Assessing Translation Quality with Multidimensional Quality Metrics (MQM)

The Multidimensional Quality Metrics (MQM) framework, developed in the EU-funded QTLaunchPad project, provides a flexible way to develop metrics appropriate for different types of translations and tie them to requester and user expectations. MQM draws upon a master listing of over 100 translation issue types that can support differing degrees of granularity and adapt to the requirements of any translation quality assessment need. **A metric (defined by completing Tasks 1–5) may be stored in a library, allowing for efficient reuse of metrics across multiple similar projects.** MQM metrics are analytic or holistic and should minimize human effort to apply.

Task 1: Complete Specifications

Specifications, based on ASTM F2575, define expectations about the translation and can serve as the basis for contractual obligations. All quality judgments should be made with respect to specifications. If a "problem" does not violate the specifications, it is either not a problem or the specifications were inadequate. (Note that a quality assessment does not have to check all specifications, but what it does check must be tied to the specifications.) Specifications are defined in terms of twenty-one "parameters" covering the translation product, project, and process.

Task 2: Select Dimensions

MQM dimensions are the high-level aspects you may check. The list of product-oriented dimensions to choose from is:

- **Fluency**: Linguistic well-formedness of target or source content on its own
- **Accuracy**: Consistency of meanings assigned to source and target content by informed users
- **Terminology**: Use of domain-specific language
- **Locale convention**: Formatting of content such as dates, times, or addresses for specific locales
- **Verity**: Appropriateness of text for the environment in which it will be used
- **Design**: Layout, formatting, and markup
- **Internationalization**: Engineering of source content to support translation/localization

Task 3: Select Method(s)

For each dimension one or more methods of assessment must be chosen. Methods include:

- **Analytic**: Identifying specific errors
- **Holistic**: Assessing the text as a whole
- **Task-based**: Assessing the text in use

Methods should also include a basis for assessment, such as the entire text, a random sample, or areas identified by automatic processes. *Task-based assessment is linked to content properties that help or hinder task performance.*

Task 4: Select Issues

For each dimension one or more issues must be chosen (the top-level issue corresponding to the dimension itself can serve as the issue if additional detail is not needed). These are used to measure that dimension in accordance with specifications. For example, if *Fluency* is to be measured, issues might include *Spelling*, *Grammar*, and *Register*. Issues that are not important to the specifications should *not* be checked (e.g., *Style* would probably not be checked for an internal technical service manual). The number of issues should be the smallest number that still provides sufficient detail to evaluate compliance with specifications while providing sufficient information to address problems in the translation product. **For validity, issues must correlate with user needs and KPIs.**

Task 5: Set Issue Weights

Weights are used primarily in analytic metrics to set the relative importance of different issues. Weights default to 1.0. To indicate that *Terminology* is more important than *Style*, *Terminology* might be given a weight of 2.0, meaning that it counts twice as heavily in the quality scores generated.

Task 6: Determine Thresholds

Thresholds set the criteria for accepting a translation. For analytic tasks these are generally expressed as percentage values based on the ratio of weighted errors to words in the text. For holistic methods, thresholds are often defined on a scale for the text as a whole. Thresholds can be set per issue or for dimensions as a whole.

Task 7: Implement a Workflow

Each composite MQM metric (consisting of dimensions, issues, method(s), and weights) plus purpose-specific thresholds must be implemented in an appropriate workflow with accompanying assessment tools. The assessment process within a translation workflow often includes "sanity checks" (i.e., texts that will obviously fail to meet checks do not need to be assessed in detail) and a set of clear and objective outcomes and decision points (e.g., send back to translator, approve, perform a detailed inspection).

CRITI Translation Quality Assessment Process for Standards

The CRITI translation quality system is based on the MQM specification specifically for the translation of standards for the Caribbean region. It is designed to provide assurance to CRITI and its customers that translations meet quality expectations. The project tasks were:

Develop Specifications (Task 1)
Specifications were drawn up based on interviews with CRITI staff about the expectations for the translated standards. Due to specific social conditions in the Caribbean region, certain deviations from the source document were expected in terms of register (standards are often written in a very formal, complex style, but this needed to be simplified somewhat for the expected educational level of users of agriculture standards).

Select Dimensions (Task 2)
Based on the specifications it was determined that only *Accuracy* and *Fluency* needed to be checked, and that these could be checked at a relatively coarse level. (CRITI's CAT tools handle formatting issues, so *Design* problems would be dealt with on a case-by-case basis.)

Select Assessment Methods (Task 3)
A complete analytic assessment of every text would be too time-consuming for CRITI staff,

so it was decided that holistic methods were more appropriate, with detailed analytic assessment reserved for marginal cases and spot-checking to confirm the holistic results.

Select Issues (Task 4)
For the holistic check, only three top-level issues are assessed, corresponding to the dimensions themselves: general *Accuracy*, *Terminology*, and *Fluency*. For analytic checking a set of eleven issues (shown at right) is used to evaluate texts for compliance to specifications. (These issues are appropriate for a technical specification text, but would not be appropriate for advertising or other style-sensitive texts.)

MQM Scorecard, an open-source tool for analytic quality assessment.

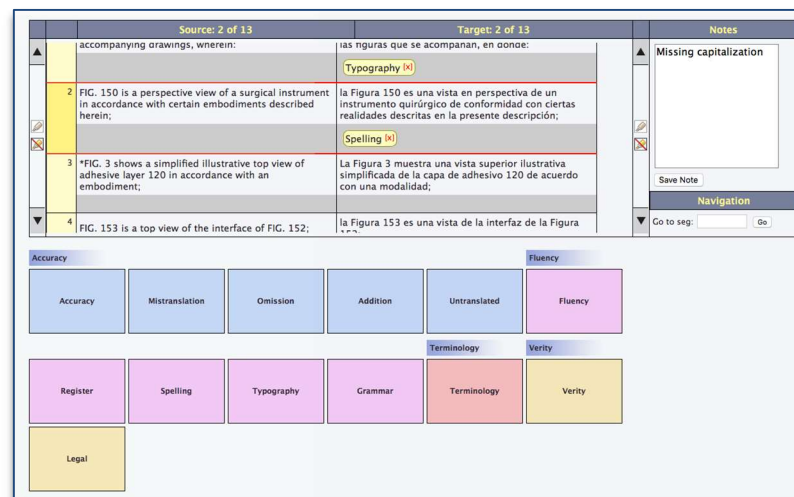*MQM issue type selection for Caribbean regional agricultural standards.*

Set Weights (Task 5) and Thresholds (Task 6)
Holistic assessment is done separately for these three dimensions on a five-point scale, from A to E: A and B indicate that the translation passes (with needed corrections for a B); C indicates that a detailed analytic evaluation is required; D and E indicate that the translation must be sent back for substantial revision. The translation must pass *all* holistic checks to be accepted.
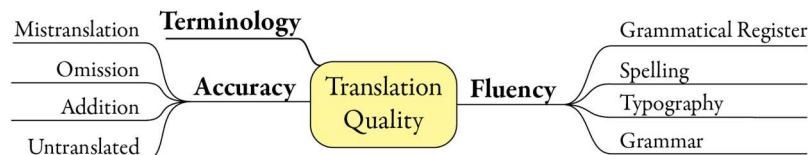
For analytic assessment, all issues have a weight of 1.0, and the overall threshold is 95% (according to the MQM scoring formula), as calculated by the MQM Scorecard tool (developed by DFKI, Brigham Young University, and LTAC Global). (These values may be revised as more data from actual assessments become available.)

Implement the Assessment Process within a Translation Workflow (Task 7)
Upon receipt from the translator, each text is evaluated holistically. If the translation passes the holistic evaluation, periodic sampled spot-checking is done with the analytic metric to confirm the results. If a full analytic evaluation is conducted, the outcome determines whether CRITI fixes problems in-house or sends the translation back for revision. Sampled spot-checking is used to assure that holistic assessors are consistent in meeting requirements. The workflow is designed to minimize the amount of analytic assessment and to simplify quality assessment while still providing positive assurance that CRITI translations meet customer requirements.