Authoritative Standards in the MT Environment

Dr. Jennifer DeCamp

jdecamp@mitre.org

March 17, 2018

Association for Machine Translation in the Americas

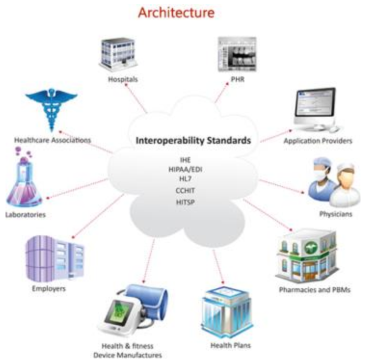
Boston, MA

**The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE''s concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.*

Approved for Public Release; Distribution Unlimited. Case Number 18-0342

How are Standards Important for MT?

1. Help with data and system interoperability
2. Implemented in software we want to use—so we deal with them whether or not we want to
3. Provide higher reliability/certainty than other methods of language ID, data exchange, and term retrieval
4. Provide guidance, replicability, and comparability in assessments
5. Sometimes cited/required in Requests for Proposal or in contracts—must show compliance
   - Direct specification
   - Minimal technical proficiency, minimal cost; technical delta; etc.



This Photo by Unknown Author is licensed under CC BY-NC-ND
This Photo by Unknown Author is licensed under CC BY-NC-ND
This Photo by Unknown Author is licensed under CC BY-NC-SA
This Photo by Unknown Author is licensed under CC BY-SA

Fragmentation, Heterogeneity, and Non-Interoperability

“Current approaches to Machine Translation (MT) or professional translation evaluation, both automatic and manual, are characterized by

- A high degree of fragmentation, heterogeneity and a lack of interoperability between methods, tools and data sets.
- As a consequence, it is difficult to reproduce, interpret, and compare evaluation results.”

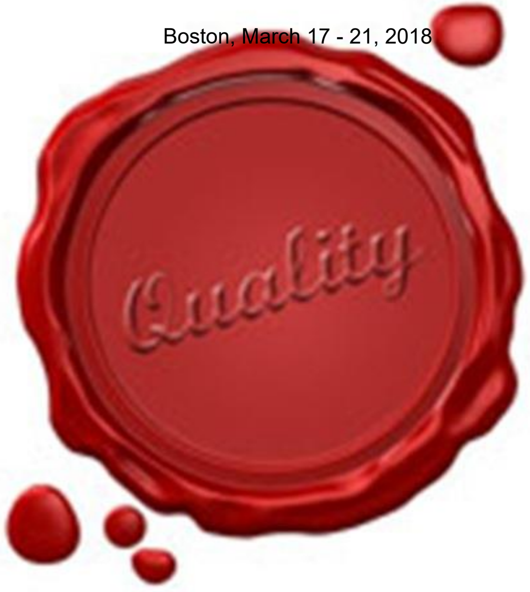
Georg Rehm, Aljoscha Burchardt, Ondˇrej Bojar, Christian Dugast,Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajiˇc, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi, Hans Uszkoreit, **Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem, Workshop held at LREC, 24 May 2016.**

Increasing Collaboration of MT and CAT

- MT increasingly used in commercial environments
- Agile configurations of MT and Computer Assisted Translation (CAT)
  - MT as an option in CAT
  - Predictive MT in CAT
  - Documents with different parts done with different methods
  - Decisions of customer or service provider of tools to use
- Need for evaluations that encompass many approaches or that are neutral to the approach

This Photo by Unknown Author is licensed under CC BY-NC-SA

Translation Quality

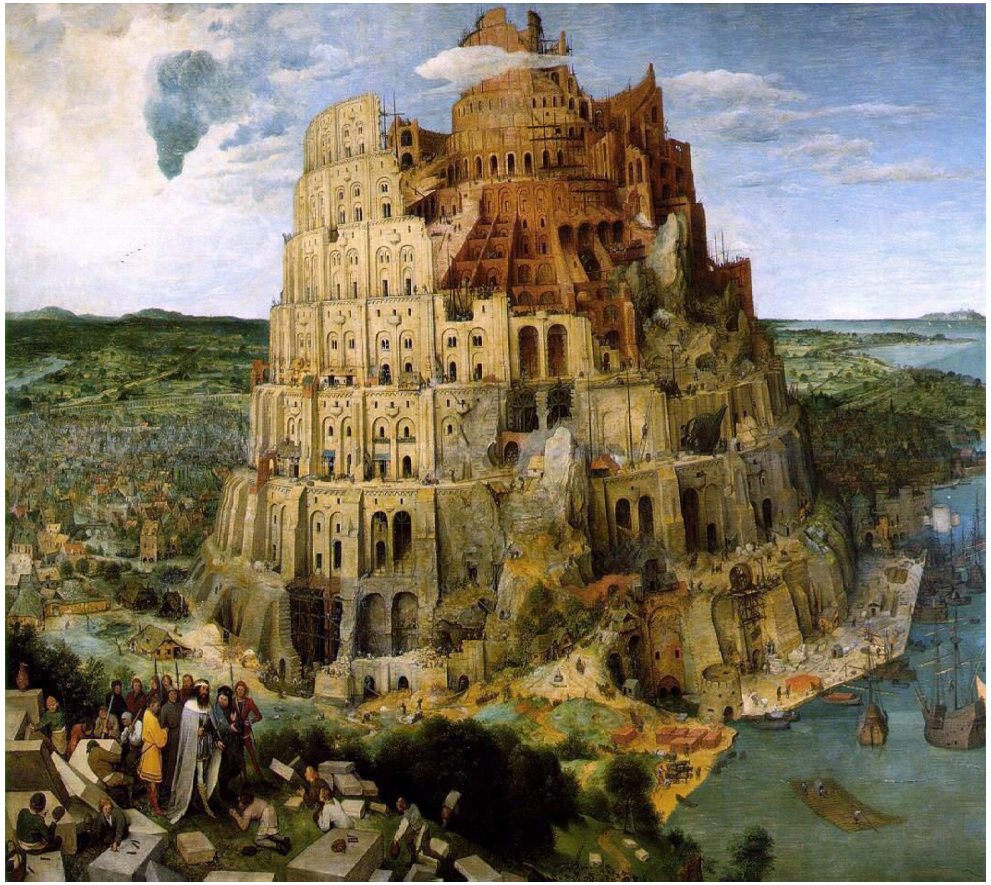


- ASTM F2575 *Standard Guide for Quality Assurance in Translation*
- ASTM WK 41374 *Standard Practice for Language Service Companies*
- ASTM WK 46396 *New Practice for the Development of Translation Q Metrics*
- ASTM Work Item (WK) 47362 *Standard Practice for Quality Assuranc Translation*
- ISO/AWI 21999 *Translation Quality Assurance and Assessment—Models and Metrics*
- ASTM WI 54884 *New Guide for Public Language Quality Assessment (LQA) Methodology*
- ISO 17100:2017 *Translation Services—Requirements for Translation Services*
- ISO 18587 *Translation Services—Post-Editing of Machine Translation Output—Requirements*
- TAUS *Multidimensional Quality Metrics* (MQM) and DFKI work
- TAUS, GALA, LT-Innovate *Translation API Class and Cases Initiative* (TAPICC)

This Photo by Unknown Author is licensed under CC BY-NC-SA

Interoperability

- ISO 639-3, Codes for the Representation of Names of Languages
- IETF BCP 47 Tags for Identifying Languages
- Translation Memory eXchange (TMX)
- ISO 21720 XML Localization Interchange File Format (XLIFF)
- ISO 24613:2008 Lexical Markup Framework
- Translation API Class and Cases (TAPICC) Initiative



This Photo by Unknown Author is licensed under CC BY-SA

How it Works Together

AMTA

ISO

W3C World Wide Web consortium

OASIS

TAUS
the language data network

ANSI
American National Standards Institute

Localization Industry Standards Association

Unicode

ASTM
ASTM INTERNATIONAL

GALA
Globalization & Localization Association

LANGUAGE RESOURCES EVALUATION CONFERENCE
LREC

ata

International Standards for
Language Engineering

TEI
TEXT ENCODING INITIATIVE

So Why This Workshop?

- Provide you with examples of how standards can affect your work with MT
- Encourage debate on the best technical approaches for achieving
  - Interoperability with data and tools
  - Comparability and replicability with evaluations
  - Best practice
- Solicit your participation in development of key standards

This Photo by Unknown Author is licensed under CC BY-ND

Participants

- **Jennifer DeCamp**
  - Chair, ATA Standards Committee
  - Member ISO, ASTM, ANSI, ILR, AMTA, and ATA
  - Chair, ASTM TAG to ISO/TC 37/SC 4
  - Principal Scientist, MITRE Corporation
- **Sue Ellen Wright**
  - Chair, ASTM U.S. TAG to ISO/TC 37
  - Chair, ISO/TC 37/SC 3
  - Member ISO, ASTM, ANSI, and ATA
  - Professor, Translation Studies, Kent State University
  - Recipient of ANSI Outstanding Achievement Award
- **David Filip**
  - OASIS XLIFF OMOS TC Chair
  - OASIS XLIFF TC Secretary, Editor, Liaison Officer
  - Spokes Research Fellow
  - ADAPT Centre
  - KDEG, Trinity College Dublin
- **Bill Rivers**
  - Secretary, U.S. Technical Advisory Group to ISO/TC 37
  - Member ASTM, ANSI, ISO, and ATA
  - Executive Dir., Joint National Committee for Languages
- **Arle Lommel**
  - Project Leader for ASTM Translation Metrics Standard
  - Senior Analyst, Common Sense Advisory
  - Member ASTM, ATA, GALA
- **Alan Melby**
  - Liaison between ATA and FIT
  - Member ISO, ASTM, ANSI, OASIS, and ATA
  - President, LTAC
  - Associate Director, BYU Translation Research Group

Agenda

• **Jennifer DeCamp**	Introduction
• **Jennifer DeCamp**	Language Codes
• **Sue Ellen Wright**	TermBased eXchange (TBX)
• **Bill Rivers**	Translation Quality Standards
• **Arle Lommel**	Translation Metrics
• **Alan Melby**	Translation API for Class and Cases (TAPICC)
• **Panel**	

Questions for Possible Discussion

- What role will standards have to play in the future?
- Are there viable and preferable alternatives to using standards?
- How can we make the standards more useful to the translation environment, particularly with MT?
- Where do we have gaps or issues?
- Where do we need additional work?
- Do we have the right organizations represented?
- Do we have the right people working on the standards?

ar ara

ISO Language Codes

Dr. Jennifer DeCamp

jdecamp@mitre.org

March 17, 2018

Association for Machine Translation in the Americas

Boston, MA

**The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE''s concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.*

Approved for Public Release; Distribution Unlimited. Case Number 18-0343'

Wherefore Language Codes?

- Demand by industry for codes for more languages
- Need for less ambiguity and overlap
- Need for linguistic rather than bibliographic orientation
  - Machine Readable Cataloging (MARC 21)
  - ISO 639-1 and ISO 639-2
  - Most commonly used system among linguists was The Ethnologue
- Need for consistency

Codes for the Representation of Language

• ISO 639-1	ar	Arabic
• ISO 639-2	ara	Arabic
• ISO 639-3	aeb	Tunisian Arabic
• ISO 639-5	ARA	Arabic, macrolanguage

- Four-letter codes for variants and registers?

• mis	Uncoded languages
• mul	Multilingual
• und	Undetermined languages
• xxx	No linguistic content/not applicable

ISO 639 Registrars and Joint Advisory Committee

PARTS

• ISO 639-1	Infoterm
• ISO 639-2	Library of Congress
• ISO 639-3	SIL International
• ISO 639-4	Joint
• ISO 639-5	Library of Congress
• ISO 639-6	TBD

- Joint Advisory Committee

This Photo by Unknown Author is licensed under CC BY-SA

International Engineering Task Force (IETF) Best Current Practice (BCP) 47 *Tags for Identifying Languages*, 2009

- zh-Hans (Chinese written using the Simplified Chinese script)
- zh-cmn-Hans-CN (Chinese, Mandarin, Simplified script, as used in China)

- sr-Cyrl (Serbian written using the Cyrillic script)
- sr-Latn-RS (Serbian written using the Latin script as used in Serbia)

Request for Comment (RFC) 5646

- Language: fr (French)
- Language-Region: de-DE (German for Germany)
- Language subtag plus Script subtag: zh-Hant (Chinese written using the Traditional Chinese script)
- Extended language subtags and their primary language subtag counterparts: zh-cmn-Hans-CN (Chinese, Mandarin, Simplified script, as used in China)
- Language-Script-Region: zh-Hans-CN (Chinese written using the Simplified script as used in mainland China)
- Language-Variant: sl-rozaj (Resian dialect of Slovenian) sl-rozaj-biske (San Giorgio dialect of Resian dialect of Slovenian) sl-nedis (Nadiza dialect of Slovenian)

- Language-Variant: sl-rozaj (Resian dialect of Slovenian) sl-rozaj-biske (San Giorgio dialect of Resian dialect of Slovenian) sl-nedis (Nadiza dialect of Slovenian)

Status

- Correlated with many other standards
- Worldwide use
- Implemented for two decades in Microsoft, depending on keyboard
- ISO 639 up for review
  - Meetings in March to discuss processes
- New ISO standards in development to supplement ISO 639
  - Variants
  - Registers

Issues

- Not coordinated with speech community
- Variable width difficult to implement with older databases
- Too few Q codes
- People repurposing codes because they like the mnemonics or because they are trying to express dialects or other information within the three-character format
- Difficult to meet requirements for new codes (although easier than it used to be!)

Why is ISO 639 important for MT?

- Automatic language identification
  - Is not available for all languages and dialects
  - Is not always possible with very small numbers of words
- Correctly tagged text needed, particularly in languages with less textual material, for
  - Identification of text
  - Application of tools
- Incorrectly tagged text can result in
  - Use of wrong tools on the data (e.g., spellchecker)
  - Use of data incorrectly (e.g., in Translation Memories)

References

- IETF BCP 47 (2009). *Tags for Identifying Languages*, 2009.
- ISO15924, International Organization for Standardization, ISO 15924:2004. *Information and documentation -- Codes for the representation of names of scripts*, January 2004.
- ISO3166-1, International Organization for Standardization, ISO 3166-1:2006. *Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes*", November 2006.
- ISO639-1, International Organization for Standardization, ISO 639-1:2002. *Codes for the representation of names of languages -- Part 1: Alpha-2 code*, July 2002.
- ISO639-2, International Organization for Standardization, ISO 639-2:1998. *Codes for the representation of names of languages -- Part 2: Alpha-3 code*, October 1998.
- ISO639-3, International Organization for Standardization, ISO 639-3:2007. *Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages*, February 2007.
- ISO639-4, International Organization for Standardization, ISO 639-4:2010. *Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages*, February 2007.
- ISO639-5, International Organization for Standardization, ISO 639-3:2007. *Codes for the representation of names of macrolanguages- Part 5: Alpha-3 code for comprehensive coverage of languages*, February 2007.