MACHINE TRANSLATION EVALUATION*

Emmanuelle Esperança-Rodier
Laboratoire LIG
Équipe GETALP
Emmanuelle.Esperanca-rodier@univ-grenoble-alpes.fr

* Adapted from H. Blanchon slides

CONTENT

- Purpose
  - Machine Translation Systems Evaluation
- Coverage
  - Human-based evaluation: Subjective Evaluation
    - Measures
    - Pros & cons
  - Efforts to formalize MT systems Evaluation
  - Program-based evaluation: Objective Evaluation
    - Measures
    - Pros & Cons
- Some proposals for improvement

OUTLINE

- Subjective evaluation foundations
- “Let’s try to formalize” efforts
- Subjective evaluation in practice
- Subjective evaluation final remarks
- Objective evaluation
- Objective evaluation final remarks
- Conclusion
- Bibliography

SUBJECTIVE EVALUATION FOUNDATIONS

IMPORTANT DATES

- 1966: ALPAC, the (In-)famous report
  - Automatic Language Processing Advisory Committee
- 1989 & 1992: JEIDA
  - Japanese Electronic Industry Development Association
- 1992 & 1994: ARPA
  - Advanced Research Projects Agency
- 2000-: NIST
  - National Industry Standards and Technology
- 2015-2018: QT21
  - Quality Translation 21

ALPAC 1966

Automatic Language Processing Advisory Committee

[ALPAC, 1966]

- An Experiment in Evaluating the Quality of Translations
- Comment
  - Poor MT performance led to cuts in MT funding in the United-States
  - Highly influential work

ALPAC

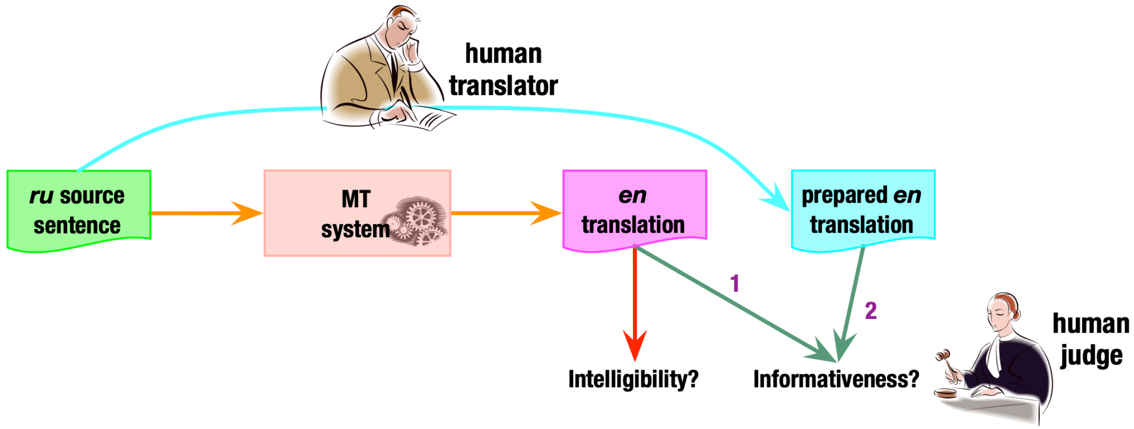
- 2 major independent characteristics of a translation
  - Its intelligibility
  - Its fidelity to the sense of the original text
- Subjective rating
  - Rating of intelligibility without reference to the source
  - Indirect rating of fidelity
    - Gather whatever possible meaning from the translation sentence
    - Evaluate the source sentence "informativeness" in relation to the understanding of the translation sentence
      - ✓ A highly informative source sentence implies that the translation is lacking in fidelity

ALPAC

- Language pair / Domain
  - Russian → English / Scientific
- Data
  - 36 sentences / 6 translations (3 human, 3 MT systems)

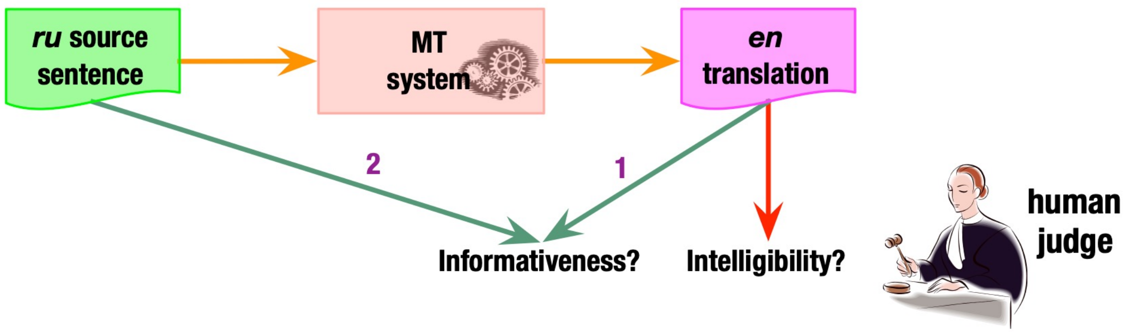
ALPAC

- 2 sets of evaluation (1/2)
  - Monolingual evaluation
    - 18 native English speakers with no knowledge of Russian and good background in science
    - Carefully prepared English translation of the source sentences (references)



ALPAC

- 2 sets of evaluation (1/2)
  - Bilingual evaluation
    - 18 native English speakers with a high degree of competence in comprehension of scientific Russian



ALPAC: Intelligibility

9– Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.

8– Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily “corrected.”

7– Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.

6– The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in a nearly acceptable form.

5– The general idea is intelligible only after considerable study, but after this study, one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present but constitute mainly “noise” through which the main idea is still perceptible.

4– Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.

3– Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.

2– Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.

1– Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

ALPAC: Informativeness

9– Extremely informative. Makes “all the difference in the world” in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation.)

8– Very informative. Contributes a great deal to the clarification of the meaning intended. Correcting sentence structure, words, and phrases, makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely.

7– (Between 6 and 8.)

6– Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader “on the right track” as to the meaning intended.

5– (Between 4 and 6.)

4– In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.

3– By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different “twist” to the meaning conveyed by the translation. It adds no new information about sentence structure, however.

2– No really new meaning is added by the original, either at the word level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended.

1– Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced.

0– The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.

ALPAC: QUOTES

- "MT presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing." → "In this context, there has been no machine translation of general scientific text, and none is in immediate prospect."
- "The reader will find it instructive to compare the samples above with the results obtained on simple, selected, text 10 years earlier (the Georgetown IBM Experiment, January 7, 1954) in that the earlier samples are more readable than the later ones."
- In the final chapter (p.32-33), ALPAC underlined once more that "we do not have useful machine translation [and] there is no immediate or predictable prospect of useful machine translation." It repeated the potential opportunities to improve translation quality, particularly in various machine aids: "Machine-aided translation may be an important avenue toward better, quicker, and cheaper translation." But ALPAC did not recommend basic research: "What machine-aided translation needs most is good engineering."

JEIDA (1989 & 1992)

- Japanese Electronic Industry Development Association
  - Jeida 1989 [JEIDA, 1989]
    - A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC.
    - 3 questions
      - ✓ What are the technological and social changes of the market since the ALPAC report?
      - ✓ According to these changes, are the conclusions of the ALPAC report still valid today?
      - ✓ If not, how should we evaluate the current state and the future of machine translation?
    - No clear answer!

JEIDA (1989 & 1992)

- Jeida 1992 [JEIDA, 1992]
  - JEIDA Methodology and Criteria on Machine Translation Evaluation
  - Several points of view using complex forms
    - Economical factors evaluation by the users
    - Technical evaluation of the systems by the users
      - ✓ "Satisfaction of the users' needs"
    - Technical evaluation of the systems by the developers
      - ✓ "Criteria to help researchers, developers, and project leaders in evaluating their systems"

ARPA (1992-1994) & NIST (2000-)

Advanced Research Projects Agency National Industry Standards and Technology

- Comparative/competitive evaluation [White et al, 1994]
  - Systems
    - Fully automatic / Human Aided MT
  - Language pairs
    - Source language: several / Target language: English
  - Domain
    - Newspaper articles about financial mergers and acquisitions
    - Professionally translated into the respective source languages or into English
  - Evaluators
    - literate, monolingual English speakers

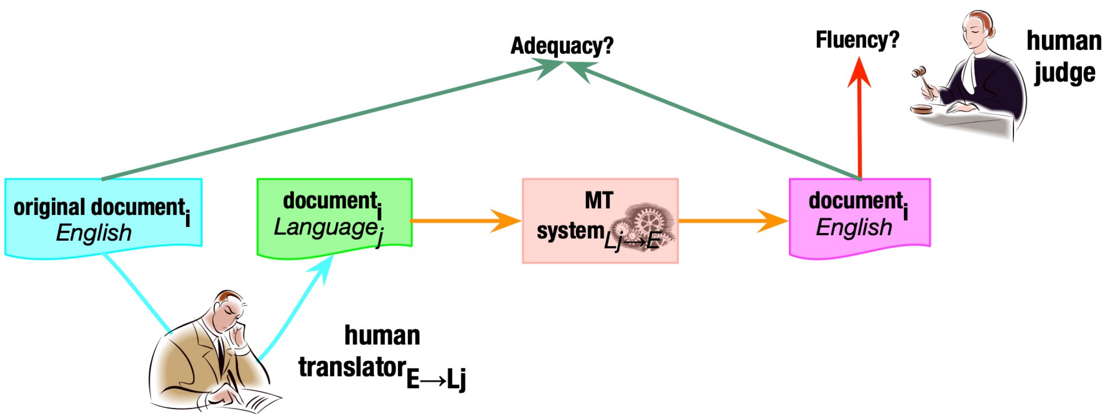
ARPA & NIST

- Criteria
  - Fluency
    - without reference to the source
- Adequacy
  - in contrast to the English original or translation

Score	Adequacy	Fluency
5	All information	Flawless English
4	Most	Good
3	Much	Non-Native
2	Little	Disfluent
1	None	Incomprehensible

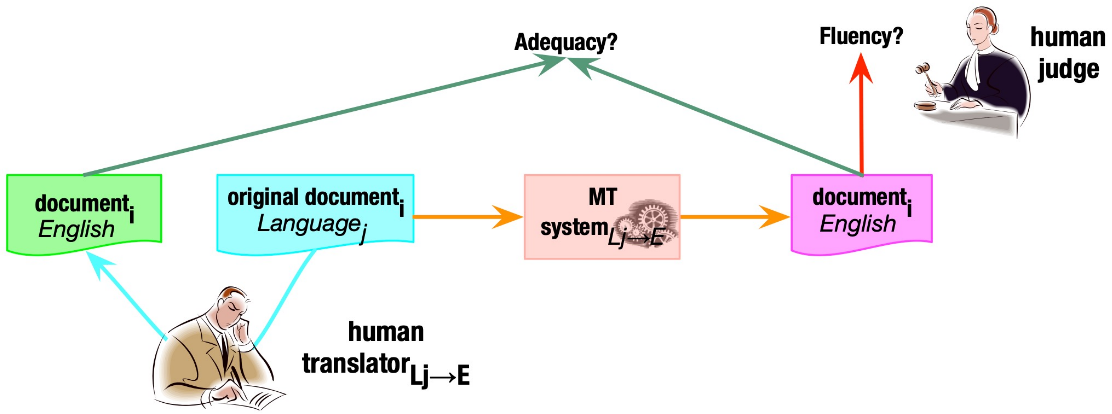
ARPA & NIST

- When source document is not available



ARPA & NIST

- When source document is available



"LET'S TRY TO FORMALIZE" EFFORTS

IMPORTANT DATES

- 1993-1996: EAGLES
  - Expert Advisory Group on Language Engineering
  - Initiative of the European Commission
  - [EAGLES-EWG, 1996] [EAGLES-EWG, 1999]
- 1999-2002: ISLE (FEMTI)
  - Framework for Machine Translation Evaluation on ISLE (International Standards for Language Engineering)
  - Joined initiative of the European Commission and National Science Foundation (NSF)
  - http://www.isi.edu/natural-language/mteval/
  - [Hovy et al., 2002] [King et al., 2003]

EAGLES

Expert Advisory Group on Language Engineering

- Goal
  - Standards for the language engineering industry
- Targets
  - Corpora
  - Lexicons
  - Grammatical formalisms
  - Evaluation
- On evaluation
  - A quality model for natural language processing tools…
  - … validated on grammar checkers,

EAGLES: A 7-STEP RECIPE

1. Why is the evaluation being done?
2. Elaborate a task model
3. Define top-level quality characteristics
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3
5. Devise the metrics to be applied to the system for the requirements produced under 4
6. Design the execution of the evaluation
7. Execute the evaluation

EAGLES: A 7-STEP RECIPE

1. Why is the evaluation being done?

- What is the purpose of the evaluation? Do all parties involved have the same understanding of the purpose?
- What exactly is being evaluated? Is it a system or a system component? A system in isolation or a system in a specific context of use? Where are the boundaries of the system?

2. Elaborate a task model

- Identify all relevant roles and agents What is the system going to be used for?
- Who will use it? What will they do with it? What are these people like?

3. Define top-level quality characteristics

- What features of the system need to be evaluated? Are they all equally important?

EAGLES: A 7-STEP RECIPE

4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3

- For each feature that has been identified as important, can a valid and reliable way be found of measuring how the object being evaluated performs with respect to that feature?
- If not, then the features have to be broken down in a valid way, into sub-attributes that are measurable.
- This point has to be repeated until a point is reached where the attributes are measurable.

EAGLES: A 7-STEP RECIPE

5. Devise the metrics to be applied to the system for the requirements produced under 4

- Both measure and method for obtaining that measure have to be defined for each attribute.
- For each measurable attribute, what will count as a good score, a satisfactory score or an unsatisfactory score given the task model (2)? Where are the cut off points?
- Usually, an attribute has more than one sub-attributes. How are the values of the different sub-attributes combined to a value for the mother node in order to reflect their relative importance (again given the task model)?

EAGLES: A 7-STEP RECIPE

6. Design the execution of the evaluation

- Develop test materials to support the testing of the object.
- Who will actually carry out the different measurements? When? In what circumstances? What form will the end result take?

7. Execute the evaluation:

- Make measurement.
- Compare with the previously determined satisfaction ratings.
- Summarize the results in an evaluation report, cf. point 1.

FEMTI

Framework for Machine Translation Evaluation on ISLE (International Standards for Language Engineering)

- Attempt to organize the various methods for MT evaluation

FEMTI

FEMTI contains

- A classification of the main features defining the context of use (type of user of the MT system, type of task the system is used for, nature of the input to the system)
- A classification of the MT software quality characteristics, into hierarchies of sub-characteristics, with internal and/or external attributes (i.e., metrics) at the bottom level.
- A mapping from the first classification to the second, which defines or suggests the quality characteristics, sub- characteristics and attributes/metrics that are relevant to each context of use.

FEMTI (TOP LEVEL CLASSIFICATION)

1 Evaluation requirements

- 1.1 The purpose of the evaluation
- 1.2 The object of evaluation
- 1.3 Characteristics of the translation task – 1.3.1 Assimilation

– 1.3.2 Dissemination

– 1.3.3 Communication

- 1.4 User characteristics

– 1.4.1 Machine translation user – 1.4.2 Translation consumer

– 1.4.3 Organisational user

- 1.5 Input characteristics (author and text)

– 1.5.1 Document type (genre, domain/field of application)

– 1.5.2 Author characteristics (proficiency in the source language, training)

– 1.5.3 Characteristics related to sources of errors (unproofed text)

FEMTI (TOP LEVEL CLASSIFICATION)

2 System characteristics to be evaluated

• 2.1 System internal characteristics – 2.1.1 MT system-specific characteristics – 2.1.2 Translation process models

– 2.1.3 Linguistic resources and utilities

– 2.1.4 Characteristics of process flow

• 2.2 System external characteristics – 2.2.1 Functionality

- 2.2.1.1 Suitability, Accuracy, Wellformedness, Interoperability, Compliance, Security

– 2.2.2 Reliability

– 2.2.3 Usability

– 2.2.4 Efficiency

– 2.2.5 Maintainability

– 2.2.6 Portability

– 2.2.7 Cost

FEMTI (SECTION 2.2.1 FUNCTIONALITY)

2.2.1.1 Suitability

– 2.2.1.1.1 Target-language only

- 2.2.1.1.1.1 Readability (or: fluency, intelligibility, clarity)

- 2.2.1.1.1.2 Comprehensibility - 2.2.1.1.1.3 Coherence

- 2.2.1.1.1.4 Cohesion

– 2.2.1.1.2 Cross-language / contrastive

- 2.2.1.1.2.1 Coverage of corpus-specific phenomena - 2.2.1.1.2.2 Style

• 2.2.1.2 Accuracy – 2.2.1.2.1 Fidelity

– 2.2.1.2.2 Consistency – 2.2.1.2.3 Terminology

FEMTI (SECTION 2.2.1 FUNCTIONALITY [CONT.])

2.2.1.3 Wellformedness

– 2.2.1.3.1 Punctuation

– 2.2.1.3.2 Lexis / lexical choice – 2.2.1.3.3 Grammar / syntax

– 2.2.1.3.4 Morphology

• 2.2.1.4 Interoperability • 2.2.1.5 Compliance

• 2.2.1.6 Security

FEMTI (2.2.1.1.1.1 READABILITY)

- Definition
  - The extent to which a sentence reads naturally.
  - The ease with which a translation can be understood, i.e. its clarity to the reader. (Halliday in Van Slype's Critical Report) .
  - This has also been called fluency, intelligibility, and clarity.
- Metrics
  - ...
  - Pfafflin (in Van Slype's Critical Report): Rating of sentences read on a 3-point scale.
  - Vanni & Miller (2001, 2002): "Do you get it?" - snap judgment rating of sentences on a scale from 0 to 3.
  - Niessen, Och, Leusch, and Ney, 2000 measure syntactic errors with an automated string edit distance metric, which according to them can also be used as a measure of readability. See also Wellformedness (2.2.1.3/186).
  - J.B. Carroll: by measuring the time spent by the evaluator in reading each sentence of the sample.
  - Pfafflin and Orr (both quoted by T.C. Halliday): by measuring the response time to a multiple-choice questionnaire.
  - H.W. Sinaiko: by measuring the time necessary for the execution of the cloze test.
- Notes
  - Readability is intended to be a metric applied at the sentence level. ...
  - Readability is the quality of the output that can be measured independently of the source language. Cloze tests can be used either at sentence level or cross-sentence level.
  - This quality has been merged with clarity, which was a separate taxon in earlier versions of this taxonomy.

FEMTI (2.2.1.2.1 FIDELITY)

- Definition
  - Subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation (Van Slype).
  - Measurement of the correctness of the information transferred from the source language to the target language (Halliday in Van Slype's Critical Report).
  - Metrics
- Metrics
  - ...
  - White and O'Connell (in DARPA 94): Rating of 'Adequacy' on a 5-point scale.
  - Bleu evaluation tool kit (in Papineni et al. 2001): Automatic n-gram comparison of translated sentences with one or more human reference translations.
  - Rank-order evaluation of MT system: correlation of automatically computed semantic and syntactic attributes of the MT output with human scores for adequacy and informativeness, and also fluency. Hartley and Rajman 2001 and 2002.
  - Automated word-error-rate evaluation (in Och, Tillmann and Ney, 1999).Notes
  - Readability is intended to be a metric applied at the sentence level. ...
  - Readability is the quality of the output that can be measured independently of the source language. Cloze tests can be used either at sentence level or cross-sentence level.
  - This quality has been merged with clarity, which was a separate taxon in earlier versions of this taxonomy.
- NOTES
  - The fidelity rating has been found to be equal to or lower than the comprehensibility rating, since the unintelligible part of the message is not found in the translation. Any variation between the comprehensibility rating and the fidelity rating is due to additional distortion of the information, which can arise from: – loss of information (silence) - example: word not translated, – interference (noise) - example: word added by the system, – distortion from a combination of loss and interference - example: word badly translated.

QT21

Quality Translation 21

- QT21 focused on MT for challenging morphologically complex and syntactically varied languages.
- QT21 has produced the largest data set available of Human Post Edits and Human Error Annotations, for four language pairs and all its software is open source and is available through its website.

QT21: CONTEXT

Often there are not enough training resources and/or processing tools. Together this results in drastic drops in translation quality. QT21 addressed this grey area by developing:

- (1) substantially improved statistical and machine-learning-based translation models for challenging languages and resource scenarios,
- (2) improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators,
- (3) all with a strong focus on scalability, to ensure that learning and decoding with these models is efficient and that reliance on data (annotated or not) is minimized.
- To continuously measure progress, and to provide a platform for sharing and collaboration (QT21 internally and beyond), the project revolves around a series of Shared Tasks, for maximum impact co-organized with WMT.

QT21: WORK ACHIEVED

- 5 language pairs, 4 with English as source (English->German, English->Czech, English->Latvian, English->Romanian) and 1 with English as the target language (German->English). In order to measure progress and compare QT21 with the international state-of-the-art (s-o-t-a), QT21 co-organises WMT 2016, 17 and 18 (the Workshop on Machine Translation http://statmt.org/wmt**) to benchmark MT technologies on shared tasks. The goal was to:
  - (1) improve statistical and machine-learning based translation models for challenging languages and resource scenarios;
  - (2) ensure that learning and decoding with these models is efficient and that reliance on data (annotated or not) is minimized;
  - (3) improve evaluation and continuous learning from mistakes, informed by human translators and post-editors, guided by a systematic analysis of quality barriers;
  - (4) provide a platform for sharing, collaboration, and evaluation (QT21 internally and beyond), QT21 revolves around Shared Tasks, for maximum impact co-organized with WMT;
  - (5) support early technology transfer, QT21 has implemented a Technology Bridge linking ICT-17(a) and (b), showing the technical feasibility of early research outputs in industry-focused environments.

QT21: MAIN RESULTS ACHIEVED

- (1) QT21 has made substantial contributions to Neural Machine Translation (NMT), pushing the state-of-the-art for NMT to comprehensively outperform the previous state-of-the-art held for many years by the family of Phrase-based Statistical MT (PB-SMT). Core technical contributions include "back translation" to produce synthetic training data, Byte Pair Encoding (BPE) to compress vocabularies of morphologically rich languages, and deeper recurrent neural networks. At the international competitions WMT16 and WMT17, QT21 systems won more than 80% of all shared tasks, outperforming large-scale commercial MT systems on En → De, En → Cz and En→Ro, the core languages of QT21.

- (2) QT21 introduced back-translation (see Objective (1)), reducing the dependency on bi-lingual data. QT21 used BPE (see Objective (1)) improving MT for morphologically rich languages by significantly compressing the representation of the vocabulary, addressing the out-of-vocabulary (OOV) issues in automatic translation. QT21 showed that multi-lingual embeddings can efficiently support transfer learning for under-resourced languages. Further, QT21 work on inter-lingual factors opens the door to translating languages not seeing during training.

QT21: MAIN RESULTS ACHIEVED

- (3) QT21 systems won all WMT16 MT evaluation metrics tasks. In addition, QT21 won the WMT16 Quality Estimation (QE) shared task on "document level quality". In the Automatic Post Editing (APE - learning from post-edits of professional translators) shared tasks. QT21 improved the baseline by 2.64 BLEU points with the 2nd best performance at WMT16 and won the WMT17 task improving the baseline by 7.6 BLEU points. A QT21 APE system that learns on-line from human post-editors further improves MT s-o-t-a by 1 to 2 BLEU points. QT21 further developed Direct Assessment (DA), showing that crowd sourcing can be a large-scale effective way of reliably evaluating MT systems. QT21 has harmonised the two major typologies for diagnostic MT error analysis, QT21's own MQM (Multidimensional Quality Metrics) and TAUS' industry standard DQF (Dynamic Quality Framework). QT21 revitalised interest in test suits in diagnostic MT evaluation.
- (4) The organisation of WMT (co-organised with CRACKER-Horizon2020 # 645357) is at the core of this objective. The +48% increase in submissions from 2015 to 2017 on the main task (News Task) and the tripling of participation in the APE task between 2015 and 2017 shows the value and recognition WMT enjoys in the community.
- (5) To implement the QT21 ICT-17 Technology Bridge, QT21 ran workshops on QT21 research outcomes and technologies with DGT (MT@EC), HimL (Horizon2020-ICT17b #644402), MMT (Horizon2020-ICT17b #645487), TraMOOC (Horizon-ICT17b #644333), and KConnect (Horizon2020-ICT15 #644753). All ICT-17(b) projects used QT21 engines and technologies. A joint QT21-HimL submission entered WMT16. DGT (MT@EC) is in the process of switching from SMT to NMT."

MQM: MULTIDIMENSIONAL QUALITY METRICS [LOMMEL, 2014]

1. Preliminary Stage: the following tasks do not need to be implemented in a specific order

- reviewing and ensuring access to the agreed-on translation specifications
- verifying (or selecting/creating, if necessary) the metric for performing the evaluation based on the translation specifications
- assigning the Threshold Value for pass/fail acceptance of the evaluation
- preparing the source text and target text for evaluation
- determining the Evaluation Word Count, usually by means of a software app such as a CAT (computer assisted translation) tool

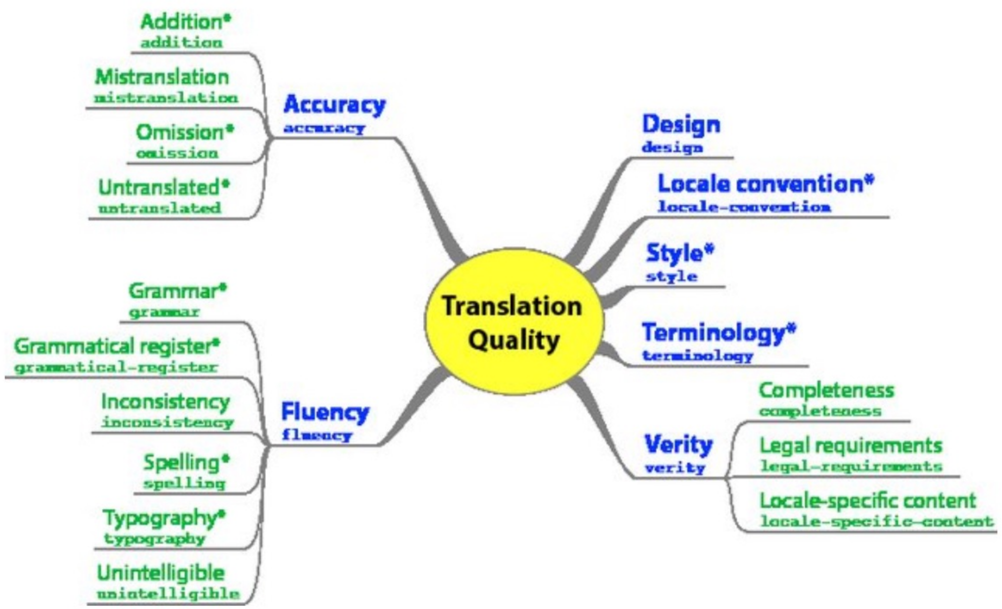
MQM

- In an established TQE system, the first three elements set forth the model that can be used in later projects as a template to be used over and over again.
- In such instances, the three elements regarding translation specifications, metrics, and Threshold Values are more like checkpoints that the implementer verifies before continuing the TQE.

MQM

2. Error Annotation Stage: during this stage, the evaluator examines the translated text against the source text and specifications, and annotates (meaning identifying, marking, and assigning error type and penalty points) errors in accordance with the metric. The 3 annotated errors generate the Absolute Penalty Total.
3. Automatic Calculation & Follow-Up Stage: during this stage, the Overall Quality Score is calculated according to the selected scoring model using the Evaluation Word Count from the Preliminary Stage and the Absolute Penalty Total from the Error Annotation Stage, then compared to the Threshold Value to assign a pass/fail rating. Other actionable items are determined as a result of the evaluation.

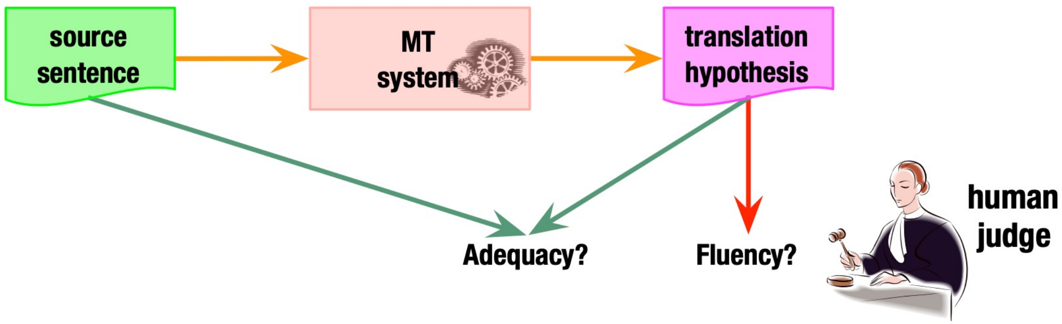
MQM


Translation Quality
Accuracy
accuracy
Addition*
addition
Mistranslation
mistranslation
Omission*
omission
Untranslated*
untranslated
Fluency
fluency
Grammar*
grammar
Grammatical register*
grammatical-register
Inconsistency
inconsistency
Spelling*
spelling
Typography*
typography
Unintelligible
unintelligible
Design
design
Locale convention*
locale-convention
Style*
style
Terminology*
terminology
Verity
verity
Completeness
completeness
Legal requirements
legal-requirements
Locale-specific content
locale-specific-content

SUBJECTIVE EVALUATION IN PRACTICE

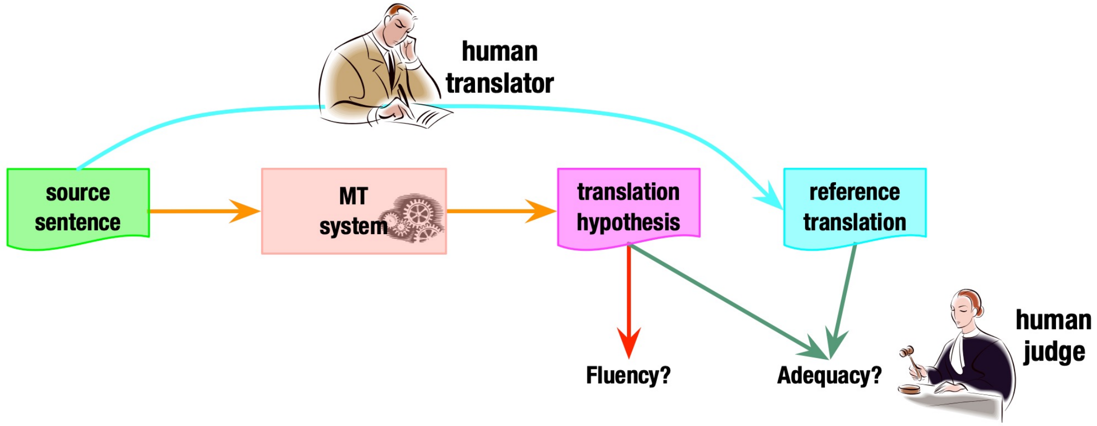
THE SETTINGS

- Bilingual Evaluation



THE SETTINGS

- Monolingual Evaluation



CONS GOLD STANDARD

Source Sentence	Reference Translation
• Mais une fiscalité insuffisante peut également produire les mêmes effets.	• Too little taxation can do the same.
• Le malaise français n'a certainement pas été induit par ces réformes.	• The French malaise has nothing to do with any of them.
• Mais quelle est la signification réelle de ces deux principes ?	• But what do solidarity and subsidiarity really mean?
• Les traités européens expriment clairement cette subsidiarité verticale.	• In the European Treaties, we find a clear expression of vertical subsidiarity.

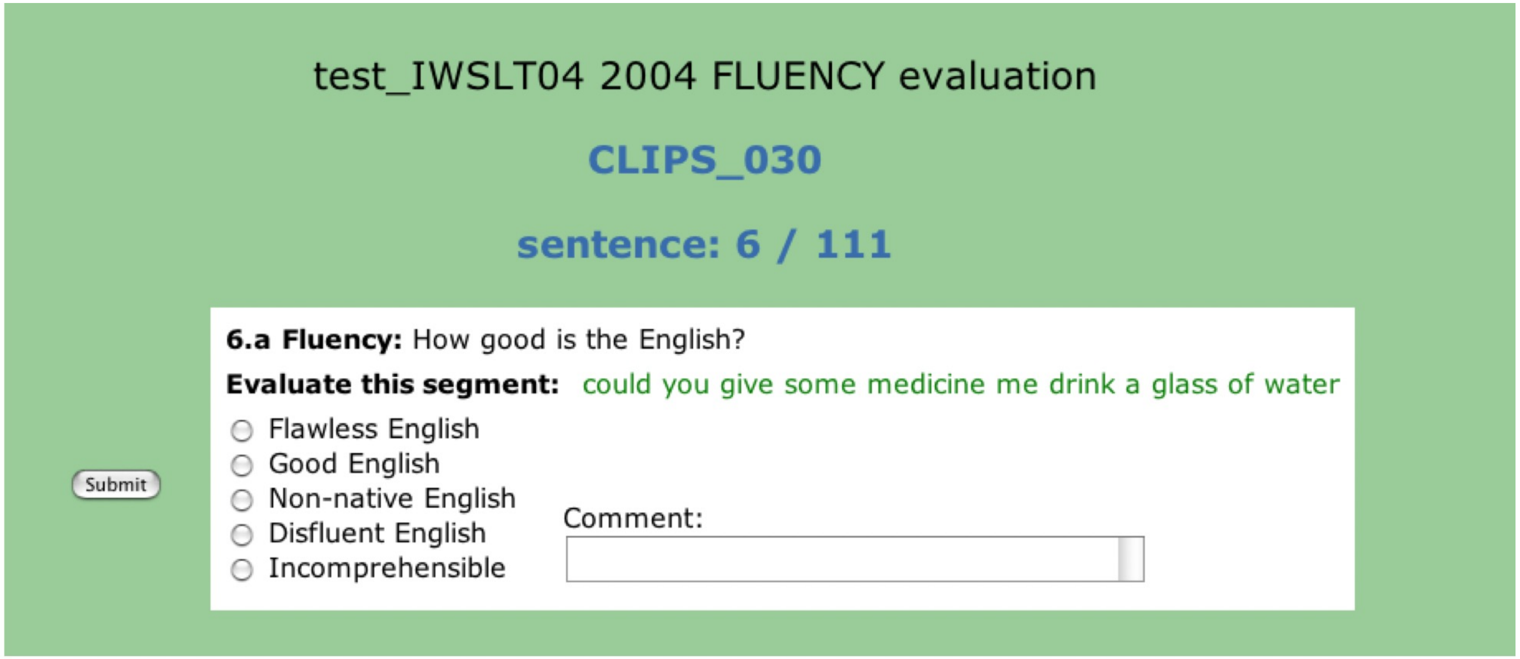
EXAMPLE: IWSLT (2004)

- Language pair
  - Japanese → English
- Domain
  - Tourism
- Evaluators
  - Native English speakers
- Evaluation criterion
  - Fluency
  - Adequacy

49

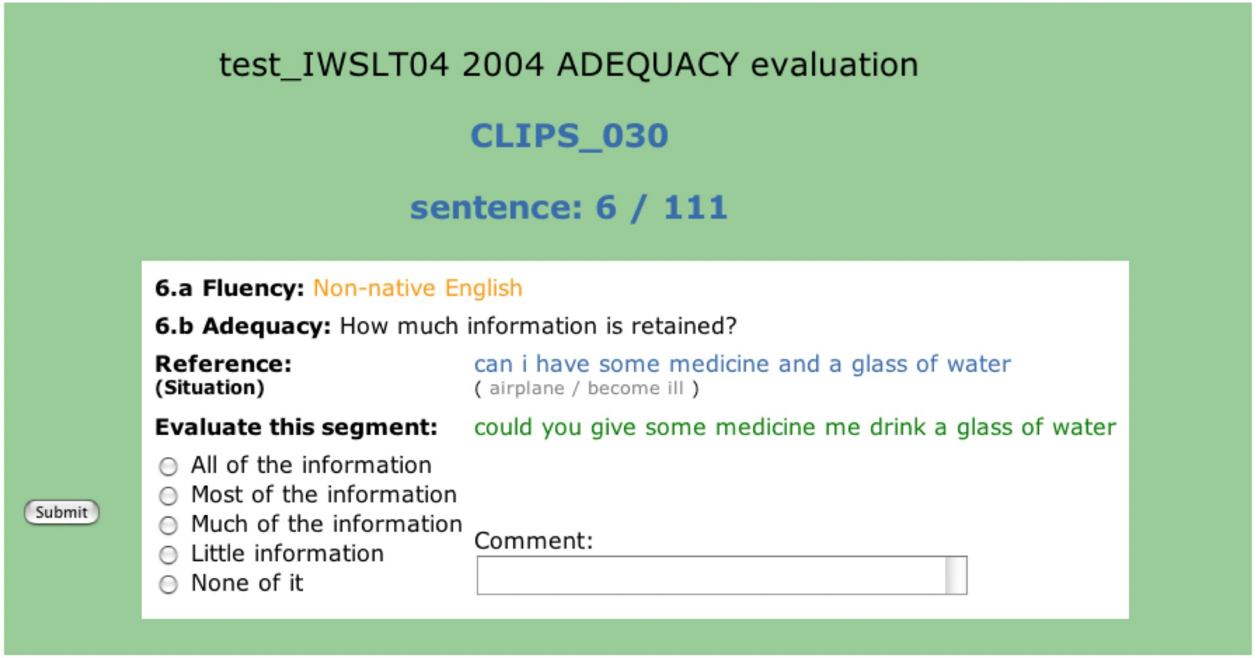
EXAMPLE: IWSLT

- Fluency



EXAMPLE: IWSLT

- Adequacy



test_IWSLT04 2004 ADEQUACY evaluation

CLIPS_030

sentence: 6 / 111

6.a Fluency: Non-native English

6.b Adequacy: How much information is retained?

Reference: can i have some medicine and a glass of water

(Situation) (airplane / become ill)

Evaluate this segment: could you give some medicine me drink a glass of water

- All of the information
- Most of the information
- Much of the information
- Little information
- None of it

Comment:

Submit

EXAMPLE: ACCOLÉ

ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. [Esperança-Rodier et al., 2019]

- Error Typologies
- Collaborative
- Aligned Corpora
- Search

EXAMPLE: ACCOLÉ

- Vilar’s Typology [Vilar, 2016]

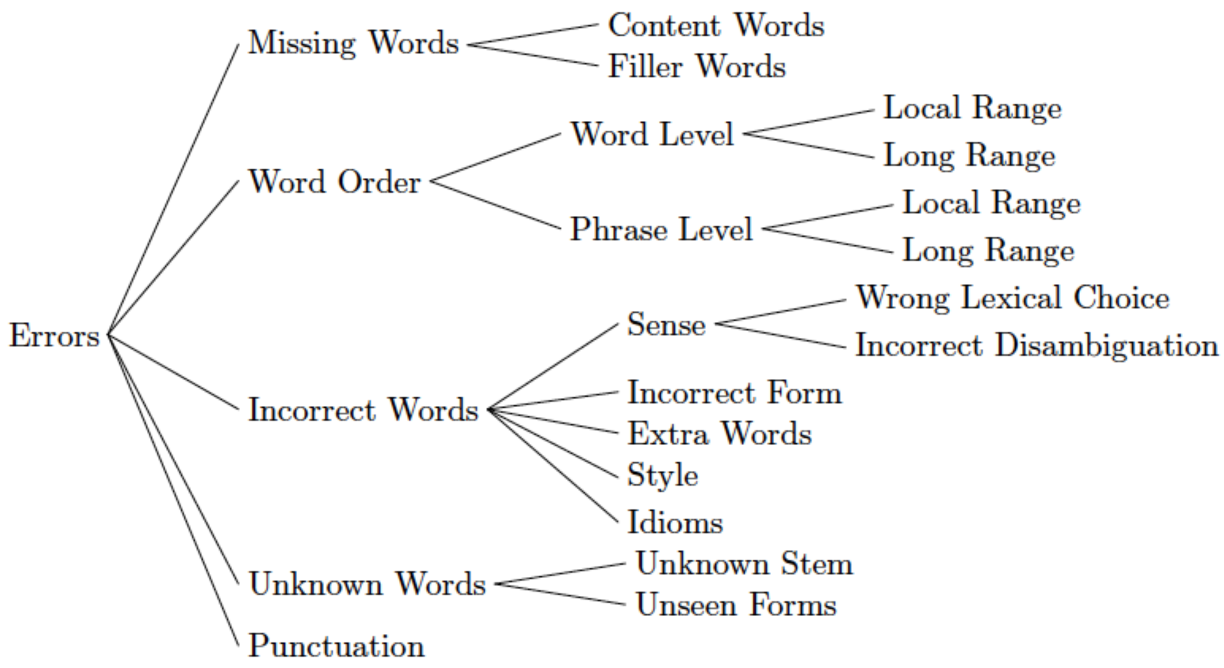
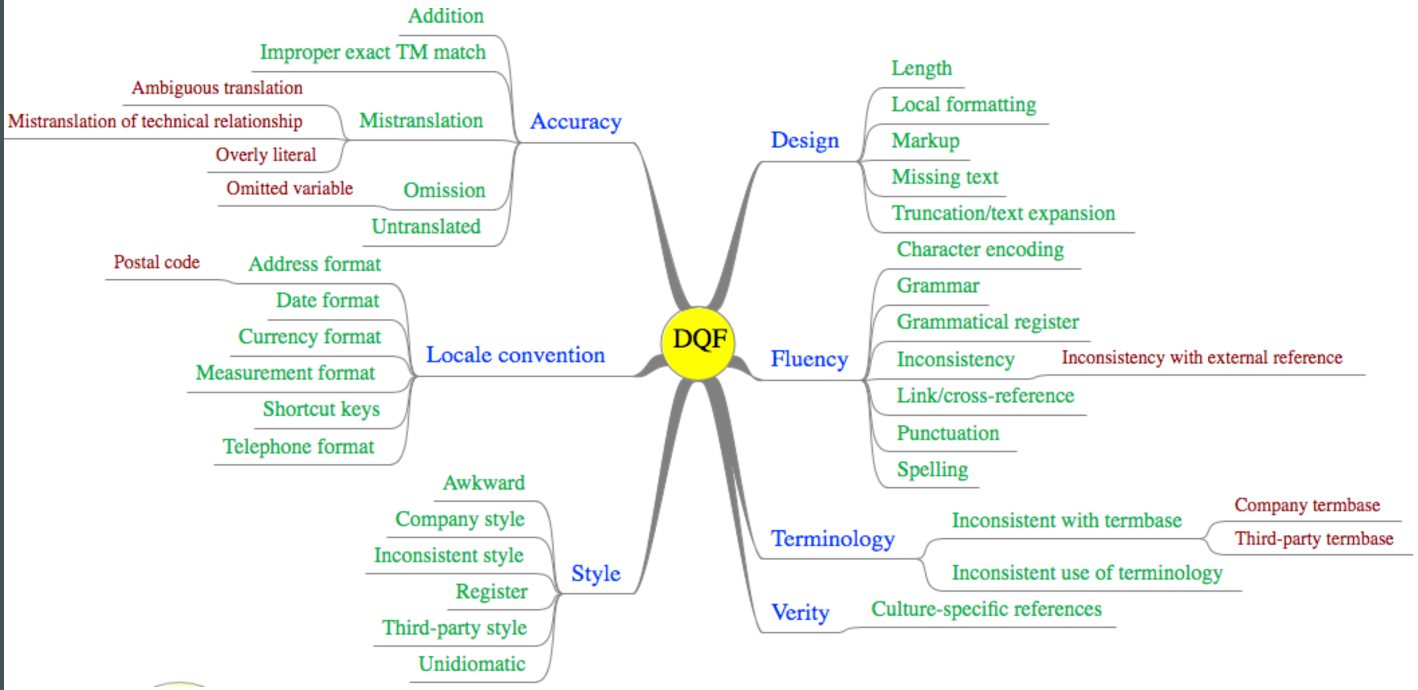


Figure 1: Classification of translation errors.

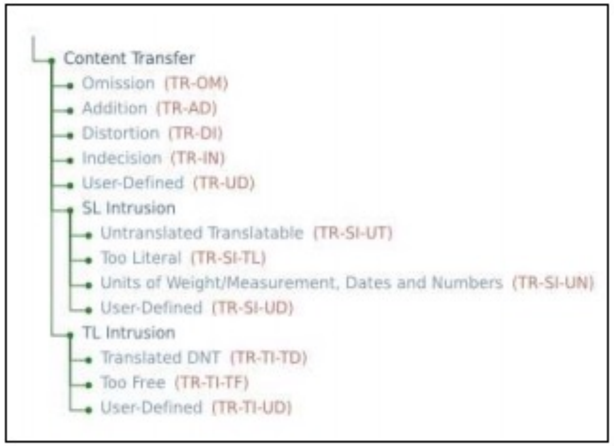
EXAMPLE: ACCOLÉ

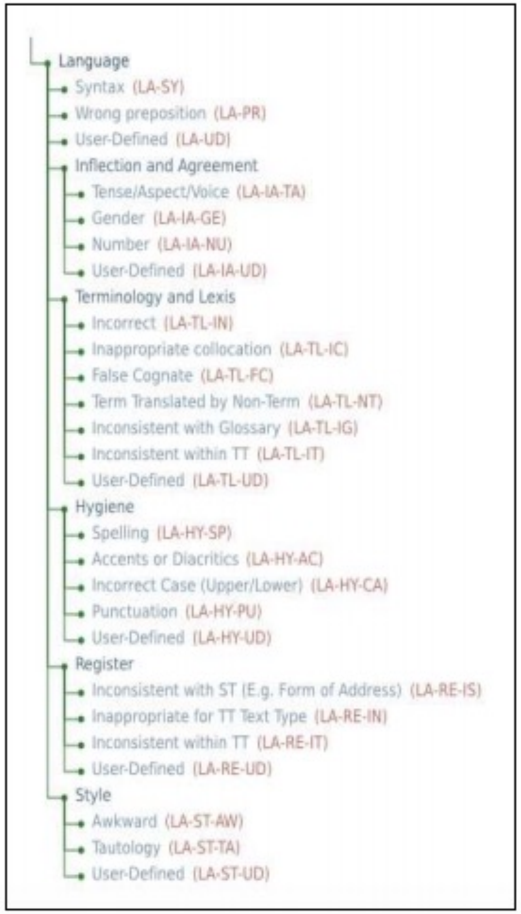
- Multidimensional Quality Metrics MQM-DQF



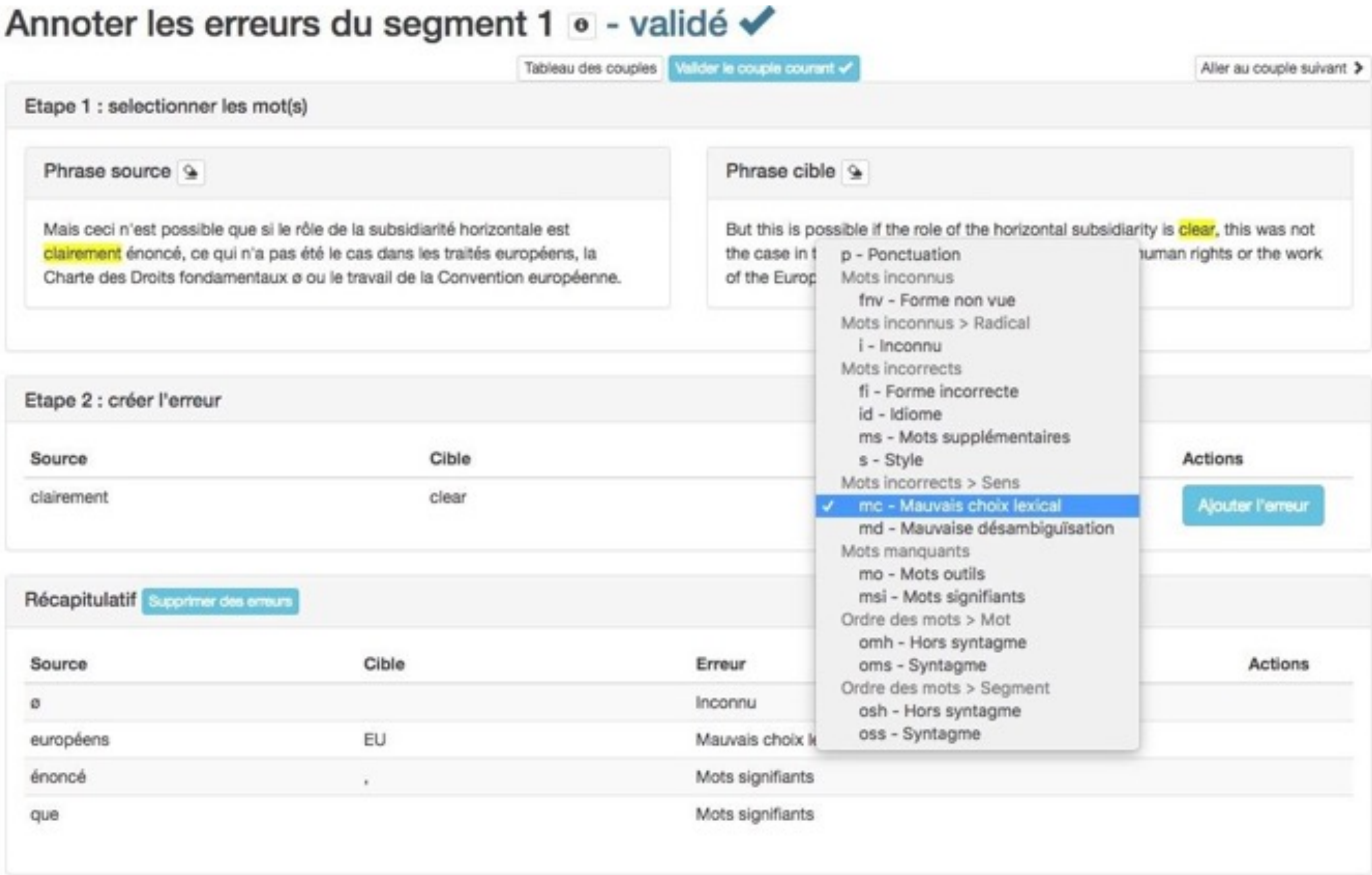
EXAMPLE: ACCOLÉ

- MeLLANGE Error Annotation Scheme (Castignoli, 2011)

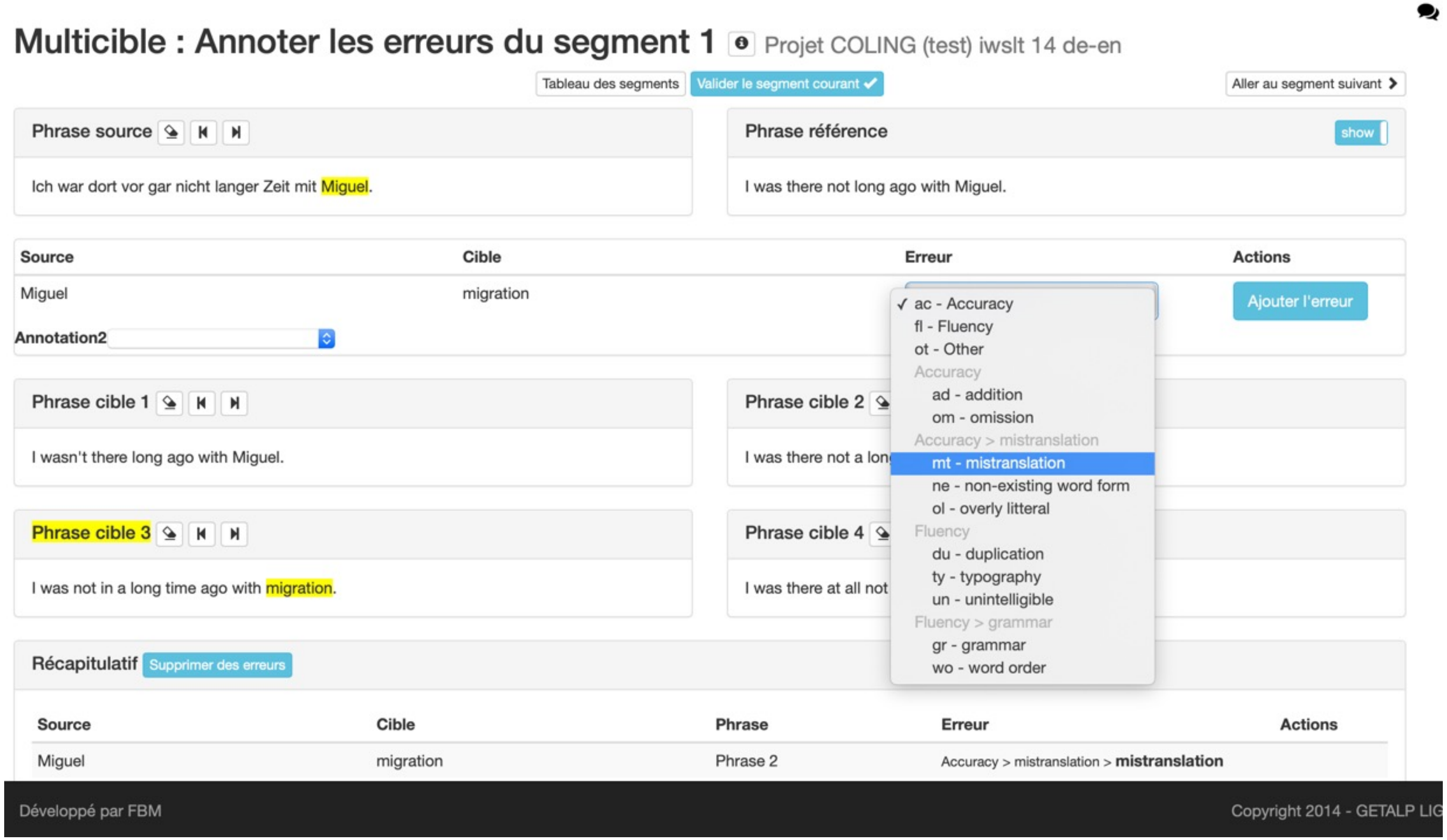




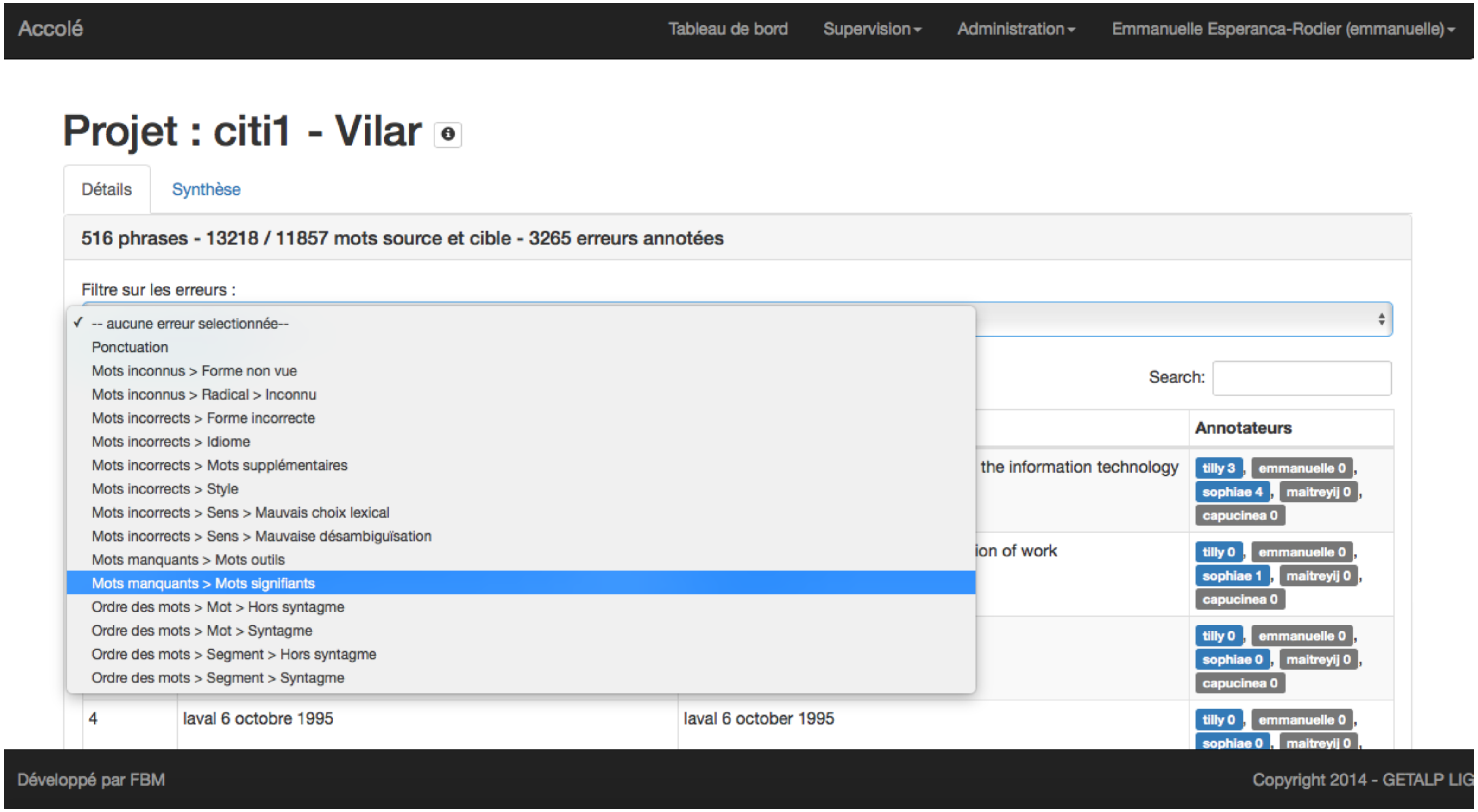
EXAMPLE: ACCOLÉ



EXAMPLE: ACCOLÉ



EXAMPLE: ACCOLÉ



SUBJECTIVE EVALUATION
FINAL REMARKS

PRO OF SUBJECTIVE EVALUATION

- Very informative

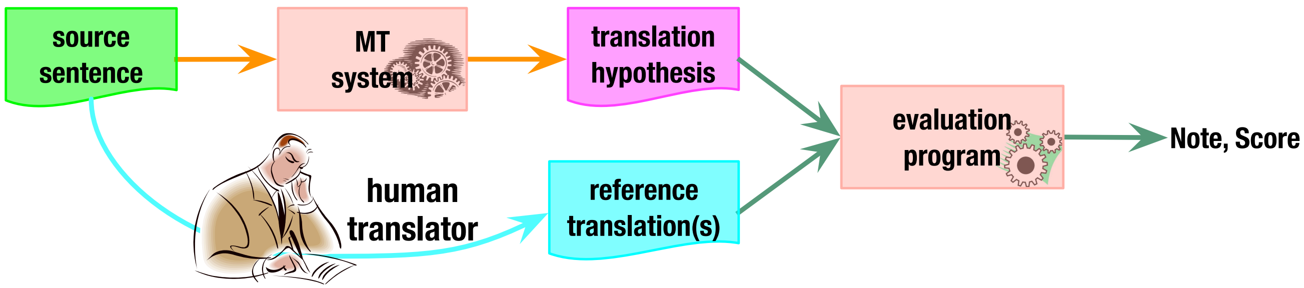
CONS OF SUBJECTIVE EVALUATION

- Labor-intensive & Time-consuming (Evaluators, Translators)
  - In practice, impossible for evaluation campaigns (subset or one-run evaluation organized as a shared task between participants)
- Not reusable
  - MT systems as dynamic components improving over time
  - Human assessment as a one-shot measure to be repeated
- Subjective
  - Evaluators' understanding of the guidelines
  - Evaluators' inter-agreement
  - Evaluators' intra-agreement
- Possibly partial
  - Mostly limited to fluency and adequacy
  - Difficulty to compare
    - E.g. fluency(SystA)<fluency(SystB) & adequacy(SystA)>adequacy(SystB) …
    - … Best(SystA, SystB) or Best(SystB, SystA)??????

OBJECTIVE EVALUATION

IDEAS

- Get rid of …
  - Subjectivity, Non-reusability, Slowness, Expensiveness
- How?
  - Take advantage of the reference(s) produced for subjective evaluation
  - Use a deterministic program to compare the hypothesis with reference(s)



IMPORTANT DATES

- 2002: BLEU [Papineni et al. 2002]
  - The beginning of objective evaluation measures
- Systems evaluation campaigns
  - 2001-: NIST Open MT
    - http://www.itl.nist.gov/iad/mig/tests/mt/
  - 2004-: IWSLT
    - Speech translation
    - http://iwslt2011.org/doku.php?id=14_related_events
  - 2006-: WMT
    - Broadcast news
    - http://www.statmt.org/wmt12/
- Metrics evaluation campaigns
  - 2008-: NIST MetricsMaTr
    - Metrics for Machine Translation Evaluation
    - http://www.nist.gov/itl/iad/mig/metricsmatr.cfm

IMPORTANT DATES

- The rough idea: lexical similarity
- Several measures*
  - Edit distance measures
    - WER, PER, TER
  - Precision-oriented measures
    - BLEU, NIST, WNM
  - Recall-oriented measures
    - ROUGE, CDER
  - Balancing precision & recall measures
    - GTM, METEOR, BLANC, SIA

*Incomplete because new measures are proposed every other day!!

EDIT DISTANCE MEASURES

The number of changes:

hypothesis → reference or acceptable translation

- **WER** (Word Error Rate) [Nießen et al., 2000]
  - Based on the Leveinstein distance: minimum number of substitutions, deletions, or insertions that have to be performed to convert the hypothesis into the reference
- **PER** (Position-independent Word Error Rate) [Tillmann et al., 1997]
  - A shortcoming of WER, PER compare the words in the hypothesis and reference without taking into account word order (bags of words)
- **TER** (Translation Edit Rate) [Snover et al. 2006] [Przybocki et al. 2006]
  - Operations performed by a post-editor to correct the hypothesis (insertion, deletion, substitution of words or sequences)

WER

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .
- Translation 4: the green house was right in front of the lake .

	WER ↓
T1	54.5455
T2	45.4545
T3	36.3636
T4	00.0000

WER

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Computation

```
REF:   the green house was right in front of the lake ***** .
HYP:   a   green house was ***** ** ***** by the lake shore .
EVAL:  S                   D     D  D     S           I
SHFT:
WER Score:   54,55 (   6,0/ 11,0)
```

WER

- Reference: the green house was right in front of the lake .
- Translation 1: the green house was by the lake shore .
- Computation

```
REF:   the green house was right in front of the lake ***** .
HYP:   the green house was ***** ** ***** by the lake shore .
EVAL:                      D     D  D     S            I
SHFT:
WER Score:  45,45 (  5,0/ 11,0)
```

WER

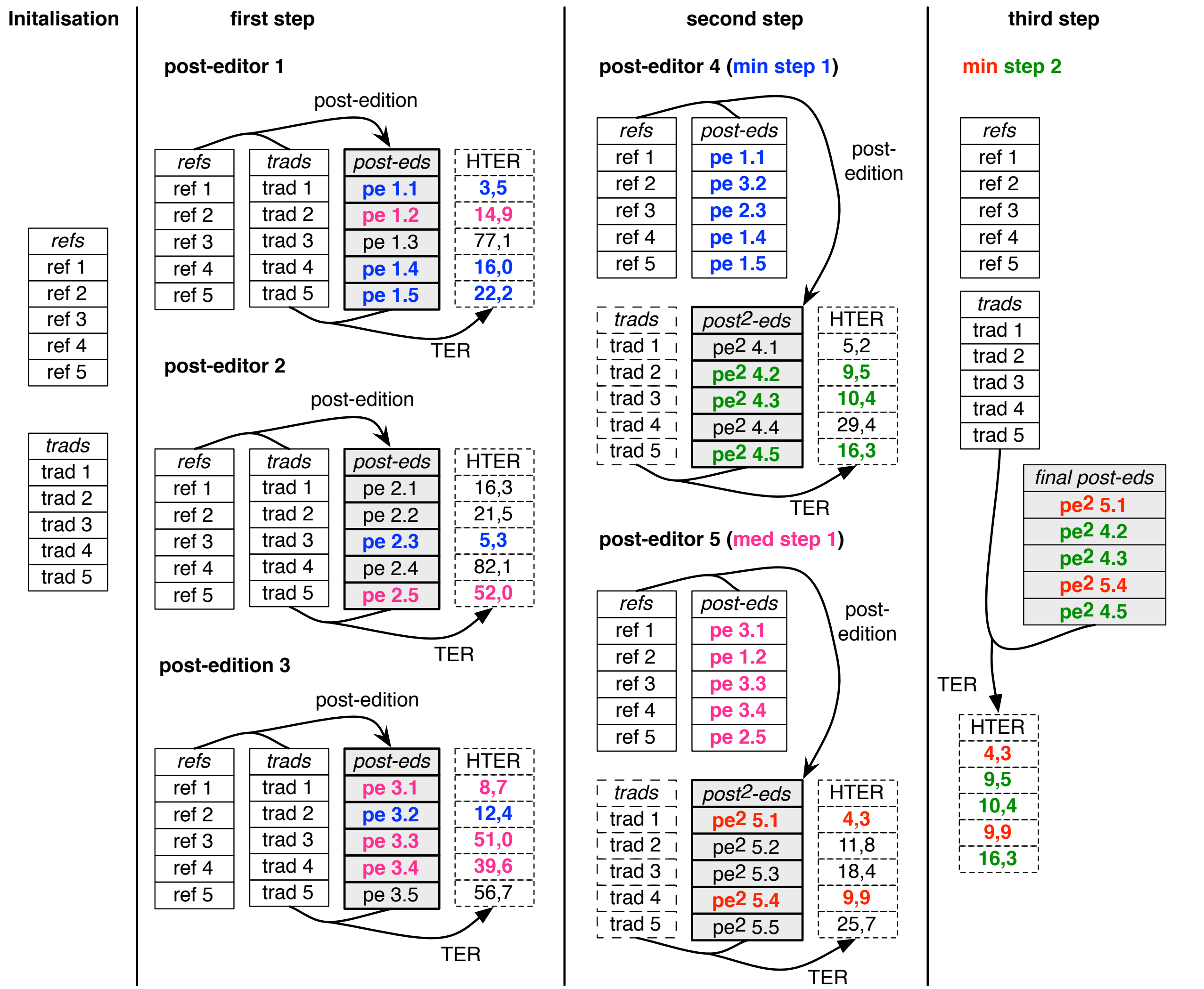
- Reference: the green house was right in front of the lake .
- Translation 1: the green potato right in front of the lake was right .
- Computation

```
REF:   the green house was    right in front of the lake *** ***** .
HYP:   the green ***** potato right in front of the lake was right .
EVAL:            D     S                                 I   I
SHFT:
TER Score:  36,36 (  4,0/ 11,0)
```

HTER

- GALE (global autonomous language exploitation) program (DARPA, 05-06)
  - develop and apply computer software technologies to absorb, translate, analyze, and interpret huge volumes of speech and text in multiple languages
  - evaluation for "go, no-go" funding
  - http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_(GALE).aspx

HTER


Initalisation
first step
second step
third step
post-editor 1
post-edition
refs
trads
post-eds
HTER
pe 1.1
pe 1.2
pe 1.3
pe 1.4
pe 1.5
3,5
14,9
77,1
16,0
22,2
TER
post-editor 2
pe 2.1
pe 2.2
pe 2.3
pe 2.4
pe 2.5
16,3
21,5
5,3
82,1
52,0
post-edition 3
pe 3.1
pe 3.2
pe 3.3
pe 3.4
pe 3.5
8,7
12,4
51,0
39,6
56,7
post-editor 4 (min step 1)
post-edition
pe 1.1
pe 3.2
pe 2.3
pe 1.4
pe 1.5
post2-eds
pe2 4.1
pe2 4.2
pe2 4.3
pe2 4.4
pe2 4.5
5,2
9,5
10,4
29,4
16,3
post-editor 5 (med step 1)
pe 3.1
pe 1.2
pe 3.3
pe 3.4
pe 2.5
pe2 5.1
pe2 5.2
pe2 5.3
pe2 5.4
pe2 5.5
4,3
11,8
18,4
9,9
25,7
min step 2
final post-eds
pe2 5.1
pe2 4.2
pe2 4.3
pe2 5.4
pe2 4.5
4,3
9,5
10,4
9,9
16,3
ref 1
ref 2
ref 3
ref 4
ref 5
trad 1
trad 2
trad 3
trad 4
trad 5

TER: EXAMPLES

- Source: a burglar broke into my room .
- Best Ref: un cambrioleur a forcé ma chambre .
- Orig Hyp: un cambrioleur est entré de force dans ma pièce .

```
REF:   un cambrioleur *** ****** ** a     forcé ma chambre .
HYP:   un cambrioleur est entré  de force dans  ma pièce   .
EVAL:                 I   I      I  S     S        S
SHFT:
```

✓ TER Score: 85,71 (6,0/ 7,0)

tercom.jar

TER: EXAMPLES

- Source: a man snatched my bag on the street .
- Best Ref: un homme a saisi mon sac dans la rue .
- Orig Hyp: un homme a saisi mon sac sur la rue .

```
REF:  un homme a saisi mon sac dans la rue .
HYP:  un homme a saisi mon sac sur  la rue .
EVAL:                          S
SHFT:
```

✓ TER Score: 10,00 (1,0/ 10,0)

TER: EXAMPLES

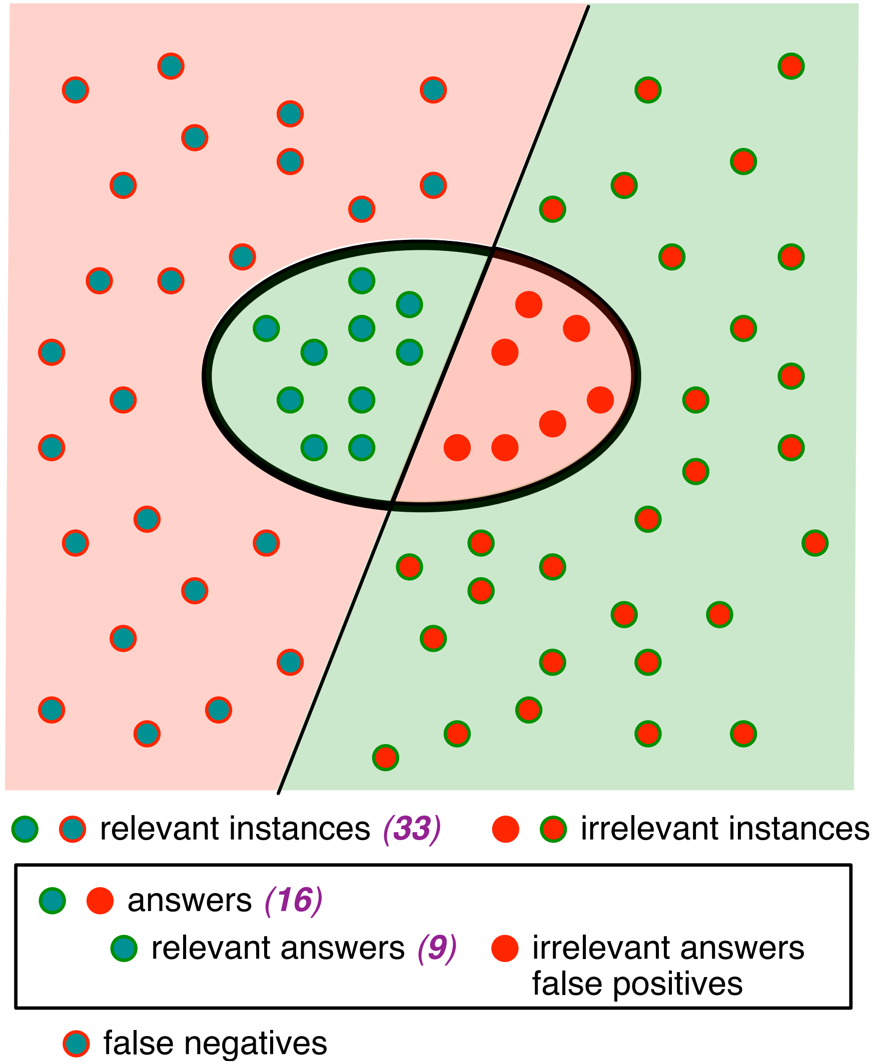
- **Source: a pickpocket took my wallet .**
- **Best Ref: un pickpocket a pris mon portefeuille .**
- **Orig Hyp: un pickpocket a pris mon portefeuille .**

```
REF:  un pickpocket a pris mon portefeuille .
HYP:  un pickpocket a pris mon portefeuille .
EVAL:
SHFT:
```

✓**TER Score: 0,00 (0,0/ 7,0)**

PRECISION AND RECALL

- Precision
  - fraction of retrieved instances that are relevant
- Recall
  - fraction of relevant instances that are retrieved
- *Example*



PRECISION-ORIENTED MEASURES

Proportion of lexical units (n-grams) in the hypothesis covered by the reference(s) translation

BLEU (Bilingual Evaluation Understudy) [Papinieni et al., 2001]

Modified precision (1 to 4 grams), geometric mean, brevity penalty

NIST [Doddington, 2002]

N-gram informativeness (1 to 5 grams), arithmetic mean, brevity penalty

WNM [Babych & Hartley, 2004]

Variant of BLEU which weights n-grams according to their statistical salience estimated out from a large monolingual corpus

BLEU: MODIFIED N-GRAM PRECISION

- Definition
  - Count the number of occurrences of each candidate n-gram in the hypothesis and count their maximum number of occurrences in the associated reference(s)
  - Clip the candidate n-gram counts by their maximum number in the associated reference(s)
  - Sum the clipped count for all n-grams and divide by the total number of candidate n-grams

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')}$$

BLEU: MODIFIED N-GRAM PRECISION

- Example 1 on unigrams
  - Hypothesis
    - it is a guide to action which ensures that the military always obeys the commands of the party .
  - References
    - it is a guide to action that ensures that the military will forever heed party commands . (2 “that”)
    - it is the guiding principle which guarantees the military forces always being under the command of the party . (4 “the”)
    - it is the practical guide for the army always to heed the directions of the party . (3 “the”)

BLEU: MODIFIED N-GRAM PRECISION

- Example 1 on unigrams (cont.)

Candidate words	Count	Max_ref_count	$Count_{clip}$
it	1	1	1
is	1	1	1
a	1	1	1
guide	1	1	1
to	1	1	1
action	1	1	1
which	1	1	1
ensure	1	1	1
that	1	2	1
military	1	1	1
always	1	1	1
obeys	1	0	0
the	3	4	3
commands	1	1	1
of	1	1	1
party	1	1	1
sum	18	/	17

$$P_1 = \frac{17}{18}$$

BLEU: MODIFIED N-GRAM PRECISION

- Example 2 on unigrams
  - Hypothesis
    - it is to insure the troops forever hearing the activity guidebook that party direct .
  - References
    - it is a guide to action that ensures that the military will forever heed party commands . (2 “that”)
    - it is the guiding principle which guarantees the military forces always being under the command of the party . (4 “the”)
    - it is the practical guide for the army always to heed the directions of the party . (2 “the”)

BLEU: MODIFIED N-GRAM PRECISION

- Example 2 on unigrams (cont.)

Candidate words	Count	Max_ref_count	$Count_{clip}$
it	1	1	1
is	1	1	1
to	1	1	1
insure	1	0	0
the	2	4	2
troops	1	0	0
forever	1	1	1
hearing	1	0	0
activity	1	0	0
guidebook	1	0	0
that	1	2	1
party	1	1	1
direct	1	0	0
sum	14	/	8

$$P_1 = \frac{8}{14}$$

BLEU: HYPOTHESES BREVITY PENALTIY

- definition
  - Hypothesis longer than references already penalized with modified precision $(\text{Countclip}/\text{Count})$
  - Need to penalize shorter hypotheses

  No penalty when the hypothesis length is the same as any reference

$$r = \sum_{\mathrm{C} \in \{candidats\}} \text{bst reference match for C}$$

  - let r be the test corpus' effective reference length

$$c = \sum_{\mathrm{C} \in \{candidats\}} \text{length of C}$$

  - let c be the total length of the hypothesis corpus
  - Brevity Penalty

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1 - r/c)}, & \text{if } c \leq r \end{cases}$$

BLEU: THE FORMULA

- BLEU is computed as follows:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

  - where
    - $N = 4$ and $w_n = 1/N$
  - $\text{BLEU} \in [0..1]$

BLEU: EXAMPLE

- Reference: the green house was right in front of the lake .
- Translation 0: the green house was right in front of the lake .

For N-Gram (green): 1
For N-Gram (house): 1
For N-Gram (was): 1
For N-Gram (right): 1
For N-Gram (in): 1
For N-Gram (front): 1
For N-Gram (of): 1
For N-Gram (the): 2
For N-Gram (lake): 1

For N-Gram (the green house): 1
For N-Gram (green house was): 1
For N-Gram (house was right): 1
For N-Gram (was right in): 1
For N-Gram (right in front): 1
For N-Gram (in front of): 1
For N-Gram (front of the): 1
For N-Gram (of the lake): 1

For N-Gram (the green): 1
For N-Gram (green house): 1
For N-Gram (house was): 1
For N-Gram (was right): 1
For N-Gram (right in): 1
For N-Gram (in front): 1
For N-Gram (front of): 1
For N-Gram (of the): 1
For N-Gram (the lake): 1

For N-Gram (the green house was): 1
For N-Gram (green house was right): 1
For N-Gram (house was right in): 1
For N-Gram (was right in front): 1
For N-Gram (right in front of): 1
For N-Gram (in front of the): 1
For N-Gram (front of the lake): 1

Precision 1-gram: 1.00 = 10/10
Precision 2-gram: 1.00 = 9/9
Precision 3-gram: 1.00 = 8/8
Precision 4-gram: 1.0 = 7/7
Weighted Precision: 1.00
Brevity Penalty: 1.00

BLEU = 1.00

BACK TO SUBJECTIVE EVALUATION

- Fluency evaluation for the 3 following translations

	Fluency
a green house was by the lake shore .	5
the green house was by the lake shore .	5
the green potato right in front of the lake was right .	5

Score	Fluency
5	Flawless English
4	Good
3	Non-Native
2	Disfluent
1	Incomprehensible

BACK TO SUBJECTIVE EVALUATION

- Adequacy evaluation given reference
  - the green house was right in front of the lake .

	Fluency
a green house was by the lake shore .	5~4
the green house was by the lake shore .	5
the green potato right in front of the lake was right .	1

Score	Adequacy
5	All information
4	Most
3	Much
2	Little
1	None

BLEU: EXAMPLE

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .

	WP	BP	BLEU
T1	0.000000	0.778801	**0.000000**
T2	0.411134	0.778801	**0.320191**
T3	0.555839	1.000000	**0.555839**

- ***Don't we have a problem!!!!***
  - T1 acceptable (one word changed compared to T2)
  - T3 wrong and nonsense

NIST: N-GRAM INFORMATION WEIGHT

- Definition
  - With BLEU all n-grams are equally important
  - NIST associate an information weight to each n-gram of the reference set

$$Info(w_1 w_2 \dots w_n) = \log_2 \left(\frac{\text{the \# of occurrences of } w_1 w_2 \dots w_{n-1}}{\text{the \# of occurrences of } w_1 w_2 \dots w_n} \right)$$

    - for a unigram w_1:

      the # of occurrences = the # of occurrences in the reference

NIST: HYPOTHESES BREVITY PENALTY

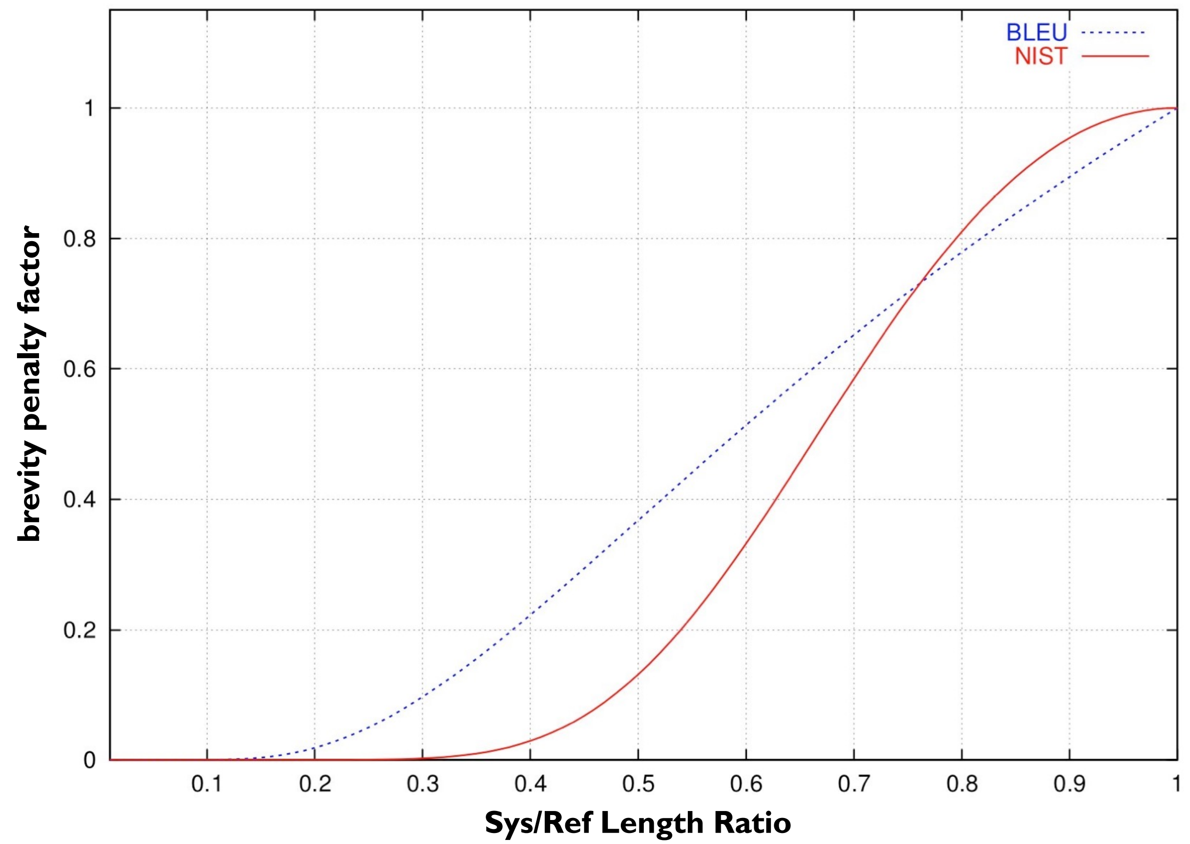
- Definition
  - New BP to minimize the impact on the score of small variations in the length of a translation
  - It reduces the contributions of length variations to the score for small variations

$$BP = \exp\left\{\beta \log^2\left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1\right)\right]\right\}$$

  - where
    - β is chosen to make the brevity penalty factor $= 0.5$ when the # of words in the system output is $2/3$ of the average # of words in the reference translation
    - $\bar{L}_{ref} =$ the average number of words in a reference translation, averaged over all reference translations
    - $L_{sys} =$ the number of words in the translation being scored

BLEU VS NIST: BREVITY PENALTY

- 0 < Hypo(Sys)/Ref Length Ratio ≤ 1



NIST: THE FORMULA

- NIST is computed as follows:

$$NIST = BP \cdot \sum_{n=1}^{N} \left\{ \frac{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{that co-occur}}} Info(w_1 \ldots w_n)}{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{in hypothesis}}} (1)} \right\}$$

  - Where
    - $N = 4$ at least

  - $NIST \in [0..+\infty[$ ($[0..15[$ *in practice*)

NIST: EXAMPLE

$$Info(w_1 w_2 \dots w_n) = \log_2\left(\frac{\text{the \# of occurrences of } w_1 w_2 \dots w_{n-1}}{\text{the \# of occurrences of } w_1 w_2 \dots w_n}\right)$$

- Reference: the green ~~house was~~ right in front of the lake . (11 1-grams)
- Translation 1: a green house was by the lake shore .
- Co-occurring n-grams
  - 1-grams: ‘the’, ‘green’, ‘house’, ‘was’, ‘lake’, ‘.’
  - 2-grams: ‘green house’, ‘house was’, ‘the lake’
    - the green ~~house was~~ right in front of the lake .
    - a green ~~house was~~ by the lake shore .
  - 3-gram: “green house was”
    - the green house was right in front of the lake .
    - a green house was by the lake shore .
- Info
  - Info(the)=$\log_2(11/2) = 2.4594$
  - Info(green)=Info(house)=Info(was)=Info(lake)=Info(.)= $\log_2(11/1) = 3.4594$
  - Info(green house)=Info(~~house was~~)= $\log_2(1/1) = 0.0000$
  - Info(the lake)= $\log_2(2/1) = 1.0000$
  - Info(green ~~house was~~)= $\log_2(1/1) = 0.0000$

NIST: EXAMPLE

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .

	NIST
T1	1.9579
T2	2.2940
T3	2.8980

- ***Don't we have a problem!!!!***
  - T1 acceptable (one word changed compared to T2)
  - T3 wrong and nonsense

MEASURES BALANCING RECALL AND PRECISION

- Precision & recall combination

F_1 score $F_1 = 2 \cdot \frac{P \cdot R}{P + R}$ F_β score $F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}$

- **GTM** (General Text Matcher) [Melamed et al., 2003; Turian et al., 2003]
  - F-measure; adjusted importance of n-grams matching
- **METEOR** [Banerjee & Lavie, 2005]
  - F-measure based on 1-gram alignment & word ordering; + stemming & synonymy through WordNet
- **BLANC** [Lita et al., 2005]
  - Family of trainable n-gram based metrics; variable size non-continuous word sequences
- **SIA** (Stochastic Iterative Alignment) [Liu & Gileda, 2006]
  - Loose sequence alignment enhanced with alignment scores, stochastic word matching and iterative alignment scheme

OBJECTIVE EVALUATION
FINAL REMARKS

PROS OF OB JECTIVE EVALUATION

- Costless
  - No! References have to be produced at some point!
- Objective
  - OK, always the same results with the same hypo & ref(s)
- Reusable
  - Always on the same test set (not a real-life situation)
  - Correlation between "translation improvement" & "score improvement"
- System optimization
  - *is it good or bad?*
- System comparison
  - *as far as they use the same development protocol! (cf. IWSLT 04)*

CONS OF OBJECTIVE EVALUATION

- System over tuning
  - When system parameters are adjusted toward the main evaluation metric
    - if it is BLEU then tune with BLEU, if it is NIST then tune with NIST
  - Several metrics are used for ranking
- Blind system development
  - When metrics are unable to capture system improvements
- Unfair system comparison
  - When metrics are unable to reflect differences in quality between MT systems
  - When systems are based on different paradigms (SMT vs. RBMT) *(cf. IWSLT 2004)*
- No utility, nor usability evaluation yet

CONCLUSION

TO BE REMEMBERED

- On BLEU [Callison-Burch et al., 2006]
  - Under some circumstances, an improvement in BLEU is *not sufficient* to reflect a genuine improvement in translation quality
  - Under other circumstances that it is *not necessary* to improve BLEU in order to achieve a noticeable improvement in translation quality

- To be transposed to all other objective metrics!

EXTERNAL VS INTERNAL MEASURES

- External measures
  - linguistic criteria: grammaticality, fidelity…
  - usage criteria: productivity, cost, delay…
  - conflict between linguistic & usage criteria
    - ex: Systran, Euratom, ISPRA: 2/20 (linguistic quality) — 18/20 (usability)
- Internal measures
  - system design: linguistic & computational architecture
  - perspectives of improvements: quality, coverage
  - ease of extension to
    - new languages
    - new document types
    - new tasks (assimilation → dissemination)

CLASSIFICATION OF EXTERNAL MEASURES

- Measures related to the task
  - High-quality written communication

    two tasks: acquisition (from one language source), diffusion (to one target language)
    - Produce a professional quality translation
    - reduction of costs (human labor) and delays
  - Spoken communication
    - Help two people to conduct a bilingual dialogue to accomplish a task
    - The accomplishment of the task
  - Comprehension, understanding of written material
    - Translate Web pages, newspapers, and e-commerce services so that end users can understand the information in foreign languages and act accordingly
    - number of purchases per visited page in e-commerce, time spent reading newspapers page (objectives measures)
    - user feedback, answers to customer questionnaires (subjective measures)

CLASSIFICATION OF EXTERNAL MEASURES

- ❖ Measures related to the task *(cont.)*
  - ➢ Comprehension, understanding of spoken material

    the typical task is to follow a monologue (speech, Parliament, etc.). or a dialogue in a foreign language (television, intelligence)

    - ■ Produce as much information as possible
    - ✧ determine the level of understanding
      - ✧ objective measure: time to complete the task, MCQ about the content
      - ✧ subjective measures: the sense of understanding, the judgment of fluidity

CLASSIFICATION OF EXTERNAL MEASURES

- Measures non-related to the task
  - with references
    - adequacy *a la* NIST
    - fidelity *a la* JEIDA or FEMTI
    - informativeness *a la* ALPAC
  - without references
    - fluidity *a la* NIST
    - adequacy through MCGQ *a la* TOEFL or TOEIC

PROPOSAL

Use only cheap task-related measures for external evaluation!

- **MT for written input**
  - **Diffusion**
    - objective usability measures
      - time spend for post-edition, correction of raw MT output
      - **Relative Efficiency**:

$$\text{Relative Efficiency}_{MT} = \frac{\text{Time}_{Human}}{\text{Time}_{MT+Human}}$$

  - an MT system may be considered efficient if its relative efficiency is > 2 (upper bound of the gain with a translation memory)

- subjective measure such as fluency or adequacy are useless and counterproductive
  - corrections made easy by the environment *(cf. "is admission fee how much?"*

PROPOSAL

- **MT for written input**
  - **Acquisition, understanding**
    - Web pages
      - compare reading time translated Web page vs reading time original Web page
        - if shorter: very bad translation
        - if longer: bad translation but usable for some understanding
        - if equal: quality OK of the use
      - Multiple Choice Questions

PROPOSAL

- **MT for spoken input**
  - **Diffusion**
    - MCQ for understanding
  - **Acquisition, Understanding**
    - MCQ but hard for dialogue

FINAL WORDS

- External methods for evaluating MT systems define various measures based on MT results and their usage.
- While operational systems are mostly evaluated since long by task-based methods, evaluation campaigns of the last years use (parsimoniously) quite expensive subjective methods based on unreliable human judgments, and (for the most part) methods based on reference translations, that are impossible to use during the real usage of a system, less correlated with human judgments when quality increases, and totally unrealistic in that they force to measure progress on fixed corpora, endlessly retranslated, and not on new texts to be translated for real needs.
- There are also numerous biases introduced by the desire to diminish costs, in particular the usage of parallel corpora in the direction opposed to that of their production, and of monolingual rather than bilingual judges.
- We propose to abandon the reference-based methods in external evaluations and to replace them with strictly task-based methods while reserving them for internal evaluations.

BIBLIOGRAPHIE

- ALPAC (1966). Language and Machine: Computers in Translation and Linguistics. n. 1416. Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Science - National Research Council. Washington, D. C. November 1966. 138 p.
- Babych, B., & Hartley, A. (2004). Extending BLEU MT Evaluation Method with Frequency Weightings. Proceedings of ACL 2004. Barcelona, Spain. July 21-26, 2004. pp. 622-629.
- Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgement. Proceedings of ACL-05, Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA. June 29, 2005. pp. 25-32.
- Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. Proceedings of ACL-2006. Trento, Italy. April 3-7, 2006. pp. 249-256.
- Castagnoli S., Ciobanu D., Kübler N., Kunz K., Volanschi A. (2011). Designing a Learner Translator Corpus for Training Purposes. Dans Corpora, Language, Teaching and Resources: From Theory to Practice. Édité par Kübler N. Bern: Peter Lang.

BIBLIOGRAPHIE

- Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proceedings of HLT 2002. San Diego, California. March 24-27, 2002. pp. 138-145.
- EAGLES-EWG (1996). EAGLES – Evaluation of Natural Language Processing Systems. Final Report EAG-EWG-PR.2, Project LRE-61-100. Center for Sprogteknologi. Copenhagen, Denmark. October, 1996. 287 p.
- EAGLES-EWG (1999). EAGLES – Evaluation Working Group. Final Report EAG-II-EWG-PR.2, Project LRE-61-100. Center for Sprogteknologi. Copenhagen, Denmark. April, 1999. 173 p.
- Esperança-Rodier, E., Brunet-Manquat, E., Eady, S. (2019) ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. Translating and the computer 41, Nov 2019, Londres, United Kingdom. ⟨hal-02363208⟩
- Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. Machine Translation, 17(1): pp. 43-75.

BIBLIOGRAPHIE

- JEIDA (1989). *A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A.* Japan Electronic Industry Development Association. Tokyo, Japan. July, 1989. 197 p.
- JEIDA (1992). *JEIDA Methodology and Criteria on Machine Translation Evaluation.* Japan Electronic Industry Development Association. Tokyo, Japan. November, 1992. 129 p.
- King, M., Popescu-Belis, A., & Hovy, E. (2003). *FEMTI: creating and using a framework for MT evaluation*. Proceedings of MT Summit IX. New Orleans, USA. September 23-27, 2003. 8 p.
- Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. Tradumàtica, pages 0455–463.

BIBLIOGRAPHIE

- Lita, L.V., Rogati, M., & Lavie, A. (2005). *BLANC: learning evaluation metrics for MT*. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, B.C., Canada. October 6-8, 2005. pp. 740-747.
- Liu, D., & Gildea, D. (2006). *Stochastic Iterative Alignment for Machine Translation Evaluation*. Proceedings of COLING-ACL. Sydney, Australia. 17-21 July, 2006. pp. 539-546.
- Melamed, I. D., Green, R., & Turian, J. P. (2003). *Precision and Recall of Machine Translation*. Proceedings of HLT-NAACL 2003 - short papers. Edmonton, Canada. May 27 - June 1, 2003. pp. 61–63.
- Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research*. Proceedings of LREC 2000. Athens, Greece. 31 May - 2 June, 2000. pp. 39-45.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of ACL-02. Philadelphia, USA. July 7-12, 2002. pp. 311-318.
- zybocki, M., Sanders, G., & Le, A. (2006). Edit Distance: A Metric for Machine Translation Evaluation. Proceedings of LREC 2006. Genoa, Italy. May 24-26, 2006. pp. 2038-2043.

BIBLIOGRAPHIE

- PrSnover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of AMTA 2006. Cambridge, MA, USA. August 8-12, 2006. pp. 223-231.
- Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP-based search for statistical translation. Proceedings of Fifth European Conference on Speech Communication and Technology (EUROSPEECH'97). Rhodos, Greece. September 22-25, 1997. pp. 2667-2670.
- Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. Proceedings of MT Summit IX. New Orleans, USA. September 23-27, 2003. pp. 386-393.
- Vilar D., Xu J., D'Haro L. F., Ney H. (2006). Error Analysis of Statistical Machine Translation Output. Actes LREC, , Genoa, Italy, 697-702
- White, J. S., O'Connell, T., & O'Mara, F. E. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons and Further Approaches. Proceedings of Technology Partnerships for Crossing the Language Barrier (the First Conference of the Association for Machine Translation in the Americas). Columbia, Maryland, USA. October 5-8, 1994. pp. 193-205.