

Of power adapters and language quality assurance

It all started with my modest desire to purchase a backup power adapter for my notebook to avoid carrying it between work and home. Reasoning that notebooks come and go, and their tips are notoriously incompatible, I decided to get myself a universal one with multiple tips and searched through amazon.com. Imagine my amazement at seeing indistinguishable electronic devices ranging between a very affordable \$7 and an impressive \$50, let alone brand adapters more intriguing was that user ratings were varied wildly for most items that anyone hardly any correlation between price and user rating (i.e. perceived quality).

By Leonid Glazychev

In the translation industry we also see significant price variations (though, luckily, on a smaller scale) and even more serious variations in review ratings. You have to wonder: Is there a real, pressing need to measure language quality and if so, is there an objective and viable method of measuring it?

The “lemon market” trap

To better understand the QA issue within the translation industry, my colleague and Logrus co-founder Serge Gladkoff has drawn a parallel to a different market: The term “lemon” is strongly associated with the US car market, where at a certain time the quality of cars was bad enough to call for a special legislation protecting consumers. The term “lemon market” was coined by the economist George Akerlof and depicts the so-called information asymmetry, which occurs when the seller knows more about a product than the buyer.

Image: Leonid Yastremskiy

The concept in brief: As far as the buyer can't reliably measure the quality of the item (a car, a power adapter or a translation), he will assume that the item is of average quality. Hence, this buyer will only be ready to pay an average price for the item, which in turn leads to a situation where items of significantly higher quality won't achieve a high enough price to make selling them worthwhile. The withdrawal (or marginalization) of high-quality items reduces the average quality of items on the market, causing buyers to revise their expectations downward for any given item. This, in turn, motivates the makers/owners of moderately good items not to sell, and so on. Thus, there will always be an incentive for sellers to pass off

low-quality goods or services as higher-quality ones and a distinct advantage for some vendors to offer low-quality goods or services to the less-informed segment of a market.

Now, let's go back to real life and start with laptop power adapters and the abovementioned user review mess. As we see, what's happening is the exact consequence of information asymmetry, i.e. we are trying to rate the quality of goods/services without being able to analyze truly important factors. When I get a power adapter I can assume with confidence that most adapters, irrespective of their price, were assembled at an obscure factory somewhere in China. I can also check some basic things: whether it has a tip that fits my laptop, whether it works at all, and whether it overheats dramatically or dies within the next several days. My review

The sad reality is that in most cases customers get a well-disguised less for less.

would therefore be based on these obvious aspects. However, my scientific background tells me that I am actually missing the main point. What does really matter in the long run, regarding the quality and reliability of my power adapter? The quality (and source) of its circuitry and parts, a proper design taking into account power dissipation, the soldering (automated rather than manual to avoid future oxidation), etc. Can I measure these factors by any means under normal conditions? No! So my guess about the device's quality/longevity will, at a closer look, remain just a wild guess.

When buying the adapter I will most likely think: \$7 seems to be marginally low, \$50 is way too much,

and most of them cost below \$25, so \$15 will probably get me a decent one. And I might be totally wrong, because this estimate might prove too low to get a reliable adapter made of robust parts at a clean, automated factory with full-scale quality control. This is a lemon market: average values are absolutely deceptive and I'm not choosing between good and bad, but between multiple lemons!

Sounds too familiar? That's because over the last ten years the translation market has evolved exactly into a market for lemons. This is due to a number of factors, including enormous pricing pressure from clients (who are typically much bigger than vendors in a B2B market), unlimited recycling (giving a new life to erroneous or obsolete translations) and abundance of poorly post-edited MT

materials, low entry cost (tempting thousands of amateurs), an incredible level of segmentation (when it becomes almost impossible for the translator to understand the context), free crowdsourcing alternatives, etc. The only difference is that we are not talking about the loss of \$15 in the worst-case scenario, but about far more important implications and consequences. When considering typical feedback related to translation quality, one can't miss another surprising parallel with power adapter reviews: While some of it is quite legitimate and substantiated, most feedback comes from in-country offices or in-house employees who happen to know the language in question, or from a randomly-chosen reviewer who was asked to take a look at the translation. Most of these people are competent in their professional area, but they are neither linguists nor translators. More than that, they are typically unfamiliar with existing terminology glossaries, unaware of inter-product compatibility or legacy-related issues, are doing their review from scratch, i.e. without any support materials or formal guidelines, and tend to introduce a strong taste-based flavor into their reviews. As a result, a significant part of the language quality feedback obtained using traditional methods produces results that may be logical, but are anything but objective. Getting back to the analogy discussed earlier, the reviewers are basing their judgment on the exterior of the power adapter and its ability to provide 19V DC alone, rather than subjecting it to a series of objective certification tests.

While the "more for less" slogan has never been more popular, the sad reality is that one typically can't get more for less without major breakthroughs, which are not too frequent.



DOKUWERK
Wir schreiben Know-how.



**We speak all languages.
Above all yours.**

Your products should be understood worldwide and this is something that requires professional translation management. Irrespective of whether you need manuals, catalogues or advertising text translated - we will translate your content into every target language. CAT tools guarantee consistency and will help you to save costs and we can offer advice on everything associated with translation processes, terminology and relevant tools. For improved corporate communication and sustained quality of translation.

Dokuwerk KG | Friedrichshafen | Germany
www.dokuwerk.de | Tel.: +49 (0) 75 41 / 37 51 4-0

In most cases, customers quite expectedly get a “well-disguised less” for less. How can one survive with dignity on the lemon market and still buy translations of good quality? Only by eliminating this information asymmetry. In our case this means getting professional, thorough and objective, independent language quality reviews that bring both confidence and peace of mind and help avoiding costly errors.

What is Language Quality Assurance?

Language Quality Assurance (LQA) refers to the assessment of the linguistic quality of materials based on international and industry-wide standards as well as the client’s standards, requirements and guidelines. Primarily, this relates to two things: terminology and style on one hand, and quality metrics and criteria on the other.

The idea that any translator or editor can become a reviewer overnight is an illusion, and a dangerous one. Such instant conversions provide the unprofessional reviews discussed earlier. In reality a lot of specific training is required, as well as a specific mindset. A reviewer can’t fix errors or improve whatever he finds necessary. Instead, he has to:

- Follow a complete set of formal rules
- Use formal feedback forms correctly
- Follow very specific and often rather peculiar guidelines for each job, as LQA requirements and guidelines might differ for various clients
- Apply strict evaluation metrics, and suppress all emotions
- Either ignore or impose style-related considerations depending on the client’s requirements
- Conduct reconciliation discussions with the translator

It’s worth mentioning that LQA is not supposed to improve materials or fix errors. It is simply expected to give us a reliable estimate of how good or bad the materials in question are. Despite this trivial definition, we’ve come across multiple cases where client expectations related to LQA were quite different:

1. The most common belief is that LQA will not only assess quality, but also fix problems, which is actually a hybrid of LQA & editing. Editing to LQA is like treatment to diagnosis. You first need to make a good



Image: Christy Thompson

In his paper “The Market for Lemons: Quality Uncertainty and the Market Mechanism” economist George Akerlof discusses the phenomenon of information asymmetry, which occurs when the seller knows more about a product than the buyer. Whether buying a used car or a translation service one problem is always the same: You don’t realize its faults until after you bought it.

diagnosis, and then start the treatment. The combination of LQA and editing is only viable short-term, but will not work as a permanent solution because if one vendor translates materials, and the other performs QA and editing, then there is no hope for quality improvement at the source. Continuing the medical analogy, if a chronic condition is the consequence of a lifestyle, treatment will not produce a permanent cure.

2. It is also tempting to combine LQA and functional testing, because it is assumed that native speakers see more errors as they have a better understanding of the language, and because both tasks seem to blend well. But in reality, a native speaker who also has a thorough technical understanding for installing or configuring the software properly, checking all cases without missing anything, reporting bugs according to standards, etc. would be an extremely rare combination of talents.

This is especially true if more than one computer is required or the setup is quite sophisticated. Goals and methods, as well as software, are quite different.

3. Finally, it may seem logical if a single vendor performs both translation & LQA. But there is a good chance that borderline LQA results would be “adjusted” (these results are often near or below the acceptable level in the lemon market case). Also, in some cases LQA results would be “covertly” taken into account during the correction stage, but not fully logged to avoid negative perception at the client end and to improve the overall picture. In other words, even if both services are separated at the administrative level, the issue of failed QAs tends to become political in large projects. And it often results in a significant share of errors either fixed “through the back door”, bypassing the QA, or simply left unlogged. Formally, everything is fixed, but it completely de-

feats the purpose, because both QA results and the initial quality assessment might be seriously skewed.

Building a comprehensive QA model

Let's now outline an approach to measuring quality. As far as language quality is concerned, all of us have heard horror stories about unfair treatment, abhorrently skewed reviews, etc., and many of them are true! Therefore, I have divided all criteria that might be used for measuring quality into three categories:

- **Objective criteria** are the ones that are universally recognized, univocal, and easily applicable. All violations/deviations can be clearly described and proof is universal, one doesn't need to know the language to understand marked errors. Typical examples include spelling and grammar, country standards, adherence to terminology and style guides, etc. Two people doing the same review based on objective criteria only will most probably come to the same results. Most objective criteria are applicable at a very low level, such as separate words or sentences.
- **Subjective criteria** cover preferential, taste-based, obscure arguments like the following: "I don't like it", "This is bad", "Poor style" "It sounds better that way"... In all such cases one can't clearly explain what's wrong and why, and the feedback is not well-structured.
- **Expert opinion-based (semi-objective) criteria** is a third, semi-objective category, based on the fact that LQA is, regrettably, not all black and white. This category includes several important expert assessments, including overall intelligibility of the text, adequacy or equivalence of translation, language fluency, etc. Is it likely that if a trained expert finds the text incomprehensible or discovers serious deviations from the original meaning, so will most other readers? Certainly. Expert opinion-based criteria are mostly applicable to bigger chunks of text, such as paragraphs or pages, rather than words or sentences.

Based on this categorization, one can offer an approach to building a comprehensive, fair and objective LQA system.

1. Select the expert opinion-based criteria and define a grading system. Since these

evaluations are not completely objective, each particular evaluation is NOT as accurate and bullet-proof as the one based on objective criteria. For this reason I would strongly discourage anyone from combining these two categories.

Let me provide a simple illustration: Intelligibility of the text as a whole (which falls under the expert-based category) is a major criterion by itself and overrides any other valuations. If the user can't understand the text, everything else becomes irrelevant. Since the evaluation according to these criteria is not objective, exact score rating might vary from expert to expert. I suggest rating it as pass/fail to eliminate incomprehensible or inadequate texts.

2. Select relevant objective criteria, clarify definitions and assign weights to different error categories to calculate an integral objective quality assessment. There are multiple models on the market, including the one developed by Logrus, and all of them use relatively similar criteria. The acceptance threshold may vary depending on the goals and expectations.
3. All subjective complaints related to style and other nuances should be consistently ignored.

Combine this recipe with representative sampling, and you get a more or less objective and robust quality model that is widely applicable.

Applying the model

First we generate a single, integral expert opinion-based criterion based on all expert-level (semi-objective) criteria, including intelligibility, fluency, adequacy etc. and set the "pass" threshold. If the text is not intelligible or adequate, it doesn't make sense to delve deeper into details. Let me remind you that expert ratings are not completely objective, and there is a grain of subjectivity in each of them. We should be thinking about this subjectivity and natural variance of results when setting the threshold. So, for example, when I am setting the threshold to 7 out of 10, it actually means that I expect the average good result to be around 8.5, and allow another 1.5 points to cover the natural imperfections related to subjectivity.

Secondly, we apply objective criteria to files that have passed the expert opinion barrier. These are already considered intelligible and adequate enough to talk about quality in more detail. As a result we get materials with two separate ratings: the integral expert rating, for example 7.8 out of 10, and the integral objective rating, for instance 8.6 out of 10.

For the files that passed the expert opinion barrier we can combine these two ratings if we really want a single indicator representing quality. One thing to remember is assigning a relatively low weight to the expert rating to avoid excessive sensitivity to subjective nuances of that rating.

Last but not least: This approach works perfectly with MT-generated materials. The only thing one needs to do is adjust thresholds, lowering them accordingly.

contact



Leonid Glacychev

graduated with honors from the Moscow Physico-Technical Institute, specializing in theoretical physics and computer

simulation. After the demise of the Soviet Union he joined one of the first software localization projects for the Russian language. In 1993 Leonid co-founded Logrus, the first professional translation and software localization company in Russia, and has served as CEO since then. Today Logrus has multiple offices and 150+ permanent employees.

leonidg@logrus.net
www.logrus.net