Towards a Dynamic Quality Evaluation Model for Translation

Sharon O'Brien Dublin City University

ABSTRACT

One of the dominant current methods for translation quality evaluation within the translation industry, i.e. the error typology, is seen as being static and unable to respond to new text types or varying communicative situations and this is leading to rising levels of dissatisfaction. This paper reports on findings from a benchmarking exercise carried out in collaboration with the Translation Automation User Society (TAUS) and eight of its member companies. Eleven translation quality evaluation models were benchmarked. A number of the companies profiled their translatable content according to the parameters of utility, time and sentiment. A review of quality evaluation models in domains related to professional translation leads to a list of alternative quality evaluation methods. The communication channel, the content profile and the parameters of utility, time and sentiment are merged to form building blocks towards a more dynamic quality evaluation model for translation. Examples are given for how such a model could be implemented.

KEYWORDS

Quality evaluation, quality model, professional translation.

1. Introduction

The assessment of translation quality has received much attention in the academic sphere (e.g. House 1997; Nord 1997; Lauscher 2000; Brunette 2000; Colina 2003). The focus here is not on the theory of translation quality assessment, nor on its implementation in translator training, but rather on the practice of translation quality evaluation in the professional sphere where translation quality is also ever topical and contentious. Quality is closely linked with customer opinion and yet quality evaluation (QE) in the translation industry is managed by gatekeepers in the supply and demand chain who work with static evaluation models, the majority of which are based on counting errors in random samples, applying penalties and maintaining thresholds with little, if any, input from customers. What's more, current QE models are predicated on a static and serial model of translation production, which is not suited to the emerging models of ubiquitous computing and 'everyware' (Cronin 2010).

Recently, TAUS, the Translation Automation User Society, registered an appetite among its members for a change in the static and normative, time-consuming modes of translation quality evaluation. It seemed that members were dissatisfied with the current 'one size fits all' approach and with the fact that little consideration was given to variables such as content type, communicative function, end user requirements, context, perishability, or mode of translation creation (i.e. whether the translation is created by a qualified human translator, unqualified volunteer, Machine Translation (MT) or Translation Memory (TM) system or a combination of these).

Kelly and DePalma (2009) have confirmed that translation QE is problematic in an even wider context. In consultation with 30 large-scale buyers of translation services, they conclude that numerous challenges exist for quality evaluation, including subjectivity, time, inappropriate use of linguistic resources, learning curve and technology. There is, therefore, some evidence for dissatisfaction with current QE methods.

The search for new ways of measuring quality is not only driven by the challenges listed above, but also by a number of recent developments:

Budgetary constraints:
- In recessionary times, companies are seeking ways to cut costs and time to market. The current quality evaluation approach is time-consuming and, therefore, costly.

New Paradigms:
- Current industry-based QE models were developed for high volume text-based translation. However, the notion of "text" itself is changing, with tweets, blog postings, multi-media and user-generated content all playing a bigger role in the translation production cycle. New content types may benefit from new approaches to quality evaluation.

New technology:
- While Machine Translation technology is not new, recent improvements have led to a noticeable increase in implementation. This development in itself demands a new approach to quality evaluation, especially when the paradigm of "fit-for-purpose" quality is gaining a foothold.

New focus:
- With the development of Web 2.0 technologies, users have taken more control over products, services and content. Companies are now generally paying more attention to the end user and this has led to an increased focus on the end users' perception of quality.

The author of this article undertook a project in collaboration with TAUS to investigate the current state of affairs in industry-based translation QE and to explore the potential of a dynamic QE model.[1] The first part of this paper reports on a benchmarking exercise, carried out in 2011, of 11 current quality evaluation models. Eight of the models included in the exercise are in active use by companies engaged in translation as clients. Six companies operate in the multi-national IT arena, one operates in the field of consumer electronics, and one in medical devices.[2] Three publicly available quality evaluation models were also included, because of their influence and popularity in industry: the LISA (Localisation Industry Standards Association) QE model (v. 3.1), the J2450 and the EN15038. The LISA model is well known in the localisation industry and many company-specific QE models have been derived from it. The J2450 is a standard generated by the Society for Automotive Engineers (SAE) and the EN15038 is a European standard, approved by CEN (Comité Européen de Normalisation) in 2006, whose purpose is "to establish and define the requirements for the provision of quality services by translation service providers" (CEN 15038: 4).

In section 2 of the paper, I reflect on quality evaluation models in areas linked with translation, such as Technical Communication. The most popular QE models in these related areas are selected and listed as possible candidates for inclusion in a dynamic translation QE model.

Following the benchmarking exercise, some of the companies engaged in a content-profiling exercise where specific types of company content were identified and then rated by internal stakeholders along the dimensions of "utility", "time" and "sentiment." Utility refers to the relative importance of the functionality of the translated content. Time refers to the speed with which the translation is required and sentiment refers to the importance of impact on brand image, i.e. how potentially damaging might it be to a translation client if content is badly translated? Part three of this paper discusses the outcome of the content profiling exercise.

In the fourth part of the paper, I set out to explain what might be the building blocks for a dynamic QE model and give examples of how specific content types and communication channels might map onto specific quality evaluation models, which are identified in section 2.

I do not claim to provide a complete and robust dynamic quality evaluation model, but provide some basic concepts and preliminary steps towards such a model.

2. Benchmarking quality evaluation models

A small sample of TAUS members were invited to collaborate on this project, resulting in the benchmarking of eleven QE models. The companies were selected by TAUS on the basis that they represented different business sectors and that they may be interested in novel methods of quality evaluation. The companies also had to be willing to share their existing quality evaluation models. The eleven models can be divided into two broad categories. The first category views translation quality from the point of view of errors in the translated product. Ten of the QE models fit into this category. The second category views translation quality in a more holistic manner, from the point of view of service provision and the competences, tools and procedures required to produce high quality translation. This is a process-oriented QE model and the EN15038 standard falls into this category. The latter does not specifically list translation errors and so cannot be directly compared with the others. However, it does list the issues that translators should attend to while translating (e.g. terminology) and these issues can broadly be equated with some of the macro error categories identified in the other ten QE models.

Each model was analysed individually, with attention being paid to the error classification, penalties applied, severity levels etc. Then a comparison was made across the models to draw conclusions about the general approach to Quality Evaluation. If any uncertainties arose in the analysis, the author had access to relevant contacts in each company who could explain the nuances in each model and how the model was applied in the translation process.

Preliminary conclusions were reported and sent to each company. Following this, a conference call was organised where company representatives had the opportunity to give their feedback on the preliminary conclusions, which were then updated accordingly.

It was found that the error-based models seek to identify errors, classify them, allocate them to a severity level and apply penalty points with a view to deciding whether or not the translation meets a specific pass mark. Two companies overtly list a positive category in their models, where outstanding translation solutions can be highlighted by the evaluator. However, no positive points are associated with this category and so outstanding translation solutions do not offset translation errors.
All but one of the error-based QE models assess quality on a segment by segment basis, giving no consideration to the larger concept of an 'information element' or 'text'. The one QE model that differs evaluates quality on a task or job basis, while also allowing reviewers to give feedback on a segment level.

2.1 Pass/fail

All error-based models utilise a pass/fail threshold. The number of words to be reviewed is specified by some models but not by others. However, the formulae typically normalise the scores per 1,000 words and the threshold for a pass/fail lies, on average, at three to four minor errors per thousand words or one major error per thousand words (see under *Error Severity, Weightings and Penalties* for an explanation of minor/major).

2.2 Content type

In addition to the main error categories, some of the QE models include specific error categories for Desk-Top Publishing (DTP) and software user interface (UI), while others do not or use the same error categories for all content types. For those where specific categories for DTP/UI were not included, the companies confirmed that translation quality testing for DTP and UI content is a separate step in the QE process. This difference between QE models makes it difficult to compare models for the DTP and

UI categories. However, for information purposes I have listed the most common DTP and UI error categories below.
The models that include error categories for User Assistance (i.e documentation and online help), DTP and UI tend to have different tolerance thresholds for each content type, with tolerance for UI translation errors being lower than for User Assistance. Other than that, however, there is little sensitivity built into the models for end user type, text function, perishability of information etc.

2.3 Language/Text type specificity

It has been suggested in the past that tolerance to language quality varies from one language community to the next, with Japanese and French being highlighted as two language communities that have low tolerance for translation errors (Kelly and DePalma 2009). Although companies tend to have language-specific style guides, none of the QE models included in this analysis take target language into account. In other words, the error categories and penalties do not change according to target language. In fact, the J2450 sets an explicit aim to be source and target language independent.

Even though source language quality is seen to have a major impact on translation quality (Kelly and DePalma 2009), none of the models make reference to source language quality as a potential cause of translation errors, although some mention the possibility that errors may be carried over from Translation Memory Exact Matches in the context where translators are instructed not to edit such matches.

As mentioned above, there is some sensitivity to content type in a limited number of the models. One model has a separate QE procedure for marketing material which is not a sentence-by-sentence review, but a review of the entire content along the dimensions of purpose, style and tone. This review is done by a marketing reviewer, as opposed to a language reviewer, and the aim is to establish the former's satisfaction. The marketing reviewers are asked to rate on a scale their satisfaction according to appropriate level of formality for the audience, whether the original purpose is conveyed, and whether the tone, style and register is appropriate for the target audience. A second model contains numerous references to suitability to target audience and readability of content as factors that are of high importance in the review process.

2.4 User focus

In the QE models reviewed there is an implicit and sometimes explicit reference to the impact an error might have on user experience. For example, one company's QE process flow for terminology errors states that "Severity: depends on impact on user experience." For the same QE model, one criterion for a critical error is if the error is "in a highly visible part of the document causing functionality/usability issues." An example from another QE model is where a significant change in meaning "means that the user is very likely to be misled." Yet another example of where utility is taken into account is in the LISA model which states that the translated documentation should "cater specifically for the consumer it is directed towards" (LISA QA 3.1). Thus, there is some evidence of a focus on users as a measurement concept in the QE models reviewed, but this is somewhat limited and implicit.

2.5 Error categorisation

All QE models contain specific error categories, the largest of which is 'Language'. Inevitably, different terms are used for the same category (e.g. 'Language' vs. 'Linguistic'), in which case what was included in the specific category was examined and decisions were made as to whether or not differently named categories corresponded. DTP and UI categories are also compared, but, as already mentioned, not all models include these categories. The EN15038 specification does not specifically include 'error categories' but is included here because it lists items that translators should attend to during translation and these corresponded well with error categories in the QE models.

2.6 'Language' errors

For errors in Language, the following sub-categories were prevalent:

- Language (10 out of 11 models)
- Terminology (10 out of 11)
- Accuracy (9 out of 11)
- Style (7 out of 11).[3]

The less prevalent categories were:

- Country standards (4 out of 10)
- Mistranslation (3 out of 10)[4]
- Consistency (3 out of 10).

It would appear from the list above that certain error categories are common across many of the QE models included in the benchmarking exercise, but a closer analysis of what is included in each category was necessary to determine whether specific error categories are comparable in terms of what they include in each macro category. An analysis of the top four error categories under 'Language' is presented below.

'Language'
Ten out of eleven QE models refer to the 'Language' category. Of these, nine include grammar in their definitions and seven include syntax. Other dominant criteria are punctuation (six references) and spelling (seven

references) although three QE models include punctuation and spelling as criteria that are separate from 'Language.' While there seems to be general consensus as to what is meant by the 'Language' category, there are some less common criteria listed too, such as 'typos,' 'fluency' and 'cultural references.'

'Terminology'
Ten out of eleven models specifically refer to terminology as an error category, though of these the EN15038 goes no further than to mention terminology as something translators must attend to. There is a general consensus across all of the QE models that include the category 'terminology' that errors in this category mean (1) lack of adherence to a client-specific glossary (or other reference materials); (2) lack of adherence to industry-specific terminology and (3) lack of consistency in term usage. Additionally, three QE models mention inappropriate use in context as a criterion for a terminology error.

'Accuracy'
Nine out of eleven QE models refer to this concept. Seven include unnecessary omissions, additions and inaccurate cross-references in their definitions for accuracy while seven also include meaning errors.

'Style'
Seven out of eleven QE models mention 'Style' as an error category, with J2450 specifically ruling it out and EN15038 again only mentioning it as something translators (and revisers) need to pay attention to. Four of the seven list 'lack of adherence to client style guide' as a criterion for a style error. Apart from this, and unsurprisingly, there is little consensus around what constitutes a stylistic error. Some criteria include tone, register, language variants, slang, literal translations and awkward syntax. Of the top four 'Language' errors, 'Style' is the one with the least consensus across models.

2.7 DTP errors

The J2450 specifically rules out any error type that is not language-related and the EN15038, while mentioning DTP, does not list error types in general. I include here the most prevalent DTP categories that were mentioned in the other QE models.

- Layout (5 out of 11)
- Formatting (4 out of 11)
- Graphics (3 out of 11)
- Index (2 out of 11)
- TOC (Table of Contents) (2 out of 11)
- Functionality (tags, variables) (2 out of 11)

2.8 UI errors

The LISA QA model (v.3.1) contains a detailed list of DTP and UI-related translation errors. As with DTP, I list here the most prevalent UI categories mentioned, but emphasise that this list is unlikely to represent all software-related QE errors.

- Integrity of localised text (3 out of 11)
- Alignment (2 out of 11)
- Layout (2 out of 11)
- Truncation (3 out of 11)

2.9 Error severity, weightings and penalties

The majority of models contain three severity levels which can be summarised as:

- minor
- major
- critical,

though some use slightly different terminology for each of these levels.
As exceptions, the J2450 and one company QE model include only two severity levels and one company specifically states that it avoids severity levels and weightings on the basis that it makes quality assessment too complex. Another company has five levels, but these are not specifically "severity" levels, but rather star ratings, with 1 being the worst and 5 being the highest.

Four QE models contain a fourth class of error severity which could be placed before 'minor' on the scale of severity and refers primarily to errors that are 'not the fault of the translator' or issues that are considered 'preferential.' This latter class tends not to have any weightings or penalties associated with it and is used as a mechanism to highlight and track issues that cannot necessarily be classified as errors but which need attention nonetheless.

There is general agreement across the models regarding the meaning for the three main categories: 'minor' errors are those that are noticeable but which do not have a negative impact on meaning and will not confuse or mislead the user. 'Major' errors are considered to have a negative impact on meaning, while 'critical' errors are considered to have major effects not only on meaning, but on product usability, company liability, consumer health, safety and behaviour.

The weightings and penalties applied for each error category are, however, not comparable across models. For example, one QE model has 'zero tolerance' for critical errors while another penalises such errors with

10 penalty points. Also, the penalties or weightings vary according to content type (e.g. UI vs. documentation) and according to error type (e.g. terminology vs. punctuation).

While the penalties and weightings are not comparable, there is general agreement on the level of tolerance to errors, which can be said to be very low. On average, one critical error will lead to a "Fail" on the quality evaluation and three minor errors in 1,000 words are tolerated, after which the translation is deemed to have failed.

The J2450 metric is the only metric which addresses the *subjective* nature of error categorisation and severity levels in its preamble by highlighting that the allocation to specific error categories is often a judgement call by evaluators. It offers two 'meta rules' for error categorisation, i.e.

1. When an error is ambiguous, always chose the earliest primary category (in the J2450 list of error categories).
2. When in doubt, always choose 'serious' over 'minor.'

2.10 Applicability to translation memory and machine translation

Only five QE models make explicit reference to Translation Memory tools in relation to translation errors. For example, one company QE model rates a terminology mismatch with a TM 100% match as a Low Severity Terminology Error. The same QE model rates changes to existing (incorrect) 100% matches as a preferential "error" category, i.e. this results in an instruction to the translator to rectify an error in a TM match that may not have been introduced by that translator and so is not penalised. Another QE model overtly includes TM matches (fuzzy and 100%) in the review sample and a third lists corrupted TM tags as a "functional" error. A fourth QE model refers specifically to 'ICEs' (In-Context Exact Matches): the translator is normally instructed not to edit these match types and so penalising them for errors in such matches is deemed to be unfair.

The only QE model that refers explicitly to Machine Translation (MT) is the J2450 where it is stated that the aim is "to establish a consistent standard against which the quality of translation of automotive service information can be objectively measured regardless of the source language, regardless of the target language, and regardless of how the translation is performed (i.e. human translation or machine translation)" (SAE J2450: 1).

2.11 Recurring errors

Only three of the QE models give instructions on how to deal with recurring errors. In two cases, the model specifically rules out the counting of repeated errors. In the third case, whether or not an error is counted more than once depends on the nature of the error; if the error results from translator negligence or lack of grammatical knowledge, the error is counted each time it occurs. If, on the other hand, the error is not the fault of the translator (e.g. the term was not included in the glossary), it is counted only once.

2.12 Feedback

The granularity of feedback enabled by the evaluation models is very fine-grained. In many of the QE models reviewed, every single error in the sample is logged with details such as Project Name, File Name, Segment Number, Source Text Segment, Translation, Proposed Fix, Comments on Error and Error Type, Severity. As QE can be a circular process, there is also a set of columns for tracking responses such as whether the fix has been implemented, vendor comments, final comments from the quality reviewer etc. This level of granularity no doubt results in a highly time-consuming task, which is duplicated across multiple languages for each translation project.

2.13 Tools

Of the models reviewed here, most of the QE process is carried out in very detailed spreadsheets, though some companies stated that they have developed their own in-house tools too. Although some QE tools exist on the market, it seems there is little uptake of such tools, with the companies in this sample preferring to manage QE either in spreadsheets or in proprietary tools.

2.14 Conclusions and questions arising from benchmarking exercise

The error-based QE models reviewed here demonstrate a relatively high level of agreement in terms of the macro error categories used for evaluating translation quality in user assistance content, with the unsurprising exception of the category "Style." The micro categories in each macro class also demonstrated significant agreement. This may be due to the influence of the LISA QA model throughout the IT industry in particular (the majority of QE models reviewed here come from IT companies). However, the penalties and weightings applied differed from one model to the next. One of the striking trends observed in this exercise is the preference for a segment-level error analysis over a holistic user-focused evaluation.

Another striking observation from the benchmarking exercise is the general lack of a holistic view of quality. Some aspects of *utility* are taken into account, though in a limited way. There is some sensitivity to content type. However, the focus is on segments and not text or even 'information elements' and the majority of models generally do not take text type, function, user requirements or perishability into account.

3. Quality evaluation in related contexts

To establish what alternative quality evaluation procedures might be available to form part of a more dynamic framework, a review was carried out of how quality is measured in the related contexts of Machine Translation (e.g. FEMTI (online); LDC 2005; Callison Burch et al. 2007; Przybocki et al. 2008; Papineni et al. 2002; Blatz et al. 2004; Specia et al. 2009), Translator Training (e.g. Nord 1997; Colina 2008; Gouadec 2010; MeLLANGE (online)), Community Translation (Ray and Kelly 2011; Rickard 2009) and (monolingual) Technical Communication (Byrne 2006; The Information Standard (online)). For economy of space the main quality evaluation methods that were noted are summarised below, but it should be emphasised that this is not an exhaustive list.

The list of evaluation models proceeds from the most controlled to the least controlled:

(1) Adherence to regulatory instruments
(2) Usability evaluation
(3) Error typology
(4) Adequacy/Fluency
(5) Community-based evaluation
(6) Readability evaluation
(7) Content sentiment rating (thumbs up/down, rating allocation)
(8) Customer feedback (Sales, Tech Support Calls etc.).

3.1 Adherence to regulatory instruments

This type of evaluation involves establishing whether (translated) content meets with the requirements of relevant in-country or continental regulatory standards. It is most likely to be carried out for translated products in the context of health and safety (e.g. medical devices) by a certified body.

3.2 Usability evaluation

Testing of (translated) content for usability can be achieved through a number of devices, e.g.

- Comprehension tests
- Questionnaires
- Participant observation
- Screen recording
- Think-Aloud Protocols
- Eye Tracking.

Usability evaluation is expensive and time-consuming and is most likely to be carried out for translated products in the context of health and safety by a usability expert.

3.3 Error typology

This involves the use of translation error typologies, such as those recorded in the benchmarking exercise above. Content (or a random sample of it) is evaluated by a qualified linguist who flags errors, applies penalties and establishes whether the content meets a pass threshold. This is the type of evaluation that is currently common in the translation sector.

3.4 Adequacy/fluency

The adequacy/fluency model is in use in Machine Translation evaluation. In the context of MT evaluation, the measurement of 'adequacy' requires comparison with a reference sentence. However, where a reference sentence is unavailable, and/or where MT was not used to generate the translation, I suggest that the concept of 'adequacy' could be adopted to mean: 'how much of the meaning in the *source text* is also expressed in the translation'? The response to this is usually given on a five-point scale of _All/_Much/_Half/_Little/_None (following the scale of Przybocki et al. 2008). The measurement of 'Fluency', simply defined as 'how fluent the translation is', does not require a reference sentence and is also rated on a five (or seven) point scale ranging from 5=Flawless to 1=Incomprehensible. This type of evaluation would be carried out by a qualified linguist but could also be done by a trained bi-lingual product specialist.

3.5 Community-based evaluation

As the name suggests, this model stems from the community translation model in which members of the community collaborate in a relatively uncontrolled way and negotiate an acceptable level of translated quality, based on agreed user needs and preferences. A Community Evaluation+ model is one in which the final say lies with the manager of the translated product (or his/her in-house staff). The traditional qualified linguist is not necessarily involved but the role of translator/reviewer could of course be adopted by a qualified linguist. Community-based evaluation could of course involve some of the evaluation models mentioned above, e.g. the Adequacy/Fluency model, the error typology, or might also entail discussion and negotiation between stakeholders until a consensus is reached based on implicit or explicit evaluation criteria.

3.6 Readability evaluation

Measuring the readability of translated content can be achieved in several ways. For example, traditional readability indices can be used (though I urge caution with these as they are generally not considered to be reliable indicators of translatability, cf. O'Brien 2010; Hvelplund 2011); end users can be asked to rate, on a scale of 1-5 or 1-7 for example, the reading ease of content; users can also be asked to participate in comprehension or recall tests, or in eye tracking studies (e.g. Doherty et al. 2010). This type of evaluation is monolingual, focuses on the target content only and could be carried out by an in-house marketing expert or by an end user. Translation accuracy is obviously not something that can be rated in this model, but appropriateness for the end user can be.

3.7 Content sentiment rating

Content sentiment rating is an evaluation model that also draws on the community translation model. Unlike the community-based model mentioned above, it is monolingual. Target language users are asked to rate the content along some parameter such as Like/Dislike, Comprehensibility Rating, Usefulness Rating, etc. We differentiate here between manual content sentiment rating and automatic sentiment analysis (see, for example O'Hare et al. 2009).

3.8 Customer Feedback

The Customer Feedback model is the least controlled QE model. Here, no specific QE is performed, but the publisher takes account of sales and the number of technical support calls or complaints received as a measure of translation quality.

3.9 Conclusions on QE in Related Contexts

We conclude from this summary of QE in contexts related to professional translation that, in addition to the error typology approach, there are numerous methods of quality evaluation which could be applied to professional translation in a more dynamic way. I return to these QE methods in the final part of the paper where I give examples of how different approaches to quality evaluation might apply to different contexts.

An important point to make here is that a dynamic approach to translation quality evaluation would not take the translated product as its starting point. Quality control of the content creation process would be the preferable starting point for a more encompassing QE model. In addition, quality control of the translation processes, along the lines of the EN15038 standard, would play an important role. Many proposals can be made for how to improve source content and translation processes, but limitations in space would not allow additional detail here. For readers who are interested in source content quality, however, Kohl's global English styleguide is a starting point.

4. Building blocks: communication channels and content profiles

The communication channel and content profile are considered to be two of the main building blocks in a more dynamic QE model. More detail on what is meant by communication channel is provided here and this is followed by results from the content profiling exercise engaged in by some of the collaborating companies.

4.1 Communication channel

There is a wide variety of communication channels in use today which could impact on quality expectations. Regardless of which communication channel is used, if content is created in a regulated environment, e.g. in the medical sector, it is subject to very specific quality guidelines and so is treated as a special channel. The communication channel might be internal to a company (e.g. internal training material), therefore this is treated as a second special channel. For external communication, it is suggested that there are at least three channels for the flow of translated content:

1. B2C: Business to Consumer
2. B2B: Business to Business
3. C2C: Consumer to Consumer.

The C2C model caters for multi-lingual user-generated content, which is consumed by other multi-lingual consumers (e.g. tweets, blogs, user forum postings etc.). The nature of the communication channel will impact on (translation) quality expectations and, ultimately, on a dynamic quality evaluation model. For example, if quality is regulated within a sector, then there will be little flexibility in the choice of translation quality evaluation model. A B2C communication channel will presumably require a stricter evaluation model than a B2B or C2C model. Quality expectations for internal communication might be lower than for external communication and so on. An additional factor is the mode of translation generation. If translation is generated by qualified and experienced translators, quality expectations will be justifiably higher than for translation generated by volunteers who are inexperienced in translation, or by a Machine Translation system. To be clear, the suggestion is not that quality *should* be compromised as one moves from one channel to another but that a dynamic QE model would take into account the varying tolerance thresholds for quality that already exist in the professional sphere. Ultimately, the quality model cannot be divorced from the communication channel and so I endeavour to build this factor into the dynamic QE model.

4.2 Content profiling

A dynamic QE model should cater for variability in content type, communicative function, end user requirements, context, perishability, or mode of translation generation. As outlined earlier, our proposal for the basic parameters in this dynamic QE model are utility, time and sentiment. To test the idea that different content types might have different requirements for utility, time and sentiment, participating companies were asked first to make a list of their distinct content types and company stakeholders were then asked to rate the content types for utility, time and sentiment according to whether they were considered to be 'Critically Important' (5 on the scale) or 'Unimportant' (1 on the scale). Five out of the eight companies responded.

The labels for the content profiles were, of course, different from company to company. Here the completed content profiles for five companies are compared (three from the IT sector, one from the medical sector, one from consumer electronics). One company had quite a granular list of content types (17), but the other four listed 9.5 content types on average (see Figure 1).

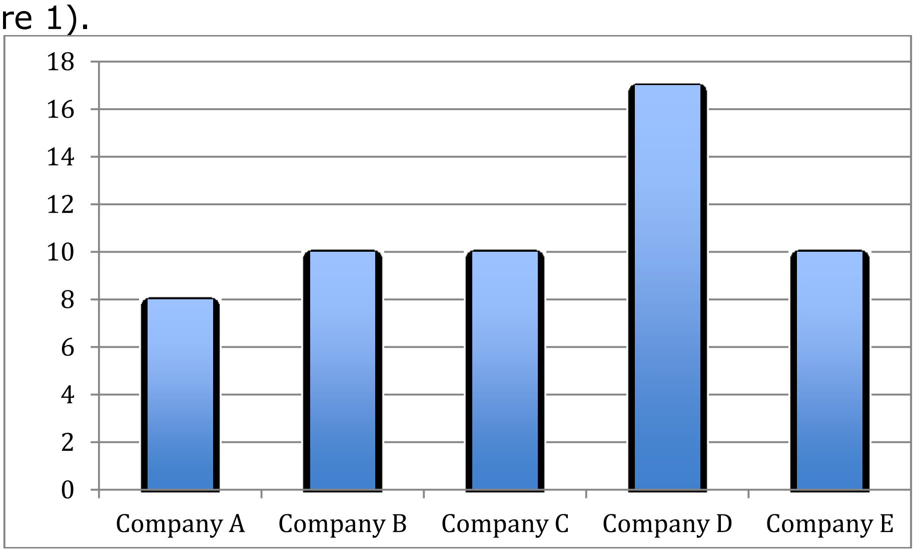

Figure 1. Quantity of Content Types

When the content types are grouped into meta-categories, there are eight:

1. User Interface Text
2. Marketing Material
3. User Documentation
4. Website Content
5. Online Help
6. Audio/Video Content
7. Social Media Content
8. Training Material.

Some companies identified specific content types that do not fall into one of the eight categories above, e.g. "Glossary," "Presentation" or "Press Release." Also, not all participating companies had content profiles in all of these categories. Figure 2 shows how often each meta-category was listed in the five companies.

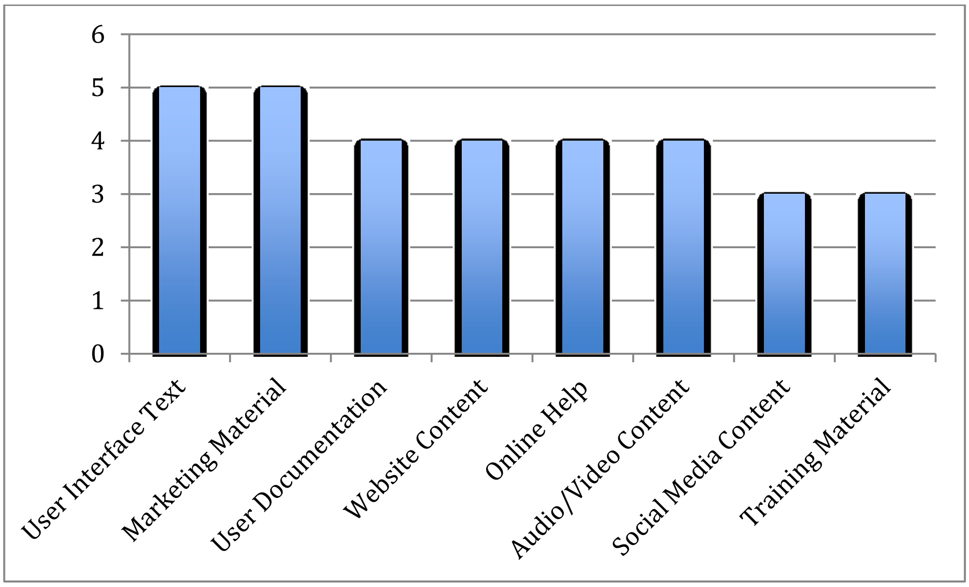

Figure 2. Common Content Types/Number of Companies.

4.3 Mapping content profiles to evaluation parameters

Our next objective was to determine how each content type was rated for utility, time and speed across each company. First, each proprietary content type was mapped to one of the meta-categories identified in Figure 2. If the content type did not fall into one of these categories, it was ignored for this particular analysis. If content was rated as Critically Important or Very Important, it was allocated to the category "High"; if it was rated as Important, it was allocated to the category "Medium;" and if it was rated as Slightly Important/Unimportant, it was allocated to "Low."[5] Thus, High, Medium and Low represent the importance allocated to utility, time and speed for each meta content type.

Figures 3, 4 and 5 show the percentage of times specific content types were allocated to High, Medium and Low. The total number of respondees across the five companies was 47. However, not all respondees rated each category, so the total does not always amount to 47. For that reason, the percentage of respondees for each meta category is calculated. Also, sometimes two or more profile types fell into one meta-category. For the sake of simplicity, only the highest scoring sub-type per category was included.

The figures are sorted according to the highest scoring content type down to the lowest. As a reminder, utility refers to the relative importance of the functionality of the translated content. Time refers to the speed with which the translation is required and sentiment refers to the importance of impact on brand image.

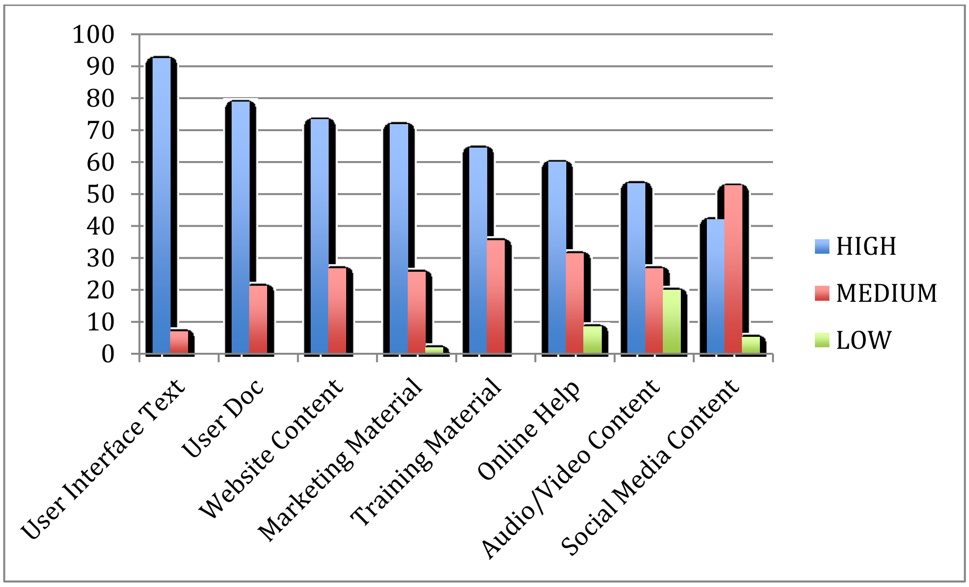

Figure 3. Importance Ratings (in %) for "Utility."

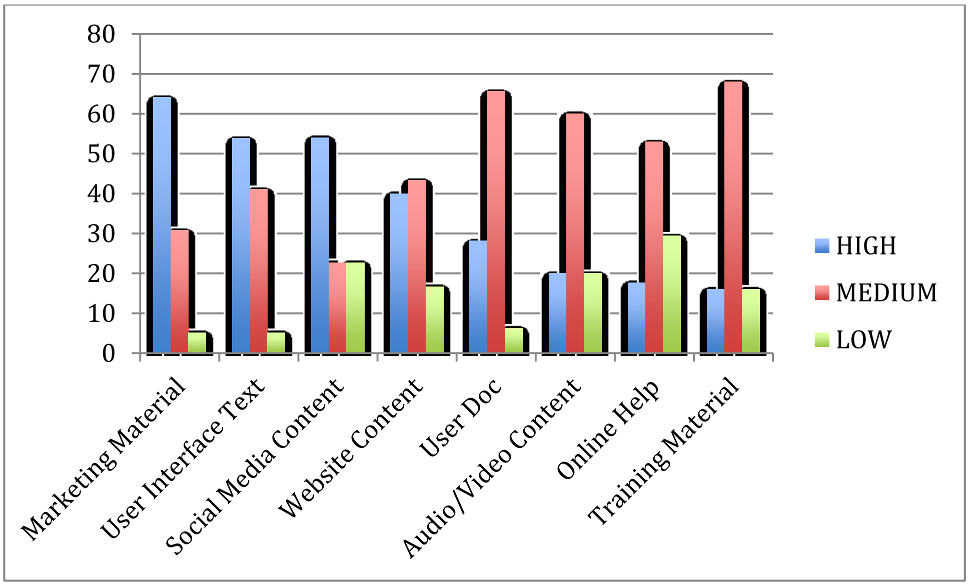

Figure 4. Importance Ratings (in %) for "Time."

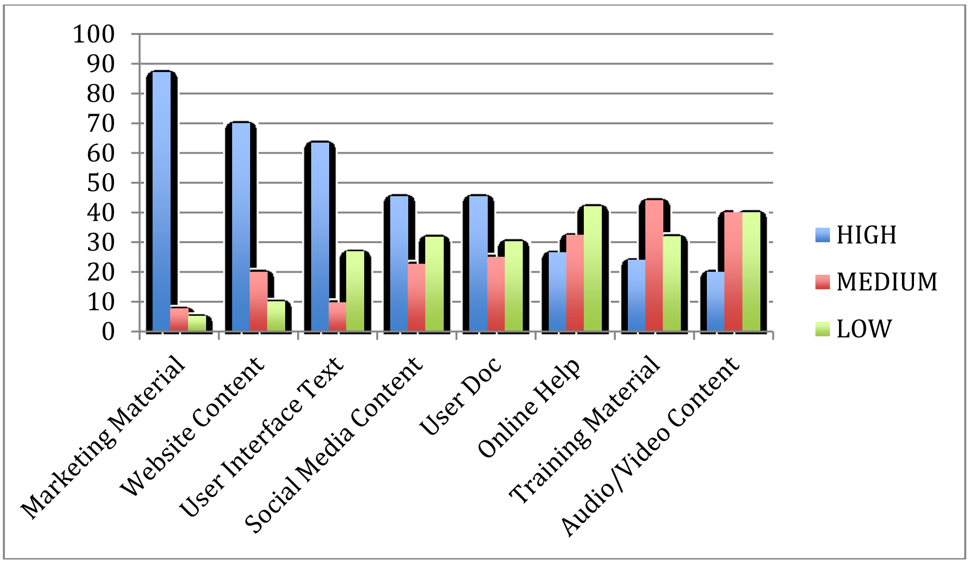

Figure 5. Importance Ratings (in %) for "Sentiment."

A summary of the conclusions from the surveys follows:

- User Interface text and User Documentation is rated highest for utility, while Audio/Video content and Social Media Content is rated lowest.
- Marketing Material and User Interface text is rated highest for time. Although Online Help and Training Material are not considered to be highly important for this parameter: they are considered to be of medium importance by a high percentage of respondees.
- Marketing material and Website Content are given highest importance for sentiment and Training Material and Audio/Video Content is rated lowest for this parameter.

The results of this small survey suggest that there are clear content differentiators for the utility and sentiment parameters, but the parameter of time is fuzzier. This is not surprising since most companies want content translated and evaluated as quickly as possible, regardless of the content type.

5. Mapping evaluation parameters to evaluation models

This section outlines a tentative suggestion for how the building blocks of communication channel, content profile and parameters for utility, time and sentiment might be brought together into a more dynamic QE model. The tables that follow are *examples* of how these building blocks might be used to determine which quality evaluation method is most appropriate. Having identified the communication channel, the profile of the content and having rated the content according to utility, time and sentiment, a number of QE models are proposed, moving from the highest level of control to a lower level. The person in charge of quality evaluation would then consider what is involved in each QE model and decide, on the basis of contextual factors, which model to apply.

Communication Channel	Content Profile	UTS Ratings	Recommended QE models in descending order of control
Regulatory	User Interface	U: *** T: ** S: *	1. Adherence to Regulatory Instruments 2. Usability Evaluation 3. Error Typology

*** = high importance, **= medium importance, *=least important

Table 1. Regulatory Channel, User Interface Content, Utility rated as highest.

Communication Channel	Content Profile	UTS Ratings	Recommended QE models in descending order of control
Internal	Training Material	U: ** T: *** S: *	1. Adequacy/Fluency 2. (Internal) Community-based Evaluation

Table 2: Internal Communication Channel, Training Material, Time rated as highest.

Communication Channel	Content Profile	UTS Ratings	Recommended QE models in descending order of control
B2C	Website Content	U: * T: ** S: ***	1. Community-based Evaluation 2. Readability Evaluation 3. Content sentiment rating

Table 3. B2C, Website content, Sentiment rated as highest.

Communication Channel	Content Profile	UTS Ratings	Recommended QE models in descending order of control
B2C	Website Content	U: * T: ** S: *** +MT	1. Adequacy/Fluency 2. Community-based Evaluation 3. Readability Evaluation

Table 4. B2C, Website content, Sentiment rated as highest, MT generated content.

Communication Channel	Content Profile	UTS Ratings	Recommended QE models in descending order of control
B2B	User Assistance	U: ** T: *** S: *	1. Adequacy/Fluency 2. Readability Evaluation 3. Content Sentiment Rating

Table 5. B2B, User Assistance, Time rated as highest.

It is unlikely that agreement can be easily reached on the sequence of recommended models proposed for each scenario above because potential users will have different levels of tolerance to errors. However, the idea is not to get agreement but to have a dynamic model that can be tailored according to preferences and tolerances.

6. Moving forward

The TAUS QE benchmarking exercise demonstrated that the preferred method for evaluating translation as a product in the translation industry is the error typology, with associated penalties and severity levels. This model, while appropriate in some contexts, cannot cater well for emerging content types, various communication channels and new needs. A more dynamic approach to QE seems to be needed by at least some members of the translation production sector.

An overview has been provided of different methods for evaluating quality in related contexts and the content profiles and utility, time and sentiment ratings from five companies were reported. On the basis of these, suggestions are made for how a dynamic QE model might be implemented.

This paper represents preliminary steps towards an alternative way of evaluating quality than that which is currently practiced by a large proportion of the translation industry. Of course there are weaknesses and limitations in what is presented here. While the sample of companies involved spans IT, the medical and consumer electronic sectors, it is not representative of all companies and sectors. The content profiles also only represent that of the companies who participated, though there is a reasonable level of confidence that it reflects content profiles in many other companies. The cost of switching from the current standard methods of QE to alternative methods has not been considered and responses to these proposals from the translation service provider community have not (yet) been commissioned. In short, much remains to be done if this proposal is to gain momentum. A risk exists that the current model, although unsatisfactory, is so deeply engrained that the pain involved in changing would be a disincentive. Time will tell.

In the meantime, TAUS is planning to develop an online dashboard that implements some of the concepts outlined here. This will be pilot tested by some of its members and, if successful, will be rolled out to a broader community. Interested parties should visit the TAUS website for updates on this initiative.

References

- **Blatz, John, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing** N (2004). "Confidence estimation for machine translation." *Proceedings of Coling 2004*: *20th International Conference on Computational Linguistics* (University of Geneva, 23-27 August). 315-321.

- **Brunette, Louise** (2000). "Towards a Terminology for Translation Quality Assessment", *The Translator* 6(2), 169-82.

- **Byrne, Jody** (2006). *Technical Translation. Usability Strategies for Translating Technical Documentation*. Dordrecht: Springer.

- **Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, & Josh Schroeder** (2007). "(Meta-) evaluation of machine translation", *Proceedings of the Second Workshop on Statistical Machine Translation* (Prague, June 23rd), 136-158.

- **Colina, Sonia** (2008). "Translation Quality Evaluation: Empirical Evidence for a Functionalist Approach", *The Translator* 14(1), 97-134.

- **Cronin, Michael** (2010). *The Translation Crowd*. http://ddd.uab.cat/pub/tradumatica/15787559n8a4.pdf (consulted 31.05.2011).

- **Doherty, Stephen, O'Brien, Sharon and Carl, Michael** (2010). "Eye Tracking as an MT evaluation technique." *Machine Translation* 24(1), 1-13.

- **Gouadec, Daniel** (2010). "Quality in Translation." in Yves Gambier and Luc Van Doorslaer (eds) *Handbook of Translation Studies, Vol. 1*. Amsterdam: John Benjamins, 270-275.

- **Hvelplund, Kristian T.** (2011). *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-Logging Study*. PhD Thesis: Copenhagen Business School.

- **House, Juliane** (1997). *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr.

- **The Information Standard**. http://www.theinformationstandard.org (consulted: 01.06.2011).

- **J2450**, *Quality Metric for Language Translation of Service Information.* http://www.sae.org/standardsdev/j2450p1.htm (consulted 31.05.2011).

- **Kelly, Nataly and Don DePalma** (2009). *Buyers Step Up Their Quality Measurement Efforts*, Common Sense Advisory Report, http://www.commonsenseadvisory.com (consulted: 24.08.2011).

- **Kohl, John** (2008) *The Global English Style Guide*. SAS Press.

- **Lauscher, Susanne** (2000). "Translation Quality Assessment: Where Can Theory and Practice Meet?" *The Translator* 6(2), 149-168.

- **LDC** (2005). *Linguistic data annotation specification: Assessment of fluency and adequacy in translations*. Revision 1.5. http://www.ldc.upenn.edu/Catalog/docs/LDC2003T17/TransAssess02.pdf] (consulted 31.05.2011).

- **MeLLANGE**, Multilingual eLearning in Language Engineering. http://corpus.leeds.ac.uk/mellange/about_mellange.html (consulted 01.06.2011).

- **Nord, Christiane** (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*, Manchester: St. Jerome.

- **O'Brien, Sharon** (2010). "Controlled Language and Readability." in G. Shreve and E. Angelone (eds) *Translation and Cognition - ATA Scholarly Monograph XV Series*, 143-165.

- **O'Hare, Neil, Michael Davy, Adam Bermingham, , Paul Ferguson, Páraic Sheridan, Cathal Gurrin and Alan F. Smeaton** (2009) "Topic-dependent sentiment analysis of financial blogs." in TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, 6 November 2009, Hong Kong, China. ISBN 978-1-60558-805-6

- **Papineni, Kishore, Salim, Roukos, Todd, Ward and Wei-Jing Zhu** (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of ACL-2002*: 40th Annual meeting of the Association for Computational Linguistics (Philadelphia, July), 311-318.

- **Przybocki, Mark, Kay Peterson and Sébastien Bronsart** (2008). "Translation adequacy and preference evaluation tool (TAP-ET)." *Proceedings of LREC 2008*: 6th Language Resources and Evaluation Conference (Marrakech, Morocco, 26-30 May), 8pp.

- **Ray, Rebecca and Nataly Kelly** (2011). *Crowdsourced Translation*, Common Sense Advisory Report. http://www.commonsenseadvisory.com (consulted 31.05.2011).

- **Rickard, Jason** (2009) *Translation in the Community*. http://www.localisation.ie/resources/conferences/2009/presentations/LRC_L10N_in_the_Cloud.pdf (consulted 31.05.2011).

- **Specia Lucia, Nicola Cancedda, Marc Dymetman, Marco Turchi and Nello Cristianini** (2009). "Estimating the sentence-level quality of machine translation systems*.*" *Proceedings the Thirteenth Annual Conference of the European Association for Machine Translation* (Barcelona, May 14-15), 28-35.

Websites

- **EN15038,** *European Quality Standard for Translation Service Providers*, http://qualitystandard.bs.en-15038.com/ (consulted 31.05.2011).

- **FEMTI**, *Framework for the Evaluation of Machine Translation*. http://www.isi.edu/natural-language/mteval/ (consulted 31.05.2011).

- **LDC** (2005). *Linguistic data annotation specification: Assessment of fluency and adequacy in translations*. Revision 1.5. http://www.ldc.upenn.edu/Catalog/docs/LDC2003T17/TransAssess02.pdf] (consulted 31.05.2011).

- **MeLLANGE**, Multilingual eLearning in Language Engineering. http://corpus.leeds.ac.uk/mellange/about_mellange.html (consulted 01.06.2011).

- **TAUS,** the Translation Automation User Society. www.translationautomation.com (consulted 22.08.2011).

Biography

Sharon O'Brien is a lecturer in Translation Studies in the School of Applied Language and Intercultural Studies at Dublin City University, Ireland, where she teaches a range of topics including software localisation, translation theory, translation practice, research methods and writing skills. She is affiliated with the Centre for Translation and Textual Studies and the Centre for Next Generation Localisation and her research centres around translation technology, localisation, and cognitive translation processes. She formerly worked in the localisation sector as a language technology specialist.

sharon.obrien@dcu.ie

[1] I take this opportunity to acknowledge the significant contributions of Rahzeb Choudhury and Jaap van der Meer (TAUS) as well as the company representatives.

[2] All company names remain anonymous throughout.

[3] J2450 specifically rules out 'style' as an error type.

[4] J2450 specifically mentions how this category is difficult to interpret and not very useful.

[5] Combining categories in this way caters for the unavoidable subjectivity involved in deciding what, for example, critically important means vs. very important.