasp the meaning of the text as a whole. Third, as Daniel Gouadec has pointed out, "There is always a risk that the most serious errors may lie outside the samples. This is especially true of the work of established translators, who are capable of dramatic, uncontrolled deviations from the meaning of the source text" (1989: 56).
- e) *Quantification of quality*: Microtextual analysis of samples has been used extensively not only because it saves time but also because it provides error counts as a justification for a negative assessment. Translation services and teachers of translation alike have developed TQA grids with several quality levels, or grades, based on the number of errors in a text of fixed length. It is felt that quantification lends objectivity and defensibility to the assessment. The problem lies with the borderline cases. Assuming that, in order to be user-friendly, such a grid does not allow for many levels of seriousness of error, it is quite possible for a translation containing one more error than the maximum allowed to be as good as, if not better than, another translation that contains exactly the maximum number of errors allowed and yet be rated unsatisfactory.
- f) *Levels of seriousness of error*: One way to circumvent the drawbacks of quantification is to grade errors by seriousness: major, minor, weak point, etc. The problem then is to seek a consensus on what constitutes a major, as opposed to a minor, error. For example, an error in translating numerals may be considered very serious by some, particularly in financial, scientific or technical material, yet others will claim that the client or end user will recognize the slip-up and automatically correct it in the process of reading.
- g) *Multiple levels of assessment*: Many authorities, including Nord (1991) and House (1997), identify a number of parameters against which the quality of a translation

should be assessed: accuracy, target language quality, format (appearance of text), register, situationality, etc. The problem is this: Assuming you can make a fair assessment against *each* parameter, how do you then generate an *overall* quality rating for the translation?

- h) *TQA purpose/function*: The required characteristics of a TQA tool built for formative assessment in a university context may differ significantly from one developed for predelivery quality control by a translation supplier. According to Hatim and Mason, "Even within what has been published on the subject of evaluation, one must distinguish between the activities of assessing the quality of translations [...], translation criticism and translation quality control on the one hand and those of assessing performance on the other" (1997: 199).

Clearly, "the devil is in the detail." It is not surprising that it has proved impossible to establish a quality standard that meets all requirements and can be used to assess specific translations. Hence, recently adopted national and international standards err on the side of caution and propose guidelines for quality control. They do not establish a standard of acceptability or levels of quality, nor do they provide TQA tools. The search for translation quality standards and measurement tools modelled on ISO quality standards and methods of industrial quality control is a worthwhile endeavour because it responds to the need for objectivity (through precise measurement and quantification) and for instruments that will enjoy widespread approval. However, a translation is an intellectual product and, as such, is a complex, heterogeneous one, not a physical unit, a piece of software or hardware, each part of which can be replicated exactly by a machine thousands of times. This is why TQA has proven to be so difficult and why TQA models have so many detractors.

4. TQA APPROACHES

A number of approaches have been proposed to resolve some, or all, of the problems of TQA. In this article, I will describe and comment on two representative models: Sical (quantitative approach) and Nord's Skopostheorie model (non-quantitative approach).

4.1. Sical

The trials and tribulations of the Canadian Government Translation Bureau's Quality Measurement System, called Sical, illustrates the limitations of a quantitative approach to quality. It was developed both as an examination tool and as a means of assessing the quality of the 300 million words of instrumental translation that the Bureau delivers yearly. Initially based on a very detailed categorization of errors — over one hundred types were identified and could be assessed by evaluators — the system evolved into a scheme based on the quantification of errors and on a twofold distinction between (1) transfer and language errors and (2) major and minor errors. In this third-generation Sical, texts were given quality ratings

according to the number of major and minor errors in a 400-word passage: A — superior (0 major errors/maximum of 6 minor); B — fully acceptable (0/12); C — revisable (1/18); and D — unacceptable. As such, it was a standards-referenced model: quality levels were defined in terms of the errors that a text of a given standard could contain.

The major error was defined as follows:

Translation: Complete failure to render the meaning of a word or passage that contains an essential element of the message; also, mistranslation resulting in a contradiction of or significant departure from the meaning of an essential element of the message.

Language: Incomprehensible, grossly incorrect language or rudimentary error in an essential element of the message. (Williams 1989: 26)

The key word is *essential*. It was left up to the quality controller to determine whether an essential element of the message was at issue.

In theory, then, a translation deemed “fully acceptable” could contain many errors of transfer. However, the designers predicated this high degree of tolerance on the statistical probability that a translation with many errors would also contain major error and would therefore not pass muster.

The typology of errors established in this context — a typology modelled on that of Horguelin (1978) — is indicative of the fact that the quality system by and large focussed on the word and the sentence, not on the text as a whole. Larose sums up the approach as follows: “The grid is based mainly on the syntactic and semantic aspects of the text, not on its discursive dimension, which lies *beyond* the statement and *between* statements” (1998: 175).

Macrotextual issues of coherence and cohesion are not apparent in the guidelines, even though the primacy of consequence of error and the relationship of major error to text message and recognition of the importance of adapting material to the reader would seem to open the door for a discourse-based approach.

For the purpose of assessing the quality of professional translators’ work before delivery to the client (quality assurance or control) or after (evaluation), evaluators were required to select one or more representative 400-word samples of texts. Thus the essential or nonessential nature of a word or passage was necessarily determined on the basis of the word or sample, not the text as a whole.

Notwithstanding the good judgment and competence of the evaluators and quality controllers involved and the flexibility and precision afforded by the system, the application of a quantified standard continued to spark debate and dissatisfaction among translators inside and outside the Bureau. Working conditions, deadlines, level of difficulty of the source text and the

"overassessment" of target language errors were regularly cited by opponents of the system. As a result, the "official" set of quantifiable quality ratings has been abandoned in favour of the industrial quality control objective of "zero defects," although the assessment of samples and quantification of errors continue.

4.2. Skopostheorie

In her **Skopostheorie** model, Christiane Nord elaborates on Katharina Reiss's premise (2000) of translation as intentional, interlingual communicative action and proposes a "translation-oriented" analysis based on the function and intention of the target text in the target culture and applicable to instrumental as much as to literary documents.

A text — and *a fortiori* a translation — is provisional until received: "Meaning is assigned to a text by the recipient in light of given function" (1992: 93). This approach necessarily relativizes the primacy of the source text and the debate over equivalence. The translation is not equivalent to the source text; it is an "offer of information about" it. However — and this is important for issues of TQA — Nord by no means downgrades the criterion of accuracy, notwithstanding the criticisms levelled at Skopostheorie. Introducing the principle of "functionality + loyalty" (1991b: 39), she sees the translator as being committed to both source and target texts and having to analyze both source text and target text in situation; depending on the function of the target text, the translator may intend to preserve all semantic and formal features of the original or adapt the source text material extensively. Thus, she envisages the possibility of establishing *grades* of required types of translation on a scale running from extreme fidelity to extreme liberty (1991b: 28).

Nord's notion of "grades" of translation quality is not comparable to those of the quantitative models. Her grades are based on a conscious decision to produce a relatively "literal" or relatively "free" translation and does not encompass a tolerance level for errors unwittingly committed by the translator.

Furthermore, Nord's approach necessarily takes the purpose of the translation and other variables into account. Recognizing the need to overcome the problem posed by recipients and possible translations, Nord presents the idea of translation instructions (specifications) specifying what type of target text is required — an approach that had already been formulated in the eighties by Simpkin (1983) and Sager (1989). It is on the basis of the instructions and the target text skopos expressed therein that the translator proceeds. The translator makes a judgment as to what in the source text is relevant to the target text skopos and produces a "grade" of translation along the fidelity/liberty continuum.

The evaluator must take the target text skopos as the starting point for TQA, assess the target text against the skopos and the translator's explicit strategies, and then do a source text/target text comparison for inferred strategies. Nord emphasizes that error analysis is insufficient: "[I]t is the text as a whole whose function(s) and effect(s) must be regarded as the crucial criteria for

translation criticism" (1991a: 166). This is a key qualification, for on the basis of a selection of relevant source text features, the translator may eliminate source text items, rely more heavily on implicatures, or "compensate" for them in a different part of the text. Indeed, as Van Leuven-Zwart points out in developing an interesting corollary of translation-oriented analysis, the "shifts in meaning" that account for many "unsatisfactory" ratings in professional translation should perhaps not be considered as errors at all, given that equivalence is not feasible (1990: 228-29). In short, microtextual error analysis is insufficient.

In the examples of translation-oriented text analysis presented to illustrate the model, Nord's judgments are generally parameter-specific (criterion-referenced), and when there is a judgment, it is not definitive. Indeed, she states that there "will be no overall evaluation of the translated texts" (1991a: 226). For example,

Neither translation gives a true impression of the ironic effect produced by the particular features of theme-rheme structure, sentence structure and relief structure. (1991a: 217).

[...] the literal translations [of a specific term] are not an accurate description of the subject matter dealt with in the text. (1991a: 222)

She does, however, make a definitive, overall judgment on the sample texts as a whole: "[N]one of [the translations] meet the requirements set by text function and recipient orientation" (1991a: 231). But how does she generate an overall assessment from the parameter-specific comparisons, particularly when her judgment is based on the nature of the errors, not their number?

5. PRESENT STATE: CONCLUSIONS AND ISSUES

Having highlighted the main characteristics of the two representative TQA models, we can draw a number of conclusions.

- a) Sical, as a quantitative, standards-referenced model, is essentially microtextual. Assessment generally operates at the subsentence level.
- b) By narrowing one's quality standard or goal down to "zero defects," an organization abandons the notion of a range of acceptable grades and the possibility of any consistent, objective, reliable measure of quality.
- c) Nord's criterion-referenced model is based not only on microtextual but also on discourse and full-text analysis and factor in text function, purpose and type. However, it does not propose clearly defined *overall* quality or tolerance levels. Nord's assessments are not related to a scale of measurable values. As

McAlester remarks, “in no case is any suggestion made concerning the amount and gravity of errors that can be tolerated for the total translation to be considered adequate” (2000: 234).

- d) Nord and other proponents of a textological approach (e.g. Larose, 1987; Bensoussan and Rosenhouse, 1990) all recognize and emphasize the interrelationship between the translation unit and the macrotext, in terms of the impact and seriousness of error. However, no definition of error gravity has been proposed on a scientific, theoretical, textological basis, and evaluators have to rely on ill-defined concepts such as "complete failure to render the meaning" and "essential part of the message." How is the "essential part" to be determined?

In the next section I propose an argumentation-centred approach as a solution to these shortcomings.

6. EVERYTHING’S AN ARGUMENT! A POSSIBLE SOLUTION

6.1. Importance of argumentative mode of discourse

Reiss and others have categorized texts according to various functions or modes. Reiss established three functional categories: *informative* (text conveying information), *expressive* (text communicating thoughts in a creative way) and *operative* (text that persuades). She maintains that the translator’s prime role is to ensure that the primary function of the source text is preserved in the target text.

However, seminal research on text content and function in the second half of the twentieth century suggests that one of these functions predominates. First, Perelman (1969) showed that Aristotle’s argumentation categories (lines of argument, or *topoi*, and figures of speech) are present in a broad range of text types. Second, Stephen Toulmin (1964; 1984) showed that all texts present an **argument macrostructure** comprising up to six key elements. This has prompted Lunsford et al. (2004) to proclaim that “Everything’s an argument!” That is, in fact, the title of their book, in which they detect and analyze arguments in every conceivable text type and medium. In essence, all texts are argumentative, even the dry, boring, informational ones, in that the writer is intending to persuade the reader to accept certain ideas and act upon them.

Accordingly, my thesis is that every instrumental source text contains an argument macrostructure and that it is this structure that the translator must preserve in the target text. This does not mean that other textual features and functions are not present and are not important. What it *does* mean is that preserving the argument macrostructure is the overriding consideration for TQA.

6.2. An argumentation-centred TQA model

The updated model proposed here (Williams, 2005, for an earlier version) draws primarily on philosopher Stephen Toulmin's analysis of argument structure and the work of some other philosophers and linguists who have focussed on issues of reasoning, coherence and cohesion in discourse (e.g. Halliday, Vignaux, Widdowson).

Our first step is to determine what an argument (reasoning) structure is. Toulmin explores arguments in a variety of areas of specialization and draws the conclusion that the components of an argument are essentially the same in all types of text and that the force of claims and assessments made in texts remain the same as we move from field to field. Objects, ideas and situations under discussion can be labelled "good," "appropriate," "satisfactory," "unsatisfactory," whatever the type, genre, purpose or area of specialization of the text. The key differences are the standards and assumptions by which judgments are made: The criteria against which an assessment or claim concerning the "goodness," "appropriateness," "effectiveness" or "correctness" of an object, idea or situation will vary from field to field (area of specialization). In short, "all the *canons* for the criticism and assessment of arguments [...] are field-dependent, while all our terms of assessment are field-invariant" (1964: 38): the generic elements of argument will be invariant while the specific kinds and content of those elements will depend on the field.

In a later work, Toulmin addresses the subject of argument (and reasoning) in terms of universal (field-invariant) and particular (field-variant) rules of procedure, and he proposes a set of elements that are required for an argument in any field — claims/discoveries, grounds, warrants/rules, and backings — and two elements that may be required — qualifiers/modalizers and rebuttals/exceptions/restrictions (1984: 25). Testing has shown that claims/discoveries and grounds are by far the two most commonly encountered and most critical components of Toulmin's model for TQA. They are defined and illustrated below.

Claim/discovery: The claim (or discovery) is the conclusion of the argument, or the main point toward which all the other elements of the argument converge. The following claims are typical of instrumental texts for translation: recommendations in a policy document or discussion paper; a request for a specific amount in a grant application; the announcement of a new health program; the judge's decision in an appeal case; the classification of a newly discovered plant as belonging to a particular order. We can see from this brief list that the areas of specialization are varied.

Grounds: Claims are not freestanding; they have to be supported by one or more pieces of information, which form the grounds of the argument. These are facts, statistics, oral testimony, matters of common knowledge, well-known truisms or commonsense observations, historical reports and so on, upon which the sender and recipient of the message can agree. The grounds for announcement of a new health program may be the observation, or report, of overcrowding

in emergency departments of hospitals. Note that a claim may be based on more than one ground. For example, the announcement may also be prompted by an infusion of new funds into the government's health budget.

6.3. Argument macrostructure and quality standards

If we accept the premise that the argumentative mode is omnipresent, then we can use Toulmin's argument macrostructure to define error not only inductively, on the basis of professional practice and empirical evidence, but also deductively, on the basis of established, widely accepted scientific, philosophical and discourse theory. Specifically, application of the macrostructure opens the door to a new concept: a translation defect that seriously undermines the usability of the translation because it impairs the core reasoning, and specifically the grounds and claims, in the text. Applying this definition, we find that not all defects considered major under the conventional models impair the argumentative macrostructure of a text.

Going one step further, we can now exploit the distinction made in classical quality control theory (e.g. Ishikawa, 1985) and add to the major and minor defect a third level of error seriousness: the critical defect. According to software developer Novell, "Critical defects are product defects that cause the performance or continued performance of any one or more mission-critical...functions to be impossible" (2009). It is our contention that any translation breach of the argument macrostructure of the source text impairs the "mission" of that text — to persuade the reader to accept and/or to do something — and thereby makes "performance" of the argumentative function impossible. Accordingly, in an argumentation-centred TQA (ARTRAQ) model,

- defects impairing translation of the argument macrostructure will be characterized as **critical**;
- other transfer defects conventionally considered major (*contresens*, *charabia*, etc.) will be characterized as **major** but will be deemed not to render the translation unusable;
- other transfer defects will be characterized as **minor**.

We now have a definition of defect (error) based on established industrial quality control theory and practice and the work of a philosophical authority, not just on empirical evidence from translation practice. The critical defect would correspond to the text-level **misinterpretation** identified by Bensoussan and Rosenhouse (1990). They consider the determination of such a defect to be potentially valuable as a means of streamlining student evaluation by focusing on the degree to which the trainee has rendered the core message(s) instead of engaging in microtextual error analysis.

More importantly, we also have an industrial and academic basis for defining an acceptable level of quality, or a minimum standard, in translation: **an acceptable translation is one that**

fully conveys the argument macrostructure of the source text and is therefore free of critical defects.

With respect to TQA procedures and the issue of full text versus sample analysis, the evaluator will save time by reading the complete translation to identify problem areas and restricting detailed TQA to passages containing argument macrostructure components.

6.4. ARTRAQ parameters

Having added a third level of defect (error) seriousness, we can now propose an evaluation grid according to two sets of parameters: core and field- or use-specific. The core (argument macrostructure) parameters will apply to all instrumental translations, whether they are produced by professionals or students, whatever their end use may be, and whatever subject field is involved. The field- or specific-use parameters will be activated at the evaluator's discretion, in light of the contract or work statement at hand, the field of specialization, and the intended use of the translation. In broad terms, parameter selection will depend on whether the translation is for information purposes or for publication.

The field/use-specific parameters typically selected by the evaluator would include terminology, figures of speech, format and target language quality. In fact, it would no doubt be appropriate to subdivide target language quality into style, usage and grammar, and typography, since their relevance and importance varies with field and use. For example, style — in which we include issues of redundancy, repetition, concision and plain language — may be of scant consequence for the translation of an administrative report but are of considerable importance in the translation of directives and instructions. Likewise, typographical errors may not be a major factor in assessing the quality of internal documents but will have major consequence in public signage. In the case of an internal report for information purposes, terminology and format may be important, but matters of target language quality may be secondary; accordingly, the translation will be assessed against the first two but not against the target language parameters.

6.5. Development of a rating scale

Having established the core and non-core parameters, we want to continue working toward our objective of a TQA model that reduces quantification of defects (errors) to a minimum. We also want to avoid the situation obtaining in the conventional quantitative model, where all types of minor error were given the same weight.

The multicriteria model proposed by Larose (1994: 369) makes it possible for the evaluator to reflect the relative importance of each parameter in the overall, final assessment. It requires that each parameter be assigned a specific weight prior to assessment and that the quality level for a specific parameter, as determined by the evaluator, be weighted accordingly in the establishment of an overall rating.

Larose draws inspiration for his model from a criteria-based analysis published by Nida, in which each of three translations of the same original is assessed against six key parameters, called “isomorphs” — referential meaning, rime, concision of lines of poetry, etc. According to Nida, “Isomorphs are essentially a way of looking at the basic problem of equivalence. But what is important about isomorphs is that they force the analyst to specify the formal and semantic features in such a way as to measure and describe the degrees of conformity. *Since isomorphs always come in sets of features, they force literary critics and translators to think in terms of patterns and not in terms of isolated resemblances and differences*” (Larose 1998: 179) [our emphasis].

Larose’s multicriteria analytical table, like Nida’s, is designed for a criterion-referenced assessment and specifically for evaluation of students and candidates, but he builds an explicit recognition of the varying importance of the parameters into his model. One proposed framework described by Larose is as shown below (1994: 369):

Criterion	1	2	3	4	5	6	Total
Weighting	10	7	7	8	9	9	
Translation							
A	7/70	6/42	8/56	7/56	4/36	5/45	305
B	6/60	5/35	9/63	6/48	6/54	7/63	323
C	4/40	8/56	1/7	7/56	5/45	9/81	285

Here, the top line, “Criterion,” represents the different parameters against which translation quality is to be assessed: transfer, terminology, typography, etc. Each criterion is then weighted from 1 to 10, depending on its importance to translation quality. The evaluator assesses each translation (A, B, C) against all criteria, giving it a mark out of 10 for each criterion and multiplying that mark by the weighting factor: thus, translation A rates a mark of 7 against criterion 1 (which has a weighting factor of 10) and thus earns 70 points (7 x 10) toward its total score, which is made up of the sum of the points earned against each criterion.

It is common practice to base weighting factors on percentages, so we can adapt Larose’s model to make it more user-friendly. Assuming that a translation with no defects is worth 100%, the evaluator would decide what portion of that 100% was represented by the selected parameters and express that portion as a decile. He would then rate translation quality for each parameter, using a rating system commonly applied in both criterion- and norm/standards-referenced assessment: 10 = excellent; 8 = very satisfactory; 6 = satisfactory; 4 = fair; 2 = poor. The resulting weighted TQA grid would cover core (argument macrostructure: grounds and claims) parameters and field- or end-use-specific ones.

6.6. Application of the model

The following passage is taken from the conclusion of an article in which the writer presents statistical evidence (“grounds” in Toulmin’s theory) for her conclusions (“claims”): (1) feminist women in Canada do not comprise a single, monolithic bloc sharing the same ideology but in fact represent a diversity of schools of thought; (2) conservative feminist women, representing a more traditional set of social values, may be more reluctant than liberal feminist women to promote feminist ideas but can nonetheless succeed in doing so.

Source text

Une première conclusion consiste à suggérer que, contrairement à l’impression d’unité qu’il projette, le féminisme n’a rien d’un mouvement unifié; il se tisse plutôt d’idées sur lesquelles les femmes *qui s’en réclament ne s’entendent pas*. En d’autres termes, il n’y a pas une seule et unique façon d’être féministe, de même que la pensée féministe ne peut se limiter à un ensemble de préceptes *fermement emboîtés les uns dans les autres, à la manière d’un jeu de logo*. Il faut dire que cette diversité au sein du féminisme n’est pas nouvelle; elle a animé le mouvement féministe de la deuxième vague, *où se côtoyaient les féministes libérales, socialistes, marxistes, et celles du courant psychanalytique, entre autres tendances*. Aussi, une seconde conclusion veut interpréter cette diversité parmi les femmes féministes non comme une limite, mais bien comme un atout pour la représentation politique des femmes. Cette diversité permet au féminisme de se manifester sous des formes diverses et selon des intensités plurielles. Bien que les féministes conservatrices *soient plus* timides que les féministes libérales dans leur engagement face à certaines idées généralement associées au féminisme, elles permettent néanmoins à ces idées d’accéder aux zones conservatrices de l’espace politique canadien — à des zones où ces idées ne pourraient peut-être même pas accéder en leur absence. Autrement dit, les messagères adaptent le message à l’environnement où il est appelé à s’exprimer. Il reste à voir maintenant si les féministes conservatrices expriment effectivement ces idées généralement associées au féminisme dans les châteaux forts du conservatisme canadien et, le cas échéant, si elles agissent en conséquence.

Target text

A first conclusion suggests that contrary to the impression of unity that it projects, feminism is not (**T — minor defect**) a unified movement; it is woven from ideas on which women *do not claim to necessarily agree* (**T — major defect**). In other words, there is no one method to being (**L — minor defect, usage**) a feminist, in the same way as feminist thought is not limited to a set of ideas *firmly entrenched in stone* (**T — major defect**). It must be said that this diversity within feminism is not new; it *has*

given life to the second-wave feminist movement *where liberal, socialist, Marxist and psychoanalytic feminists border on other trends* (**T — critical defect**). A second conclusion interprets (**L — minor defect, over-personification**) these diversities amongst feminist women not as a liability, (**T — minor defect, punctuation**) but as an asset with a view to (**T — minor defect**) the representation of women in politics. This diversity allows feminism to manifest itself in different shapes with varying intensities (**L — minor defect, usage**). Although conservative feminist women *are no more* (**T — critical defect**) hesitant than liberal feminist women in (**L — minor defect, propositional usage**) supporting certain ideas generally associated with feminism, they have allowed (**T — minor defect**) these ideas to enter conservative areas in the Canadian political arena, areas where these ideas could not even exist if they were absent (**T — minor defect, ambiguity**). In other words, the messengers adapt the message to the environment in which it will be expressed. It remains to be seen whether conservative feminist women will (**T — minor defect**) effectively express those ideas generally associated with feminism within very conservative areas and if so, to see whether they will then (**T — minor defect**) act on them.

The passage contains the two claims of the article. The translator fails to exploit the clear logic of the statistical evidence–claim arguments established and reiterated by the author. In the second sentence, for example, the author uses the propositional function of elaboration (Widdowson, 1978) to develop her claim (conclusion 1) of feminist diversity, arguing by means of the *logo* (read *Lego*) metaphor that feminist ideas are not necessarily interrelated, and defining the diversity of the earlier, second-wave feminist movement as the coexistence of a number of feminist schools of thought described in detail at the beginning of the article. The translator renders neither the idea of interrelatedness nor that of coexistence, and the author’s elaboration of the diversity claim is thus severely diminished. In addition, the error in tense (“has given life”) gives the reader the impression that the second-wave movement is a present-day phenomenon.

In addition, the translator’s misunderstanding of the Lego metaphor generates a concomitant failure to render an argument from definition (division) bearing on the composition of the feminist movement.

The author’s elaboration of the second conclusion or claim — that conservative feminist women are more reluctant to communicate and promote feminist ideals than are liberal feminist women but can succeed in doing so — is also compromised in the target text because of the mistranslation “are no more hesitant,” which contradicts the intent of the source text at the microtextual level and the argument macrostructure developed by the researcher.

We therefore assess two critical defects: one for *where liberal, socialist, Marxist and psychoanalytic feminists border on other trends* and the other for *are no more hesitant*. The reader cannot use the co-text to “reconstitute” the source text claims. The two defects assessed

as major — *do not claim to necessarily agree and firmly entrenched in stone* — would be considered just as serious as the two critical defects were it not for the fact that they would not, in and of themselves, prevent the reader from understanding the claims conveyed by the source text.

We now use the weighted grid to give the translation a quantitative score. We have also assessed a number of minor transfer and target language defects and, since the translation is to be published, it must achieve a high rating not only against the argument macrostructure parameters but also against target language parameters, which must be included in the grid and given considerable weight. In addition to requiring the maximum score for transfer of the source text claims, the evaluator may require a minimum non-weighted score of 8/10 for other elements of transfer and target language parameters. Thus, if we give argument macrostructure a weight of 4, other elements of transfer a weight of 1, usage and grammar a weight of 4, and typography a weight of 1, we obtain a minimum acceptable score of $(4 \times 10) + (1 \times 8) + (4 \times 8) + (1 \times 8) = 88\%$.

Article for Publication: Weighted ARTRAQ Grid

Parameter	Weight (/10)	Quality (/10)	Minimum weighted score (/100)	Actual score (/100)
Grounds/Claims	4	2	40	8
Other elements of transfer	1	4	8	4
Usage/Grammar	4	6	32	24
Typography	1	10	8	10
Total	10	(/100)	88	50
Rating	Unsatisfactory			

Note: In this revision of the initial model (Williams, 2005), I have replaced the broad core parameter “Argument schema” with explicit reference to the two “critical” argument macrostructure features. In addition, I have dropped specific reference to other argumentation parameters such as “Conjunctives” and “Arguments.” Testing has shown that defects in transfer of these features are generally captured in the analysis of argument macrostructure (schema). Accordingly, in the interests of providing a more user-friendly grid, I have replaced them with “Other elements of transfer.” This does not mean that such argumentation parameters should not be incorporated in the grid in other circumstances: again, parameter selection is at the evaluator’s professional discretion.

The ARTRAQ model is flexible and modular: it can be adapted to different end uses, text types and areas of specialization. For example, had the translation been for information purposes only and not for publication, more weight might have been given argument macrostructure and other

elements of transfer at the expense of target language parameters. Yet the rating would not change — a satisfactory rating is predicated on full transfer of the argument macrostructure. Had the translator met that requirement, then the translation “for information purposes” would have been satisfactory because of the lower weights attributed to target language parameters.

But how does the assessment of student or trainee translations fit into the model? The target quality level necessarily depends on the learning objectives to be attained, and the accurate rendering of all manifestations of argument macrostructure may be too lofty a goal to reach, even for graduating students. This does not prevent us, however, from incorporating the macrostructure into the student translation rating grid and weighting it differently from industrial translations. In short, a flexible, comprehensive model must include a number of “standard” grades, and the narrative rating (e.g. satisfactory/unsatisfactory, acceptable/unacceptable) must be determined in relation to the relevant standard.

6.7. Argumentation-centred quality standards

In light of the foregoing, it is now possible to propose the following set of standards adapted to context and end use:

Maximum/publication standard

The text renders all grounds and claims and satisfies all non-core parameters. It contains no critical or major defects. (This standard would be applied to the translation in the example.)

Minimum professional standard

The text accurately renders all grounds and claims. It contains no critical defects. (This standard would be applied to translations for information purposes.)

Student standard

The text renders grounds and claims and satisfies non-core parameters **in accordance with assigned learning objectives.**

Substandard

The text fails to render grounds and claims and fails to satisfy one or more of the non-core parameters relevant to its specific function or the assigned learning objectives.

7. CONCLUSION

ARTRAQ (revised) offers solutions for the shortcomings of other models:

- a) It reflects assessment of both microtextual and text-level features.
- b) While eschewing the idealism of “zero defects,” it nonetheless provides for

TQA against a very high standard **where required in light of the end use of the translation.**

- c) Unlike the two illustrative models, it is both criterion-referenced and standards-referenced, successfully incorporating parameter-specific assessment in an overall quality rating.
- d) Because it is modular and adaptable, it makes it possible to focus the assessment on the criterion or criteria of interest and ensure the validity of assessments across the various conditions of production. In the specific context of translator training, for example, it seems to offer greater potential for **content validity** — i.e., for targeting the broad range of skills necessary for achieving translation competency — than does the conventional quantitative grid.
- e) With the revised definition of the critical defect to cover the critical components of argument macrostructure, the model provides for the application of standards based on generally accepted industrial and academic theory and practice and, in that sense, it ensures a more “objective” TQA.

REFERENCES

- Bensoussan, Marsha, and Judith Rosenhouse. "Evaluating Students translations by Discourse Analysis." *Babel* (1990): 65-84
- Chesterman, Andrew. "From 'Is' to 'Ought': Laws, Norms and Strategies in Translation Studies." *Target*, 5, 1 (1993): 1-20
- Dyson, Steve. "'Horses for Courses' or Meeting Customer Needs, Both Implicit and Explicit." Catriona Picken, ed. *Quality—Assurance, Management and Control* (1994): 87-94
- Gouadac, Daniel. *Le traducteur, la traduction et l'entreprise*. Paris: Afnor, 1989.
- Halliday, M.A.K., and Ruqaiya Hasan. *Cohesion in English*. London: Longman, 1976.
- Horguelin, Paul. *Pratique de la révision*. Montreal: Linguattech, 1978.
- House, Juliane. *A Model for Translation Quality Assessment*. Tübingen: TBL Verlag Gunter Narr, 1977.
- . *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag, 1997.
- Ishikawa, K. *What is Total Quality Control? The Japanese Way*. Englewood Hills, N.J.: Prentice-Hall Inc., 1985.
- Kingscott, Geoffrey. "Providing Quality and Value." *The Translator's Handbook*, 3rd edition (1996): 137-46
- Larose, Robert. *Théories contemporaines de la traduction*. Sillery: Presses de l'Université du Québec, 1987.
- . "Qualité et efficacité en traduction : réponse à F. W. Sixel." *Meta*, 34, 2 (1994): 362-73
- . "Méthodologie de l'évaluation des traductions." *Meta*, 43, 2 (June 1998): 163-86
- McAlester, Gerard. "The Evaluation of Translation into a Foreign Language." *Developing Translation Competence*, edited by Christina Schäffner and Beverly Adab (2000): 229-42

Nord, Christiane. *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*. Amsterdam-Atlanta: Rodopi, 1991a.

———. "Scopos, Loyalty, and Translational Conventions." *Target*, 3, 1 (1991b): 91-109

———. "Text Analysis in Translator Training." Dollerup, Cay, and Anne Loddegard, eds. *Teaching Translation and Interpreting*. Amsterdam: John Benjamins, 1992: 39-48

———. "The Relationship Between Text Function and Meaning." Thelen, Marcel, and Barbara Lewandowska-Tomaszczyk, eds. *Translation and Meaning*, 2. Maastricht: Rijkshogeschool, 1992. 91-96

Novell. "Software Defect Report," <http://support.novell.com/additional/bugreport.html>

Perelman, Chaïm, and Lucie Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. Translated by J. Wilkinson and P. Weaver. Notre Dame, Ind.: University of Notre Dame Press, 1969.

Picken, Catriona, ed. *The Translator's Handbook*. 2nd ed. London: Aslib, 1989.

Reiss, Katharina. *Translation Criticism — The Potential and Limitations*. Translated by Errol F. Rhodes. Manchester: St. Jerome Publishing, 2000.

Sager, Juan. "Quality and Standards — The Evaluation of Translations." *The Translator's Handbook* (1989): 91-102

Scriven, Michael. *Hard-won Lessons in Program Evaluation*. San Francisco: Fossey-Blass, 1993.

———. "Key Evaluation Checklist" (KEC)."
http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf

Simpkin, Richard. "Translation Specifications." *The Translator's Handbook* (1989): 129–139

Thelen, Marcel, and Barbara Lewandowska-Tomaszczyk, eds. *Translation and Meaning*, 1. Maastricht: Rijkshogeschool, 1990.

Toury, Gideon. "Translated Literature: System, Norm, Performance: Toward a TT-Oriented Approach to Literary Translation." *Poetics Today*, 2, 4 (Summer/Autumn, 1981): 9-27

- . *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins, 1994.
- . “A Handful of Paragraphs on ‘Translation’ and ‘Norms’ in Translation Studies.” *Translation and Norms*, edited by Christina Schäffner (1999): 9–31
- Toulmin, Stephen. *The Uses of Argument*. Cambridge: Cambridge University Press, 1964.
- , et al. *An Introduction to Reasoning*. 2nd edition. New York: Macmillan, 1984.
- Van Leuven Zwart, Kitty. “Shifts in Meaning in Translation: Do’s and Don’t’s?” *Translation and Meaning, I* (1990): 226–233
- Viaggio, Sergio. “The Limitations of a Strictly Socio-Historical Description of Norms: A Response to Theo Hermans and Gideon Toury.” *Translation and Norms*, edited by Christina Schäffner (1999): 122–28
- Vignaux, Georges. *L’Argumentation : essai d’une logique discursive*. Geneva: Droz, 1976.
- Widdowson, H. G. *Explorations in Applied Linguistics*. Oxford: Oxford University Press, 1978.
- Williams, Malcolm. “Creating Credibility out of Chaos: The Assessment of Translation Quality.” *TTR*, 2, 2 (1989): 13–33
- . “The Application of Argumentation Theory to Translation Quality Assessment.” *Meta*, 46, 2 (June 2001): 326–44
- . *Translation Quality Assessment: An Argumentation-Centred Approach*. Ottawa: University of Ottawa Press, 2005.