A New Model of Translation Quality Assessment

Hans Uszkoreit

joint work with Arle Lommel and Aljoscha Burchardt at DFKI,
Lucia Specia and her group at U Sheffield,
Josef van Genabith and his group at CNGL,
Serge Gladkoff and others at GALA,
Alan Melby and others at FIT

DFKI and Saarland University

Funded by the 7th Framework Programme of the European Commission through the contract 296347.

Outline

- The project QT Launch Pad
- The reasons for a metric shared by HT and MT
- The notion “translation quality”
- The basic concepts and components of the metric
- Conclusion
- Next steps
Consortium

- DFKI – Hans Uszkoreit (Coord.)
  Aljoscha Burchardt Project Manager, Stephan Busemann Administr. Coord.
- CNGL DCU – Josef van Genabith
- U. Sheffield – Lucia Specia
- ILSP Athena – Stelios Piperidis

- Subcontractor: GALA
  Hans Fenstermacher, CEO and former President
  Laura Brandon, GALA Managing Director
  Serge Gladkoff, GALA Standards Director

Main Objectives

- The support action will prepare the grounds for a new type of collaborative MT research dedicated to overcoming existing quality barriers.

- To this end, QTLaunchPad will...
  - assemble and provide needed data and tools including specialised translation corpora, test suites and tools for quality assessment,
  - create a shared quality metrics for human and machine translation, improve automatic translation quality estimation,
  - extend an existing platform for resource-sharing to the needs of quality MT research,
  - define strategies and challenges and then plan and launch a large-scale research and innovation action for a breakthrough in quality translation technology.
is the working name of the grand endeavor, standing for

Quality Translation Technology for the 21st Century

It should serve as a trail blazer for a European push toward overcoming language barriers in HORIZON 2020 and maybe also CEF.

One scenario: A lead project funded by FP 7 and Satellites funded by national programmes, maybe also by regions or industry
Research Approaches

- highly specialized quality services,
- an analytical systematic fight against quality obstacles,
- improved utilization of shared data and tools,
- the integration of translation providers in research, deployment, testing and business models,
- and a concerted competitive model of technology evolution.

A New Approach

We view the planned new mode and approach of research as a paradigm shift. Future emerging research paradigms cannot be safely predicted but the envisaged mode of research should increase the chances for the development of completely new methods.
We view the planned new mode and approach of research as a paradigm shift. Future emerging research paradigms cannot be safely predicted but the envisaged mode of research should increase the chances for the development of completely new methods.
For the planning process, QTLaunchPad has assembled:

- the best European centres of MT research,
- the most sophisticated and enthusiastic large-scale users of translation technology including enlightened potential users,
- technophile and quality-conscious translation service providers,

soon to be extended by:

- providers of language tools and resources,
- and experienced technology integrators.

In addition to the consortium:

- Jan Hajic of U. Prague,
- Stephan Oepen of U. Oslo,
- Philipp Koehn of U. Edinburgh,
- Alex Waibel of Karlsruhe KIT,
- Marcello Federico of FBK Trento,
- Mikel Forcada of U. Alicante,
- Hermann Ney and Volker Steinbiss of RWTH Aachen,
- Nuria Bel of U. Barcelona,
- Joseph Mariani of LIMSI Paris,
- Johann Roturier of Symantec Ireland,
- Spyridon Pilos of EC DGT Luxembourg,
- Serge Gladkoff, Kim Harris and Hans Fenstermacher of GALA.
- Andrejs Vasiljevs, Tilde
Problem: No Shared Quality Metric

- Quality measured by BLEU, NIST, METEOR etc. does not indicate the type of quality problems.
- These metrics are also better suited for measuring progress in the “ugly” and “bad” sectors of the quality spectrum.
- Even the human evaluations usually by ranking, often done by CS researchers and students, do not help the human translators.
- The LISA QA Model, EN-15038 and current ISO work on a successor, are not known and not used in MT research, neither ISO/TS-11669 the 20 criteria for translation projects.

Keeping Separate Things Separate

Current QA methods confuse distinct things:

Product Quality

- Fluency: *How well does the text read, independent of how it was produced?*
- Accuracy (or adequacy): *Does the target say the same thing as the source?*
- End-user or purpose adequacy: Does the text (source or target) meet its communicative purposes and does it reflect the real-world conditions around it?

Process: *How is the translation produced? effectiveness, QA*

Project: *How satisfactory is the project overall? processes, times, costs, results*
Cooperations

- Cooperation with the Standards Committee of GALA
- Cooperation with ISO project 14080 (Translation Quality)
- Cooperation with the Standards Committee of FIT (Fédération Internationale des Traducteurs / International Federation of Translators)

- Hopefully in the future also cooperation with TAUS

What Is Translation Quality?

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.

– Alan Melby
On quality

◆ Quality is relative.
  ▪ It is an ideal target that depends on several given factors such as purpose, audience and budget. The target can be explicitly specified before translation. A useful quality metric must be suited for such quality specification.

◆ Quality is multidimensional, composed of several components that each may be measured by several criteria.
  ▪ There are three types of criteria: the quality of the resulting text (textual quality, fluency), the quality of the translation (felicity, accuracy, adequacy), and formal quality (formatting, tags, fonts, etc. often also subsumed under fluency).

◆ Quality is also gradient in nature. Most criteria are scalar, they can be approximated.
  ▪ The specified target usually is some composite approximation of an ideal target.
  ▪ For many criteria, the distance to the ideal of minimal target can be measured by the frequency and severity of errors (error density).
  ▪ In addition to a composite target there may minimum targets for the individual criteria.

Goal: Simplicity + Sophistication

◆ For different tasks in research and practical QA, different levels of sophistication are required. Sophistication translates into complexity.

◆ Nobody wants to deal with unneeded complexity

◆ On the other hand, any standard that does not accommodate the degree of sophistication needed for research and for complex diagnostic assessment will be dismissed by some of the most relevant players

◆ Thus our philosophy: Have a complex model, of which everyone can use just the parts required for some actual purpose.
More Requirements

- We build as much as possible on existing models such as LISA QA or SAE J2450 and experience through initiatives by LISA, ISO, FIT, GALA, TAUS, etc.
- Foresee and realize a tool for human analytical assessment that automates everything that can safely be automated.
- Build in hooks for future software tools that may automate more than possible today

Even More Requirements

- allow the recording of automatic or composite scores
- allow the embedding of the metrics into a translation QA model that also evaluates entire projects and workflows thus going beyond simple translated texts
- allow the combination of the metrics with QA schemes for measuring monolingual text quality or document quality and QA measures for evaluating information development projects (e.g. in the techdoc world).
Quality Criteria

◆ Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

◆ Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking
Orthography

◆ spelling
◆ capitalization
◆ diacritics
◆ word segmentation
◆ character encoding
◆ disallowed characters/patterns

Quality Criteria

◆ Language
  ▪ Lexical Choice/Terminology
  ▪ Orthography
  ▪ Grammar
  ▪ Meaning/Accuracy
  ▪ Style
  ▪ Punctuation

◆ Document
  ▪ Doc Structure
  ▪ Layout
  ▪ Fonts/Styles
  ▪ Objects
  ▪ Marking
Grammar

- morphology (word forms)
- part of speech
- agreement, case, congruence
- structure
- valence
- ellipsis
- word order
- coordination

Quality Criteria

- Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking
Style

- stylistic adequacy
- register
- clarity (ambiguity)
- cohesion/coherence
- special conventions
  - numbers
  - dates/times
  - measurements
  - names/titles
  - adresses
  - formulae

Quality Criteria

- Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking
Document

- footnotes/side notes
- index
- references/crossreferences broken links
- lists
- captions,
- image titles
- tables
- pagination,
- numbering
- toc

Quality Criteria

- Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking
## Quality Criteria

### Source Document
- **Language**
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- **Document**
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking

### Target Document
- **Language**
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- **Document**
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking
Quality Criteria

Source Document

- Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking

Target Document

- Language
  - Lexical Choice/Terminology
  - Orthography
  - Grammar
  - Meaning/Accuracy
  - Style
  - Punctuation

- Document
  - Doc Structure
  - Layout
  - Fonts/Styles
  - Objects
  - Marking

Quality Specification

- In the absence of any explicit quality specs, the requirements are implicitly specified by the source text and by the knowledge of conventions in the target language

- explicit quality specs can contain requirements for each quality criterion
  - use VW automotive terminology
  - translate titles of persons where possible
  - use imperial measures in English target text
  - address customers with informal you in German target text

- They can also contain for each criterion maximally tolerated error scores (error rates)
**Quality Requirements (QRs)**

- Existing quality assessment methods have tended to assume a “transcendent” or absolute model of quality: all “errors” are counted as errors, regardless of the intended use of the translation.
- Adaptation to needs has been handled by setting quality thresholds.
  
  E.g., a user who needs a quick translation might expect a quality score of 70% (relative to a theoretical ideal translation) while a publisher might require 99.5%.
- Mismatch of expectations (QRs) is a major cause of business disputes
- QRs are addressed by the 21 “parameters” of ISO/TS-11669.

**Process Quality**

- Process is well addressed in standards like EN-15038 (and current ISO work on a successor).
- Does not address the product directly, but instead assumes that controlling to process will result in a better product.
- Existing specifications (particularly the LISA QA Model) may mix process and product concerns.

- **NB:** Process Quality is not addressed in QT LaunchPad.
Project Quality

- Project quality measures satisfaction with the entire project and addresses issues such as payment, timeliness of completion, ease of working between parties, and so forth

- NB: Project Quality is not addressed in QT LaunchPad.

Keeping Separate Things Separate (2)

- QTLP proposes a set of issues that break quality assessment down into its constituent aspects: fluency, accuracy, and end-user adequacy.

- Metrics are based on dimensions of the translation: aspects corresponding to a subset of ISO/TS-11669.
A translator can improve the text

- Any useful quality metric must account for low-quality source
- Often translators are able to fix errors or to improve readability

An Honest Formula for TQ

\[ TQ = (A_t - A_s) + (F_t - F_s) + (E_t - E_s) \]

with respect to \( QR_t \)

where:
- \( Q \) = translation quality score
- \( A \) = accuracy score
- \( F_t \) = fluency score for target text
- \( F_s \) = fluency score for source text
- \( E_t \) = end-user adequacy score for target text
- \( E_s \) = end-user adequacy score for source text
- \( QR_t \) = quality requirement specifications for the translation

All can be thought of as percentages. If \( F_s \) or \( F_t \) values are unknown, 100 (i.e., perfect quality) should be assumed
The Formula vs. the Definition

\[ TQ = (A_t - A_s) + (F_t - F_s) + (E_t - E_s) \]

with respect to \( QR_t \)

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.

Example

- A translator is asked to provide an English translation of a German source text. The source text has some fluency problems and is written in a fashion that is somewhat confusing. The translator is able to improve on the fluency of the translation but the end-user acceptability of the translation is slightly lower than for the source, resulting in the following scores:
  - \( A = 99\% \) (accuracy is very good)
  - \( F_s = 92\% \); \( F_t = 96\% \)
  - \( E_s = 85\% \); \( E_t = 83\% \)
  - \( Q = 99 + (96 - 92) + (83 - 85) = 101\% \)

- Yes, quality score of over 100\% are possible if the translator fixes problems present in the source!
No single metric can address all needs.

Multiple needs:

- Diagnostic quality applications (i.e., why is this translation bad and how can I resolve the root cause?)
- Business quality applications (i.e., can this translation move forward in my process or does it need revision?)

- Need a way to distinguish scenarios and provide situationally relevant input to the quality formula.
- Scores for one scenario may not match those for another (e.g., a perfect “gist” translation will be a poor publication translation).
- Dimensions provide a way to qualify a translation’s intention and understand its quality with respect to that intention.
Profile Dimensions (2)

Derived from ISO/TS-11669 (see [http://www.ttt.org/specs](http://www.ttt.org/specs))

- Language/locale
- Subject field/domain
- Textual Characteristics
  - Text Type
  - Audience
  - Purpose
  - Register
  - Style
- Content correspondence (e.g., full localization, summary translation...)
- Output modality (e.g., printed text, subtitle, text-to-speech...)
- File format (e.g., InDesign, XLIFF, resource file...)
- Production technology (e.g., MT, TM...)

TQAM Structure
Other Elements in the TQAM

◆ Any Quality Requirement Specification and any Quality Assessment Record are instances of the model.

◆ At any node in the tree, the following meta-data elements are foreseen
  - a documentation of the dimension or criterion
  - a default procedure for measuring or judging (e.g. error counting)
  - actually selected different procedures — optional — including omission
  - a field for a list of requirements (qualitative or quantitative) — optional
  - a field for a list of subjective ratings (ordered pairs of rating and source)
  - a field for a list of automatically determined scores (ordered pairs of scores and sources)

Next Steps

◆ Finish the first design by end of February
◆ Implement a first version of an assessment tool
◆ Start a feedback process in LSP industry and professional associations
◆ Do some sample assessments
◆ Revision Process on the Basis of Experience and Feedback
◆ Finish a revised version still in 2013
- morphology (word forms) “Fröhlicheit, unmaterial”
- part of speech “He ate slow” “Electric costs too much”
- agreement, case, congruence
- structure and valence
  “Er spielte die Westküste” from “He played the West Coast”
- coordination and ellipsis
- word order “Am Morgen er kam zu spät.”