Big wrong data: Insights from developing and using an automatic translation revision tool with error memories

Bert Wylin
KU Leuven and Televic Education, Belgium

Gys-Walt van Egdom
Universiteit Utrecht, The Netherlands

Abstract

Since August 2017, more than 40 institutions (and over 800 translators) piloted the translationQ revision platform. This paper will not focus on the development of but shows some lessons learned from translationQ's error/revision memories: how to recycle the errors ('Big Wrong Data') into useful insights for translation, translation evaluation and translation teaching.

The translationQ project (a joint initiative of KU Leuven and Televic Education) was developed to automate and speed up evaluation processes in both translator education and the translation profession. The tool works for bilingual or monolingual text productions. The program's core is an error or revision 'memory': it allows the system to recognize errors in new translations and to suggest corrections and feedback automatically; the program leaves room for human intervention. Still, the bulk workload of retyping the same corrections and feedback time and again is now done automatically by the tool, leading to more rapid and more consistent revision feedback and scoring. Revision memories can be shared and reused with new texts and with new trainees.

The reporting module of the platform allows the profiling of translators (strengths and weaknesses) in an objective way.

Introduction

The domain of translation technology can be described as a hive of activity. In a survey, Translation Automation User Society (TAUS) canvassed no less than eighty 'technological profiles' for Language for Specific Purposes (LSP) (2016). We believe that, with the latest advances in machine learning, this number is likely to continue to increase. A survey amongst translator trainers would yield different results. Although modern translation environment tools (TEnTs) can serve to aid some of the ergonomic and general qualitative problems in translator training, one will be hard-pressed to name but one tool that is geared to the specific needs of translator trainers. By ergonomics, we understand the combination of psychological and technical principles in the product design, with the goal being to reduce human error, to increase productivity and to enhance the wellbeing of the people involved. At first blush, it is quite remarkable that little to no attention has been paid to technological means to remove the drudgeries from tasks like trainer-to-trainee revision and student translation evaluation, and, thereby, speed up these processes, and boost quality.

With its official release in April 2018, the cloud-based tool translationQ has come to occupy an exceptional position in translator training. The tool, designed by Televic, was developed in close collaboration with KU Leuven, and, as a result of this co-creation, is primarily attuned to translator trainers' and trainees' needs. When designing the product, researchers and developers concluded that what is needed in trainer-to-trainee revision is 1) authenticity, 2) objectivity, and 3) efficiency.

Authenticity

First of all, the tool offers the potential to enhance authentic experiential learning, as it bears resemblance to the virtual working environment of TEnTs, not only in the translation mode but also in the revision and the post-revision mode. A good case in point is the lay-out of a translation assignment: the source text and the target text are segmented and appear in two columns (Figure 1).

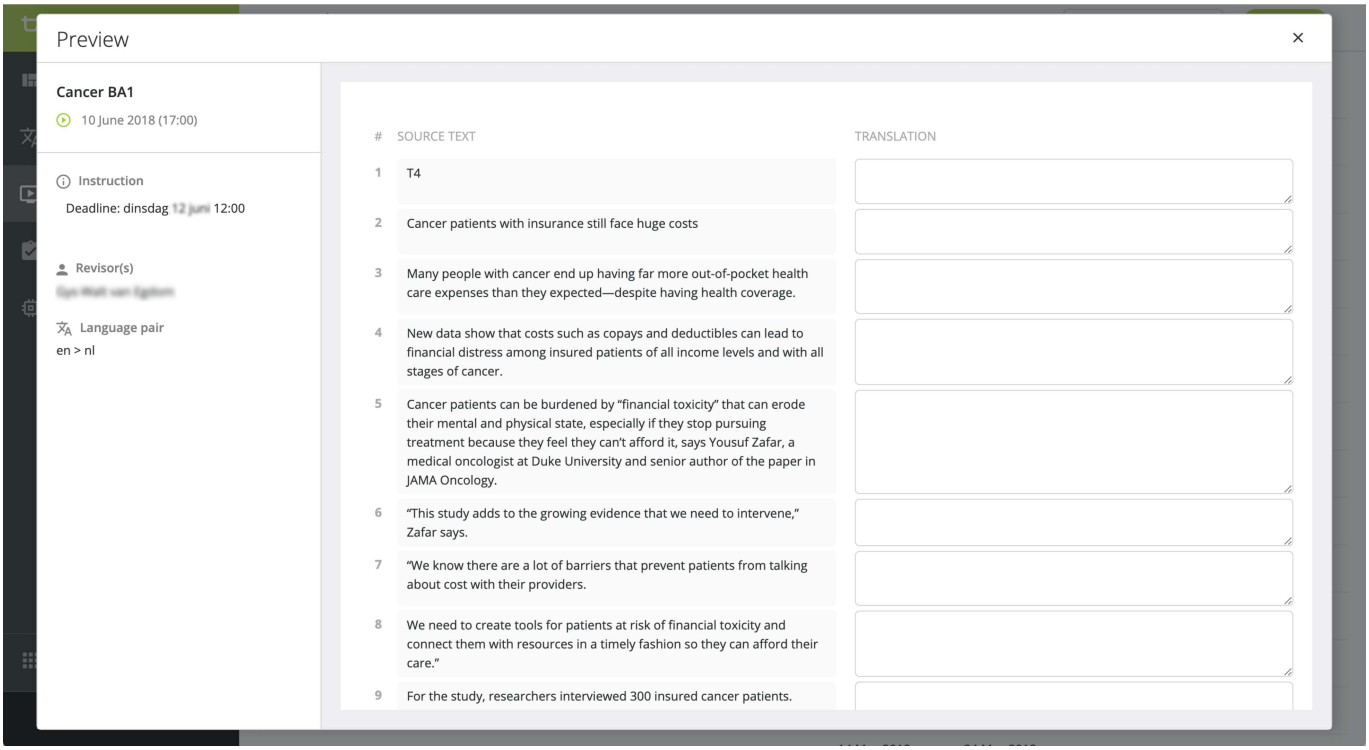

Figure 1 Screenshot of traditional two column source text target text layout for trainees to enter their translation

Trainees and trainers benefit from this layout: there is no need for them to divide attention between a source file (or even a paper sheet with the source text) and a target file. Furthermore, revision processes are authenticated through the use of error categories that have become common currency in the language industry, more specifically, the Dynamic Quality Framework or DQF categories (TAUS, 2015), as well as through error weighting, which means that a distinction can be made between minor and major errors (Figure 2).

Objectivity

By employing the analytical categories of DQF, translationQ developers have not only sought to authenticate trainer-to-trainee revision processes, but also to provide a fillip to objectivity in revision. The tool allows for trainers to rename the DQF categories and use their own preferential error labels. This renaming can be beneficial to individual trainers but pose a threat for cross-trainer portability of the categories or error memories. For that reason, renamed categories are still (silently) linked to the original DQF categories, thus permanently allowing the sharing and comparing of error memories of all trainers and institutions. However, we must not be too hasty to assume that the common use of DQF categories allows for complete objectivity in revision. From the criticism garnered against analytical testing, one can infer that, despite clear definitions, ' it is often a matter of subjective judgement [. . .] whether an error falls into one category or another' (Saldanha & O'Brien, 2014, pp. 101–102). Furthermore, subjectivity looms large in error weighting: it is up to the reviser to decide whether an error is minor, major or even critical. By sharing and exchanging revision memories, revisers can make sure they use the exact same penalties

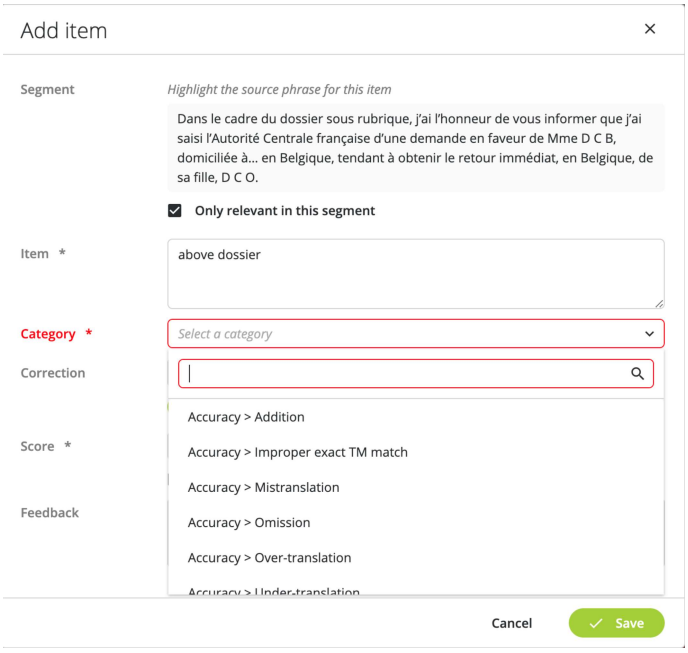

Figure 2 Screenshot of how to add and categorise a translation error in translationQ

and weights for identical errors, which will improve the inter-rater consistency. Of course, the lack of inter-rater reliability cannot be *fully* remediated in the tool; as has been pointed out by a great many specialists, there will probably never be a panacea for subjectivity in translation revision and evaluation (e.g. Bowker, 2001; Holmes, 1988; Thelen, 2016). Still, inter-rater evaluation is made more consistent, as the error categories are institution-dependent and serve as a basis for an intersubjective error memory. What is more, the tool does seem to provide a solution to problems with intra-rater reliability, problems that are not very often admitted in revision literature but that are perhaps the most salient problems in trainer-to-trainee revision processes. In translator training, it is incredibly difficult to keep tabs on all revisions of a single assignment. Common human reviser errors that cause low intra-rater reliability are:

- identical errors are sanctioned in one student version and overlooked in the other
- identical errors are sanctioned by the reviser, but the error made by one student is classified differently from that of another student
- identical errors are sanctioned by the reviser, but the error weighting differs from one version to another.

In translator training, these inconsistencies are frequently brought to light in translation courses, since students not only have to complete similar tasks, they are often in close contact and tend to compare their versions. The consequences of low intra-rater reliability are far-reaching, as blatant inconsistencies undermine the trainer's expert position and might even discourage students (Hönig, 1998, p. 15). Drawing upon basic insights in corpus-based approaches to translation and computational linguistics, the developers of translationQ have provided a way for the translator trainer to revise in a consistent manner: the algorithm of the tool goes in search of, detects and flags identical errors (in the same segment of other students' versions) and similar errors (in different segments); and the trainer can decide to accept or reject computed revision suggestions (Figure 3).

Efficiency

This brings us to the final issue that the translationQ project members have sought to address: efficiency. In keeping with current technological trends in the language industry, the tool has been developed with a view to making lighter work of professional practices, in this case: translation revision. This means that a great deal of time has been dedicated to enhancing user experience and, ultimately, meeting the ergonomic threshold set by potential users. The tool has been made accessible for the less IT-literate as well as challenging for trainers with a knack for technology. By this, we mean that it allows for project creation and correction in a few simple steps; however, it also provides functionalities for trainers who wish to get the most out of a revision tool (authentication, elaborate revision, revision data ('rich wrong data')).

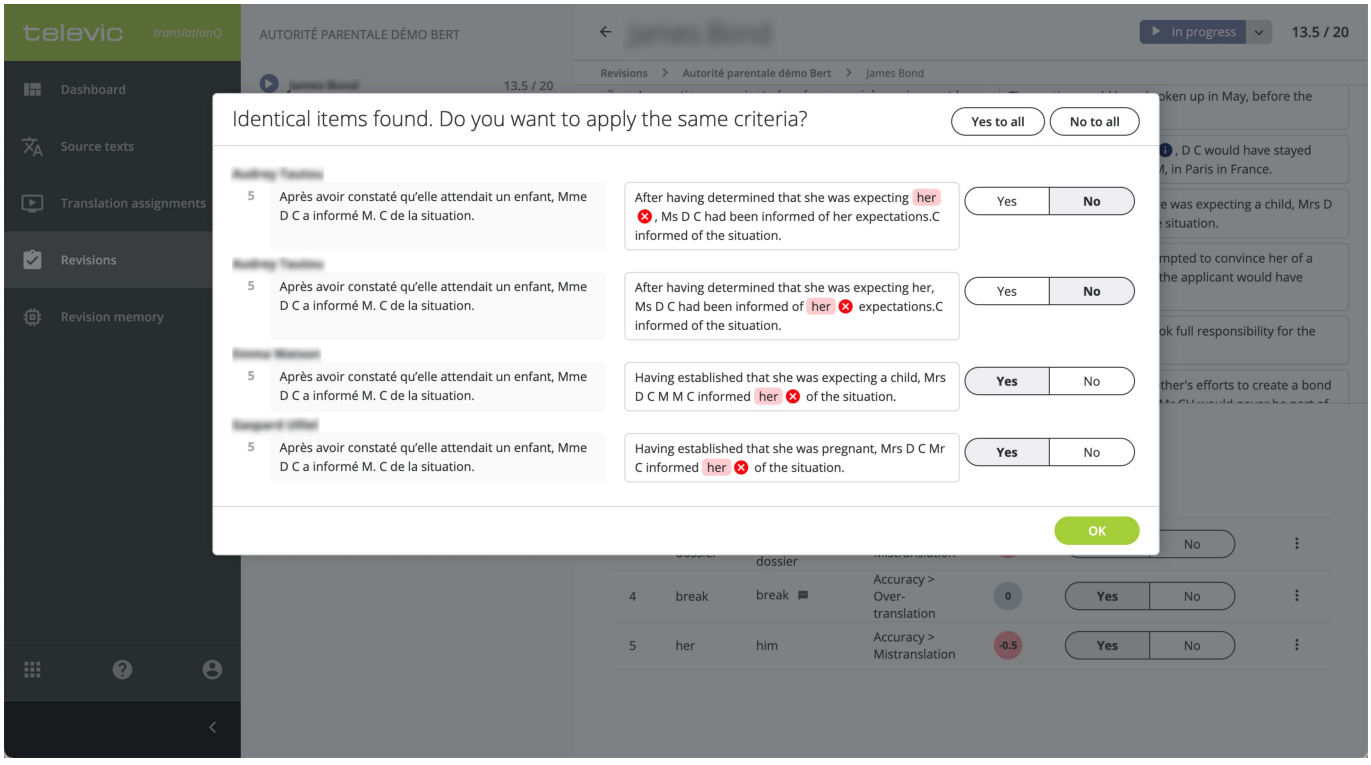

Figure 3 Screenshot of translationQ error algorithm

In order to be able to meet the ergonomic threshold, special heed was to be taken of factors that make trainer-to-trainee revision a tedious and tiresome task. Trainer-to-trainee revision can be aptly described as donkey work, primarily because of its repetitive nature. In trainer-to-trainee revision, it is imperative that the reviser work his way through every student version and examines the text in detail for its suitability to the purpose as laid out in the assignment. In practice, this means that the reviser gets caught up in a vicious cycle, in which identical mistakes are detected, labelled and, ideally, corrected time and again. As mentioned earlier on, translationQ removes this drudgery from the revision process: the algorithm flags identical (and similar) errors in the same student version and in others' versions and copies the comments and suggestions for improvement onto the flagged fragments that are deemed relevant by the reviser.

The ergonomic quality is further increased through the revision memory function. The functioning of the algorithm is not limited to one sole project; translationQ automatically stores identified errors (along with the error types, comments and suggestions) and incrementally builds up a revision memory. By gradually building up a revision memory, translationQ reduces the reviser's cognitive load and speeds up revision practices considerably.

This is not all there is to say about the added ergonomic value of translationQ. The software also provides automated solutions for the following tasks:

- calculating scores for formative assessment
- publishing feedback
- archiving student translations
- adapting didactic strategies to student needs.

The feedback provided by trainers in a training context is intended to give students an idea of (the types of) errors made. However, students seem to feel somewhat in the dark when interpreting the number of errors; they often wonder whether the quality of their translation would be sufficient to obtain a passing grade in a summative context or, perhaps even, how far they are removed from a professional level of translation competence. As a solution to this problem, trainers can treat the assignment as if there is something at stake; they count and weigh errors and calculate a score for each translation. Unfortunately, the calculation of scores usually takes quite a bit of time. In translationQ, points are automatically subtracted from the total score of the assignment. Trainers can publish the automatically calculated scores along with the feedback for each student with a simple click of a button. Published revisions are then stored in the 'User Reports' tab (which can be consulted via the Dashboard, another feature that is built in to foster authentic experiential learning). Through the user reports, the trainer can glean an idea of student performance on specific tasks, and get an overview of student and group progress and error frequency over time (Figure 4).

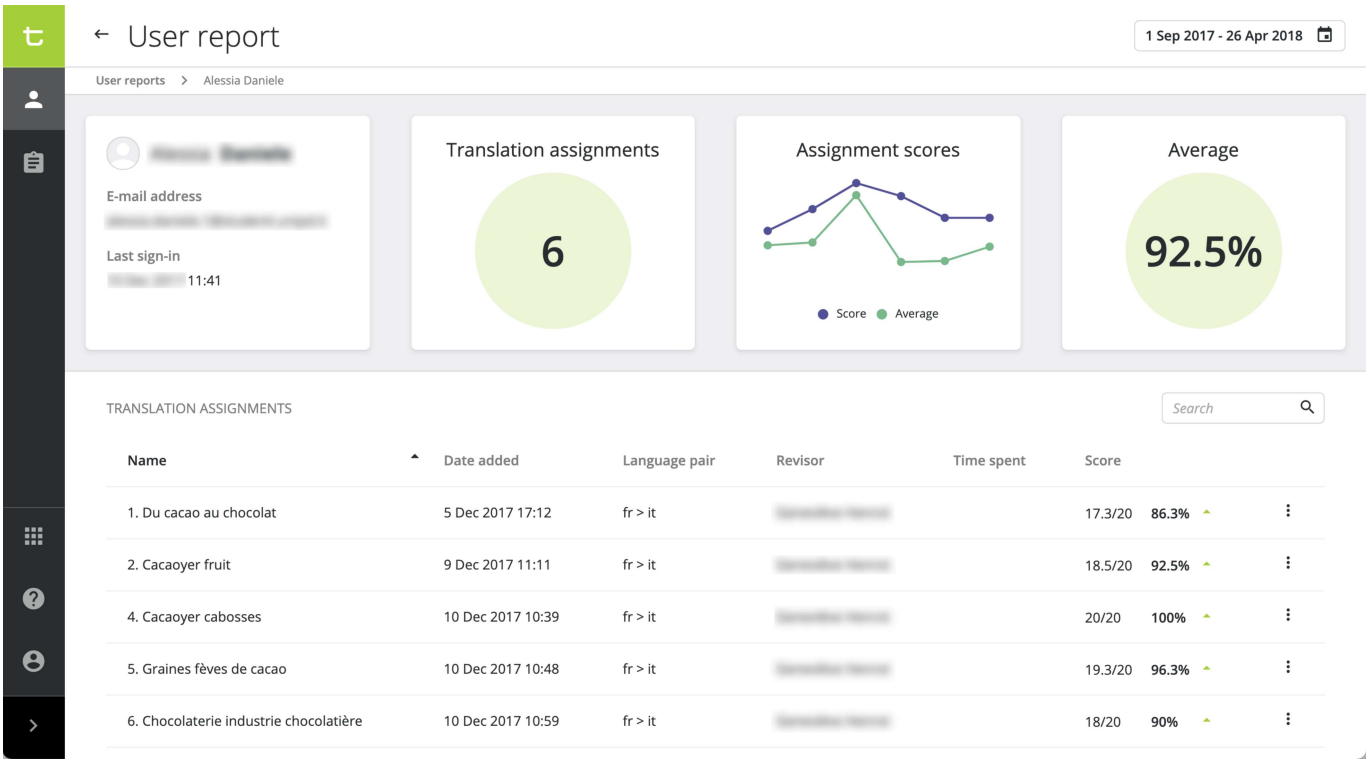

Figure 4 Screenshot of user report

The trainer can also delve deeper into the 'big wrong data', by exporting the data in an Excel sheet and experimenting with pivot tables. By pivoting the data in the tables, insights can be gained and visualised that seem outside the trainer's grasp in traditional revision practices. The tables offer insight into: score distribution per error category (answering the question: does the trainer give the same or other weights to errors within the same category?); general error frequency (answering the question: which are the most frequent errors incurred in training?); group profiles (answering questions like: which error categories are more frequent in a group and require the trainer's special attention?); and trainee profiles (answering key questions like: which are the specific weaknesses of a student?). In the same 'User report' tab, the student too gets a good idea of their performance, as the tool displays error frequency, tasks scores over time and it sets individual scores in opposition to group means. It stands to reason that this formative feature is likely to add a powerful impetus to social-organisational ergonomics. Information yielded through user reports can be used not only to better address the needs, abilities and limitations of the student group, but they also serve to identify the specific needs, abilities and limitations of the individual, steering didactic practices in the desired direction. In terms of social and organisational ergonomics, this effect is highly desirable, as didactics not only inform assessment, but assessment also come to inform didactics.

References

Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. *Meta: Journal des Traducteurs*, *46*(2), 345–364.

Holmes, J. S. (1988). *Translated! Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi.

Hönig, H. G. (1998). Positions, power and practice: Functionalist approaches and translation quality assessment. In C. Schäffner (Ed.), *Translation and Quality* (pp. 6–34). Clevedon: Multilingual Matters.

Saldanha, G., & O'Brien, S. (2014). *Research Methodologies in Translation Studies*. London: Routledge.

TAUS. (2015). *Quality Dashboard White Paper June 2015*. Retrieved from: www.taus.net/insights/reports/taus-quality-dashboard-white-paper

Thelen, M. (2016). Quality and objectivity of testing and assessing in translator training: Dilemmas and suggestions. In M. Thelen, G. van Egdom, D. Verbeeck, & B. Lewandowska-Tomaszczyk (Eds.), *Translation and Meaning* (pp. 79–98). New Series volume 1. Frankfurt am Main: Peter Lang.