revista.tradumàtica
tecnologies de la traducció

430-436

Training Quality Evaluators

Ignacio García
University of Western Sydney

ABSTRACT

In the present age of fluid and prolific content, notions about the quality (or fitness) of texts are changing. We may now contemplate a text continuum, with the enduring and critical at one extreme, and the ephemeral and inconsequential at the other. Having focussed on the former, traditional translation and translator education approaches struggle to keep pace with the 'fast and fit' imperative of the latter. Accordingly, there is a drive towards pinpointing both translation quality and purpose in a more transparent and consistent way. New approaches to formal assessment and the corresponding application tools are in development. Graduates who encounter them early in the classroom will be better prepared to allocate their time, self-assess their output, and revise that of others. Thus equipped, professional translators could assert their standing by designing custom solutions, writing scopes of work, and signing off at completion using the relevant QA procedures.

Keywords: translation quality evaluation; quality assurance; dynamic quality framework; multidimensional quality metrics; translation specifications.

RESUM *(La formació d'avaluadors de la qualitat)*

En aquests moments en què la creació de continguts és a l'abast de tots, el concepte tradicional de qualitat no sempre serveix. No tots els textos tenen el mateix interès: uns són duradors i de vital importància; altres, efímers i trivials. Amb la vista sempre posada en els primers, tant els traductors com els que ensenyen o avaluen traduccions troben ara difícil decidir què és qualitat (o què és l'apropiat en cada cas) amb la rapidesa que exigeixen els segons. És per això que estan apareixent nous criteris per avaluar i noves eines per ajudar a aplicar aquests criteris amb objectivitat i transparència. L'estudiant de traducció que s'hagi familiaritzat en les seves classes amb aquests criteris i eines estarà més ben preparat per distribuir el seu temps, autoavaluar el seu treball i revisar el treball d'altres. Aquesta formació millorarà l'estatus professional del traductor ja que el capacitarà per dissenyar solucions a mida, a cercar termes de referència o a procedir a l'aprovació final d'un treball després d'haver aplicat els controls de qualitat pertinents.

Paraules clau: avaluació de la qualitat de la traducció; control de qualitat; *dynamic quality framework*; *multidimensional quality metrics*; especificacions de traducció.

RESUMEN *(La formación de evaluadores de la calidad)*

En estos momentos en los que la creación de contenidos está al alcance de todos, el concepto tradicional de calidad no siempre sirve. No todos los textos tienen el mismo interés: unos son duraderos y de vital importancia; otros, efímeros y triviales. Con la vista siempre puesta en los primeros, tanto los traductores como los que enseñan o evalúan traducción encuentran ahora difícil decidir qué es calidad (o qué es lo apropiado) con la rapidez que exigen los segundos. Es por ello que están apareciendo nuevos criterios para evaluar y nuevas herramientas para ayudar a aplicar esos criterios con objetividad y transparencia. El estudiante de traducción que se haya familiarizado en sus clases con

tales criterios y herramientas estará mejor preparado para distribuir su tiempo, auto evaluar su trabajo y revisar el trabajo de otros. Tal formación mejorará el status profesional del traductor al capacitarle para diseñar soluciones a medida, especificar términos de referencia o proceder a la aprobación final de un trabajo tras haber aplicado los controles de calidad pertinentes.

Palabras clave: evaluación de la calidad de la traducción; control de calidad; *dynamic quality framework*; *multidimensional quality metrics*; especificaciones de traducción

Quality under Scrutiny

Apart from proven past performance, translation training is, with certification, the most reliable indicator for buyers that a given translator will be competent for a given task. Paradoxically, quality has been something of a silent presence in training programmes, perhaps because it has been considered an indispensable and omnipresent ideal. However, in the face of new industry imperatives, it is now receiving increased attention. As global content volume escalates, resource allocation acquires greater importance, and there is a corresponding impetus for finding ways of defining and even dimensioning quality. This article will explore some ways in which those issues are being dealt with and will consider how to bring them to the classroom.

Besides the business-to-customer texts typical of the localisation era, we now have peer-to-peer communication typical of our social media era. Depending on perspective, the content in social media is not necessarily trivial: its writers might want fast cheap translation to engage internationally with their blogs or tweets; institutional translation consumers may explore the same material for deeper social and economic insights. Since expert human translation can be a time consuming and expensive endeavour, cost/benefit assessment is acquiring increasing relevance. If translation quality (and effort) is not pitched accurately at the ultimate purpose, then the exercise can be uneconomical, and at worst futile and valueless.

Therefore, in order to prepare work-ready graduates, translator educators need to understand the new environment in which translation takes place and how translations will be assessed in the real world. Furthermore, it would help graduates if the way that academia teaches quality were to mirror, as far as possible, the way in which end-users view or measure it. Graduates need to understand what 'quality' effectively means in a particular context, and use that understanding to efficiently check their own work or revise that of others; moreover, it even can even underpin a quality assessment (QA) framework to help employers ensure that translations meet their expectations for a given moment and situation.

In short, the conventional assessment models in translation training may need an overhaul similar to the one now taking place in industry. Accordingly, we will outline below the challenges that the industry is facing and that academia needs to respond to.

Measuring Quality in the Social Media Era

The conventional translation model, the so called translate-edit-proofread (TEP) model, is slow and expensive. Such detailed treatment may in particular cases be well justified, but as a rule the web prefers translation that's fast and affordable. Since conventional translation could not provide it, alternative ways have emerged: machine translation (MT), raw and post-edited; crowdsourcing.

The quality evaluation strategy developed in the nineties, the SAE J2450 and the LISA QA being the reference (Drugan 2013: 95), was based on choosing a sample, analysing it for errors and setting a pass threshold. It suited the TEP localisation era, when the web was static, content was mostly business-to-customer, and the only tools in the translation assembly line were memories and glossaries.

Now, user's feedback evaluation has proven that raw MT can be good enough for knowledge base content aimed at technicians (Gerber 2008), or community-based evaluation for crowdsourced translation in social media platforms (Little 2008). For some content, it makes sense to publish first, then edit, if required – editing in the web is inexpensive. Only that content on which the risk of getting it wrong the first time will be too high to consider will no need the full TEP treatment.

Two new approaches to measuring quality evaluation are now in development trying to respond to today's needs: the TAUS Dynamic Quality Framework (DQF) and the QTLaunchPad Multidimensional Quality Metrics (MQM). Both are being covered elsewhere in this issue (Görög 2014 and Lommel et al. 2014). Suffice to point out here what both share: that evaluation metrics must be adaptable, with "dynamic" or "multidimensional" basically meaning that on size will not fill all cases; that they should be affordable, able to be shared across industries, suitable for benchmarking; and, most importantly, should be objective, in as much as it can. Hopefully, both will converge, and those principles will be taken on board by translation buyers as well as language service providers. Training should prepare graduates to deal with this emerging scenario.

Teaching Translation Quality

Quality in itself is a concept notably difficult to define, but quality as applied to translation offers special challenges. Ten expert translators will produce ten different renditions of the same source, and ten expert markers are likely to produce ten different results even using the same metric. Errors of form, from spelling to syntax, can be detected by contrasting against a specific set of rules that every language speaker learns, and are easy to agree with; being rules, machines can often learn them too. Discourse elements (rhetoric, style, register, ambiguity, implication, allusion, metaphor) or textual properties (coherence, cohesion, however, do not lend themselves to being codified into precise rules. It is a reasonable proposition that higher translation quality correlates with fewer errors.

Yet, measuring quality should be based on more solid grounds than a subjective judgement of what constitutes an error. Translation educators have delved into the discipline of translation studies in search for help. The equivalence paradigm has been influential since Vinay and Darbelnet published their seminal work in 1958 (1995). The initial focus was on ensuring that the content of the source is transferred to the target in full and that the target respects the conventions of the new language, even in the understanding that the achievement of complete accuracy will make full fluency elusive, and vice versa. From sentence-level equivalence, later scholars (House 1997 in particular) went on to add textual and pragmatic equivalence, enriching quality evaluation with concepts of domain, register and text type. Functionalist theories, as popularised by Nord (1997), shifted attention from the textual (translation happens between texts) to the social: translation happens between people, and for a purpose. The concept of 'evaluation' has itself evolved across disciplines in general, with the principles of psychometrics, sampling, validity, reliability, informing the quest for objectivity while keeping the task manageable (Mitchell 1999).

Yet, these principles have rarely trickled down to the typical translation classroom setting. Here the normal learning process consists in commenting on the difficulties a particular source presents and the strategies for overcoming them, with students grasping how quality is assessed by the way the tutor marks their work (Muzii 2013). It is on this same basis that they will learn to evaluate their own work, and assess the translations produced by others. By and large, educators keep pursuing the absolute ideal of capital-Q quality, measuring it on the basis of professional experience and judgement. This can often be supported by error deduction strategies (sometimes adapted from guidelines developed for certification) and, more recently, with the often even less precise rubrics – neither being overly transparent or consistent, particularly in borderline cases.

The DQF and MQM initiatives mentioned above constitute serious attempt on the part of public organisations (including the European Union), research bodies, translation buyers and language service providers, to overhaul current, inefficient systems and develop better ones. Translators' organisations and certification bodies should also participate in the discussion, and so too should translation educators. The next sections outline some steps to begin moving in that direction.

DQF and MQM to the Classroom

The digital revolution has already assailed the walls of academia and forced translation programs to react. Over the past decade, the main response consisted in bringing computer-assisted translation (CAT) tools into the classroom. CAT of course went hand-in-hand with localisation and a new paradigm of non-sequentialness and segmentation. This is quite at odds with the top-down text tradition, and most generalist programs incorporated CAT through a separate unit, so that the upheaval it brought ('chunkiness', reuse, challenges to ideas of authorship) was largely quarantined from the time-honoured academic approach.

More recently, academia's new besiegers are MT and crowdsourcing, particularly in regard to quality: if the translator is merely asked to produce light post-editing, then thorough attention to detail becomes superfluous and even counterproductive. But the demand exists, and the choice is to ignore it or meet it; if the latter, then changes in both attitude and practice will be required.

The first lesson to apply is that translation is an interplay between requester, provider and end-user. This supposes a different kind of quality assessment than transfer excellence alone. To achieve an appropriate target text, the translator needs to consider not just the source text, but also the instructions received from the requester and the needs of the final user. An appropriate starting point could be a common definition of quality attributed to Alan Melby and often repeated in MQM circles:

A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications taking into account end-user needs.

There is nothing wrong with training that challenges the student to offer uncompromising excellence: complete pragmatic transference in masterful, idiomatic language. That is the right kind of preparation for graduates who want to access the high-end market where no expense or revision cycles will be spared. By the same token, a client who wants something quick and useful might be better served by light post-editing. Regardless, if we concede that quality is relative to the purpose of the requester and the needs of the end-user, then these are difficult criteria to satisfy consistently if guided by solely subjective appraisals. The task becomes much easier of one can learn and apply valid, reliable and practical metrics, based on standardised concepts and values shared by requesters and translators alike.

Translators have learned to use traditional memories and glossaries as a matter of course. Equally, modern graduates should familiarise with next generation tools that help in assessing the quality of an MT engine, or its suitability for post-editing; or in transparently categorising and reporting translation errors, whether made by machines or humans. Such tools form the subject of the following section.

Tools for Evaluation Training

During the past decade, most translation programs at universities have been training students in the use of CAT tools, including the QA features now available in all of them. Harking back to our earlier section on errors, most of these features deal with rule-based

forms, whether linguistic (use of prescribed terms) or engineering (all tags transferred to target).

More recently, as post-editing gains ground and new quality evaluation approaches (e.g. the DQF and MQM models) achieve prominence, new tools are now in development. QA tools first appeared as stand-alones (QA Distiller, ApSIC Xbench, ErrorSpy), and were eventually incorporated as features in commercial CAT suites. The same may happen with the emerging tools described below. While some may not be industry-ready, they could nonetheless be used for classroom exercises - say within the context of a typical technology-focused unit that already has a quality evaluation component. All should be suitable also for more research-like environments, when student attention can be applied over an extended period of time.

We will firstly consider DQF tools, followed by others aimed at better capturing error typology, post-editing productivity, and automated metrics for MT evaluation. The list is by no means exhaustive. Most, if not all, would be available to educators who seek appropriate permission from the developer.

DQF Tools

Developed by TAUS within its DQF agenda and launched in February 2014, the package bundles tools with a series of written articles. These writings succinctly explore, apart from error deduction, other ways of assessing quality: adequacy and fluency, adherence to regulatory instruments, community-based evaluation, customer feedback, readability, and usability. A 'content profiling' feature helps users determine which model or models better suit a given task depending on text type, intended usage and resources (for an underlying rationale, see O'Brien 2012).

The tools, Ranking Engines, Quality Evaluation (including Adequacy, Fluency and Error Typology), and Productivity Testing, while aimed at MT, can also be used for human translation. The ranking tool enables evaluators to assess de quality of up to three different sources. The adequacy and fluency tools rate source-content transfer and target grammar on a scale of 1-4. They are straightforward to use. The productivity tool can be set to a Post-Edit + Translate mode (one segment post-edited and the next translated and so on) or a Post-Edit only mode. It too could become an effective teaching aid once the editing window is enlarged to accommodate the full segment - to date, only a single line displays. The error typology needs more development.

POST-EDITING Productivity Tools

A once rare creature, MT post-editing is becoming quite common for some content types and between some language pairs. Tools are being developed to gauge productivity (time spent on post-editing vs. time spent on conventional translation) and edit distance (changes introduced by the post-editor in the machine generated baseline). Ultimately such data may serve to apply the fuzzy match 'discount' principle of TM to MT matches (Claverie 2014). Some of these tools – PET (Aziz et al. 2012), CASMACAT (Ortiz-Martinez et al. 2012) – are geared towards research, while others – MateCat, iOmegaT – seem oriented towards production. Just as conventional CAT tools are already incorporating MT plugins, it seems inevitable they will offer post-editing metrics: memoQ is already moving in that direction.

MateCat is being developed by Translated - the LSP behind the MyMemory massive online database – with some EU funding. Its thrust appears to be toward machine learning algorithms (Federico et al. 2014). Of more immediate relevance here is the capturing of time and edit distance data, shown as stats in the Edit log. The user can import and apply a memory and a glossary too. Although MateCAt is not an industry-ready, full-strength CAT tool as yet, it already seems potentially instructive as a teaching aid.

iOmegaT, built onto the open-source CAT tool OmegaT, is being developed by Welocalize and the Center for Next Generation Localisation in Ireland. It lets users optionally (de)activate MT for different segments, and offers another user-friendly aspect too. If translators in CAT environments commonly review and edit their segments, MT post-editors would more logically be inclined to do the same. Yet productivity measuring tools tend to lock segments after translation. Interestingly, iOmegaT provides for re-editing – and naturally, (re)captures the relevant data (Moran 2012).

Error Typology Tools

As explained elsewhere in this issue, the MQM, provides a way to define metrics that can span all possible 'dimensions' of a translation (ten are stated). It invokes a 'full' account of issue types, which may also involve the source if appropriate. Based on this shared vocabulary, specifications will then be tailored to a given task, building on the 'core' to make the conditions as complex as needed - the metric is extensible. Then, score cards can be used, with severity levels and a set threshold.

Even better than score cards would be an actual tool that helped apply such specifications in a transparent and consistent way. That tool exists in the form of translate5, conceived as a front-end for MQM. It works by highlighting a sub-segment (or a full segment), and picking one category (or more) from the MQM pre-defined typology. Once configured, the tool enables category-based error filtering and export of all results in report form so as to identify problem issues and take remedial action (Sikes 2013).

It is arguably the most elegant solution yet to the problem of feasibly avoiding errors, while providing consistent feedback and avoiding subjectivity. Being open-source, it is by definition 'work in progress'. Thus far, changes and comments made in translate5 can be imported into SDL Trados Studio only. But it is not difficult to visualise this way of capturing feedback becoming another inbuilt CAT tool feature: XTM Cloud seems to be engaged on it already, as does memoQ.

Translators at The Helm

A worthy goal for any translation training program is for its trainees to learn the corresponding theory and practice with a view to competently producing their own work and assessing and revising that of others. Their acquired competence should enable them to make a living, should they pursue translation professionally, by meeting peer standards of performance and satisfying client expectations. If that is the case, universities must now provide work-ready graduates who can not only translate in the old-fashioned sense, but also read (or write) appropriate job specifications and assess quality outcomes (theirs or others') based on a series of generally accepted QA procedures.

Hopefully, initiatives such as DQF and MQM will converge, standardized procedures will be shared by translation buyers and LSPs, and tools will be developed to assist in applying them in a transparent and consistent way. So far the demands of streaming content and social media have practically penalised painstaking translation in favour of 'anything (and anyone) goes'. Once quality can be 'objectively' pinpointed, translation buyers will be able to distinguish between work that requires rigorous specifications that only trained translators can meet, and what can be done by the semi-skilled ones or bilingual volunteers.

With such standardised procedures in place, universities can respond with training programs that yield graduate translators who understand quality as much in terms of cost/benefit for their future clients as an absolute academic ideal. In either case they will be equipped to identify the appropriate service level, and if it has been provided. After all, tools may help with the task, but QA will always require a competent translator in control.

References

Asia Online (2011, November). New Release of Language Studio Pro –V3.0. Language Studio Newsletter.

Aziz, W., Castilho, S., and Specia, L. (2012, May). PET: a Tool for Post-editing and Assessing Machine Translation. In LREC: 3982-3987.

Claverie, A. (2014). Post-editing MT: Is it worth a discount? Multilingual 25 (3): 40-41.

Drugan, J. (2013). Quality in Professional Translation: Assessment and Improvement. London and New York: Bloomsbury.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M. and Germann, U. (2014). The MateCat tool. COLING 2014, 129.

Gerber, L. (2008, November). Recipes for Success with Machine Translation: Ingredients for Productive and Stable MT deployments. ClientSide News: 15-17.

Giménez, J., & Marquez, L. A. (2010). An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. The Prague Bulletin of Mathematical Linguistics, 94: 77-86.

Görög, A. (2014) Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. Revista Tradumàtica 12: 443-454. Translation and quality.

House, J. (1997). Translation quality assessment: a model revisited. Tubingen: Narr.

Little, C. (2008). Facebook in Translation. The Facebook Blog.

Lommel, A., Uszkoreit, H., Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Revista Tradumàtica 12: 455-463. Translation and quality.

Michell, J. (1999). Measurement in Psychology. Cambridge: Cambridge University Press.

Moran, J. (2012). Experiences instrumenting an open-source CAT tool to record translator interactions. Expertise in Translation and Post-editing-Research and Application.

Muzii, L. (2013), Translation education: A three-legged table. Multilingual 24(8): 22-24.

Nord, C. (1997). Translating as a Purposeful Activity. Manchester: St Jerome.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. The Journal of Specialised Translation, 17: 55-77.

Ortiz-Martınez, D., Sanchis-Trilles, G., Casacuberta, F., Alabau, V., Vidal, E., Benedı, J. M., and González, J. (2012). The CASMACAT project: The next generation translator's workbench. In Proceedings of the 7th Jornadas en Tecnologıa del Habla and the 3rd Iberian SLTech Workshop (IberSPEECH): 326-334.

Sikes, R. (2013). translate5: A new approach to translation review. Multilingual 24(6): 18-22.

Vinay, J. P, and J. Darbelnet. (1958/1995). Comparative Stylistics of French and English. A Methodology for Translation, trans. Juan C. Sager and M.-J. Hamel. Amsterdam and Philadelphia: John Benjamins.