

Quality Evaluation using Adequacy and Fluency Approaches

TAUS Signature Editions

Table of Contents

Why are TAUS industry guidelines needed?	4
Adequacy/Fluency Best Practice Guidelines	5

Why are TAUS industry guidelines needed?

Adequacy and/or Fluency evaluations are regularly employed for assessing the quality of machine translation. However, they are also useful for evaluation of human and/or computer assisted translation in certain contexts. These methods are less costly and time consuming to implement than an error typology approach and can help to focus on assessing quality attributes that are most relevant for specific content types and purposes.

Providing guidelines for best practices will enable the industry to:

- Adopt standard approaches, ensuring a shared language and understanding between translation buyers, suppliers and evaluators
- Better track and compare performance across projects, languages and vendors
- Reduce the cost of quality assurance

Adequacy/Fluency Best Practice Guidelines

Establish clear definitions

Adequacy

"How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation" (Linguistic Data Consortium).

Fluency

To what extent the translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" (Linguistic Data Consortium).

Clearly define evaluation criteria and rating scales, using examples to ensure clarity.

Adequacy

On a 4-point scale rate how much of the meaning is represented in the translation:

- o Everything
- o Most
- o Little
- o None

Fluency

Rate on a 4-point scale the extent to which the translation is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker:

- o Flawless
- o Good
- o Dis-fluent
- o Incomprehensible

Data segments should be at least sentence length.

- The chosen evaluation data set must be representative of entire data set/content
- For MT output, at a minimum two hundred segments must be reviewed
- For MT output, the order in which the data/content order is presented should be randomized to eliminate bias

Human evaluator teams are best suited to provide feedback for Adequacy and/or Fluency.

- For periodic reviews there should be at least four evaluators per team
- The level of agreement between evaluators and confidence intervals must be measured
- Evaluators must rate the same data

To ensure consistent quality human evaluators must meet minimum requirements.

- Provide training materials, screening tests, and guidelines with examples
- Evaluators should be native or near native speakers, familiar with the domain of the data
- Evaluators should ideally be available to perform one evaluation pass without interruption

For adequacy evaluation evaluators will need to be able to understand the source language. This requirement can be overcome if evaluators are given a gold reference segment for each translated segment. Note that the quality of the gold reference should be validated in advance.

Determine when your evaluations are suited for benchmarking, by making sure results are repeatable.

- Define tests and test sets for each model and determine minimal requirements for inter-rater agreements
- Train and retain evaluator teams
- Establish scalable and repeatable processes by using tools and automated processes for data preparation, evaluation setup and analysis

Capture evaluation results automatically to enable comparisons across time, projects, vendors.

- Use color-coding for comparing performance over time, e.g. green for meeting or exceeding expectations, amber to signal a reduction in quality, red for problems that need addressing.

Implement a CAPA (Corrective Action Preventive Action) process.

- Best practice is for there to be a process in place to deal with quality issues - corrective action processes along with preventive action processes. Examples might include the provision of training or the improvement of terminology management processes.
- Adequacy/Fluency review may not identify root causes. If review scores highlight major issues, more detailed analysis may be required, for example using Error Typology review.

If evaluations follow these recommendations you will be able to achieve reliable and statistically significant results with measurable confidence scores.

Further resources:

For information on when to use adequacy and/or fluency approaches, conditions for success, step-by-step process guides, ready to use templates and guidance on training evaluators, please refer to the TAUS Academy.

Our thanks to:

Karin Berghoefer (Appen Butler Hill) for drafting these guidelines.

The following organizations for reviewing and refining the Guidelines at the TAUS Quality Evaluation Summit 15 March 2013, Dublin:

ABBYY Language Services, Capita Translation and Interpreting, Crestec, Intel, Jensen Localization, Jonckers Translation & Engineering s.r.o., Lingo24, Logrus International, Microsoft, Palex Languages & Software, Symantec, and Tekom.

Consultation and Publication

A public consultation was undertaken between 11 and 24 April 2013. The guidelines were published on 2 May 2013.

Feedback

To give feedback on improving the guidelines, please write to dqf@taus.net

About TAUS

TAUS is a think tank and resource center for the global translation industries. We develop communities through a program of events and online user groups and by sharing knowledge, metrics and data that help all stakeholders in the translation industry develop a better service.

The shared knowledge helps TAUS members decide on effective localization strategies. Shared metrics support more efficient processes and the normalization of quality evaluation. Shared data leads to improved translation automation.

TAUS develops APIs that give members access to services like DQF, the Quality Dashboard and the TAUS Data Cloud through their own translation platforms and tools. TAUS metrics and data are already built into most of the major translation technologies.