Evaluation of outsourced translations. State of play in the European Commission's Directorate-General for Translation (DGT)

Ingemar Strandvikd

Thursday 30th November, 2017

1 Introduction

The European Commission is the executive body of the European Union. It implements the European policies, it proposes new legislation and monitors that EU law is applied correctly by the Member States. All these activities are carried out and communicated in 24 official languages with an equal status. This means that multilingualism is at the heart of the EU. The resulting massive translation demand was formerly met almost exclusively through in-house translation. However, after the number of official languages has increased over the years, with successive enlargements, in particular the "big bang" enlargement in 2004, when 9 new official languages were added to the then 11, and as translation volumes have continued to grow, more and more translations are now outsourced, both in pursuit of cost-efficiency and due to insufficient internal capacity. All in all, the number of in-house translators in the pre-enlargement languages has been reduced by almost 50 per cent over the last twenty years, but with the arrival of the new languages, the total number of translators in the Directorate-General for Translation (DGT) of the European Commission has remained roughly the same. Today, with 24 official languages, there are some 1,500 in-house translators and DGT translates over 2 million pages per year for the European Commission, roughly a third of which are supplied by external contractors via outsourcing.

This article aims at describing how DGT has organised its outsourcing operations. In particular, it focuses on evaluation principles and practices and some of the challenges involved.

2 Outsourcing and evaluation

To outsource these considerable volumes, DGT relies on multiple framework contracts with a dynamic ranking system. The system features a tendering procedure where the quality/price ratio has been put at 70/30. It also features systematic evaluation, where a 10 per cent sample of each translation is revised, assessed and marked using a five-grade scale.[1] The mark affects the contractor's position in the dynamic ranking, which in turn influences how assignments are distributed.

As the proportion of outsourcing has increased considerably over the years, streamlining outsourcing has become a real issue, to achieve both cost-efficient work organisation and equal treatment of hundreds of external contractors. Moreover, since DGT today outsources all types of documents, including draft legislation and high-profile policy documents, it has become crucial to ensure that

[1] "Very good" (10), "Good" (8), "Below standard" (6), "Insufficient" (4) and "Unacceptable" (0).

outsourcing does not have a negative impact on quality. To this end, the tender specifications of the most recent call for tenders for outsourced translations (OMNIBUS-15, in place since 1 July 2016) included the following quality requirements:

> The quality of the translations must be such that they can be used as they stand upon delivery, without any further formatting, revision, review and/or correction by the contracting authority. To this end, the contractor must thoroughly revise and review the entire target text, ensuring inter alia that:
>
> - it is complete (without unjustified omissions or additions);
> - it is an accurate and consistent rendering of the source text;
> - references to documents already published have been checked and quoted correctly;
> - the terminology and lexis are consistent with any relevant reference material and internally;
> - appropriate attention has been paid to the clarity and register and text-type conventions;
> - it contains no syntactical, spelling, punctuation, typographical, grammatical or other errors;
> - the formatting of the original has been maintained (including codes and tags if applicable);
> - any specific instructions given by the authorising department are followed; and
> - the agreed deadline (date and time) is scrupulously respected.

Evaluation plays a key role to ascertain whether these quality requirements have been complied with. To evaluate linguistic and textual quality, all outsourced translations are assessed on the basis of the evaluation grid in table 1 below.

Table 1: Evaluation grid currently used by DGT

Error type	Code	Low	High	Error type	Code	Low	High
Mistranslation, unjustified addition	SENS			Reference docs/material not used; norm sources or job-specific instructions not adhered to	RD		
Unjustified omission or non-translation	OM			Wrong or inconsistent EU usage or terminology	TERM		
Clarity, register and text-type conventions	CL			Punctuation	PT		
Grammar	GR			Spelling	SP		

Identified errors are further classified according to their severity as 'low-relevance' or 'high-relevance' errors. A high-relevance error is defined as an error that seriously impairs the usability of the text for its intended purpose. Moreover, evaluators assess whether the product delivery (including translation memories, etc.) is complete, whether DGT's instructions have been followed and whether the formatting requirement and set deadlines have been complied with. If this is not the case, separate penalties apply.

The evaluation is carried out by the in-house translators, who are expected to possess the competence needed to evaluate outsourced translations. Evaluations, just like translations and revisions, are assigned on the basis of the competence profiles of the available staff.

3 Issues in the past

Under the former framework contract GEN-11 – which applied from 1 July 2012 to 1 July 2016 – the system worked rather well and the performance improved over the contract period, partly because of the feedback given to the contractors to clarify DGT's needs and expectations as regards quality. Having said that, some issues related to evaluation were considered to be problematic. The main issues identified were consistency of evaluations and the high cost of administration and contract management.

Consistency of evaluation practices and results is inevitably a challenge when 1,500 in-house translators are expected to be able to carry out translation quality assessments in a uniform and supposedly repeatable manner. Translation is constant decision-making. It is about constantly making choices. As **Pym1992** puts it, there are binary translation errors (choices that are correct or incorrect) and non-binary ones (choices that are not necessarily right or wrong but more or less appropriate). The problem for translators, revisers and evaluators alike is that most quality issues are of the non-binary type. To assess quality consistently you need to be clear about why some choices are better than others. When operating in an institutional translation setting of DGT's scale, this obviously becomes an issue.

Moreover, it is a fact that freelance markets are different. The freelance markets in the German language area with a population of 100 million people, the Estonian with around a million, or the Maltese with some 450,000 are not the same in terms of capacity, specialisation and maturity. This inevitably has an effect on consistency in the approaches to how to interact with the markets.

As to the management and administration costs, in principle, according to the

outsourcing framework contract, the translations received were supposed to be usable as they stood upon delivery, without any further intervention from DGT, other than the evaluation of the 10 per cent sample applied to all outsourced texts. Despite this, two thirds of the outsourced pages were further quality controlled in-house i.e. beyond the 10 per cent. This appeared to be a failure cost, considering that almost 95 per cent of the translations still received the pass marks "very good" or "good". It was asked why DGT should spend in-house resources to revising texts that had already been revised by the contractor and that were marked "very good" or "good", which should mean they are usable as such.

The inquiries into why this happened showed several things. First, that the time allocated for the task of evaluating a page amounted to 10 per cent of the time allocated for the task of translation, while the conversion rate for the task revision was 40 per cent of a translated page. Since evaluation consists of *revising* a 10 per cent sample, this meant that carrying out a thorough evaluation, mechanically led to spending more time than what was accounted for, thereby lowering internal productivity and further increasing the difference in costs for internally and externally produced pages. This led to instances where evaluation was based on a less thorough revision of the sample, giving the external translator the benefit of the doubt, applying an overly lenient marking.

Second, it appeared that often the additional revision was done because of the type of document concerned and the risks involved. When higher-risk documents such as strategic communications, articles for publication, or draft legislation were outsourced, language departments did not dare to rely solely on a spot check. For the sake of comparison, it could be mentioned that the translation department of the European Court of Justice contends that when they outsource the translation of court rulings, they revise the entire outsourced text, even if it has been revised by the contractor according to the contract, because it produces legal effects and the translation departments need to ensure that the legal effects are correct. If only parts of the document are revised, this cannot be guaranteed.

Third, it was also found that in many language departments it was the *usefulness* of the translation that was assessed rather than its *usability*. As mentioned above, the overall quality requirement according to the tender specifications is that the text delivered should be usable as it stands. However, even in cases where the entire text needed revision, modifications and corrections, outsourced translations were regularly considered to be clearly *useful* for the finalisation of the text – and therefore "very good" or at least "good". Finally, instances were also identified where evaluators awarded good marks as a reflection of their empathy with freelance translators and their (assumably) less favourable working

conditions.

4 Developing quality guidelines and the notion of quality

Traditionally in the EU context, when someone passes a recruitment test – a competition – for a post as translator, it is taken for granted that he or she has the competences needed to translate, revise, evaluate translation quality and carry out terminological work. At the same time, for many languages translators were recruited without formal studies in translation, since in many cases such studies did not exist at the time of accession of their country (**Biel2011**, **Strandvik2014**). If we add to that the sheer number of the people involved, it is clear that a major challenge has always been – and is likely to always be – to ensure that the institution speaks with one voice, not only when translating and revising, but also when evaluating outsourced translations. What has been done to address this issue?

In a major quality management project called *Programme for Quality Management in Translation – 22 Quality Actions* (**DGT2009**), **DGT2009** set up a number of working groups to analyse 22 quality-related topics and processes relevant for the quality of the translation services provided. Several of these actions were related to outsourcing and evaluation, for instance actions aiming to improve translation briefs and feedback for freelancers and develop standards for the evaluation of freelance translations (including specific training, error quantification and tools for evaluation).

As a result of these initiatives, apart from a series of language-specific revision workshops and quality control guidelines, common guidelines for evaluation of outsourced translation were issued in 2009 (cf. **DGT2013**). Moreover, a quality assessment tool based on the LISA QA model (cf. **Doherty2013**) and attributing penalty points for errors was introduced, on the basis of the widely spread belief that translation quality had to be measured with analytical, and not holistic quality assessment. However, at the time, only five language departments found the error quantifying tool useful. Not surprisingly, one of the main objections was that in order for quality assessment to be consistent (so called inter-rater reliability), there needs to be a common understanding of the principles for evaluation and of the error categorisations and severity levels used. Otherwise, the objectivity of the assessment tool is reduced to an objective calculation of error points resulting from a subjective identification of errors.

Indeed, over the years, in different internal contexts, there has been a growing awareness about the fact that a pre-requisite for any institutional attempts

to speak with one consistent voice in translation, revision and evaluation is that there is a shared understanding of what is actually meant by quality. Around 2012, it appeared that while everybody agreed to **DGT2014**'s mission statement that **DGT2014** should provide the Commission with high-quality translation, there was no common definition of what DGT meant by high-quality translation. Time was ripe to come up with such a definition and develop a more structured approach to quality management. This resulted in *DGT's Quality management framework* (**DGT2014**), a steering document for quality management in which quality is defined as fitness-for-purpose and key processes are described. The definition adopted reads:

> A translation is fit for purpose when it is suitable for its intended communicative use and satisfies the expressed or implied needs and expectations of our direct customers (requesting DGs), our partners in the other EU institutions, the end-users and any other relevant stakeholders.
>
> Consequently, fitness for purpose means high quality in the abovementioned sense. It should not be mixed up with the good-enough quality concept used by the software industry and in the machine translation context. The fitness for purpose concept is at the core of DGT's internal quality control (QC) guidelines (Consolidated guidelines on quality control) and of the Service Level agreements (SLAs) DGT has signed with other DGs.

With this definition, DGT boldly aims at reclaiming the fitness-for-purpose concept to mean suitability for the intended purpose, in line with the logic of all professional standards, and not "good enough quality" as it has been defined for example by **TAUS2017dqf** (TAUS EUG Resolution #2).

To operationalise the fitness-for-purpose principle, common translation quality guidelines (**DGT2015a**) were then issued. They describe the different purposes of different types of EU documents, explain potential risks caused by deficient quality and provide text type specific instructions for translation and quality control, based on risk assessment. These developments and DGT's reference model for quality management (Figure 1) are described in detail in **Strandvik2017** and **DruganEtAlforthcoming**).

Figure 1: DGT's Reference model for translation quality management.

5 From fidelity to fitness-for-purpose

During the last 15 years, a pragmatic, functionalist approach to specialised translation has made its way into the standards of the profession. Successively, the German **DIN1998**, the European **EN2006**, the American **ASTM2014**, and **ISO2012**, **ISO2015** all clearly state that extra-linguistic aspects such as specifications (or briefs) are key for quality, revision is defined as assessing a translation as to its suitability for the intended purpose, which is to say that the purpose and the specifications are the yardstick against which you determine the appropriateness of the translation choices. Indeed, in service provision, quality is defined as compliance with requirements. Translation service provision is no exception: translation quality is compliance with requirements, it is not just faithfulness to the original. The reason why the functionalists made their way into the standards is that their theories work in practice and make sense, not only for translators but for all the stakeholders involved.

This move from fidelity to fitness-for-purpose has taken place not only in DGT but also in the other EU institutions (see for instance the contribution from the Council in this volume). This is logical if translation is approached as professional drafting and as communication acts. Any text can be improved. Most texts drafted for professional purposes contain imperfections and even errors, without being unfit for their purpose. The same applies to professional translation.

As spelled out in the DGT Translation quality guidelines, the European Commission has issued a number of drafting guidelines to explain to drafters how it wants to communicate and what it wants to achieve when communicating through different text types. This communicative intent is not limited to the source text and should be fulfilled also through the 23 translated official language versions. Therefore, translators should be familiar with these guidelines to apply them to the extent possible when translating. This is all the more important as today there is hardly ever any in-depth editing after translation, not even for legal acts (**Guggeis2012**, **Strandvik2014**). The translated texts should stand on their own. According to constant case law, once an EU legal act is adopted, there are no originals and no translations, only equally authentic language versions. And as **Husa2012** puts it (**Husa2012**), what matters in legal translation is not what the texts say linguistically, but what they say legally.

DGT has witnessed this evolution also in its evaluation practices. Formerly, the severity level "high relevance" was defined with a reference to a change in meaning (a high relevance spelling error was a spelling error that changed the meaning), whereas now, as explained above, a high relevance error is defined as

an error which "seriously impairs the usability of the text". Exactly the same error can be of high or low relevance not because it affects the meaning but because it affects the usability of the text differently. A spelling mistake in a 15-page text is likely to be a non-issue, whereas if it appears on a poster in big letters it could be fatal. A wrong date appearing on page 55 in a report could be insignificant, whereas the date of entry into force of a legal act or the date and time of a meeting are crucial, etc. A mistranslation in the enacting terms of a legal act is likely to affect the usability of the text, whereas exactly the same mistranslation in the explanatory memorandum is less likely to have that effect. Formerly, the quality requirements in tender specifications stated that the contractor should provide a faithful rendering of the source text and eliminate any discrepancies between the source and the target text. Now, as quoted above, they state that the text should be an "accurate and consistent rendering of the source text". Discrepancies is an unclear concept. Discrepancies as to denotation, connotation, text-type convention, pragmatics or form? Discrepancies are, in fact, sometimes required to comply with the formal style guides for legislative drafting for different languages, or to make a web text read smoothly, or to make a text fit to a button on screen.

Moving towards an understanding of quality that could be shared and embraced by 1,500 translators from 28 different national contexts with very different professional and educational backgrounds, working in 24 different language departments, is a challenge. If we scratch the surface, there are still different conceptual understandings of what translation is. One which embraces the functionalist approach to translation (fidelity to the purpose of the communication) seeing the translators as active and competent drafters of the equally authentic translated language versions of texts with a function, and another which embraces the idea of faithfulness (fidelity to the source text's surface structure) as the main criterion for translation quality, seeing the translators as "just translators", where their task is limited to the faithful rendering of the "original" in the target text.

These perceptions seem to be deeply anchored in beliefs and values. It would be interesting to explore this further: Is it a divide between experienced and unexperienced translators? Or between translators with and without formal studies in translation? Is it linked to age? Is it the accuracy requirements of legal translation that contaminate all other aspects of translation? Are there different national translation cultures? Does it have to do with administrative culture and institutional power relations affecting the translators' agency? Some translators naturally interact with requesters and national experts for clarifications, whereas

some rather do not. The latter, do they "hide" behind the source text?

MelbyEtAl2014 and **KobyEtAl2014** address this issue in an interesting way, with reference to discussions at FIT's World Congress 2014 on the relation between localization/transcreation and translation, suggesting a distinction between different beliefs on what translation is (**MelbyEtAl2014**) and what translation quality is (**KobyEtAl2014**).

6 Recent developments and further challenges

To cope with the increase in outsourcing, and to ensure a streamlined and consistent workflow, a new Outsourcing framework was adopted in 2016. Furthermore, in view of the new framework contract OMNIBUS-15, which entered into force on 1 July 2016, new evaluation guidelines were drafted to address the abovementioned issues.

The *DGT Outsourcing framework* (**DGT2016b**) puts emphasis on supplier management. It aims to improve the quality of outsourced translations through improved communication: via meetings with the suppliers, better specifications linked to the Translation quality guidelines and systematic and more harmonized feedback. Even if the quality requirements in the new framework contract (quoted above) have remained the same as before, the new evaluation guidelines, *DGT Guidelines for evaluation of outsourced translation* (**DGT2016a**) introduced some novelties: the link between evaluation and the quality requirements of the tender specifications was made clearer. Definitions were added to the marks. It was also clarified that the evaluation is above all a contractual obligation for payment clearance, not as such a reliable quality control measure for risk mitigation. In other words, its result which is based on a 10 per cent spot check does not guarantee the intrinsic quality of the entire text. To address the issue of additional quality control applied after outsourcing, it was decided that the Translation quality guidelines apply to all translation, whether produced externally or internally. The result of the evaluation therefore feeds into the global risk assessment. A poor evaluation result is likely to trigger extended quality control, according to the escalation principle, whereas a very good result could lead to stopping the effort after the evaluation of 10 per cent. At the same time, depending on the risks involved, it can be decided that regardless of the result of the evaluation, the entire document, for instance speeches and binding legislation, should undergo full revision. Moreover, to ensure a consistent approach to marking, systematic *validation* of all marks was introduced, with a limited number of validators checking all evaluation results for consistency (but not re-doing

the evaluations).

Current evaluation challenges further include sharing practices across languages on where to draw the line between the two severity levels (high and low relevance errors) and where to put the thresholds between the different marks, and how to harmonise feedback comments in a way that is consistent with the tender specifications and the definitions of the marks. Another challenge is to finetune the sampling practices. In the EU institutions and in industry practices range from industrial sampling based on the **ISO2006** standard to full in-house revision. We are still lacking empirical evidence as to the reliability of quality assessment based on different sample sizes. Is 10 per cent reliable? Is 20 per cent more reliable?

With the new quality management structure, and a more common understanding of quality, time has also been deemed ripe for a new attempt to consider introducing a tool to further streamline the quality assessment of outsourced translations. **DGT2016a** is currently testing different existing tools and assessment models and follows closely the ongoing standardisation initiatives of ISO and ASTM as well as the EU funded QT21 project, and the resulting Multidimensional quality metrics (MQM).[2] The outcome of those initiatives are likely to lead to an updating of the error categorisation and of the weightings currently used.

7 Conclusions

The experience gained in **DGT2016a** over the years shows that we cannot translate, revise or evaluate translation quality in a consistent way, if we do not have a shared understanding of translation quality. With so many actors, it is important to state the quality requirements explicitly, to avoid misunderstandings and miscommunication. The reference model for translation quality management recently put in place in DGT is a useful step on the long and winding road towards this long-term objective.

In this endeavour, **DGT2016a** is increasingly relying on international standardisation efforts. The very purpose of standardisation is to identify and define key concepts, to ensure seamless communication, and to establish and prescribe workflow steps, so that all stakeholders know what to expect from each other when interacting in relation to the standardised activity. Standards represent the distilled wisdom of the profession. Even if DGT doesn't need translation standards for certification purposes, referring to them for benchmarking purposes has become a means to improve working methods and communication.

[2] http://www.qt21.eu/quality-metrics/

One emerging key question in that context is what kind of competence profile is needed to be able to evaluate translations. Is it the same as for revision, e.g. in terms of subject matter competence? A very important related question is: How much can we outsource? Is there a tipping point after which the European Commission will no longer be in control of its communication and legislative drafting because it no longer has the in-house domain competence to assess and ensure quality?

The administration of contracts is expensive. Ideas to elicit savings often end up being costly, creating hidden costs that put strain on the in-house staff. Attempts to apply industrial (and much cheaper) sampling methods have so far been problematic and not given satisfactory results. As with any service provision, what really matters is to specify the quality requirements. What text quality does the European Commission need and who is responsible for ensuring this quality? The question is perhaps not whether the European Commission can afford to quality control outsourced translations but rather whether it can afford not to do it.