Tailor-made Quality-controlled Translation

Sergio Penkale *Andy Way*

Lingo24
Greenfield
Greater Manchester, UK

{sergio.penkale, andy.way}@lingo24.com

Abstract

Traditional 'one-size-fits-all' models are failing to meet businesses' requirements. To support the growing demand for cost-effective translation, fine-grained control of quality is required, enabling fit-for-purpose content to be delivered at predictable quality and cost levels. This paper argues for customisable levels of quality, detailing the variables which can be altered to achieve a certain level of quality, and showing how this model can be implemented within Lingo24's Coach translation platform.

1. Introduction

For businesses to improve their position in an increasingly competitive international marketplace, the need for fast, reliable, cost-effective translation is greater than ever. However, traditional 'one-size-fits-all' models are failing to meet businesses' requirements.

In recent years, the explosion in online content has created new challenges for global companies. Only a tiny fraction of content that could be translated currently is. It's clear that high-quality professional human translation is not a feasible option in every case. At the same time, Machine Translation (MT) technology is improving rapidly, and now broadens the alternatives, with or without human reviewing and editing.

These recent developments mean traditional translation workflows are becoming supplemented with new ones. Instead of providing a limited choice of service levels, translation providers need to move to a more flexible approach, offering clients a guaranteed quality of output, which need not be 'perfect' human quality. Instead of a single method for evaluating quality, the concept of a "fit-for-purpose" translation will become more widespread.

The required quality depends on factors such as the lifespan and volume of content, its purpose and target audience, and its urgency. For example, while style and fluency are vital for a press release, they're less important in a technical manual (although accuracy is essential). In some situations, such as real-time conversations, speed is the main priority.

Given these imminent paradigm shifts in the translation process, it is no surprise that Language Service Providers (LSPs) and Language Tools providers are focusing on researching and developing new technologies which support these new use-cases. In this paper we describe Lingo24's Coach translation platform. Coach has been designed with customisable quality levels in mind. As described in Section 4, the tool allows enforcing or disabling of automatic QA checks on a project-by-project basis, which both aids linguists in their task and ensures that the quality level required for each type of content is achieved. The tool also allows for a range of different workflows, such as traditional translation and editing, but also post-editting of customised MT output, as well as raw MT, where appropriate.

The remainder of this paper is organised as follows. Section 2 argues for customisable levels of quality. In Section 3 we detail the many variables that can affect translation quality, explaining how different use-cases can be accommodated by fine-tuning each of them. Section 4 presents our translation platform Coach, which instantiates most of these use-cases by allowing different quality

levels to be ensured on a per-project level. We finish in Section 5 with our conclusions.

2. The Case for Customisable Levels of Quality

In Way, (2013), our companion paper, we describe a number of traditional and emerging use-cases where both raw MT and post-edited MT output lead to perfectly acceptable quality for the task at hand. Assuming that ever more use-cases will emerge in the near future, it cannot be the case that where quality is concerned, a 'one-size-fits-all' definition suffices; there must be at least two levels of quality given the existence of both light and full post-editing services.

However, there are some who disagree with this basic tenet:

> *One recent trend is the offering of various 'quality levels', something professional translators cannot and will not do. For us there is only one quality level: professional, publication-ready quality.* (Rose Newell)[1]

As argued by Way (2013) and Bota et al. (2013:313), (some) translators are quite disparaging towards the work carried out by others in their profession; Bellos (2011:328) – a very well-regarded translator in his own right – refers specifically to criticisms such as "bad translators, 'servile', 'mechanical', second-rate translators". We're not sure who Newell is referring to by 'us' in the above quote, but it is clear that not all translators would agree with this statement; Bellos observes (op cit., p.335) that "not all [translators] are great at their job", so the whole notion that there is "one quality level" is inherently flawed. Indeed, rather than trying to be critical of "just about every bulk translation agency",[2] Newell is instead being dismissive of the PEMT work that many thousands of her fellow translators perform – clearly these professional translators are more than capable of offering different levels of quality – especially when one considers the PEMT 'light' service, where the output is less than "professional, publication-ready quality", but is nonetheless fit for purpose.[3]

Indeed, one of the main themes emerging from LocWorld 2013 in London was "that the old quality models may no longer be the answer when applied to post-edited output used for new content delivery models",[4] thus rendering Newell's comments even further out-of-date.

Sharon O'Brien – another trained translator, note – argues that we are (or at least should be) moving toward a dynamic quality evaluation model for translation.[5] She notes that with respect to translation quality evaluation, TAUS 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, MT or TM system or a combination of these)." Much like we argue here, O'Brien states that new ways of measuring quality are required not only because "numerous challenges exist for quality evaluation, including subjectivity, time, inappropriate use of linguistic resources, learning curve and technology", but also due to recent developments such as budgetary constraints, new paradigms ("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 technology, and a new focus from companies on the end-user.

Jost Zetzsche agrees with O'Brien (see URL in footnote 2), but as we've alluded to above, he observes that "translation quality will remain a contentious topic of discussion, maybe more so than as a matter of implementation." He gives specific examples of MT-ed and human-translated Help files on Microsoft's knowledge-base, and notes that "a translator who compares the translation quality of the two articles will immediately have a visceral response: one has 'good quality' and the other seems to scream out its 'poor, machine-translated quality.' *But the users? They find them both (virtually) equally helpful*" [our emphasis]. He concludes by saying that "the

1 lingocode.com/translation-isnt-getting-cheaper/
2 Zetzsche observes in this regard that there is an increasing dichotomy between translators who work with LSPs and those who do not: thebigwave.it/technology-here-and-now/interview-with-jost-zetzsche/
3 For reasons of space, we leave aside here a discussion of whether it is appropriate for the same translators to perform both light and full post-editing, given the differing levels of quality required for both services.
4 Olga Beregovaya: www.welocalize.com/mt-here-to-stay/
5 www.jostrans.org/issue17/art_obrien.php

perception of quality needs to be a lot more dynamic. There is certainly room for quality metrics and standards, but we need to accept that these don't apply to everything. And some of the translation buyers have long figured that out."

3. Quality Levels in Human and Machine Translation

Where translation is concerned – whether we're talking about human translation or MT – we need to make sure the output is good enough to fulfil the requirements of its intended use-case.

In Way (2013), we have already commented on the fact that translators can be pretty critical of each other's work. Even among equally well-trained professional translators there will be disagreements over the quality of a translated text. We also pointed out in that paper that whether it's the case or not, in evaluating MT quality, we usually compare its output against a 'perfect' human reference. This is done using automatic MT evaluation metrics such as BLEU (Papineni et al., 2003) to measure the number of matches of words and phrases in the MT output against a (single, usually) 'gold standard' human reference; the more closely the MT output resembles the human reference, the better it is deemed to be. While BLEU scores and the like have been used for other purposes than originally intended by Papineni et al. (2002), this is principally of use to MT developers in judging whether incremental improvements to the MT system can be seen by the incorporation of other techniques and data sources.

Another way in which MT quality is judged is by measuring the fluency and adequacy of the MT output on Likert scales compared to the source sentence. Note that in the automatic evaluation method just described, the source sentence plays no role at all, which despite clearly being a major flaw in how MT developers operate, allows them to quickly use the same objective methodology no matter what the target language.

Some of the early work carried out measuring fluency and adequacy of MT systems was performed by Van Slype (1979). Here fluency (or 'intelligibility' in Van Slype's terminology) and adequacy ('fidelity') were measured using Likert scales using different numbers of points depending on the exact method chosen.[6]

Other ways in which MT quality is measured involve task-based evaluation (e.g. White & Taylor, 1998), or ranking (e.g. He & Way, 2009), where native speakers are presented with multiple MT outputs (usually from different systems) and asked to rank them according to which one best covers the meaning of the provided source string. Note also that performance as good as – if not better than – so-called 'experts' can be achieved here, by availing of 'non-experts' using Mechanical Turk (Callison-Burch, 2009), a finding we can confirm from our own experiments.

Given the above, and the myriad use-cases cited by Way (2013), it is clear that there are potentially as many ways of measuring MT quality as there are ways in which MT can be used. In each case, the MT output needs to be evaluated using the most appropriate metric, bearing in mind at all times the importance of fitness of purpose given the task at hand.

When it comes to an end-to-end translation job, where MT is just one step in the pipeline, there are many variables that can affect not only the end quality of a finished translation, but also its final cost, and the time that it takes to complete. As previously mentioned, we believe the notion of quality must take the purpose of the translation into account, and that there are many scenarios where a compromise in (say) fluency is preferred to compromising on delivery time. Accordingly, the instructions given to the linguist can vary from use-case to use-case. If a linguist is instructed to ensure the usage of specific terminology, this requirement will impact in both the traditional sense of quality as well as in cost. The same goes for fluency, style, TM adherence, etc. Having access to a tool that supports all these different use-cases by allowing these variables to be altered is essential to ensure that translations are delivered at the expected quality level, time and cost. We present one such a tool in the next section.

6 See http://www.isi.edu/natural-language/mteval/refs.html for a useful bibliography of work done on MT evaluation in the ISLE project: http://www.issco.unige.ch/en/research/projects/isle/.

4. Customisable Quality Levels in Coach

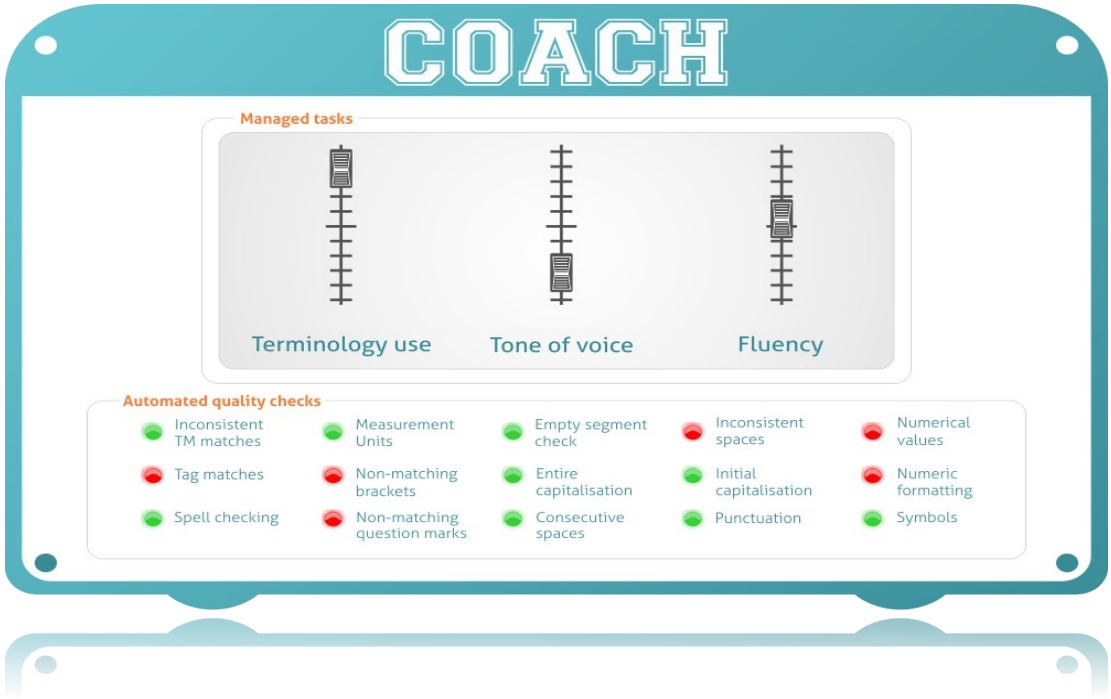

Figure 1: Customisable levels of Quality in Coach

As we demonstrated in Bota et al. (2013), tools such as Lingo24's Coach exist already which help clients to stipulate precisely which QA checks should be performed for their particular use-case. Coach redefines requirements by offering clients the freedom to determine their own quality levels. Translation is viewed as a series of granular tasks, where quality can be customised to balance volume, budget and turnaround times. MT can be incorporated into the workflow, but instead of preserving the old dichotomy of 'light' vs. 'full' post-editing levels, these can be tailored to fit the client's needs for each type of content.

For each project, a wide range of automatic and manual checks can be switched on and off, while the importance of factors such as tone of voice and fluency can be defined on a sliding scale (cf. Figure 1). The workflow can be adapted to include additional review stages, as well as terminology and style checks. In some cases, every segment will need to be checked by specialist linguists; in other cases a 5-10% sample will be enough to ensure reliability.

The screenshot in Figure 2 illustrates how individual automatic automatic QA checks can be enabled or disabled. When a QA check is enabled it will aid the translator or reviewer in their work, by drawing their attention to potential problems. The most usual sources of quality problems can thus be detected early, reducing the translator's workload (and thus making them more productive) and easing the burden on the reviewer, if such a stage is present in the pipeline. However, a distinction is made between checks which are mandatory, and those which are optional. While the former checks are there to aid the linguists, they can choose to ignore these checks, and they will not be penalized for doing so. In contrast, when a QA check is set to mandatory, the linguists will be prevented from signing off their work until all potential problems have been resolved, or until they have justified why they have not done so. While such strict enforcement will clearly impact somewhat on turnaround times, they help in achieving a predictable quality level. Enabling and disabling these checks allows the balance between price, turnaround time and required quality to be fine-tuned.

Coach's ability to automatically highlight potential sources of quality problems shifts the role of QA in quite a unique way. As is the case with traditional tools, these can be used to generate reports after a project has finished, or to give instructions to a reviewer after the translation task is done. However, unlike any other tool we are aware of, these can also be used in real time while the translator is working, before the translation part of the pipeline has been delivered. Coach allows users to filter the content they see according to a range of criteria (cf. Figure 3). This allows linguists to focus their attention on the tasks that are more relevant to them at any specific point in time.

The available checks that can be enabled in Coach range from spell checking, formatting checks such as double spacing and capitalisation, and balanced brackets, to checking proper TM and terminology adherence. As regards the latter, Coach can identify occurrences of client-supplied terms in a source segment, and ensure that the translator adheres to the translation stipulated by the client's glossary. A range of language-dependent QA checks are also available. These ensure that numbers, dates, currencies, punctuation, etc. have been correctly expressed according to the target-language rules.

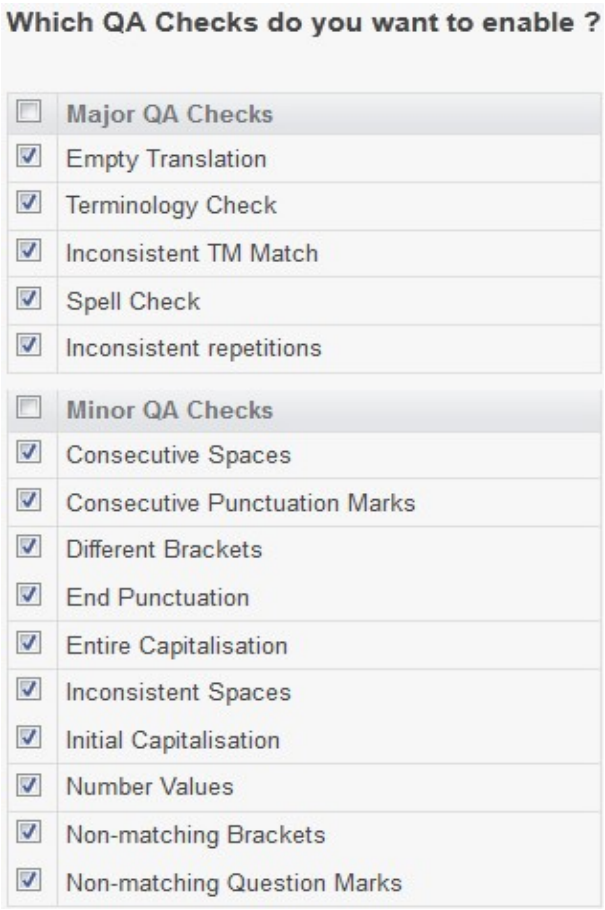

Figure 2: Enabling automatic QA checks in Coach

However, Coach's ability to tailor translation projects to the client's need goes well beyond these automatic quality checks. The Coach platform uses automated algorithms to break down workflows into specific tasks. It is flexible enough to bring in translators of varying levels of linguistic proficiency. In some cases, parts of a text can even be checked by a monolingual native speaker for style and fluency. Coach has been designed with translators in mind, and can be used simultaneously as a learning management tool. This will help address problems with the supply chain in the industry, and ease pressure on supplier rates. Highly experienced professional translators can focus on tasks where their expertise is needed most, earning more per hour than they currently do. It also enables new translators to enter the industry, starting with basic tasks and gradually progressing to more sophisticated tasks.

As previously mentioned, Coach allows the incorporation of state-of-the art MT into translation workflows. While generic services such as Google Translate[7] or Bing Translator[8] are supported, Coach also benefits from tight integration with Lingo24's own MT technology. This allows users to gain access to vertical engines specialised in specific subject matters, such as Finance, IT, Tourism, etc, which demonstrably outperform generic engines when translating each kind of subject matter, in terms of automatic MT evaluation metrics such as BLEU, but also in terms of measured translation efficiency gains. We should also note that where sufficient translation assets such as TMs and glossaries are available from the client, Lingo24 is also able to build MT engines which are tuned to the client's specific terminology and style, achieving even further quality and efficiency gains. In some cases, this customisation is possible even when only monolingual assets are available. In the systems we build in Lingo24, if required, we can guarantee adherence to glossary items. This is one area where MT can improve over human translation quality, as checking that terminology has been translated consistently is easy to ensure automatically, whereas a human can easily translate a term in different ways especially in longer documents. We can provide this

7 translate.google.com/
8 www.bing.com/translator

guarantee whether building customised engines for specific clients, or when using the aforementioned industry-specific vertical engines.

At its core, Lingo24's MT uses the Moses Statistical MT toolkit (Koehn et al., 2007), which is enhanced by a range of custom-built pre- and post-processing steps which perform Natural Language Processing tasks such as tokenization, lowercasing and recasing, compound splitting, segmentation, etc. Special attention is dedicated to the handling of in-line tags. Since most of the content translated nowadays makes uses of tags to specify formatting information such as boldface or italics fonts, it is of critical importance that MT can correctly deal with tags. Accordingly, a modified version of Moses is used which can ensure the well-formedness of the tag structure produced by the MT engine, thus avoiding technical difficulties when regenerating target files, and retaining the formatting present in the source.

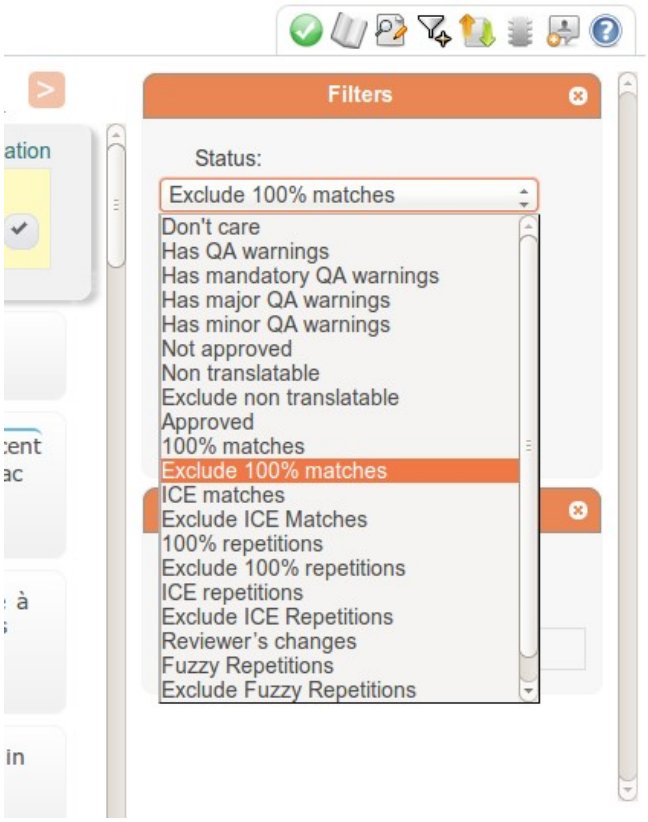

Figure 3: Filtering content in Coach

Finally, the software can be combined with an Application Programming Interface (API). This allows the client to automatically send new content for translation, with quality levels adjusted depending on pre-defined criteria. For example a customer might choose to specify different quality requirements for press releases, product information and user-generated reviews. An API request will trigger the translation pipeline and the sourcing of appropriate suppliers, streamlining the translation process and reducing both turnaround times and cost by automating some of the steps in the translation workflow. This is in some scenarios facilitated by our custom-built Subject Detection technology, which is able to assign a label to an input document indicating the subject matter that it covers. This automatic classification ensures that the document is routed through the most appropriate MT engine, and aids in sourcing the necessary linguistic suppliers.

5. Conclusions

The traditional notion of 'one-size-fits-all' when it comes to quality is shifting towards the notion of fitness for purpose. If the translation industry is to take advantage of all the current emerging use-cases, tools that enforce a specific level of quality are required.

We believe that Coach will revolutionise the translation process for new and experienced users alike, by facilitating novel workflows to cater for an ever-increasing number of translation use-cases. It enables translation providers to offer more solutions to clients at different price points, expanding the scope of what the industry can achieve and making it applicable to more people. We expect Coach to become the digital marketplace for quality controlled translation in the next few years.

References

David Bellos. 2011. *Is that a fish in your ear: translation and the meaning of everything*. Particular Books, Penguin Group, London.

Laura Bota, Christoph Schneider and Andy Way. 2013. COACH: Designing a new CAT Tool with Translator Interaction. In *Machine Translation Summit XIV, Main Conference Proceedings,* Nice, France. pp.313—320.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.286—295.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL 2007: proceedings of demo and poster sessions, Prague, Czech Republic,* pp.177—180.

Yifan He and Andy Way. 2009. Metric and Reference Factors in Minimum Error Rate Training. *Machine Translation* **24**(1):27—38.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL-2002*: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, pp.311—318.

Georges Van Slype. 1979. Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk.

Andy Way. 2013. Traditional and Emerging Use-Cases for Machine Translation. In *Proceedings of Translating and the Computer 35*, London, UK.

John White and Kathryn Taylor. 1998. A Task-Oriented Evaluation Metric for Machine Translation. In *First International Conference on Language Resources & Evaluation*, Granada, Spain.