Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality

Aljoscha Burchardt, Arle Lommel & Hans Uszkoreit (DFKI)

Funded by the 7th Framework Programme of the European Commission through the contract 296347.
PROBLEMS IN ASSESSING QUALITY
95% of professionally translated content at one major LSP is never evaluated for quality.
“Translators are the garbage collectors of the documentation world” – Alison Toon, HP
“I know it when I see it”
Machine Translation quality scores (BLEU, NIST, etc.) are totally different from human evaluation.
MT methods require **reference translations**: Cannot be used for production purposes
Change the reference translation(s) and the score changes
The problem with BLEU

No substantial improvement for human use
The problem with BLEU

**Substantial human improvement but no BLEU improvement**
Human quality assessment takes too much time and is not principled.
Most sampling today is random but errors are not random.
Wait a minute...

What do you mean by quality?
Quality: A New Definition

A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs.

Source: Alan Melby
Sounds simple, right?
It’s actually quite radical and it drags translation kicking and screaming into the modern world of quality management.
Multidimensional Quality Metrics
Why not use a single metric everyone shares?
OK.

Which one?
LISA QA Model
SAE J2450
SDL TMS
Acrocheck
ApSIC XBench
CheckMate
QA Distiller
XLIFF:Doc
EN15038...
All of them disagree* about what is important to quality

*The only thing they agree on is terminology
(Probably because there is no single set of criteria that applies to all kinds of translation)
There is no one-size-fits-all metric
MQM provides a catalog of issue types suitable for various tasks
The “full” MQM (“Spaghetti”)
Zooming into the "full" MQM
Wait! Weren’t we trying to improve things?
The MQM Core

Issue Types

Accuracy
- Terminology*
- Mistranslation
- Omission*
- Addition*
- Untranslated*
- Completeness
- Legal requirements
- Locale applicability

Verity

Fluency
- Register*
- (Content)
- Style*
- Inconsistency
- (Mechanical)
- Spelling*
- Typography*
- Grammar*
- Locale violation*
- Unintelligible

The MQM Core
Issues marked with a star (*) can be automatically detected.
Accuracy and Fluency

What’s Verity?
Verity provides a way to deal with the text in relation to the real world.
Locate the ground wire (bare copper)
Finden Sie das Erdungskabel *(blankes Kupfer)*?
You don’t use all of MQM (or its core): you use the parts you need.
MQM for MT Diagnostics
**SAE J2450**

- **Terminology** (=Wrong term)
- **Omission** (=Omission)
- **Other** (=Miscellaneous error)
- **Spelling** (=Misspelling)
- **Typography** (=Punctuation Error)
- **Grammar** (=Syntactic Error)
- **Morphology (word form)** (=Word structure or agreement error)
MQM lets you declare your quality metric in a shared vocabulary.
Dimensions help you decide what to check (and also help you communicate with your LSP)
No more assuming what the parties want or how to check it
12 Dimensions
(from ISO/TS-11669)

1. Language/locale
2. Subject field/domain
3. Terminology (source/target)
4. Text type
5. Audience
6. Purpose
7. Register
8. Target text style
9. Content correspondence
10. Output modality
11. File format
12. Production technology
Open-source tools* to demonstrate MQM

*translate5 source code is published. Other tools’ code will be published in 2014
An Online Tool for Building Dimensions and Metrics

QTLaunchPad Infrastructure: Edit Dimensions

Step 2: Define Dimensions

Section 1: Language/locale (source) and (target)
The language or locale (combination of language with geographical location) of the source and target texts. E.g. the text was written in Egyptian Arabic.

Language

<table>
<thead>
<tr>
<th>Source:</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region:</td>
<td>RUSSIAN FEDERATION</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target:</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region:</td>
<td>UNITED STATES</td>
</tr>
</tbody>
</table>

*Please verify the correct source and target languages

Does the target language pose any particular grammatical or stylistic difficulties of which you are aware?

- [ ] no
- [ ] yes
- [ ] unknown

Does the target language use a different writing system than the source?

- [ ] no
- [ ] yes
- [ ] unknown

Was the author a native writer of the source language?

- [ ] no
- [ ] yes
- [ ] unknown

Was the translator a native speaker of the target language?

- [ ] no
- [ ] yes
- [ ] unknown

Is the source text already a translation of another text (e.g., a pivot language translation)?

- [ ] no
- [ ] yes
- [ ] unknown

Notes:

[ ]

Section 2: Subject field/domain

The subject field(s) to which the text pertains (e.g., legal, pharmaceutical, heavy equipment maintenance)
Section 2: Subject field/domain

The subject field(s) to which the text pertains (e.g., legal, pharmaceutical, heavy equipment maintenance)

Subject Fields:
Is the subject field in a regulated industry where legal compliance is mandated?  
  no  yes  unknown
Is the subject field a technical field where particular terminology is expected?  
  no  yes  unknown

Notes:

Section 3: Terminology (source/target)

Any special terminology that needs to be noted in the source or target texts

Terminology (list or reference):
  volga.tbx

Notes:

Section 4: Text type

The type/genre of the content (e.g., novel, technical user manual, marketing)

Text type:
  service manual

Is the text type likely to require special attention to style or accuracy?  
  no  yes  unknown
Will the text type have any other implications for other aspects of the translation?  
  no  yes  unknown

Notes:
Tabular Scorecard

<table>
<thead>
<tr>
<th>Issue Type</th>
<th>Weight (default=1)</th>
<th>Minor</th>
<th>Major</th>
<th>Critical</th>
<th>Raw</th>
<th>Adj</th>
<th>Target Subscore</th>
<th>Severity Multipliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>1.0</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>1</td>
<td>1</td>
<td>99.7%</td>
<td></td>
</tr>
<tr>
<td>Terminology</td>
<td>1.0</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>1</td>
<td>1</td>
<td>99.7%</td>
<td></td>
</tr>
<tr>
<td>Mistranslation</td>
<td>1.0</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>1</td>
<td>1</td>
<td>99.7%</td>
<td></td>
</tr>
<tr>
<td>Omission</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Untranslated</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Accuracy Subtotal</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td>99.3%</td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unintelligible</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>99.5%</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconsistency</td>
<td>1.0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>14</td>
<td>95.3%</td>
<td></td>
</tr>
<tr>
<td>Register</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Mechanical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spelling</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Grammar</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Fluency Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95.3%</td>
<td></td>
</tr>
<tr>
<td>Verily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal requirements</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Verity Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>16</td>
<td>94.7%</td>
<td></td>
<td>2</td>
<td></td>
<td>99.9%</td>
<td></td>
</tr>
</tbody>
</table>

http://www.qt21.eu/MQM
Ergonomic Scorecard

http://www.qt21.eu/MQM
Currently clarifying the (complementary) relationship between MQM and the TAUS DQF Error Typology
Questions?