AMTA 2018
March 17 - 21, 2018
Boston, MA, USA

The 13th Conference of
The Association for Machine Translation
in the Americas

www.conference.amtaweb.org

TUTORIAL
March 17, 2018

MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)

Presenters: Arle Lommel *(CSA Research)*, Alan K. Melby *(LTAC Global)*

Tutorial | MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)

Presenters:

Arle Lommel (Common Sense Advisory, Inc.)

Alan K. Melby (LTAC Global)

Agenda

1. Introductions
2. Some basic terminology
3. Typology of translation quality metrics
4. Overview of MQM-DQF & Key Features
5. Market adoption
6. Detailed case studies
7. Validity and reliability

Some basic terminology

Quality Management				
Quality Planning Design Designing Systems	Quality Assurance Auditing Auditing Procedures	Quality Control Real-Time Monitoring Monitoring Processes	Quality Evaluation Post-Production Appraisal Evaluating Products	Quality Improvement Prevention Preventing Variation

Quality Management	The integration and coordination of management activities focused on ensuring the organization fulfils stakeholder requirements predictably, consistently, and reliably. *Note: Quality Management comprises quality planning, quality assurance, quality control, quality appraisal, and quality improvement.* *Note 2: Development of stakeholder requirements for particular translation projects is defined in ASTM F2575-14, Section 8 (Specifications)*
Quality Planning	Quality management activities for designing a system of policies, processes, and procedures to be followed capable of producing products that will meet stakeholder requirements.
Quality Assurance	Quality management activities of auditing processes and procedures to provide confidence to management, customers, and third parties that stakeholder requirements can be fulfilled. *Note: Quality assurance is often used as a synonym for quality appraisal in industry, but this conflation creates a strong source of confusion, and shall not be used in this fashion.*
Quality Control	Quality management activities for monitoring and assessing process and performance in real time in order to verify that stakeholder requirements are being fulfilled and that quality measures are being maintained within proscribed limits. *Note: In quality control, data collected in real time is analyzed and used during production (vs. being stored only for future quality assurance audits).*
Quality Evaluation	Quality management activities for validating that stakeholder requirements have been fulfilled through inspection, examination, and testing. *Note 1: Quality evaluation comes after production and prior to delivery to the consumer or requester.* *Note 2: This activity is sometimes called "quality appraisal" in industry segments outside of translation.*
Quality Improvement	Quality management activities for preventing variation from stakeholder requirements in the product by eliminating sources of variation in the process. *Note 1: Continuous improvement of the process will have benefits across products and over time.* *Note 2: Sources of variation in the process include improperly designed policies, poor resources, or inconsistent application of procedures.*

The Problem: What Is Quality?

- No industry agreement about what constitutes quality ("I know it when I see it")
- How can we achieve what we can't define?
- European approach (ISO 17100) is *process*-oriented: can't tell you for sure whether the *product* is good
- Most translated content is accepted based on trust (95% of text from one major LSP is never checked)
- Many different systems/standards claim to solve the problem, but they disagree about what to measure and how

What is translation quality?

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

Measure vs. metric

- A measure determines some property of an item:
  - This table provides 74 cm clearance
  - This house is 200m^2
- A metric is a measurement with a purpose:
  - We are measuring tables to determine which ones will allow a wheel-chair to slide beneath them
  - We are determining whether the house is big enough for a family with six children
- Thresholds are the criteria we use to determine whether something measured with a metric meets requirements:
  - The table must have at least 77 cm of clearance and no more than 79.
  - The house requires 25 m^2 per family member

A typology of translation quality metrics

Typology of translation quality metrics

- Holistic vs. analytic
- Fine-grained vs. coarse
- Reference-based vs. reference-free
- Objective vs. subjective

Holistic metrics

- Look at the entire text to provide a single result
- E.g.,
  - This translation (as a whole) has a ____ score of 96.5.
  - 76% of users rated the translation as “useful”

Analytic metrics

- Measure multiple qualities and allow *decomposition* of any single score
- E.g.,
  - This translation has a score of 96.3, with 100 for accuracy, 98 for fluency, and 92 for style (a *composite* metric)
  - 76% of users rated the translated text as good using a three-section rating scale that covers readability, technical accuracy, and ease of use.

Fine-grained vs. coarse

- Varying degrees of analytic metrics: Some identify individual issues and allow decomposition down to individual errors
- A coarse metric: Accuracy and Fluency
- A fine-grained metric:
  - Accuracy
    - Addition
    - Mistranslation
    - Omission
  - Fluency
    - Grammar
    - Spelling
    - Typography…

Which is better?

Reference-based vs. reference-free

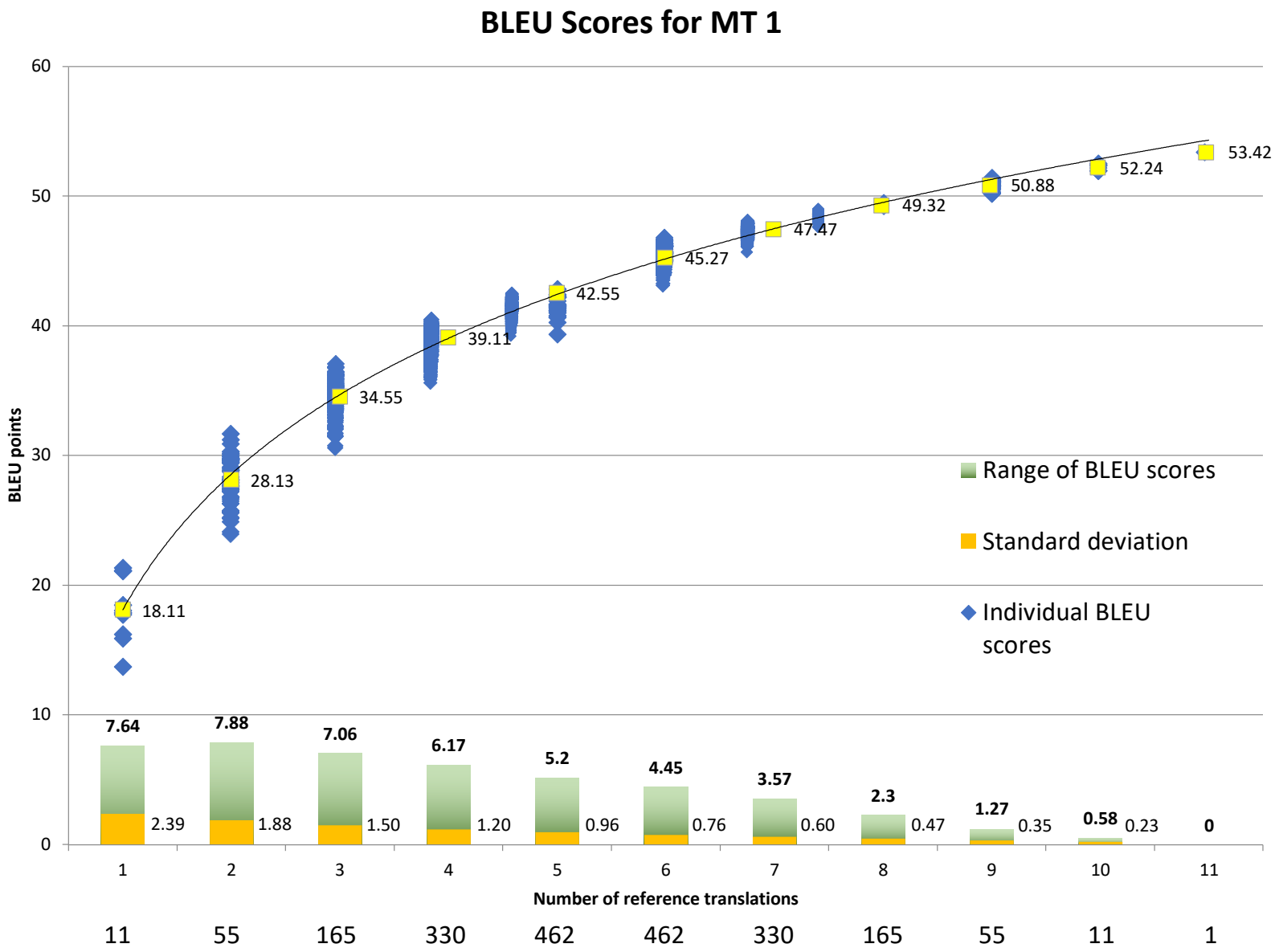
- Reference-based: Comparison against a “gold standard”
- Reference-free: No point of comparison

Objective vs. subjective

- Objective: Metric is based on observable facts that are – in principle – not dependent upon the individual applying the metric
- Subjective: Metric is based on the reaction of the individual and depend on taste or other non-objective factors
- Which of these is possible with translation?

Exercise: Categorize various metrics

- LISA QA Model
- BLEU
- Customer feedback survey
- Post-editing distance
- Adequacy and fluency rating
- Output ranking
- Compliance with terminology list


BLEU Scores for MT 1
BLEU points
60
50
40
30
20
10
0
18.11
28.13
34.55
39.11
42.55
45.27
47.47
49.32
50.88
52.24
53.42
Range of BLEU scores
Standard deviation
Individual BLEU scores
7.64
2.39
7.88
1.88
7.06
1.50
6.17
1.20
5.2
0.96
4.45
0.76
3.57
0.60
2.3
0.47
1.27
0.35
0.58
0.23
0
1
2
3
4
5
6
7
8
9
10
11
Number of reference translations
11
55
165
330
462
462
330
165
55
11
1

Overview of MQM/DQF & Key Features

MQM-DQF

- The intersection of MQM and DQF (just one part of each)
- Focuses on *product* quality
- Analytic (error typology) focus:
  - Identify the *nature* of problems with a goal of *preventing* or *correcting* them
  - Relate problems to a list of known issue type
- Divides issues into high-level issues
- MQM defines a superset of issues checked in industry and provides a way for tools to declare what they check and compare it with other tools

Dimensions

Terminology Issues

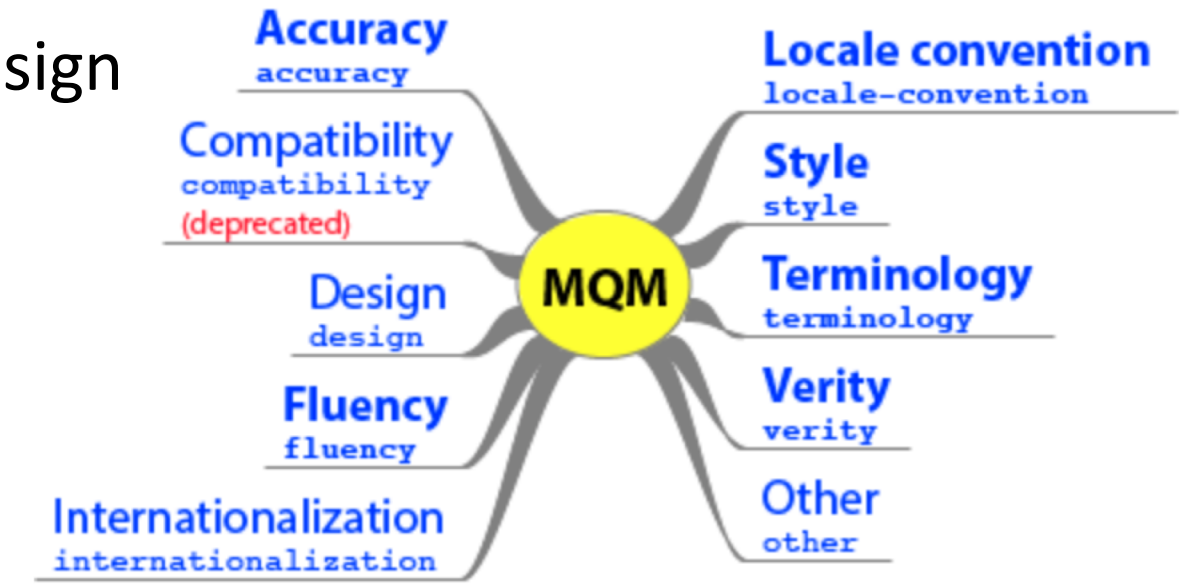
- Terminology Issues (bilingual and monolingual)

Bilingual

- Accuracy
- Locale convention
- Internationalization
- Verity/Locale suitability

Monolingual

- Fluency
- Style
- Design



Terminology

- Bilingual:
  - Not using a specified termbase
  - Not using established terminology for a domain
- Monolingual:
  - Inconsistent use of terms within a document for the same concept

Accuracy

- Does the target text convey the same information that the source text does?
- Can be determined only by comparison to the source text.
- Not identical to *adequacy*

Locale convention

- Mechanical aspects of localization such as representation of dates and times.
- Note 1: Locale-convention issues are often identified by software in the category QA Checkers.
- Note 2: A few other mechanical aspects of localization (Locale-convention issues) involve conversion of units of measure such as meters vs yards or degrees Fahrenheit vs. Celsius or euros vs. Canadian dollars: based on specified source-locale vs. target-locale)

Internationalization

- Issues related to whether or not the source content has been created to facilitate subsequent localization
- Note: An internationalization error is not a translation error as such, but lack of proper internationalization is typically manifested in a translation error or failure for software to function as expected.
- For example, if insufficient space is allowed for a string (such as a menu item or a message), the translation of that string might be truncated if it is longer than the source string.

Verity/Locale suitability

- Aspects of localization other than locale-convention, that is, those requiring human detection and judgement, namely, target-text suitability issues because of differences between source and target locale).
- Note: This dimension includes adjustments for differences in culture, usages, laws, or even physical aspects of the geographical region such as the shape of electrical plugs. It is sometimes called Verity (in the sense of “accordance with fact”) because it concerns whether something matches the facts of the target locale.

Fluency

- Is the text linguistically well-formed?
- Can be assessed without consulting a source text
- Includes items often called “language errors”

Style

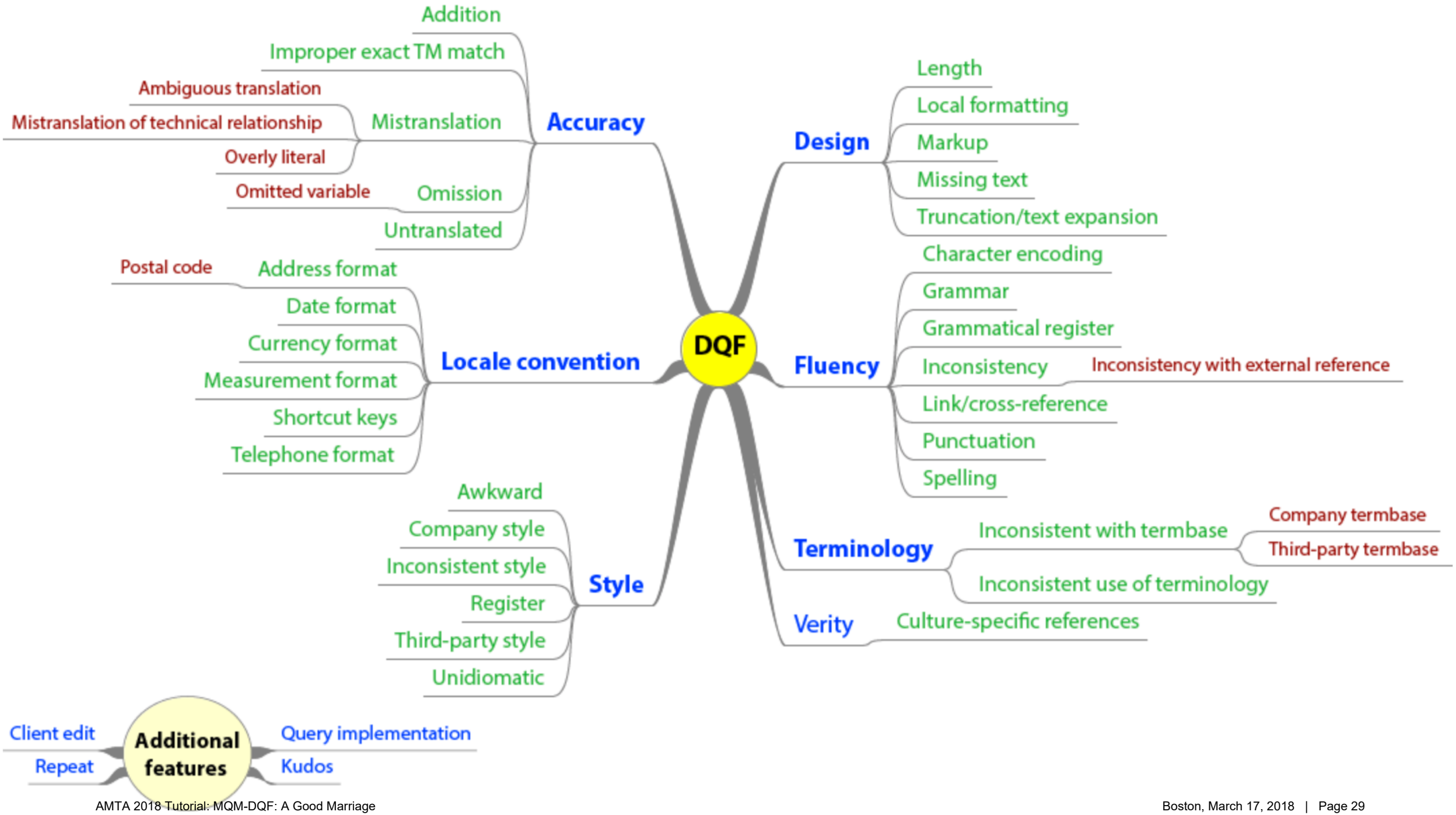
- Monolingual style manual
- Use of other specified target-language resources such as relevant reference documents in target language, and other style issues such as those regarding register, collocations, and structural awkwardness.
- Style issues should be as clearly specified as possible to avoid subjectivity and hyper revision

Design

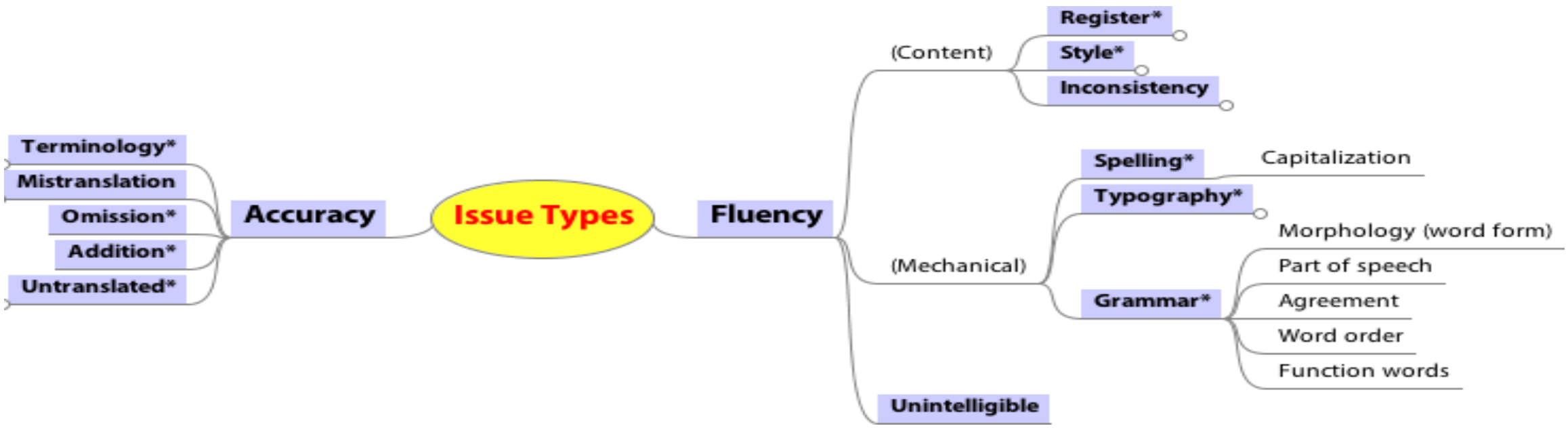
- Monolingual layout, formatting, and markup issues, not explicitly covered in a specified style manual
- Appearance of text (i.e., the accuracy and fluency are OK, but the text looks wrong)

Quality = spaghetti?

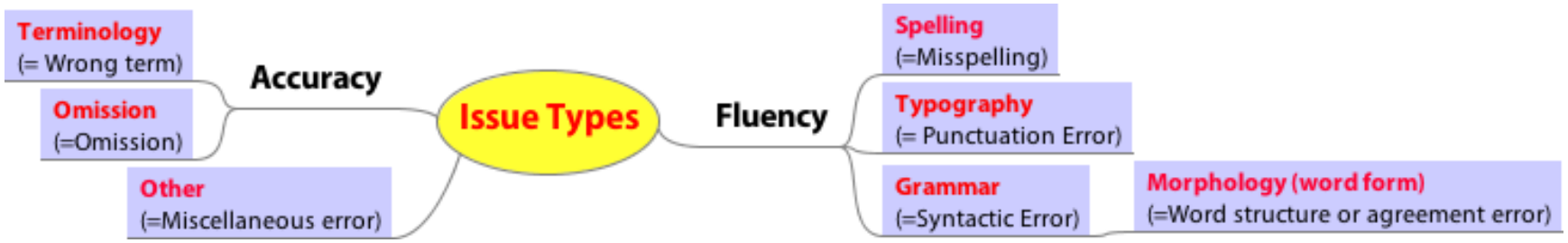
- Too many issues, but…
- Somebody checks everything we have in the master MQM set
- Based on an analysis of existing quality specifications (public and private)
- Can be overwhelming
- So… Use the DQF subset of MQM


DQF
Accuracy
Addition
Improper exact TM match
Mistranslation
Ambiguous translation
Mistranslation of technical relationship
Overly literal
Omission
Omitted variable
Untranslated
Locale convention
Address format
Postal code
Date format
Currency format
Measurement format
Shortcut keys
Telephone format
Style
Awkward
Company style
Inconsistent style
Register
Third-party style
Unidiomatic
Design
Length
Local formatting
Markup
Missing text
Truncation/text expansion
Fluency
Character encoding
Grammar
Grammatical register
Inconsistency
Inconsistency with external reference
Link/cross-reference
Punctuation
Spelling
Terminology
Inconsistent with termbase
Company termbase
Third-party termbase
Inconsistent use of terminology
Verity
Culture-specific references
Additional features
Client edit
Repeat
Query implementation
Kudos

Another subset: For MT analysis



Representing other standards: SAE J2450



Note that SAE J2450 does not consider accuracy in much detail.

Market adoption

DQF-MQM

- Has emerged as de facto market standard
- Entered formal standards process in ASTM

Adoption of MQM-DQF

- **LSPs and Enterprises**
  - Dell-EMC
  - eBay
  - LDS Church
  - Lionbridge
  - Microsoft
  - Moravia
  - Mozilla
  - Seprotec
  - Synergium
  - Tableau
  - Welocalize
- **Tools**
  - ContentQuo
  - MemSource
  - SDL (plug-in)
  - XTM
- **European projects**
  - QTLaunchpad
  - QT21
- Academia
  - Various projects
- **In process**
  - Argos Translations
  - Booking.com
  - CA Technologies
  - Capita
  - Crestec
  - Daimler
  - Intuit
  - John Deere
  - Nike
  - TNT-Fedex

Detailed Case Studies

Caribbean NGO

Decision tree for severity

In your assessment, you will be asked to flag errors in the translation. When you encounter an issue in the text, you have to decide whether it is an error or not (Question 1) and, if it is an error, how severe it is. If an issue would not be considered an error per the specifications, it should not be marked as such in the text. In questionable cases, please make a note of the issue.

MQM has three severity levels. Select the level based on the following criteria. Note that severity is determined with respect to the translation specifications. For example, a stylistic issue that would be critical for a high-visibility marketing piece might be considered a minor error (or even not an error at all) for an internal service manual.

1. Is the issue a violation of the translation specifications or of general professional translator practice that would be expected for a project of the type in question?

- **YES**: *Go to Question 2.*
- **NO**: The issue is not considered an error.[Add note in Appendix]

2. For the issue, do any of the following?

(a) render the product unusable,
(b) expose the user to potential physical or legal harm,
(c) expose the content creator to potential legal liability,
(d) potentially harm the content creator's brand
(e) would directly result in the intended user needing to contact technical support
(f) otherwise render the project unfit for purpose?

- **YES**: The issue is **CRITICAL (100 point penalty)**.

NOTE: Any critical issues MUST be repaired prior to acceptance of the translation. The presence of a single critical error renders the project unfit for purpose. If you feel that an error is CRITICAL and it does not fit the above criteria, please provide an explanation. Otherwise, any issue that does not meet the above criteria is either MAJOR or MINOR.

Examples:

- A translation of a contract misstates an amount that must be paid, and would result in the requester owing money that should not be owed.
- A legal contract omits a *not* in a list of obligations, thereby subjecting the requester to a legal obligation that was not intended.
- A French translation of a report on an NGO's activities systematically states that the NGO made an investment in *La Guyane* (French Guiana) instead of *Le Guyana* (British Guiana), leading to confusion about where it was active.
- A translation of technical instructions in a standard is incorrect and as a result the standard they are in cannot be used as intended.
- A translation omits negation at a crucial step in a process and therefore instructs the user to carry out a step that can damage a product or result in injury or death to the user.
- A translation contains a phrase that could be considered obscene and that conflicts with the brand image of the content creator.
- A translation uses terminology (including names of products) from a competitor of the content creator instead of ones from the content creator, thereby causing harm to the content creator's brand.

- **NO**: *See question 3.*

3. Is the issue one that prevents the intended user from correctly understanding the intended meaning of the text but does not render the text unfit for purpose?

- **YES**: The issue is **MAJOR (10 point penalty)**.

NOTE: Major errors must be repaired prior to acceptance but do not, individually, render the text unfit for purpose.

Examples:

- A legal contract provides an incorrect local phone number for one party. It would not result in invalidity, but cannot be easily corrected by the reader.
- An NGO's report mistranslates one item in a description of materials provided to refugees. It does not invalidate the work or threaten harm to anyone, but readers will not easily know that it is incorrect.
- A list of authors for a technical standard leaves the English title *Mr.* untranslated instead of rendering it as *Dhr.*, thereby indicating the person is an attorney (*Mr.* is used in Dutch for attorneys). Although the wrong information is conveyed, the mistranslation not impact the standard's usability, so this counts as a major error. (For another text, such as a contract, where the title might be seen as making a false claim about qualifications, this same issue could be critical.)

- **NO**: *See question 4.*

4. Is the issue primarily cosmetic in nature or one that is easily corrected by the intended user (perhaps without them evening noting its presence) without any loss of information?

- **YES**: The issue is **MINOR (1 point penalty)**.

NOTE: Minor errors should be noted, but if they are present in small numbers would not result in rejection of a translation; if there are sufficient numbers of minor errors to cause the translation to miss thresholds, then they must be addressed sufficiently to bring it to thresholds.

Examples:

- A word is misspelled in running prose in a NGO's annual report but the meaning is clear and unambiguous. (Note that a misspelling in a headline or title, however, might be MAJOR or CRITICAL, depending on how it would impact perception of the content creator)
- A translation of a contract makes a common mistake and uses the wrong form of a relative pronoun, but does not change the meaning.
- A translation of a treaty omits a period at the end of a sentence.
- A translation from English into Dutch uses an Anglicism that is not grammatically or stylistically ideal, but which is nevertheless fully understandable.

- **NO**: If none of the questions apply, the issue should be considered a preferential change rather than an error. It may be noted, but shall not impact acceptance or use of the translation.

Note: If in doubt about any issues, their severity, or whether they apply to a given translation, please make a note of the issue and provide an explanation.

1a. Do the specifications adequately represent the requirements for the translation?

- **YES**: Do not count the issue as an error. It does not violate specifications. It may be considered a preferential change for the future, but should not be counted against the translator.

Example:

- If the specifications for a *technical standard* state that Style is not important, these specifications would match general industry practice for technical standards. Therefore a passage that is fully understandable but stylistically awkward should *not* be counted as an error.

- **NO**: The specifications must be revised. Any issues that do not violate the specifications as provided to the translator, but which would violate adequate specifications, should be noted and addressed per requirements, but should not be counted against the translator. *Go to Question 2.*

Example:

- If the specifications for an *NGO's annual report* state that Style is not important, these specifications are likely to be inadequate because the projection of corporate image is crucial in such documents. Therefore the specifications should be revised and violations of the revised specifications noted. However, the translator should not be held responsible for Style problems in this instance because he/she was told in advance that Style is not important.

2.1. Holistic metric for standards texts

The following questions should be answered by the reviewer after reading the text (or a sample thereof for longer texts).

1. How well does the translation meet specifications with regard to the following aspects?

	A: Perfectly, with no problems	B: There were MINOR problems that *did not impact usability*.	C: There were MAJOR problems that *impact correct understanding* of the text, but leave it usable.	D: Problems were so serious that the translation is *not fit for purpose.*
Following TERMINOLOGY guidelines	❑	❑	❑	❑
ACCURACY of the translation	❑	❑	❑	❑
Linguistic FLUENCY of the target text	❑	❑	❑	❑
Appropriate REGISTER	❑	❑	❑	❑

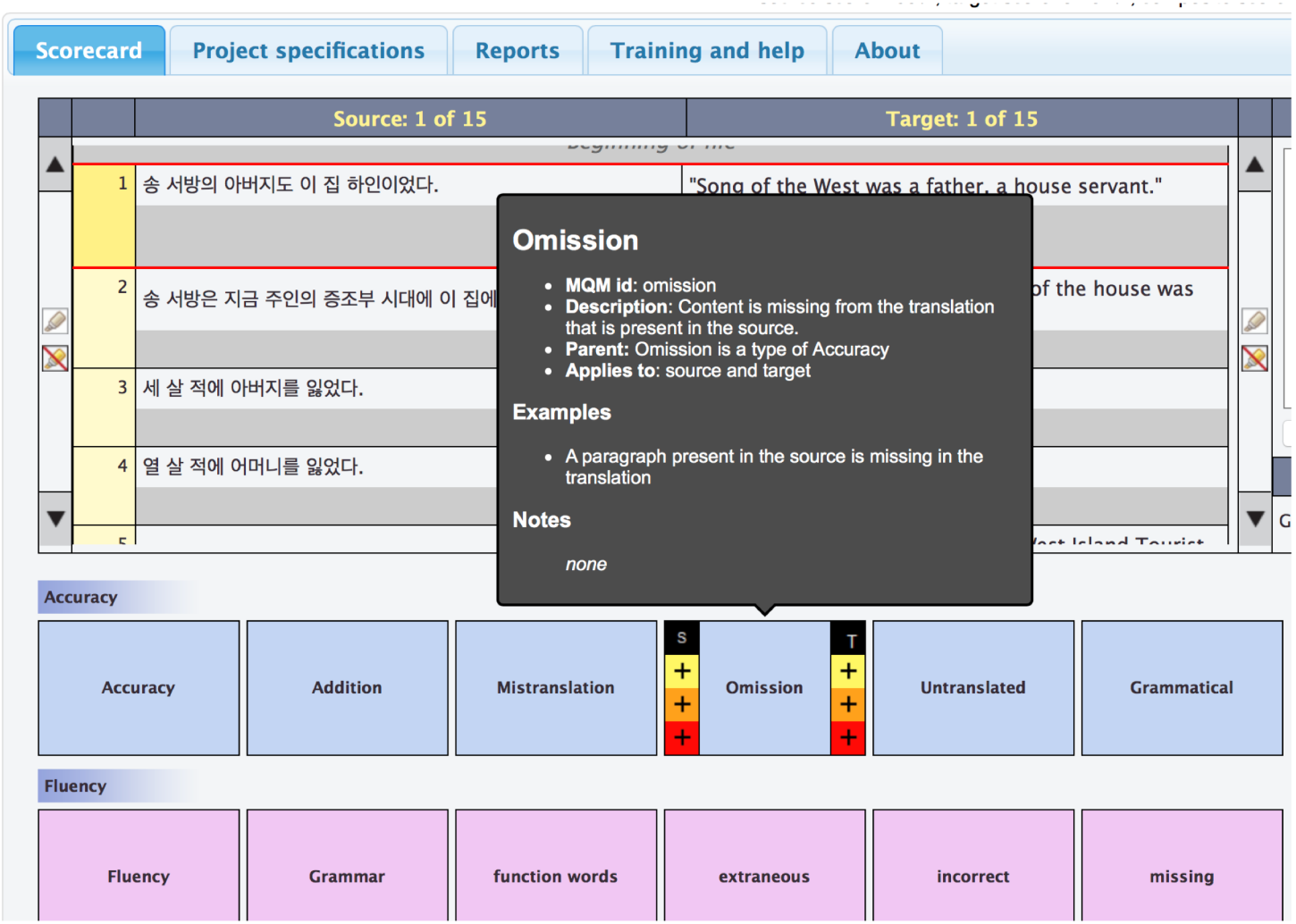
Considering the lowest mark (A is highest, D is lowest), the following course of action is recommended:

A. If all marks are an A, the translation is accepted as is.
B. If all marks are B or higher, the translation should be accepted as is with issues revised in house if time allows. (Minor errors, by definition, do not impact understanding or usability, and may be tolerated in small numbers, depending on specifications.)
C. If all marks are C of higher, the translation cannot be accepted as is and MUST be revised. At the discretion of [redacted] it may be sent back to the translator or revised in house. **If there is more than one major error per thousand words noted in the holistic review, the translation should be reviewed using the analytic metric to determine whether it meets acceptable levels and to ensure that all issues are identified.**
D. If any marks are at the D level the translation cannot be accepted and MUST be revised. It does not meet specifications. The translation may be sent back to the translator or may be sent to another translator so that problematic portions can be *retranslated*. If a translator consistently returns translations that receive a D mark in any aspect, he/she should be removed from [redacted]'s list of qualified translators.

2. Are there any other aspects of the translation that do not meet requirements?

- If not, then the decision from Question 1 applies.
- If there are other problematic aspects, [redacted] needs to decide how to resolve them. If such problems occur consistently and cannot be accounted for in the holistic metric, they should be evaluated for addition to the holistic and analytic metrics.

Scorecard tool allows tagging by issue and severity



Analytic drill-down

Issue	Source				Target				Total
	Minor	Major	Critical	Subtotal	Minor	Major	Critical	Subtotal	
Accuracy									
Accuracy	-	-	-	-	0	0	0	0	0
Addition	-	-	-	-	0	0	0	0	0
Mistranslation	-	-	-	-	0	1	0	1	1
Omission	-	-	-	-	0	0	0	0	0
Untranslated	-	-	-	-	0	0	0	0	0
Subtotal	-	-	-	-	0	1	0	1	1
Fluency									
Fluency	0	0	0	0	0	0	0	0	0
Grammar	0	0	0	0	0	0	0	0	0
Incosistency	0	0	0	0	1	0	0	1	1
Spelling	0	0	0	0	0	0	0	0	0
Typography	0	0	0	0	0	0	0	0	0
Punctuation	0	0	0	0	0	0	0	0	0
Subtotal	0	0	0	0	1	0	0	1	1
Style									
Style	0	0	0	0	0	0	0	0	0
Company style	0	0	0	0	0	0	0	0	0
Subtotal	0	0	0	0	0	0	0	0	0
Terminology									
Terminology	0	0	0	0	0	0	0	0	0
Subtotal	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	1	1	0	2	2

Validity & Reliability

Validity

- Does the metric measure what it is supposed to?
- Are the qualities appropriate to the goal?
- Does the metric determine whether specifications have been met?
- Examples:
  - Using Style to evaluate a support manual
  - Using Accuracy and Verity to evaluate a support manual

Reliability

- Can the metric consistently – across time and across assessors – deliver the same results?
- Tolerance
- Inter-annotator agreement
- Can multiple evaluators agree upon the same result?