A Harmonized Approach to Translation Quality Assessment

Arle Lommel (DFKI),
Attila Görög (TAUS),
Hans Uszkoreit (DFKI)
A little history

- Translation quality assessment was subjective
- 1990s: move to vendor-specific checklist-based criteria for quality assessment based on error counts
- Next step: shared metrics
  - LISA QA Model
  - SAE J2450 (automotive industry)
  - Often heavily modified
- Attempt in ISO to implement a universal metric for all translation (cancelled in 2013)
Why no universal metric?

• Little agreement on precisely what constitutes a good translation at the detailed level
• Different requirements
• Different types of translation (MT vs. HT)
  – Incompatible methods (reference-based versus error count)
  – Different types of errors
Two efforts, two approaches (1)

- Multidimensional Quality Metrics (MQM)
  - EU project-based (QTLaunchPad)
  - Intended to unify approaches for (diagnostic) evaluation of MT with HT approaches
  - Designed as a non-strict superset of prominent metrics (LISA QA Model (3 models), SAE J2450, ISO 14080 (cancelled), SDL TMS Classic, ApSIC XBench, Okapi CheckMate, xliff:doc, Yamagata QA Distiller, ATA certification
  - 100+ issue types
  - Intended as a “master vocabulary” for describing task-specific metrics; heavily customizable
  - Very early version standardized in ITS 2.0
  - Issue types in a (top-down) hierarchy
Two efforts, two approaches (2)

- Dynamic Quality Framework (DQF) – Error typology
  - Bottom-up approach based on industry feedback
  - Designed to address common needs at relatively low granularity for simple use
  - Six error categories + four additional features
  - Implemented in online dashboard + vendor-specific tools
The Problem

- Despite differences, these two specifications had largely similar function.
- Confusion from the LSP and technology community about which one to use: why have two specifications?
- The two were not compatible: they used different structures.
- Both DFKI (lead developer of MQM) and TAUS were working in EU projects together: reviewers kept asking why we had two things and made it a requirement for new projects that they be merged.
MQM structure (2014)

• Hierarchy with five dimensions:
  – Accuracy
  – Fluency
  – Design
  – Verity
  – Internationalization (underdefined)

• Primarily focused on error-count metrics
MQM Structure (2014), cont…
DQF Structure (2014)
DQF Structure (2014), cont…

- Only highest-level nodes were counted in software:
  - (Translation-specific)
    - Language
    - Terminology
    - Accuracy
    - Style
  - (Localization-specific)
    - Country standards
    - Layout
  - Others (not issues)
    - Query implementation
    - Client edit
    - Repeat
    - Kudos
• Additional types functioned as examples, but the plan was to expand the software to allow them to be optional types in checking
• “Others” provide a way to mark specific items for attention that are not considered errors
Harmonization process

• Started in September 2014 (in anticipation of future projects) with work on harmonizing terminology
• Continued February–May 2015 with focus on methods and hierarchy
• Weekly phone conferences and exchange of ideas
• Substantial changes on both sides
On the MQM side

• Increased the number of “dimensions” by adding:
  – Terminology (moved terminology-related issues from other dimensions)
  – Locale-convention (corresponds to DQF *Country standards*)
  – Style (split off from *Fluency*)

• This shift made the MQM high-level branches match the six core DQF issue types

• Added additional issue types to cover DQF subtypes (e.g., “Improper exact TM match”, “Mistranslation of technical relationship”) – prepared MQM to become a superset of DQF
On the MQM side

- Expanded Locale-convention to cover all TAUS *Country standards* types
- Moved some issue types to more closely match DQF
- Additional changes
  - Added Internationalization subtypes (unrelated to harmonization, but happened at same time)
  - Added some issue types to support LQA initiative (in ASTM F43)
On the DQF side

- Adopted MQM naming conventions
- Moved some sub-issues that did not properly fit under parent issues or that did not match MQM hierarchy
- Dropped two sub-issues that were not properly translation issues (they had to do with functionality and had been listed under *Layout*)
- Kept “Others” as additional features in the software (will not be added to MQM)
MQM hierarchy (5)

Locale convention

- Postal code
  - postal-code
- Address format
  - address-format
- Calendar type
  - calendar-type
- Currency format
  - currency-format
- Date format
  - date-format
- Name format
  - name-format
- National language standard
  - national-language-standard
- Number format
  - number-format
- Shortcut keys
  - shortcut-keys
- Telephone format
  - telephone-format
- Time format
  - time-format

Locale-specific punctuation
- locale-specific-punctuation

Measurement format
- measurement-format

Quote mark type
- quote-mark-type
MQM hierarchy (6)

Awkward
awkward

Company style
company-style

Inconsistent style
inconsistent-style

Register*
register

Third-party style
third-party-style

Unidiomatic
unidiomatic

Variants/slang
variants-slang

Style
style
MQM hierarchy (8)

- Lists
  - incomplete-list

- Procedures
  - incomplete-procedure

- Completeness
  - completeness

- Culture-specific reference
  - culture-specific-reference

- End-user suitability
  - end-user-suitability

- Legal requirements
  - legal-requirements

- Locale-specific content
  - locale-specific-content
New MQM-Compliant DQF hierarchy
Three of the four (Client edit, Repeat, and Query implementation) can be implemented using the new MQM *none* severity level (0 penalty) with an explanatory note.

Remaining item—kudos—remains outside of the scope of MQM, but can be marked and counted as positives in scoring. *Question: should this feature be added to MQM?*
DQF Hierarchy

• Six main categories (same as before, but with updated names) – compatible with MQM Core (next slide)

• With subcategories:
  – 42 issues
  – Adds Verity dimension for one issue type
  – Extends to three levels in MQM hierarchy
  – Supports more detailed requirements
  – Simpler than full MQM and tailored to general industry requirements
  – Not contained in MQM Core, but reducible to Core
Present state

• DQF error typology is now a subset of MQM
• DQF can be implemented in MQM tools (translate5, Scorecard, XTM, etc.)
• Mapping for DQF to ITS 2.0 (only problem is that Fluency maps to Other in ITS 2.0, since it has no general category corresponding to Fluency)
• Only remaining DQF feature that cannot be represented in MQM is kudos.
Future plans

• Harmonization to be maintained by mutual agreement

• MQM to move to standards body with DQF as an officially recognized profile
  – This will allow joint, open maintenance of MQM.
  – Will certainly involve further changes.

• Joint metric to be used in QT21 analysis, with comparison to additional DQF assessment methods
Advantages for the end user

• You no longer have to choose between MQM and DQF
• Users of DQF gain an industry-derived metric suitable for use in the localization industry while still being able to use MQM tools
• MQM can be used for scenarios that require other issues than those in DQF, but the differences are immediately analyzable due to the shared vocabulary
• Somewhat analogous to TBX vs. TBX-Basic
More information

- Latest version of MQM is maintained at http://qt21.eu/mqm-definition
- Tables on mapping to SAE J2450 and ITS 2.0 can be found at this URL
- This document is currently open for a round of public review (through the end of June 2015) and we welcome feedback (send to arle.lommel@dfki.de and attila@taus.net)