Linguistics and Literature Studies 5(2): 122-131, 2017 http://www.hrpub.org
DOI: 10.13189/lls.2017.050208

Reliably Measuring Something That Isn't Completely Objective: The Quality Triangle Approach to Translation Quality Assurance

Leonid Glazychev

Logrus IT, USA

Abstract The article introduces the first hybrid language quality assurance (LQA) methodology for translated texts, the Quality Triangle. This methodology was developed to assess original translated materials, with no available reference translations. It combines holistic and analytical approaches and provides the most flexible solution for a wide variety of applications and subject areas. Special attention is paid to the proposed approach detailing how holistic quality factors can be taken into account given the challenge they represent: these factors need to be assessed, but are at the same time not completely objective by definition. The resulting Quality Triangle methodology provides a reliable model for incorporating semi-objective, holistic factors into quality assurance metrics and can serve as the foundation for building models applicable in real production processes and serving a multitude of purposes.

Keywords Language Quality Assurance, Translation, Holistic Approach, Analytical Approach, Atomistic Criteria, Acceptance Threshold, Objective Factors, Semi-objective Factors

1. Introduction

While topics like quality issue classification or particular quality metrics have attracted plenty of attention for years, the methodology of Language Quality Assurance (LQA) of translated texts has received significantly less focus. In other words, much was said about concrete issues and metrics and very little about how to build these metrics (i.e., what is essential, which misconceptions people might typically have, the most frequently made errors, etc.)

This paper addresses the abovementioned cornerstone question and discusses general principles of building LQA metrics. It also explains why language quality measurement cannot be reduced to a single rating assigned to the text (however convenient it appears on paper) without significant distortion, proposing, instead, the Quality Triangle methodology. This methodology relies on three independent quality apexes: Holistic Readability, Holistic Adequacy, and Atomistic Quality.

2. General Considerations for Building an LQA Model

When talking about general principles of building LQA models, it makes sense to start with a list of essential expectations that the methodology needs to address. In my opinion, this list includes at least the following four line items:

1. Reflecting the Perception and Priorities of the Target Audience

- Concentrating on factors producing the strongest impression and
- Separating global (holistic) and local issues, understanding that the former are typically more conspicuous and play a bigger role in the initial and/or overall impression

2. Universal Applicability

- Covering the whole spectrum of potential uses, subject areas, and materials
  - From slightly post-edited MT to ultra-polished manual translations
- Common approach
  - Same approach to knowledge bases and marketing leaflets
  - Only adjusting acceptance criteria/thresholds based on expectations

This may appear simple, but is not always easily accepted in practice. The truth is, we are all humans and, irrespective

of what exactly we are looking at, be it a restaurant menu or prescription label, we make our first judgment about text quality using exactly the same criteria. We do not need a different approach or a completely new metric for each subject area or type of content. In reality, the only thing that requires adjustment is tolerance level. We are ready to accept a barely comprehensible menu translation, but expect perfect clarity and lack of ambiguity in the medical area. In technical terms, this means that we are still measuring the same thing, i.e. readability/clarity, but with different expectations, and this approach applies to all other criteria.

3. Viability of the Methodology

- Clear, not overly complicated
- Process-oriented, i.e. reasonably economic and applicable in the real world

4. Flexibility of Approach

- Concentrating on methodology rather than particular cases/uses
- Criteria/issue classification is not an inalienable part of the methodology, but rather an add-on component. It can be taken from elsewhere, for instance:
  - Based on any publicly available source
  - Based on legacy criteria used/provided by the client
- Weights assigned to particular issues are expected to vary within a wide range depending on the goals set, subject matter, type of material, etc. Certain issues might simply prove irrelevant for the job or area of focus, which results in zero weights being assigned to these issues.

3. What Matters Most: The 3D Nature of the Approach

3.1. Quality Cornerstones: Holistic Adequacy and Readability

The idea that two major criteria for any translation, irrespective of its origin, target audience, brand impact etc., are holistic Adequacy and Readability (Fluency) first appeared as early as the 1960s in a venerable ALPAC report, Computers in Translation and Linguistics[1]. Expecting translated content to be readable and convey the meaning of the original adequately is absolutely essential, independent of the subject area, type of material, or anything else, for that matter. Both holistic factors are discussed in more detail in the next section.

While it is hard to dispute the universal nature of readability and adequacy, expectations as such might vary dramatically for both of these factors. These expectations are reflected in an answer to a different question: HOW readable or adequate the translated content should be.

3.2. Closing the Quality Triangle: Atomistic Quality

First of all, let me say that the term "atomistic" might sound somewhat out of place, but, in this context, it serves as an opposite to holistic. It describes all quality issues encountered at and limited to the "atomistic" level of the content, i.e. sentences, strings, translation units, etc. These issues are numerous and include such things as incorrect or inconsistent terminology, style guide deviations, incorrect formatting, broken tags or missing placeholders, atomistic-level adequacy and readability issues, etc. One can find an example of a comprehensive issue framework that includes both classification and definitions in the Multidimensional Quality Metrics (MQM) document[2].

3.3. Introducing the Quality Triangle

The Quality Triangle is shaped by all three criteria taken together: holistic adequacy, holistic readability and atomistic quality (error level). Prior to advancing, there are a few things I want to clarify:

- The Quality Triangle represents a 3D approach to translation quality.
- Holistic criteria are by no means "quick, preliminary evaluations" that can be later replaced by a more detailed, in-depth atomistic quality analysis
  - The relationship is a more sophisticated one between the whole and its constituents, as discussed in the next section
  - All three measurements are independent, they represent different "quality dimensions" and complement each other
  - For instance, while each sentence may be perfect in isolation, the translated text *overall* can completely lack coherence or fail to convey a message as a whole. On the other hand, the text as a whole may be easily readable and make sense, but at the same time be riddled with annoying atomistic-level errors (spelling, grammar, corrupt markup, etc.)
  - Real-life experiments with 3D, hybrid quality metrics confirm this assumption. While detailed discussion goes beyond the scope of this paper, we have discovered that there is little or no correlation between holistic and atomistic quality measurements. Poorly translated, hardly comprehensible or incoherent texts were often close to perfect when it comes to atomistic quality, and vice-versa.
  - In one especially interesting case, one and the same content translated into two flavors of Spanish (for Spain and Mexico) produced similarly high holistic quality ratings for both adequacy and readability, but diametrically opposed atomistic error levels (almost no errors for Spain and abundance of minor defects for Mexico).
- As a rule, none of the three criteria is more important than the other. At the same time, the expectations for

each are strongly dependent on particular circumstances, the area, etc.

3.4. Where to Start in Real Life

As mentioned earlier, in the world of commercial translation cost and approach flexibility and viability, are all key issues.

- Atomistic-level review always involves considerably more effort compared to a holistic one.
- While the process of detecting many technical errors (like corrupted markup), and even some content-based translation errors and shifts, can be automated to a certain degree, manual effort is required to localize issues within sentences (strings, units) or to distinguish actual errors from false positives[3]. Moreover, each issue needs to be not only discovered, but also classified, logged, and described with sufficient clarity, let alone reviewer training, providing all required ancillary materials, etc.
- Automated evaluation methods like BLEU[4] demonstrate a certain level of success when it comes to training, fine-tuning or comparing MT engines, but all of these methods require reference human translations that are unavailable in the real world, when new texts are translated for the first time.
- A significant share of translated content is not originally designed to be perfect or even good. For instance, due to budget considerations (huge volumes; continuous updates) expectations for most knowledge bases or in-depth Q & As are quite modest, based on the idea that having "good enough", i.e. comprehensible, more or less adequate content in local language is much better than having nothing at all. Hence MT with limited post-editing is the approach of choice.
- Under these circumstances spending effort on logging most atomistic-level issues like typos or grammar errors makes little sense. Showstopper-level errors (see a separate section below) are the only exception.
- Reasonable holistic adequacy and readability are essential and need to be checked for any contiguous content that is expected to make sense as a whole. Moreover, if in a particular case these evaluations produce unsatisfactory results it simply doesn't make sense to carry out atomistic evaluation at all. Why waste effort and go through sophisticated technical details or error counts if the text as a whole is either unreadable (incomprehensible) or inadequate (inaccurate) and requires major rework anyway?

For all the reasons listed above, I strongly recommend beginning with holistic evaluations in all cases where these are relevant, as far as we can measure them with sufficient objectivity (discussed in the next section). Moreover, for projects with severe timing or budget limitations, I recommend using a simplified metric that does not include atomistic-level review at all. (We have developed and successfully tested such a metric at our company, Logrus IT, but it goes beyond the scope of this paper.)

4. Semi-objective Nature of Holistic Criteria

4.1. Defining Readability and Adequacy

Without pretending to present something that is perfect or carved in stone, I would like to define both fundamental notions of readability and adequacy based on their intuitive meaning, and keep these definitions as concise and simple as possible. We will need them to better define the model.

Readability of translation defines how easy it is to read and understand the target text (sentence, string, etc.). In other words, readability measures how clear, unambiguous and well-written the target text is. Zero readability means that the text is unreadable or incomprehensible, i.e. it represents a senseless sequence of words. Perfect readability means that you can easily read the text without stumbling over words or phrases, the meaning of the text is absolutely clear and unambiguous, and no additional pondering is required to get to this meaning.

Readability goes far beyond translation quality as such, and formally applies to any text, including monolingual text in any language.

Important! The mere fact that you can easily read and comprehend the text does not mean that its meaning is correct. That is why we also need to measure adequacy.

Adequacy of translation defines how closely the translation follows the meaning of and the message conveyed by the source text (sentence, string, etc.). In other words, adequacy measures whether the translation process resulted in any discrepancies between source and target texts (except for intended ones), including plain translation mistakes, omissions or additions. Zero adequacy means that after translation the meaning of the source text was distorted beyond recognition. Perfect adequacy requires that both the meaning of and message conveyed by the source text were preserved in their entirety, without any deviations.

Adequacy only applies to a combination of both source and target texts (units, strings, sentences, etc.). You will need a bilingual resource to assess it. Analyzing translation would reveal a certain number of potential issues, but they cannot be verified without access to the source, let alone cases where translated text is smooth enough, but incorrect.

Important! It is worth noting that neither of the fundamental quality concepts described above comprises sub-concepts. Each is a standalone, "elementary" semi-objective quality factor.

4.2. Emphasis on Holistic Assessment

It is important to reiterate that we are talking about global

content characteristics dealing with the perception of translated text (piece of software, website, leaflet, etc.) as a whole. That is, any potential reader/consumer is primarily interested in holistic readability and adequacy of the whole piece, and only then in readability or adequacy of particular sentences (units). The latter is, of course, important as such, but not at high level. It's one of the universal laws of nature: The whole is always more important to us than its constituents, and its properties can't be fully revealed or described based on these parts alone.

There is a direct material analogy here: When I need a hammer, I am primarily concerned with whether the object in question resembles a hammer and can be used as one, not concentrating on various manufacturing imperfections.

Below, please find some more translation-related reasons:

- Natural human perception. A translated tourist guide article on local restaurants can serve as a good illustration of holistic adequacy and readability dominating everything else.

As far as the piece provides necessary information about local landmarks, it is much better than the complete lack thereof in a foreign country with an unfamiliar language, even if translations are imperfect and some pieces are incomprehensible. Particular mistranslations (like recommending the "cancer shakes" at a certain restaurant in St. Petersburg, Russia – meaning "crabmeat") would certainly be considered errors, but local ones, limited to a particular part of the text and not seriously affecting the perception of the document as a whole.

If the overall number of errors is not excessive, we would probably come up with a conclusion like this one: "The translation is useful and relatively acceptable, but there is a certain number of errors that need to be fixed to improve overall perception". On the other hand, if one couldn't make sense of the text as a whole, particular mishaps would not be as essential any longer, because the high-level diagnosis would be quite different: "The translation is incomprehensible. It does not make sense to discuss individual errors as far as the whole piece requires complete retranslation from scratch."

- Patchy, out-of-context translation of small standalone pieces is the fact of life in the modern world, and it's becoming more and more ubiquitous. Separate translated segments might look perfect by themselves, but taken together often create clumsy, controversial, incongruous texts.
- The "cunning translator" phenomenon.

Anyone with experience in the industry has observed it multiple times. When a translator does not completely understand the source text due to limited subject knowledge, lack of context or some other problem (this, regrettably, is not an exception given the number of subject areas), he/she often tries to make translated sentences (segments) as "round", ambiguous and fluent as possible. Each particular sentence sounds nicely and might pass the editing/proofing stage without corrections when dealt with on a sentence-by-sentence basis (that's exactly what they are counting on). But taken together the piece often makes little or no sense whatsoever.

One important consequence: Quality assurance cannot be complete or accurate if there is no way of making holistic evaluations. There is a certain similarity here with trying to make a judgment about the object based on its atomistic (molecular) structure alone: Even if it's just iron, that doesn't tell us anything about its state (liquid or solid) or the object's shape (does it look like a hammer?). That's exactly the reason why, for instance, simply going through software strings or sentences one by one is absolutely insufficient to make general conclusions about overall usability or quality of a software product or a web portal. On the other hand, looking at screen shots representing essential parts of the functionality would result in a much more accurate overall evaluation.

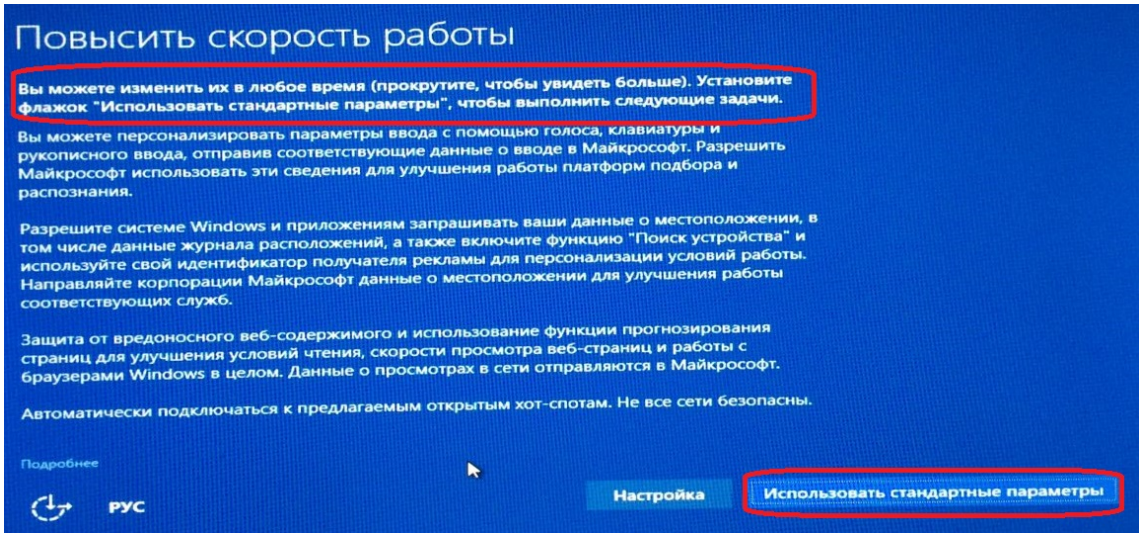

Figure 1. Windows 10 (Russian) – One of the first setup screens

Not completely convinced? Let's take a quick look at one of the first setup screens for the Microsoft Windows 10 OS localized into my native Russian, for example. There isn't a single correct translation in the entire dialog. At the same time, all issues but one revealed while analyzing the screenshot as a whole will pass unnoticed during sentence-by-sentence review. The first paragraph (outlined in red) contains three serious translation errors and the rest of the paragraphs in the screenshot do not logically follow the first one.

1. The first sentence (reverse translation from Russian) reads: "You can change the them [these] at any time (scroll to see more)". This is a perfectly fine standalone sentence, and the reverse translation fully corresponds to the original, which reads "Change these at any time…"
2. In English, however, the word "these" in this context actually means "the following settings/parameters". In Russian, the sentence makes zero sense because "them/these" must necessarily be preceded by some sort of definition/explanation regarding what the actual subject is (what it is, exactly, that you can change), which is not provided here.
3. The second sentence (as a reverse translation from Russian) reads: "Check the 'Use Express settings' checkbox to perform the following tasks." This functions fine as a standalone sentence, and its translation represents one of the possible interpretations of the original that reads "Select 'Use Express settings' to…" The Russian translation needs to be more verbose due to the specifics of the language. For instance, Russian requires the use of a noun specifying the control used (checkbox, button, etc.) as well as the action verb and the subject on which the action is performed. This is why the words "checkbox" and "tasks" got added.
4. Still, within the actual context, the original meaning was completely distorted.
   - It turns out that 'Use Express settings' is not a checkbox to check, but a button to click (see the screenshot bottom)
   - What's worse, the original sentence really means that if the user selects the easy way (to 'Use Express settings'), Windows 10 will use default settings as described in the subsequent paragraphs. But the Russian translation says that we need to set the checkbox to perform certain tasks that are listed below. In reality, these are not tasks, but settings, and no tasks are going to be carried out at all… Ultimately, the translation is completely incorrect contextually.
5. Instead of describing default (express) settings, subsequent paragraphs interpret the text as something one can do (first paragraph), something Windows suggests you to do (second), etc. All of these make sense as standalone pieces, but are completely unsuitable in relation to the first or other paragraphs.
6. Additionally, the title is also translated incorrectly (US source: "Get going fast"; reverse translation: "Increase work speed"), but this is simply an incorrect translation (looks like unedited MT) that will be immediately discovered during a sentence-level review.

4.3. Semi-objective Nature of Holistic Criteria, and How to Deal with It

Neither of the two major translation quality criteria are completely objective. In an ideal world, where we could hire a whole expert panel to assess each piece, there would still be an inevitable variation in ratings assigned to one and the same text by individual panel members. This would happen even in a case when all of these people have similar backgrounds, undergo similar training, and use the same reference materials, guidelines and instructions. There is an unavoidable tint of subjectivity to these assessments.

At the same time, ratings produced by the panel would not be completely arbitrary either. In reality, they typically produce a normal opinion curve around the average rating for each document rather than white noise.

A couple of real-life examples are presented in the charts below. As many as 17 professional translators rated one and the same translated portal, assessing overall holistic readability and adequacy of translation. The horizontal axis represents rating values (on a scale between 0 and 10), and the vertical axis reflects the number of reviewers who came up with each particular rating. Bigger overall numbers of reviewers produce more reliable statistical results that are following the normal distribution.

Even in this case, with relatively few evaluators (by statistical measure), standard deviations in ratings are reasonable (much smaller than average values), which serves as additional proof of these holistic ratings being far from arbitrary. That is why I am calling both criteria (and associated grades) semi-objective, and we always need to remember that they are NOT too accurate by design.

The question is, how can we deal with this lack of complete objectivity in a real-world scenario, when no reference translations are available, there is a single reviewer who can only look at a certain percentage of the overall content, and we still need to evaluate and grade translated texts?

Your particular reviewer could either be good-natured and relatively tolerant, and would assign a near-perfect rating to the text (content), or turn out to be extremely fastidious or simply in bad humor on that day… The very same content could potentially get a rating anywhere under the dome of the Gaussian curve. And, in the real world, we do not even know up front how demanding this or that reviewer is… All of the above considerations are true, but it doesn't make the whole situation hopeless.

Figure 2. Readability and Adequacy Sample Charts

The solution lies in developing a clear, well-defined scale for holistic evaluations to minimize subjectivity and in evaluating the two major holistic criteria (readability and adequacy) separately on a PASS/FAIL basis to accommodate the natural variance in individual approaches.

Being realistic and judging based on existing statistical data, we can more or less safely assume that extremely low or high ratings deviating very far from the median value are a rarity (bad luck). In the majority of cases where reviewers use an elaborate, clear scale, are qualified enough, unbiased and trained properly, actual holistic ratings for both readability and adequacy will not be completely uniform, but are rather concentrated within a limited range around the average value.

Thus, the logical thing to do is create two separate evaluation scales for adequacy and readability respectively and establish an acceptance threshold that would correspond to the lower end of the statistical range described above. A rating below that threshold would mean that the translation is unsatisfactory for our needs.

That way, we can take into account the natural and unavoidable variance in semi-objective ratings. The range above the threshold would accommodate the majority of potential expert opinions (the bulk of the normal distribution curve). In other words, when setting expectations relatively high, for instance around 8 out of 10 on both scales, we should remember about the variance in ratings assigned by reviewers: A substantial share of reviewers might rate the text that "deserves" an 8 as a 7 or even a 6.

Set the threshold too high, and you run the risk of failing a significant share of good translations just because the reviewer in that particular case was too strict or didn't get all instructions/guidelines. (The risk of accepting a number of so-so translations just because the reviewer was lax is always there, irrespective of the approach).

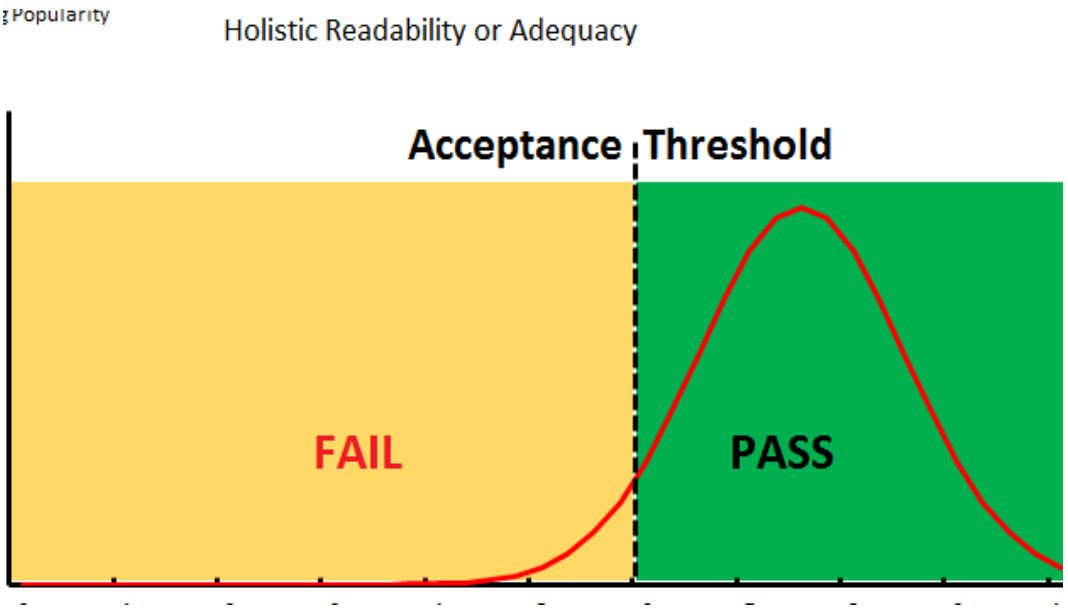

Figure 3. Acceptance Threshold Illustration

One important and direct consequence of this approach is that the scale used for holistic translation ratings should be at least between 0 and 10, and by no means smaller. Otherwise, it will prove too narrow to accommodate the real-life rating variance (the Gaussian curve will simply not fit), and there will be no choice left for setting acceptance thresholds depending on the requirements, subject matter, specifics, etc.

Assuming that a 10-point scale is used (0 meaning unreadable/completely inadequate and 10 meaning a perfect text), we can consider the following typical scenarios:

- For a marketing text, one would expect to have acceptable grades between 8 and 10. We do allow some variance, but any reviewer needs to consider the content well-translated (even though it could be either an 8 or a perfect 10).
- For a knowledge base, our requirements would become considerably more moderate. It is normal to set the acceptance threshold at 5.
- In each scenario, the acceptance threshold is defined by the area, visibility of materials, time constraints, target audience, etc.

As mentioned earlier, it is crucial to develop a clear, well-defined scale to guarantee a certain level of objectivity and make sure that each reviewer understands the difference between ratings. In other words, every person involved needs to know what a rating of 5 (or any other rating) means, and how 5 is different from 6. At Logrus IT we have developed both scales for holistic adequacy and readability based on the principles outlined in the ALPAC publication[1]. Currently, both original scales are part of Logrus IT's 3D, hybrid translation quality metric family. Publishing both scales in their entirety goes beyond the scope of this publication; these scales are also original work that currently constitutes part of our "corporate trade secret." Here, I can provide a sample that gives an idea of what is expected.

- Adequacy = 5: Translation has introduced minor misapprehensions about the general meaning of the text or the meaning of individual sentences or words. It has also changed or distorted a certain amount of information about the paragraph or sentence structure and syntactical relationships.
- Adequacy = 6: Translation has given a slightly different "twist" to the meaning conveyed by the original and has somewhat changed or distorted a couple of potentially critical meanings, primarily at the word level. It didn't, however, alter anything at the sentence or paragraph structure.
- Readability = 6: The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. An additional round of editing could leave this in nearly-acceptable form.

Experiments staged using real-life quality assurance projects and multiple reviewers grading the same content have demonstrated high level of result consistency when our original scales were used. Holistic ratings from different reviewers for the same text using Logrus IT scales were either identical or extremely close.

It is important to understand that each of these two major criteria should be evaluated separately:

- Accurate but hardly readable texts are as useless as fluent but inadequate ones.
- One can't simply summarize or combine these two factors by any means – these are two independent "coordinates on a holistic quality plane". Depending on the circumstances, one might easily have different and independent expectations for each of the holistic criteria. For instance, we might tolerate marginal readability, but

still expect very high adequacy for sophisticated technical content targeted at experts only.

For the reasons stated above, it is simply impossible to combine readability and adequacy requirements into a single unified criterion or formula. Any such attempt results in significant quality assessment distortions.

5. Atomistic Quality

Continuing the material analogy, at the atomistic level we no longer consider the hammer's shape or basic functionality, but rather concentrate on alloy structure and purity, handle quality, etc. This analysis complements holistic usability/quality evaluation and makes it possible to answer questions about the tool's potential durability and internal structural defects, disposition towards rusting, and other such things.

Besides having an essentially local nature, atomistic quality issues, unlike holistic ones, are mostly objective, because cumulative atomistic quality ratings are expected to be very similar irrespective of the reviewer's personality. A typo is still a typo, an error in country standards is still an error, and all reviewers will notice and classify all such issues in a similar way, given proper training and background. Everything depends on issue classification and the weighting system applied. The only potential sources of discrepancy are attention lapses on the part of the reviewer or minor differences in assessing readability or adequacy of particular translated sentences (units).

There is a price for achieving objectivity in atomistic quality evaluation. Atomistic LQA results will only be uniform if reviewers are professionals in the area and were specially trained for the job (leaving emotions aside and adhering to guidelines), issue classification is comprehensive and clear, and all ancillary materials are provided, including a complete and approved terminology glossary, style guides, special requirements, etc. As far as any of these components are missing, evaluation objectivity starts to vanish. This is a typical case, for instance, when client representatives who know the language start reviewing translations without access to glossaries, guidelines and TMs…

For professional LQAs issue categories can be based on MQM[5, 6] or other public source, on legacy client-sourced criteria or other criteria. The resulting atomistic quality rating for any text (content) is calculated based on a simple and straightforward approach. For each issue category/type the number of issues found is multiplied by the relative weight assigned to that issue type. Resulting values are summarized across all issue types, and the sum is then divided by the number of words reviewed for normalization purposes: $Q_A = \sum_i (N_i * W_i) / V$, where Q_A is the atomistic quality rating, N_i is the number of issues of type *i* found in the text, W_i is the relative weight assigned to this type of issues, and V is the volume reviewed.

6. The Fourth Apex: Showstopper Problems

In some cases, it makes sense to use an additional quality dimension to the three described earlier. **Showstopper problems** are discovered at the atomistic level, but stand out due to their overall impact. This subcategory comprises issues that could result in dramatic distortions in the text meaning, serious factual errors or political incorrectness, use of pejorative text, etc. The issues as such can belong to any branch in the issue classification, but need to be treated separately, with utmost attention, because they could seriously and negatively affect overall user perception and/or result in incorrect user actions.

In technical terms, relative weights of showstopper problems are determined not by their category within the issue framework, but by their negative impact, and that's exactly the reason why this additional quality criterion is required. For example, typos typically have a relatively low weight in any translation quality metric compared to other issues, such as country standards violations, tag corruption, etc. But a typo in a major headline on a news website, especially the one that can result in complete distortion of the meaning, would definitely need to be treated in a completely different way.

Typically, showstoppers do not need to be separated from other atomistic-level issues. It is sufficient to assign a huge weight (by far exceeding weights of regular issues) to each of them to guarantee an LQA failure. Separating this category makes a lot of sense for projects with huge volumes, limited budgets or tight deadlines, where the primary goal is to guarantee that the material to be published does not contain anything terrible or outrageous (no showstoppers), while logging minor or medium atomistic-level errors goes beyond the project's scope. In the latter case, it is recommended to evaluate overall atomistic quality at a holistic level (perceived concentration, severity and annoying effect of atomistic issues encountered in the content) and separating it from showstoppers, each of which is logged separately.

Under normal conditions, no showstopper errors are allowed in any content, irrespective of the volume. These need to be eliminated at the editing/reviewing stage, prior to publication.

7. Building the Quality Assurance Metric

The universal methodology suggested in this paper allows building any number of metrics suitable for various subject areas, types of content and expectation levels with minimal adjustments.

The whole process of creating a metric looks like this:

1. Select the quality model type (Quality Triangle or Quality Square) depending on your needs.

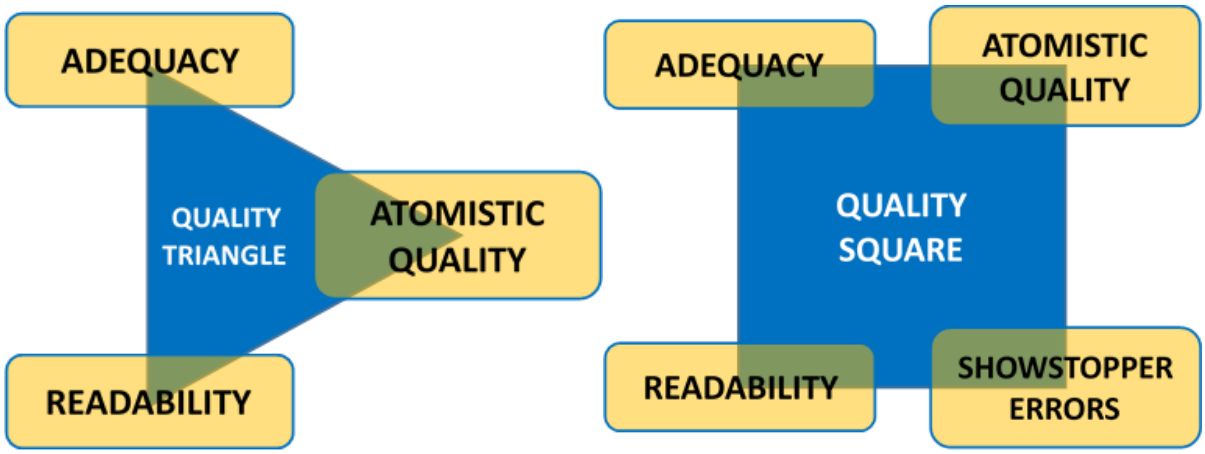

Figure 4. Quality Triangle and Square Model Illustration

2. Select issue classification/framework for measuring atomistic quality (for instance, MQM).
3. Select acceptance thresholds for the two holistic factors (adequacy and readability) and one for the atomistic quality, weights for each issue within the framework, and the scale for the ratings (such as 0 to 10, 0 to 100, etc.). The choice depends on multiple factors, including market segment, subject area, type of material, target audience, time limitations, brand impact, overall expectations, budget, etc.

One can find multiple examples that are publicly available. Typically, all issues are divided into three to four severity categories, and the same weight is used for all issues belonging to a certain severity category. For example, a serious error can be two or three times more "important" than a minor one, a major error might be two times more important than a serious one, etc.

Each set of acceptance thresholds and issue weights represents a so-called "quality vector" that completely describes and defines our expectations in each case. You can create

- One quality vector for marketing content (extremely high expectations in all areas, limited number of potential issues),
- One for software (high expectations, bigger number of issues, including software-specific ones),
- One more for user assistance and web content (reasonable expectations, adding content-specific issues),
- An additional one for knowledge bases and user forums (very limited expectations), etc.

A limited number of quality vectors will easily cover the whole spectrum of translated materials. As far as the chosen model and issue classification used in each case are the same (which is highly recommended), quality vectors will be the only components differentiating one metric from the other.

The methodology fully covers all types of translated content, including those produced using MT and/or MT + post-editing.

Applying the Created Language Quality Metric

Schematically, the LQA process based on the metric created will include the following:

1. **Selecting the quality vector** prepared earlier as part of the metric and applicable for the selected type of content.
2. **Applying semi-objective holistic quality criteria on a pass/fail basis**.

- The assessment as such is relatively quick and economic.
- Often scanning through the text is sufficient, especially so when quality is really low.
- Each text receives two separate quality ratings (readability and adequacy) that are compared to acceptance thresholds defined earlier as part of the quality vector.
- Failing texts are sent back for improvement or retranslation, saving the unnecessary effort of logging and counting atomistic-level errors.

Important. As mentioned earlier, these two quality ratings cannot be combined into a single integral factor as far as they are dealing with two content characteristics that are not only independent, but also not too accurate by definition.

Important. One needs to be bilingual or have a bilingual expert ready just in case

3. **Applying objective atomistic quality criteria. Only content that passes on both holistic accounts is further analyzed for technical and language imperfections, which removes the unnecessary workload of marking numerous errors in already disqualified texts.**

- Each text gets the atomistic quality rating calculated based on the number of issues of each type found and their relative weights.
- The atomistic quality rating is compared to the acceptance threshold defined earlier as part of the quality vector.

- The complete list of issues can be very long and detailed, but not all of them are taken into account. Some issues are irrelevant within the context and have zero weight.
- Showstopper errors are counted separately if required. Their presence typically means that the text failed the QA and needs to be fixed before publishing. Alternatively, these errors can be counted as regular atomistic errors, but with a higher relative weight.

Important. Each QA result includes three independent quality ratings (or four in case when we also separate showstopper errors) that present a complete, 3D "lossless" quality picture. Semi-objective, holistic ratings cannot be combined with the atomistic one due to their incompatible nature. Any formula combining these ratings would produce a highly unstable result that is too much dependent on the reviewer's personality due to the natural variance in holistic quality evaluations.

I will present a simplified quality metric built using the Quality Square methodology and also the accompanying process developed specifically for LQAs carried out through crowdsourcing in a separate publication.

REFERENCES

[1] Computers in Translation and Linguistics, a report by the Automatic Language Processing Advisory Committee (ALPAC), Publication 1416, National Academy of Sciences, National Research Council, Washington, D.C., 1966: http://www.nap.edu/openbook.php?record_id=9547

[1] Multidimensional Quality Metrics project (Primary contact: Dr. Aljoscha Burchardt, DFKI GmbH):http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

[2] Éric André Poirier, A method for automatic detection and manual localization of content-based translation errors and shifts. Journal of Innovation in Digital Ecosystems, Volume 1, Issues 1–2, December 2014, Pages 38–46

[3] K. Papineni, S. Roukos, T. Ward, W.J. Zhu. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2002), pp. 311–318

[4] Multidimensional Quality Metrics (MQM) Definition (Editors: Arle Lommel, Aljoscha Burchardt, Hans Uszkoreit; Contributors: Kim Harris, Alan K. Melby, Attila Görög, Serge Gladkoff, Leonid Glazychev): http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

[5] Multidimensional Quality Metrics (MQM) Issue Types (Editors: Arle Lommel, Aljoscha Burchardt, Attila Görög, Hans Uszkoreit, Alan K. Melby; Contributors: Serge Gladkoff, Leonid Glazychev, Kim Harris, Dale Schultz, Jean-François Vanreusel): http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html