Context-based Evaluation of MT Systems: Principles and Tools

Tutorial @ MT Summit XI, Copenhagen, 10.9.2007

Paula Estrella, Andrei Popescu-Belis and Maghi King

ISSCO, School of Translation and Interpreting
University of Geneva

UNIVERSITÉ DE GENÈVE

ISSCO
University of Geneva

Evaluation of MT software: two views

- MT researchers and developers
  - focus on the core functionality of their system, i.e. quality of MT output, in a given domain
- MT users / buyers
  - consider also other qualities of MT output
    - e.g. terminological correction
  - are also sensitive to a larger range of qualities
    - core functionality remains important
    - plus: speed, adaptability, user-friendliness, etc.

    ➔ indicators of quality depend on the context of use

Goal of this tutorial

- Outline a model for context-based evaluation, applied to MT systems

- Introduce a tool that automates the design of context-dependent evaluation plans: FEMTI

- Apply the model and tool to design a simple evaluation plan for a given scenario of use

Overview of the tutorial

1. Principles

   1.1. Role of the context of use in MT evaluation

   1.2. ISO standards for software evaluation:
   terminology and role of context of use

   1.3. FEMTI guidelines: theoretical model

2. Tools

   2.1. Implementation of FEMTI interfaces

   2.2. Use of FEMTI by evaluators & evaluation experts

3. Practical application

The increasing importance of « utility » for the evaluation of MT systems (1/3)

- "Good applications for crummy MT" (Church & Hovy 1993)
  - quality of MT systems can have various aspects
    - e.g., translation speed or ease of dictionary update
  - quality of MT output (translated text) can itself be decomposed
    - e.g. translation of technical terms, correctness of punctuation
  - the relative importance of these parameters varies with the intended use of an MT system

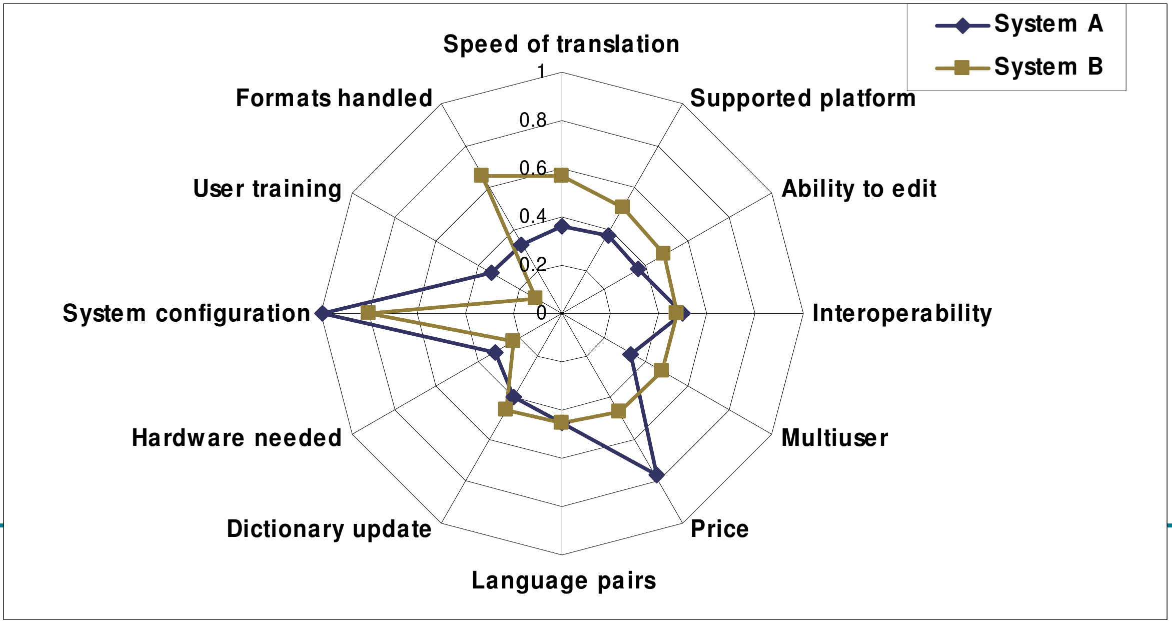
- OVUM Report (Mason & Rinsche 1995)
  - comparison of commercial MT systems by decomposing "quality" on a dozen dimensions

The increasing importance of « utility » for the evaluation of MT systems (2/3)

- Task-based quality metrics (White & Taylor 1998)
  - quality of MT systems for a given use of their output can be measured by assessing the performance of humans using MT output to accomplish a specific task
    - e.g. automatic summarization, or document classification
  - required quality levels vary with the task

The increasing importance of « utility » for the evaluation of MT systems (3/3)

- JEIDA Report (Nomura & Isahara 1992)
  - objective: to characterize the *intended context of use* and the *performance* of an MT system
  - two radar charts with 7 dimensions
    - can be matched to indicate how relevant is an MT system in a given context



ISLE Evaluation Work Group

- ISLE Project : *International Standards for Language Engineering*
  - EU, Switzerland, USA (1999-2002)
  - Evaluation Work Group
    - http://www.issco.unige.ch/projects/isle/ewg.html

- Achievements
  - apply the EAGLES guidelines for NLP evaluation to MT
  - normalize MT evaluation in a comprehensive framework
  - ensure compatibility with the ISO/IEC standards for software evaluation

➔ First proposal of FEMTI

- *Framework for the Evaluation of Machine Translation in ISLE*

Overview of the tutorial

1. Principles
   - 1.1. Role of the context of use in MT evaluation
   - 1.2. ISO standards for software evaluation: terminology and role of context of use
   - 1.3. FEMTI guidelines: theoretical model
2. Tools
   - 2.1. Implementation of FEMTI interfaces
   - 2.2. Use of FEMTI by evaluators & evaluation experts
3. Practical application

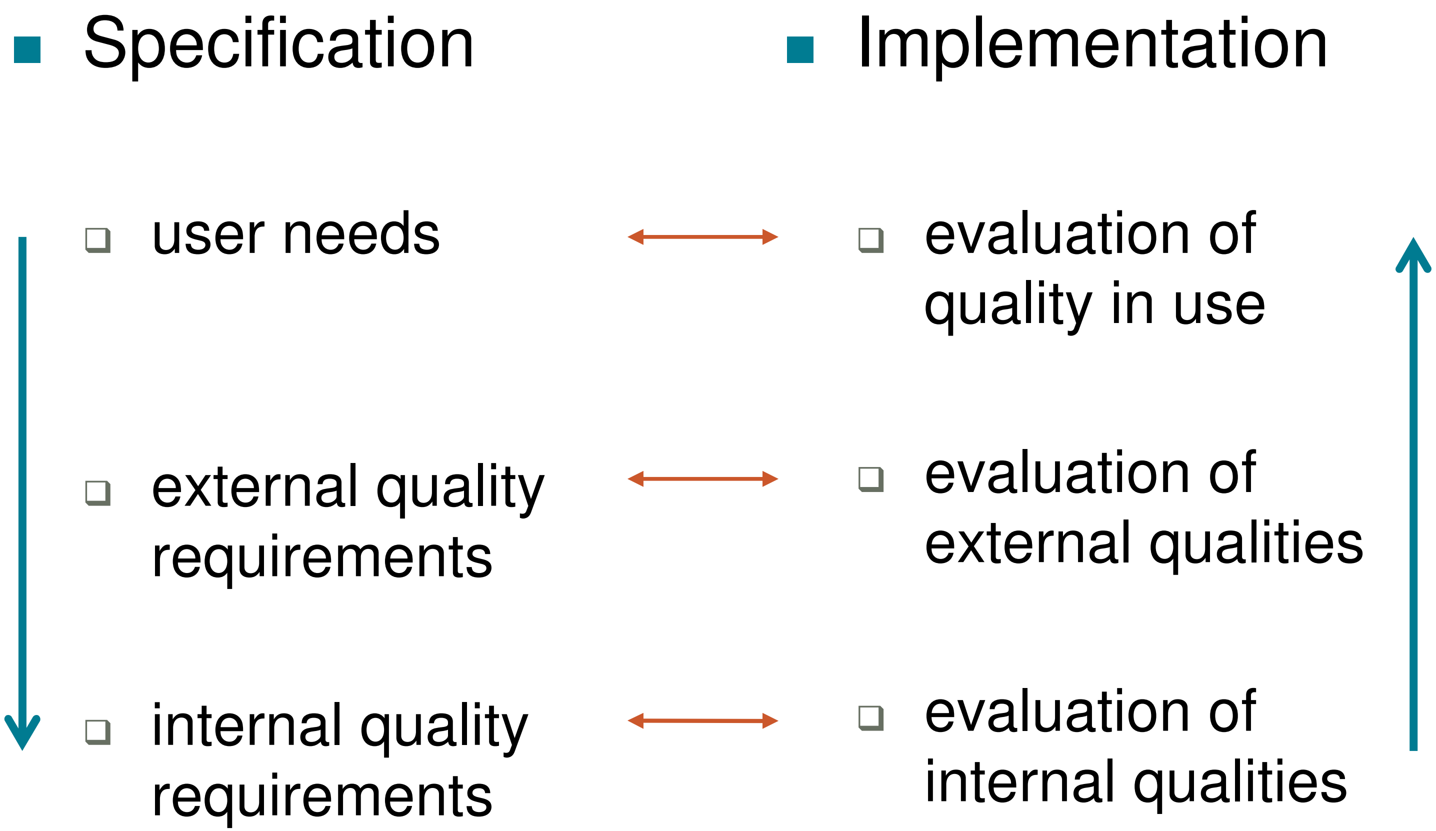
ISO/IEC Standards on software evaluation

- ISO 14958 – product evaluation process
  - quality in the software life cycle
  - process for developers, acquirers and evaluators
- ISO 9126 – product quality
  - model for software product quality
  - defines six main quality characteristics
    - functionality, reliability, usability, efficiency, maintainability, portability
  - further subdivided into subcharacteristics
    - terminal nodes of this hierarchy (*quality model*) can be measured using internal or external metrics

ISO/IEC Standards on software evaluation: list

- ISO/IEC 9126 – quality models
  - first version : 1991
  - 9126-1 (2001) : quality models overview
  - 9126-2 (2003) : internal quality characteristics
  - 9126-3 (2003) : external quality characteristics
  - 9126-4 (2004) : characteristics of quality in use
- ISO 14958 – evaluation process
  - 14958-1 (1999) : overview
  - 14958-2 (2000) : planning and management of the process
  - 14958-3 (2000) : process for developers
  - 14958-4 (1999) : process for acquirers
  - 14958-5 (1998) : process pour evaluators
  - 14958-6 (2001) : documentation of process

Software development life cycle: role of quality and evaluation



- Specification
  - user needs
  - external quality requirements
  - internal quality requirements
- Implementation
  - evaluation of quality in use
  - evaluation of external qualities
  - evaluation of internal qualities

Influence of the intended context of use on evaluation procedures in ISO/IEC standards

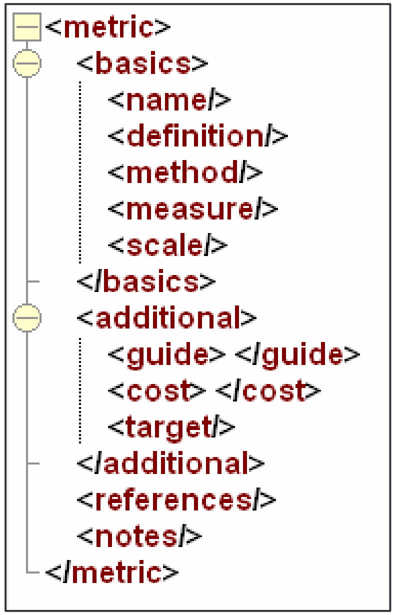
- Visible in standards for quality in use, measured in context

- ISO 14598-4
  - required system integrity: the higher this is, the more complete the evaluation should be

- ISO 14598-5
  - "evaluation levels" should be related to level of risks (4-point scale) resulting from system malfunction
    - risks to environment, safety of persons, installations, data
    - more demanding evaluation procedures should be applied when the level of risk is higher

Evaluation metrics in ISO/IEC standards (1/2)

- ISO/IEC 14598:
  - “a measurement is the use of a metric to assign a value (i.e., a measure, be it a number or a category) from a scale to an attribute of an entity”
- ISO/IEC standards provide
  - explanation of how to apply metrics
  - a basic set of metrics for each sub-characteristic
  - examples of application during software lifecycle
- Normalized description of metrics
  - name, method, measure, scale, target audience, notes



```
<metric>
  <basics>
    <name/>
    <definition/>
    <method/>
    <measure/>
    <scale/>
  </basics>
  <additional>
    <guide> </guide>
    <cost> </cost>
    <target/>
  </additional>
  <references/>
  <notes/>
</metric>
```

Evaluation metrics in ISO/IEC standards (2/2)

- Internal metrics (9126-3:2003)
  - Measure quality of intermediate deliverables
  - example: reliability – maturity: number of mistakes removed during design/coding
- External metrics (9126-2:2003)
  - Measure derived indirectly from its behaviour
  - example: reliability – maturity: number of mistakes removed during testing
- Quality in use metrics (9126-4:2004)
  - measure whether product meets specifications by user
  - (not covered below by FEMTI)

Overview of the tutorial

1. Principles
   - 1.1. Role of the context of use in MT evaluation
   - 1.2. ISO standards for software evaluation:
     terminology and role of context of use
   - 1.3. FEMTI guidelines: theoretical model
2. Tools
   - 2.1. Implementation of FEMTI interfaces
   - 2.2. Use of FEMTI by evaluators & evaluation experts
3. Practical application

FEMTI: Context-based evaluation

- FEMTI: Framework for the Evaluation of Machine Translation in ISLE
  - ISLE : *International Standards for Language Engineering* European project (1999 – 2002)
  - collected & structured knowledge from the MT community
  - 100+ evaluation metrics, from over 30 years
  - Developed, based on ISO/IEC standards
    - classification of contexts of use
    - classification of quality characteristics

    → Context based evaluation guidelines

ISLE workshops on MT evaluation

(URLs available at http://www.issco.unige.ch/projects/isle/ewg.html)

- "Hands-on MTEval" @ AMTA 2000
  - first presentation of ISLE taxomonies
- "MTEval: An invitation to get your hands dirty!“ @ UniGe/ETI, avril 2001
  - experiments with the taxonomies
- "MTEval: Who did what to whom?" @ NAACL, June 2001
  - experiments with task-based evaluation
- "MTEval at MTSummit VIII" @ MT Summit VIII, September 2001
  - reports and analyzes of previous evaluations
- "MTEval: human evaluators meet automated metrics“ @ LREC, May 2002
  - experiments with correlations between metrics on human translations and MT output
- "MTEval: expert sessions" @ USC/ISI, février 2003
  - update and stabilize the taxonomies
- MTEval @ MTSummit IX, September 2003
  - presentations of work based on the FEMTI guidelines

Classifications based on ISO/IEC standards

ISO generic quality characteristics

- functionality
- reliability
- usability
- efficiency
- maintainability
- portability

Quality characteristics particular to MT

- Functionality
  - Suitability
    - Accuracy
      - Fidelity
      - Consistency
      - Terminology
      - … … …

Definitions adapted for MT from ISO/IEC 9126

- Context of use
  - Environment where the system is to be used
    - FEMTI: classification of the characteristics of the translation *task / user / input* (along with the *purpose* and *object* of the evaluation)
    - Examples: document routing, email translation, information extraction

- Quality characteristics
  - Attributes that constitute *software quality*
    - FEMTI: classification of *MT software* quality characteristics
    - Examples: fidelity, readability, terminological correctness, speed

- Quality model
  - Quality characteristics + related metrics
  - Depends on the intended context of use

Contents of the FEMTI guidelines

- Part 1 : **contexts of use** + relation to quality models
  - defines types of tasks, users, input data
  - helps to specify the needs for an MT system and relates them to required qualities
- Part 2 : **quality models** + relation to metrics
  - particularizes the six ISO/IEC quality characteristics
    - functionality, reliability, usability, efficiency, maintainability, portability
  - suggests metrics for each quality characteristic
    - references to studies of metrics
    - recommendations for choosing metrics

Design of a context-based evaluation plan using the FEMTI guidelines

1. Define the intended context of use (Part 1) of the system
   - ❑ task, user, input data

2. Select the relevant quality characteristics (Part 2) and attributes
   - ❑ among those that apply to the system's type and task
   - ❑ including relative importance
   - ➔ quality model

3. Select appropriate metrics for each attribute
   - ❑ drawn from the literature, or new ones
   - ❑ define what counts as an acceptable score

- ■ NB: This is a part of the larger ISO/IEC evaluation process

Example 1: contextual evaluation of an MT system for instant messaging

- Task
  - Communication
    - Synchronous
- User
  - Non specialist
  - No knowledge of TL
- Type of input
  - Document type
    - colloquial messages
    - not domain-specific

→

- Functionality
  - readability
  - fidelity
  - grammar
  - punctuation
- Efficiency
  - speed
- Reliability
  - (low) crashing frequency

« Part 1 » « Links » « Part 2 »

Example 2: contextual evaluation of an MT system for routing of multilingual patents

- Task
  - ❑ Assimilation
    - Document routing
- User
  - ❑ Specialist
  - ❑ Knowledge of TL
- Type of input
  - ❑ Doc. type
    - patent-related doc.
  - ❑ Author type
    - domain specialist

- Functionality
  - ❑ accuracy
    - terminological correctness
  - ❑ readability
  - ❑ style
- Amount of linguistic resources
  - ❑ size/type of dictionaries
- Maintainability
  - ❑ Changeability
    - Ease of dictionary updating

« Part 1 » « Links » « Part 2 »

Overview of entire evaluation process: based on ISO/IEC + role of FEMTI (1/2)

1. Preliminary considerations **[FEMTI 1.1]**
   - objective of evaluation
   - stakeholders – do they all have the same objective?
   - object of evaluation – how is it accessible?
2. Define the context of use of the system **[FEMTI 1.2-1.4]**
   - what are the tasks the system is aimed for?
   - who are its potential users?
   - which types of texts will have to be translated?
3. Define the required quality characteristics **[FEMTI 2]**
   - i.e., the quality models resulting from (1) and (2)
   - list quality characteristics with their relative importance
     - possibly decomposed into elementary characteristics

Overview of entire evaluation process: based on ISO/IEC + role of FEMTI (2/2)

- 4. Specification of the evaluation [FEMTI 1+2]
  - decompose all qualities into elementary ones (attributes)
  - select metrics for each attribute
  - define assessment criteria for each metric: how will the measured value be transformed into a score? what are the acceptable values? how will the scores be aggregated?
- 5. Design of the evaluation
  - write *evaluation plan*: summarize previous points, state how metrics will be applied [FEMTI 2 + literature], assign responsibility to persons
- 6. Execution of the evaluation
  - follow the evaluation plan, then write preliminary version of evaluation report
- 7. Conclusion
  - formulate results in response to evaluation objectives
  - write final version of report

Overview of the tutorial

1. Principles

    1.1. Role of the context of use in MT evaluation

    1.2. ISO standards for software evaluation:
    terminology and role of context of use

    1.3. FEMTI guidelines: theoretical model

2. Tools

    2.1. Implementation of FEMTI interfaces

    2.2. Use of FEMTI by evaluators & evaluation experts

3. Practical application

Requirements for a support tool

- Help evaluators characterize the intended context of use as a set of features
  - list known context characteristics

- Help evaluators define a quality model for evaluation
  - list known quality characteristics, attributes and metrics
  - **depending on the selected context characteristics, suggest relevant quality characteristics**

- Generate a summary of context and quality model

How are the relevant quality characteristics computed from context characteristics?

- FEMTI stores an a priori list of relevant quality characteristics for each context characteristic

  ➔ GCQM = generic contextual quality model

- When an evaluator selects a list of context characteristics, all the relevant quality characteristics are combined and ranked

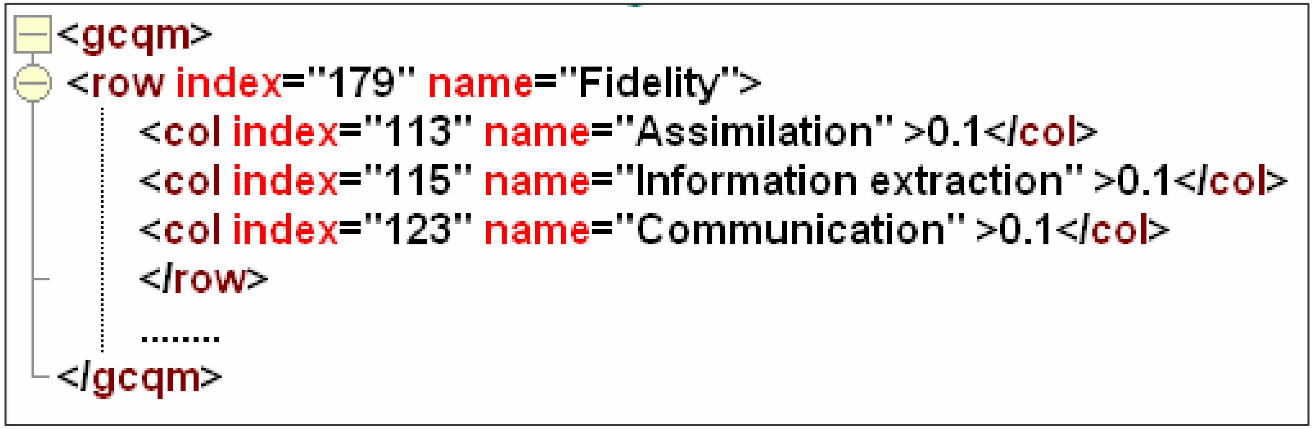
  ➔ FEMTI proposes a quality model,
  to be adapted by evaluators

How is the GCQM constructed?

- Expert interface
  - allows experts to build and view individual GCQMs
  - experts indicate for each context characteristic what are the most relevant quality characteristics
- The global FEMTI GCQM
  - synthesis of individual GCQMs

Generic Contextual Quality Model (GCM)

- Defined as matrix of weights for each couple (CC_i, QC_j)
  - Initially set with equal weights, no dependency between context characteristics
    - Ex. Translation task / MT user



```
<gcqm>
 <row index="179" name="Fidelity">
     <col index="113" name="Assimilation" >0.1</col>
     <col index="115" name="Information extraction" >0.1</col>
     <col index="123" name="Communication" >0.1</col>
     </row>
     ........
</gcqm>
```

Overview of the tutorial

1. Principles

   1.1. Role of the context of use in MT evaluation

   1.2. ISO standards for software evaluation:
   terminology and role of context of use

   1.3. FEMTI guidelines: theoretical model

2. Tools

   2.1. Implementation of FEMTI interfaces

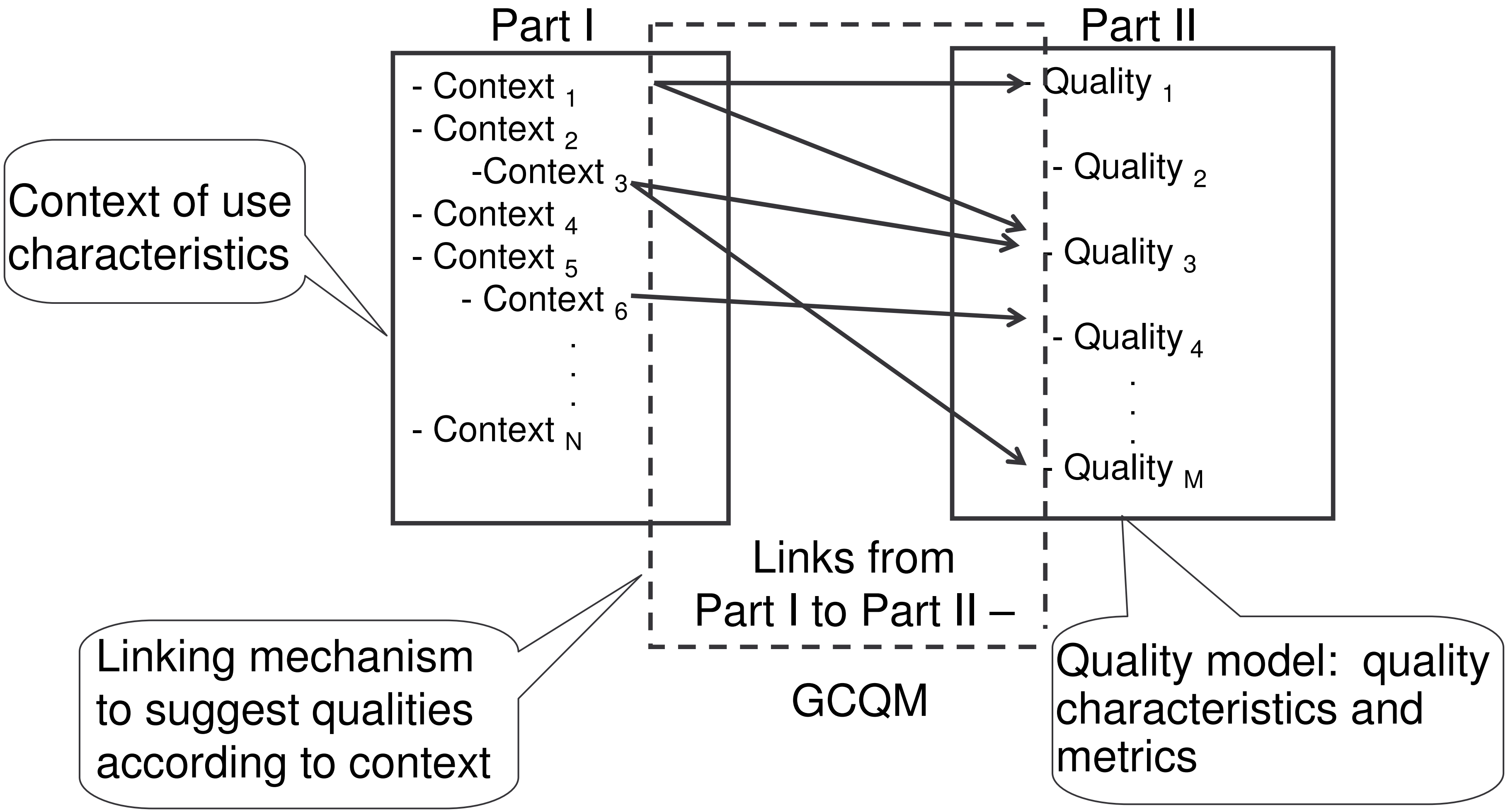
   2.2. Use of FEMTI by evaluators & evaluation experts

3. Practical application

Recent developments of FEMTI

- «Quality models and resources for the evaluation of MT»
  - Swiss National Science Foundation Project (2004 - 2006)
    - Continued 2006-2008
- Converting FEMTI guidelines into a tool
  - Automatic linking between contexts and metrics
  - Flexible implementation: XML → HTML, PDF, RTF
    - Continuous development not affecting online service
    - Dynamic (vs. static) generation of documents
  - Experts' interface
    - Create links between Part I and Part II
    - Put weights on the links (relevance: high, medium, low, n/a)

Developments of FEMTI (2/2)



Overview of the tutorial

1. Principles

   1.1. Role of the context of use in MT evaluation

   1.2. ISO standards for software evaluation:
   terminology and role of context of use

   1.3. FEMTI guidelines: theoretical model

2. Tools

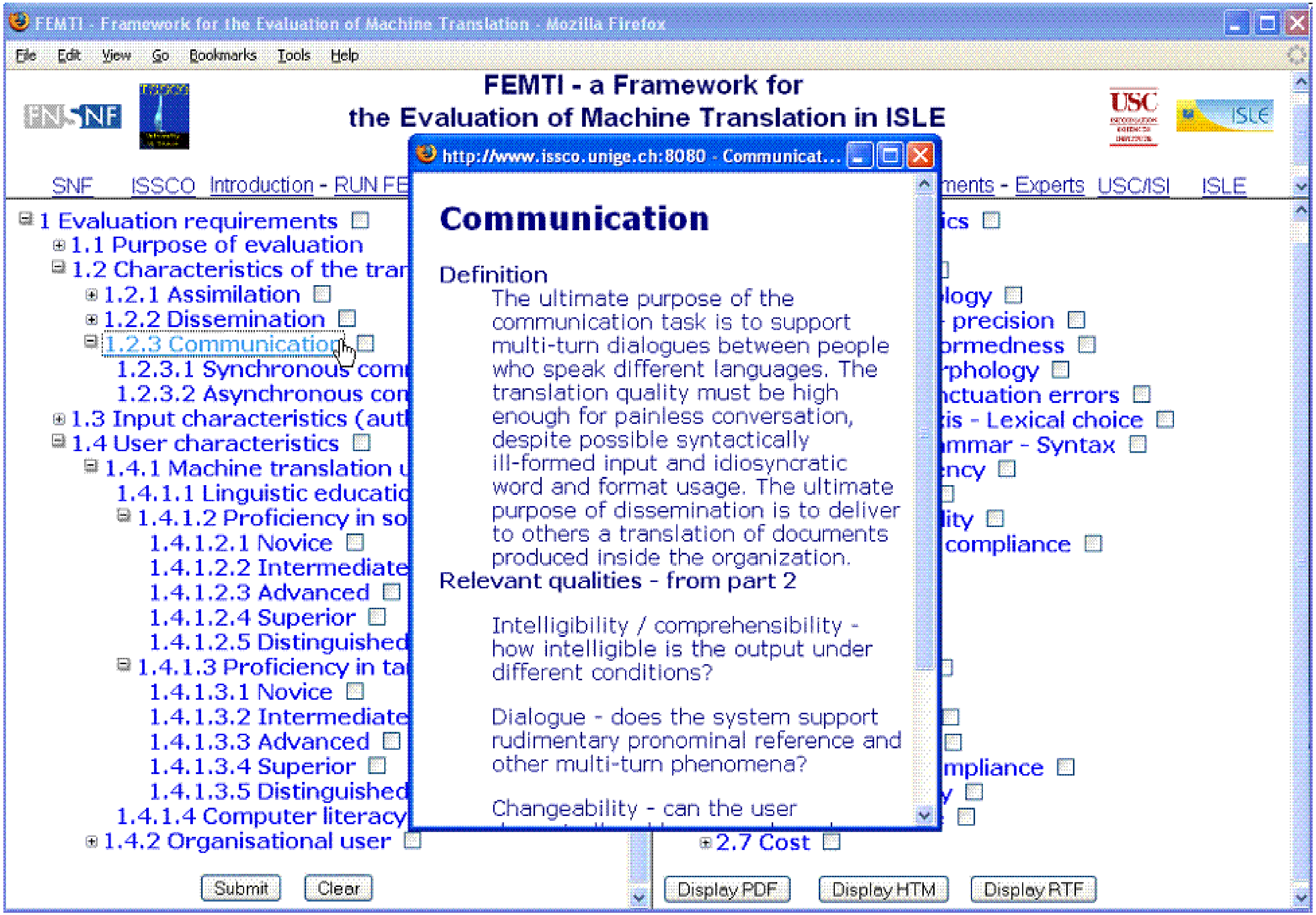
   2.1. Implementation of FEMTI interfaces

   2.2. Use of FEMTI by **evaluators** & evaluation experts

3. Practical application

Using FEMTI as evaluator

- Content accessible though pop-up windows


FEMTI - Framework for the Evaluation of Machine Translation - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
FNSNF
ISSCO
USC
ISLE
SNF ISSCO Introduction - RUN FE
ments - Experts USC/ISI ISLE
http://www.issco.unige.ch:8080 - Communicat...
Submit
Clear
Display PDF
Display HTM
Display RTF

FEMTI - a Framework for
the Evaluation of Machine Translation in ISLE

⊟ 1 Evaluation requirements □
⊞ 1.1 Purpose of evaluation
⊟ 1.2 Characteristics of the tran
⊞ 1.2.1 Assimilation □
⊞ 1.2.2 Dissemination □
⊟ 1.2.3 Communication □
1.2.3.1 Synchronous comm
1.2.3.2 Asynchronous con
⊞ 1.3 Input characteristics (autl
⊟ 1.4 User characteristics □
⊟ 1.4.1 Machine translation u
1.4.1.1 Linguistic educatio
⊟ 1.4.1.2 Proficiency in so
1.4.1.2.1 Novice □
1.4.1.2.2 Intermediate
1.4.1.2.3 Advanced □
1.4.1.2.4 Superior □
1.4.1.2.5 Distinguished
⊟ 1.4.1.3 Proficiency in ta
1.4.1.3.1 Novice □
1.4.1.3.2 Intermediate
1.4.1.3.3 Advanced □
1.4.1.3.4 Superior □
1.4.1.3.5 Distinguished
1.4.1.4 Computer literacy
⊞ 1.4.2 Organisational user □

ics □
logy □
- precision □
ormedness □
rphology □
nctuation errors □
is - Lexical choice □
mmar - Syntax □
ency □
ity □
compliance □
mpliance □
y □
⊞ 2.7 Cost □

Communication

Definition

The ultimate purpose of the communication task is to support multi-turn dialogues between people who speak different languages. The translation quality must be high enough for painless conversation, despite possible syntactically ill-formed input and idiosyncratic word and format usage. The ultimate purpose of dissemination is to deliver to others a translation of documents produced inside the organization.

Relevant qualities - from part 2

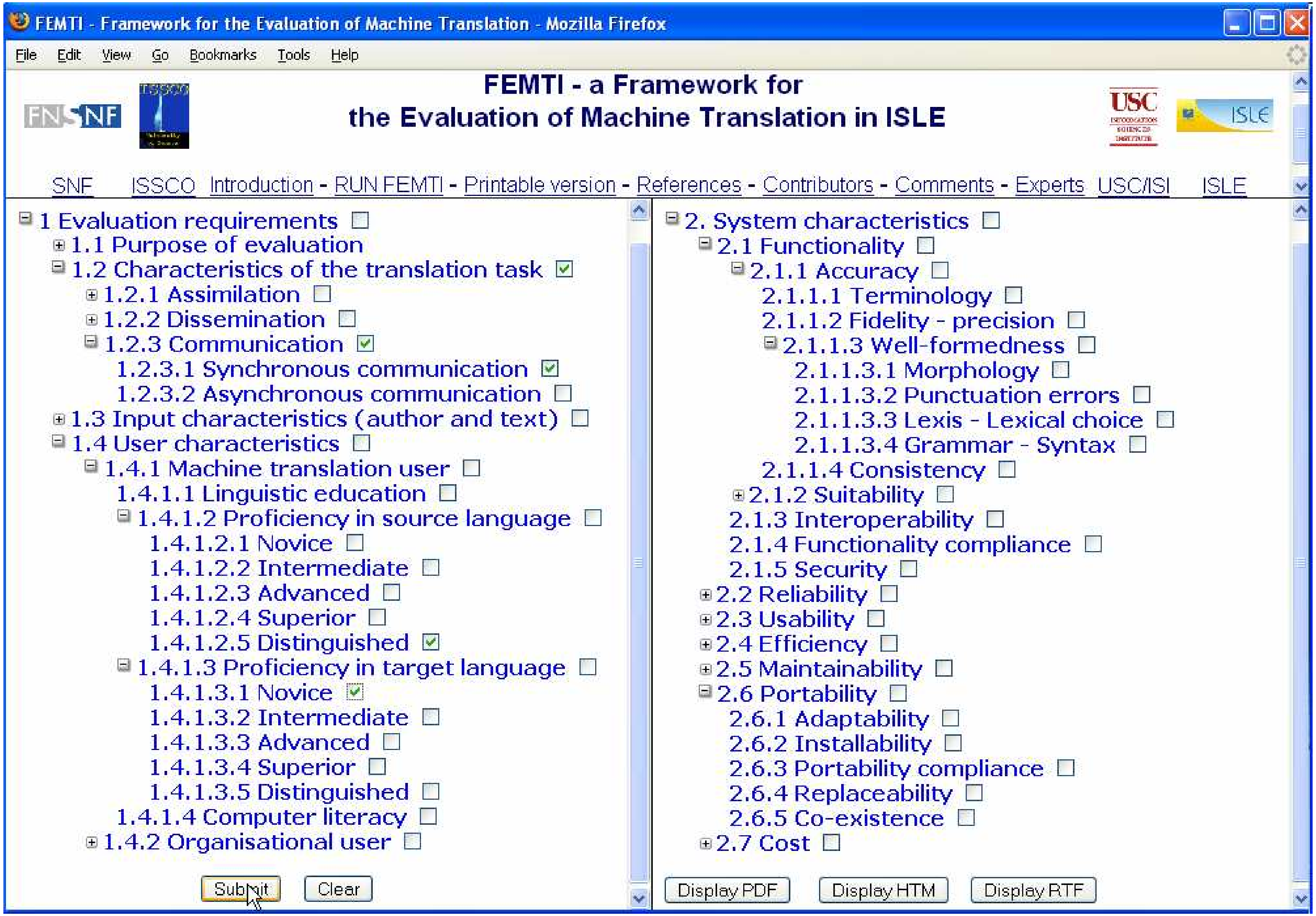
Intelligibility / comprehensibility - how intelligible is the output under different conditions?

Dialogue - does the system support rudimentary pronominal reference and other multi-turn phenomena?

Changeability - can the user

Using FEMTI as evaluator

- Content accessible though pop-up windows
- Part I: context characteristics
  - Define context of use


FEMTI - Framework for the Evaluation of Machine Translation - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
FNSNF
ISSCO
USC
ISLE

FEMTI - a Framework for the Evaluation of Machine Translation in ISLE

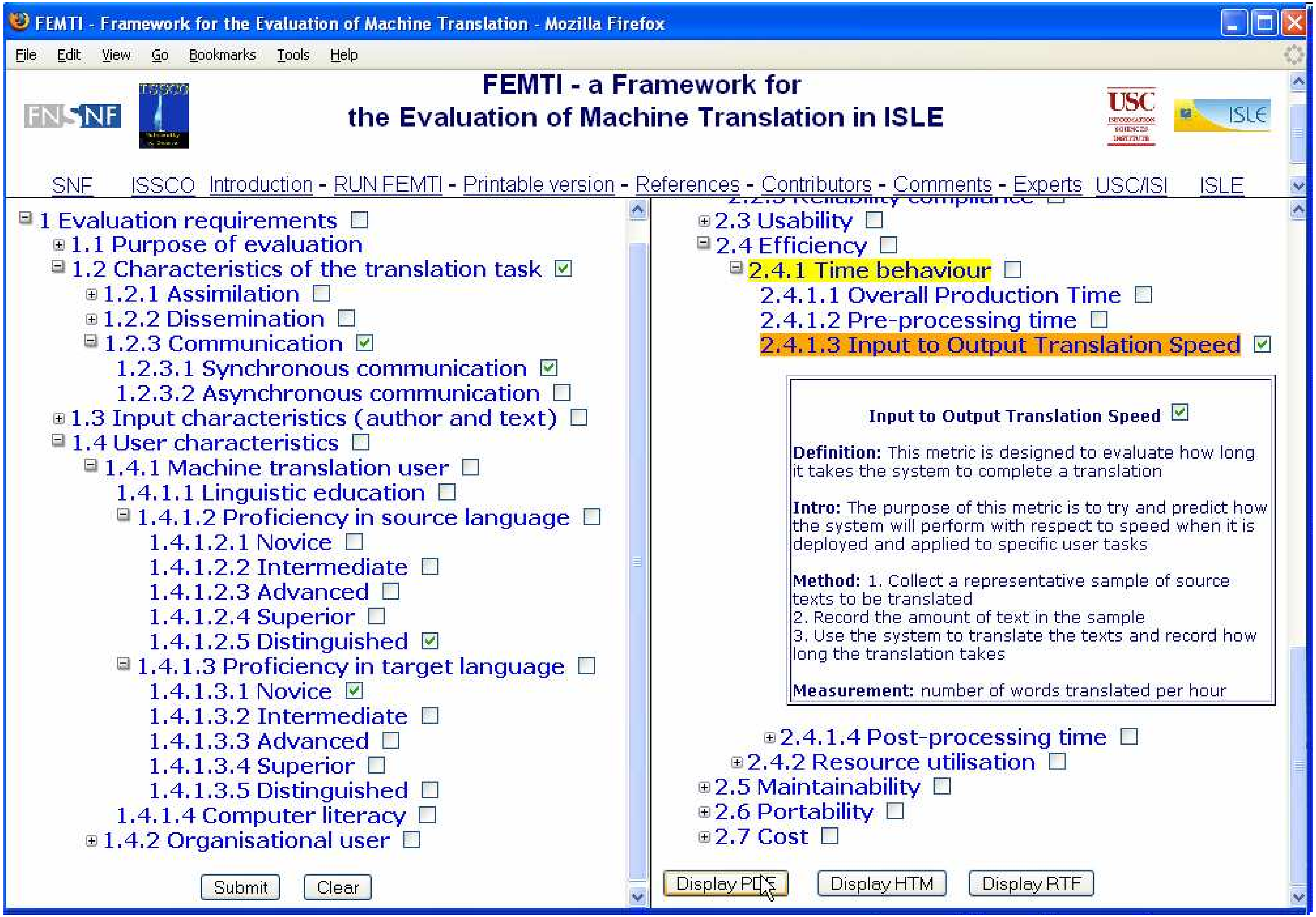
SNF ISSCO Introduction - RUN FEMTI - Printable version - References - Contributors - Comments - Experts USC/ISI ISLE

- ⊟ 1 Evaluation requirements □
  - ⊞ 1.1 Purpose of evaluation
  - ⊟ 1.2 Characteristics of the translation task ☑
    - ⊞ 1.2.1 Assimilation □
    - ⊞ 1.2.2 Dissemination □
    - ⊟ 1.2.3 Communication ☑
      - 1.2.3.1 Synchronous communication ☑
      - 1.2.3.2 Asynchronous communication □
  - ⊞ 1.3 Input characteristics (author and text) □
  - ⊟ 1.4 User characteristics □
    - ⊟ 1.4.1 Machine translation user □
      - 1.4.1.1 Linguistic education □
      - ⊟ 1.4.1.2 Proficiency in source language □
        - 1.4.1.2.1 Novice □
        - 1.4.1.2.2 Intermediate □
        - 1.4.1.2.3 Advanced □
        - 1.4.1.2.4 Superior □
        - 1.4.1.2.5 Distinguished ☑
      - ⊟ 1.4.1.3 Proficiency in target language □
        - 1.4.1.3.1 Novice ☑
        - 1.4.1.3.2 Intermediate □
        - 1.4.1.3.3 Advanced □
        - 1.4.1.3.4 Superior □
        - 1.4.1.3.5 Distinguished □
      - 1.4.1.4 Computer literacy □
    - ⊞ 1.4.2 Organisational user □

- ⊟ 2. System characteristics □
  - ⊟ 2.1 Functionality □
    - ⊟ 2.1.1 Accuracy □
      - 2.1.1.1 Terminology □
      - 2.1.1.2 Fidelity - precision □
      - ⊟ 2.1.1.3 Well-formedness □
        - 2.1.1.3.1 Morphology □
        - 2.1.1.3.2 Punctuation errors □
        - 2.1.1.3.3 Lexis - Lexical choice □
        - 2.1.1.3.4 Grammar - Syntax □
      - 2.1.1.4 Consistency □
    - ⊞ 2.1.2 Suitability □
    - 2.1.3 Interoperability □
    - 2.1.4 Functionality compliance □
    - 2.1.5 Security □
  - ⊞ 2.2 Reliability □
  - ⊞ 2.3 Usability □
  - ⊞ 2.4 Efficiency □
  - ⊞ 2.5 Maintainability □
  - ⊟ 2.6 Portability □
    - 2.6.1 Adaptability □
    - 2.6.2 Installability □
    - 2.6.3 Portability compliance □
    - 2.6.4 Replaceability □
    - 2.6.5 Co-existence □
  - ⊞ 2.7 Cost □

Using FEMTI as evaluator

- Content accessible though pop-up windows
- First define context of use
  - ❑ Part I - context characteristics
- Select relevant aspects of quality
  - ❑ Part II: quality characteristics
  - ❑ Relevant QC from GCQM highlighted
  - ❑ Select metrics to apply


FEMTI - Framework for the Evaluation of Machine Translation - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
SNF ISSCO Introduction - RUN FEMTI - Printable version - References - Contributors - Comments - Experts USC/ISI ISLE
Submit
Clear
Display PDF
Display HTM
Display RTF

FEMTI - a Framework for the Evaluation of Machine Translation in ISLE

- 1 Evaluation requirements □
  - 1.1 Purpose of evaluation
  - 1.2 Characteristics of the translation task ☑
    - 1.2.1 Assimilation □
    - 1.2.2 Dissemination □
    - 1.2.3 Communication ☑
      - 1.2.3.1 Synchronous communication ☑
      - 1.2.3.2 Asynchronous communication □
  - 1.3 Input characteristics (author and text) □
  - 1.4 User characteristics □
    - 1.4.1 Machine translation user □
      - 1.4.1.1 Linguistic education □
      - 1.4.1.2 Proficiency in source language □
        - 1.4.1.2.1 Novice □
        - 1.4.1.2.2 Intermediate □
        - 1.4.1.2.3 Advanced □
        - 1.4.1.2.4 Superior □
        - 1.4.1.2.5 Distinguished ☑
      - 1.4.1.3 Proficiency in target language □
        - 1.4.1.3.1 Novice ☑
        - 1.4.1.3.2 Intermediate □
        - 1.4.1.3.3 Advanced □
        - 1.4.1.3.4 Superior □
        - 1.4.1.3.5 Distinguished □
      - 1.4.1.4 Computer literacy □
    - 1.4.2 Organisational user □

2.2.5 Reliability compliance □

- 2.3 Usability □
- 2.4 Efficiency □
  - 2.4.1 Time behaviour □
    - 2.4.1.1 Overall Production Time □
    - 2.4.1.2 Pre-processing time □
    - 2.4.1.3 Input to Output Translation Speed ☑

Input to Output Translation Speed ☑

Definition: This metric is designed to evaluate how long it takes the system to complete a translation

Intro: The purpose of this metric is to try and predict how the system will perform with respect to speed when it is deployed and applied to specific user tasks

Method: 1. Collect a representative sample of source texts to be translated
2. Record the amount of text in the sample
3. Use the system to translate the texts and record how long the translation takes

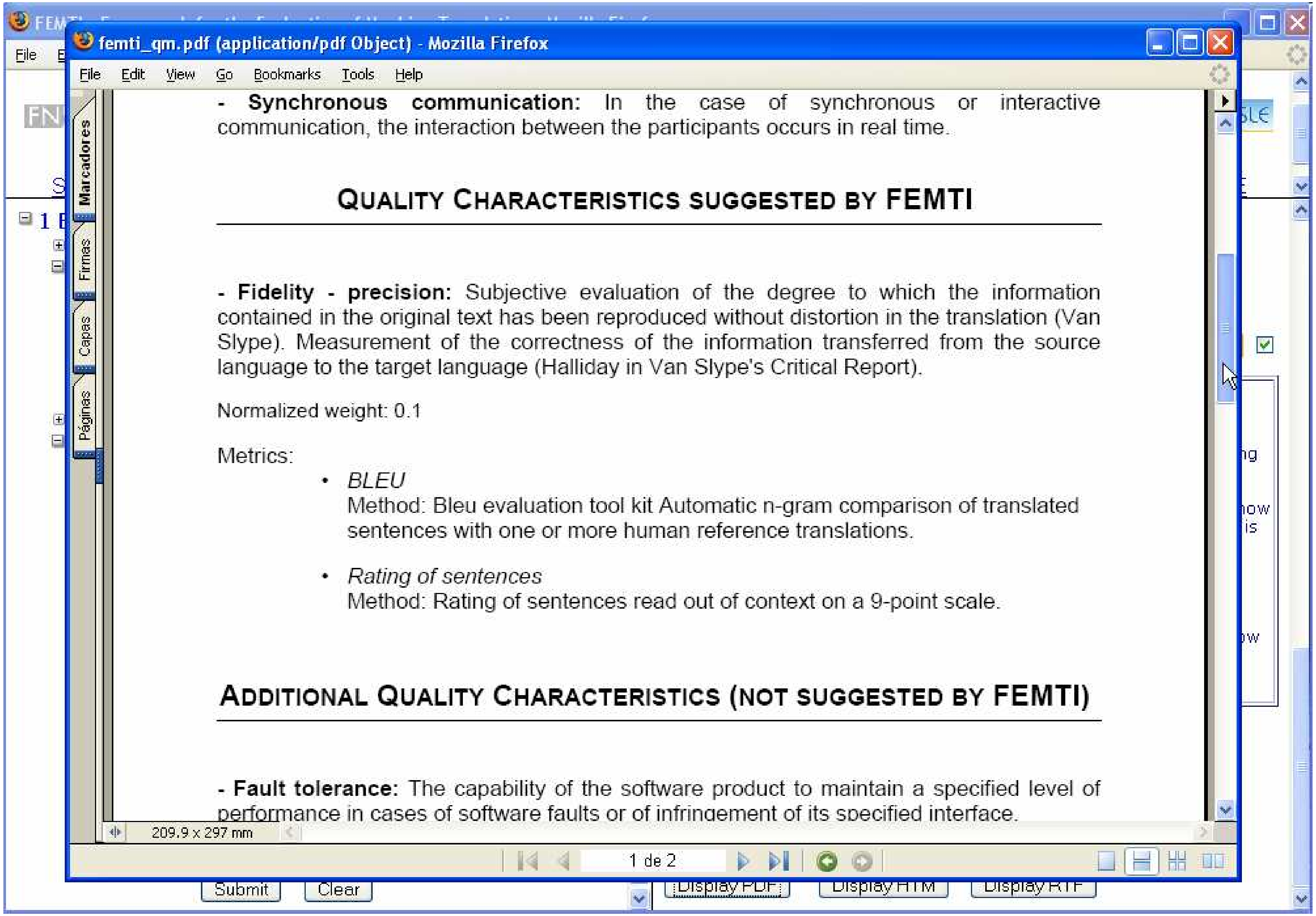
Measurement: number of words translated per hour

    - 2.4.1.4 Post-processing time □
  - 2.4.2 Resource utilisation □
- 2.5 Maintainability □
- 2.6 Portability □
- 2.7 Cost □

UNIVERSITÉ DE GENÈVE
ÉCOLE DE TRADUCTION ET D'INTERPRÉTATION

Using FEMTI as evaluator

- Content accessible though pop-up windows
- First define context of use
  - Part I - context characteristics
- Select relevant aspects of quality
  - Part II: quality characteristics
  - Relevant QC from GCQM highlighted
  - Select metrics to apply
- Save evaluation plan



femti_qm.pdf (application/pdf Object) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Marcadores
Firmas
Capas
Páginas
209.9 x 297 mm
1 de 2
Submit
Clear
Display PDF
Display HTM
Display RTF

- **Synchronous communication:** In the case of synchronous or interactive communication, the interaction between the participants occurs in real time.

QUALITY CHARACTERISTICS SUGGESTED BY FEMTI

- **Fidelity - precision:** 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).

Normalized weight: 0.1

Metrics:

- *BLEU*
  Method: Bleu evaluation tool kit Automatic n-gram comparison of translated sentences with one or more human reference translations.
- *Rating of sentences*
  Method: Rating of sentences read out of context on a 9-point scale.

ADDITIONAL QUALITY CHARACTERISTICS (NOT SUGGESTED BY FEMTI)

- **Fault tolerance:** The capability of the software product to maintain a specified level of performance in cases of software faults or of infringement of its specified interface.

Overview of the tutorial

1. Principles

    1.1. Role of the context of use in MT evaluation

    1.2. ISO standards for software evaluation:
    terminology and role of context of use

    1.3. FEMTI guidelines: theoretical model

2. Tools

    2.1. Implementation of FEMTI interfaces

    2.2. Use of FEMTI by evaluators & **evaluation experts**

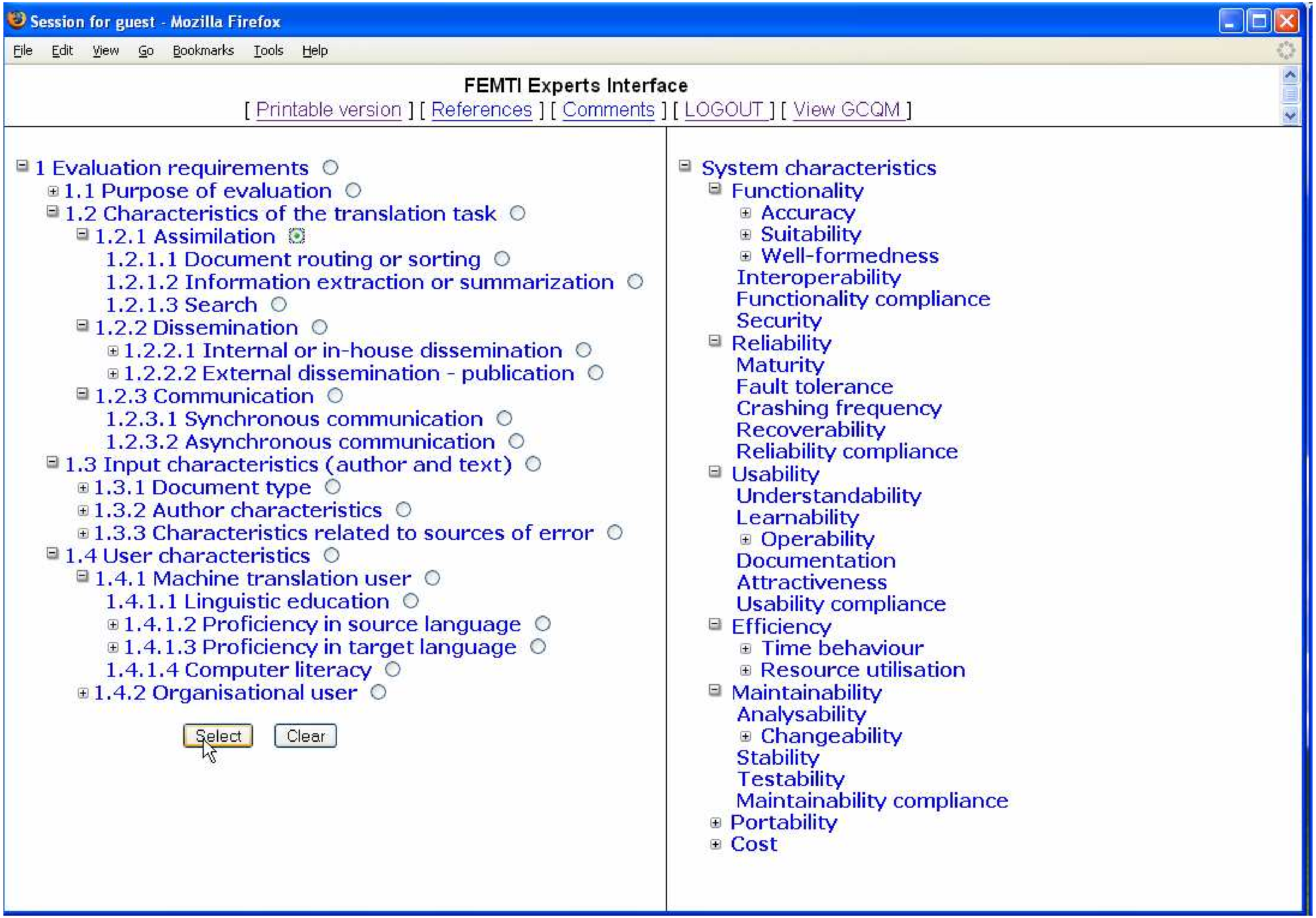
3. Practical application

Evaluation experts: access & modify GCQMs

- Used to suggest relevant QC given a context of use
- Expert interface
  - allow experts to build and view individual GCQMs
  - experts indicate for each context characteristic what are the most relevant quality characteristics
- Merge several GCQMs
  - generate the global FEMTI GCQM

Using FEMTI as expert

- Similar interface as before
  - Experts work on 1 CC at time
- First select 1 context characteristic to work on



FEMTI Experts Interface

[Printable version] [References] [Comments] [LOGOUT] [View GCQM]

- ⊟ 1 Evaluation requirements ○
  - ⊞ 1.1 Purpose of evaluation ○
  - ⊟ 1.2 Characteristics of the translation task ○
    - ⊟ 1.2.1 Assimilation ◉
      - 1.2.1.1 Document routing or sorting ○
      - 1.2.1.2 Information extraction or summarization ○
      - 1.2.1.3 Search ○
    - ⊟ 1.2.2 Dissemination ○
      - ⊞ 1.2.2.1 Internal or in-house dissemination ○
      - ⊞ 1.2.2.2 External dissemination - publication ○
    - ⊟ 1.2.3 Communication ○
      - 1.2.3.1 Synchronous communication ○
      - 1.2.3.2 Asynchronous communication ○
  - ⊟ 1.3 Input characteristics (author and text) ○
    - ⊞ 1.3.1 Document type ○
    - ⊞ 1.3.2 Author characteristics ○
    - ⊞ 1.3.3 Characteristics related to sources of error ○
  - ⊟ 1.4 User characteristics ○
    - ⊟ 1.4.1 Machine translation user ○
      - 1.4.1.1 Linguistic education ○
      - ⊞ 1.4.1.2 Proficiency in source language ○
      - ⊞ 1.4.1.3 Proficiency in target language ○
      - 1.4.1.4 Computer literacy ○
    - ⊞ 1.4.2 Organisational user ○

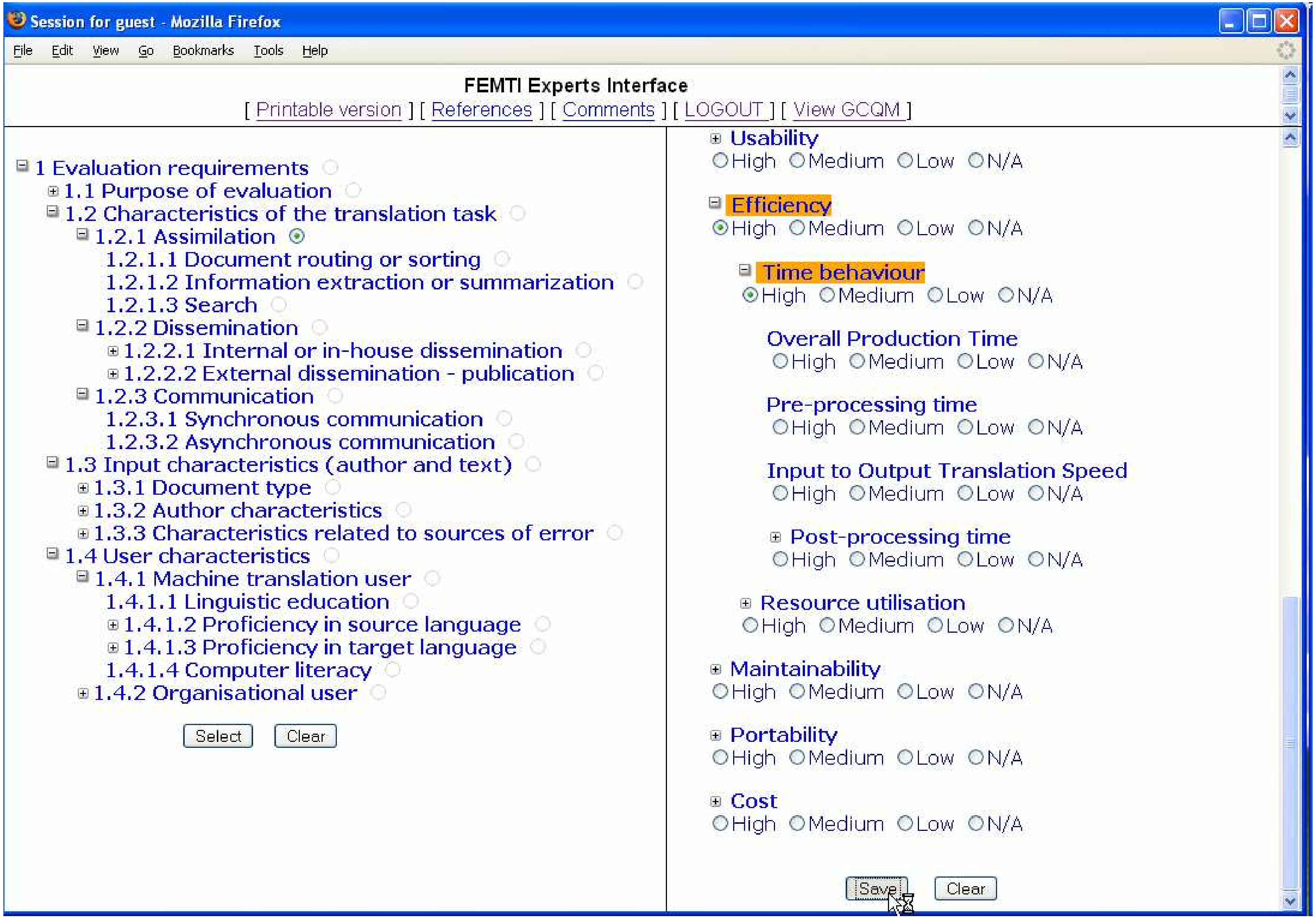
- ⊟ System characteristics
  - ⊟ Functionality
    - ⊞ Accuracy
    - ⊞ Suitability
    - ⊞ Well-formedness
    - Interoperability
    - Functionality compliance
    - Security
  - ⊟ Reliability
    - Maturity
    - Fault tolerance
    - Crashing frequency
    - Recoverability
    - Reliability compliance
  - ⊟ Usability
    - Understandability
    - Learnability
    - ⊞ Operability
    - Documentation
    - Attractiveness
    - Usability compliance
  - ⊟ Efficiency
    - ⊞ Time behaviour
    - ⊞ Resource utilisation
  - ⊟ Maintainability
    - Analysability
    - ⊞ Changeability
    - Stability
    - Testability
    - Maintainability compliance
  - ⊞ Portability
  - ⊞ Cost

Using FEMTI as expert

- Similar interface as before
  - Experts work on 1 CC at time

- First select 1 context characteristic to work on

- Create links to QC
  - by selecting them in Part II
  - If possible, indicate relevance of the link


Session for guest - Mozilla Firefox
File Edit View Go Bookmarks Tools Help

FEMTI Experts Interface

[Printable version] [References] [Comments] [LOGOUT] [View GCQM]

- ⊟ 1 Evaluation requirements ○
  - ⊞ 1.1 Purpose of evaluation ○
  - ⊟ 1.2 Characteristics of the translation task ○
    - ⊟ 1.2.1 Assimilation ●
      - 1.2.1.1 Document routing or sorting ○
      - 1.2.1.2 Information extraction or summarization ○
      - 1.2.1.3 Search ○
    - ⊟ 1.2.2 Dissemination ○
      - ⊞ 1.2.2.1 Internal or in-house dissemination ○
      - ⊞ 1.2.2.2 External dissemination - publication ○
    - ⊟ 1.2.3 Communication ○
      - 1.2.3.1 Synchronous communication ○
      - 1.2.3.2 Asynchronous communication ○
  - ⊟ 1.3 Input characteristics (author and text) ○
    - ⊞ 1.3.1 Document type ○
    - ⊞ 1.3.2 Author characteristics ○
    - ⊞ 1.3.3 Characteristics related to sources of error ○
  - ⊟ 1.4 User characteristics ○
    - ⊟ 1.4.1 Machine translation user ○
      - 1.4.1.1 Linguistic education ○
      - ⊞ 1.4.1.2 Proficiency in source language ○
      - ⊞ 1.4.1.3 Proficiency in target language ○
      - 1.4.1.4 Computer literacy ○
    - ⊞ 1.4.2 Organisational user ○

⊞ Usability
○ High ○ Medium ○ Low ○ N/A

⊟ Efficiency
● High ○ Medium ○ Low ○ N/A

⊟ Time behaviour
● High ○ Medium ○ Low ○ N/A

Overall Production Time
○ High ○ Medium ○ Low ○ N/A

Pre-processing time
○ High ○ Medium ○ Low ○ N/A

Input to Output Translation Speed
○ High ○ Medium ○ Low ○ N/A

⊞ Post-processing time
○ High ○ Medium ○ Low ○ N/A

⊞ Resource utilisation
○ High ○ Medium ○ Low ○ N/A

⊞ Maintainability
○ High ○ Medium ○ Low ○ N/A

⊞ Portability
○ High ○ Medium ○ Low ○ N/A

⊞ Cost
○ High ○ Medium ○ Low ○ N/A

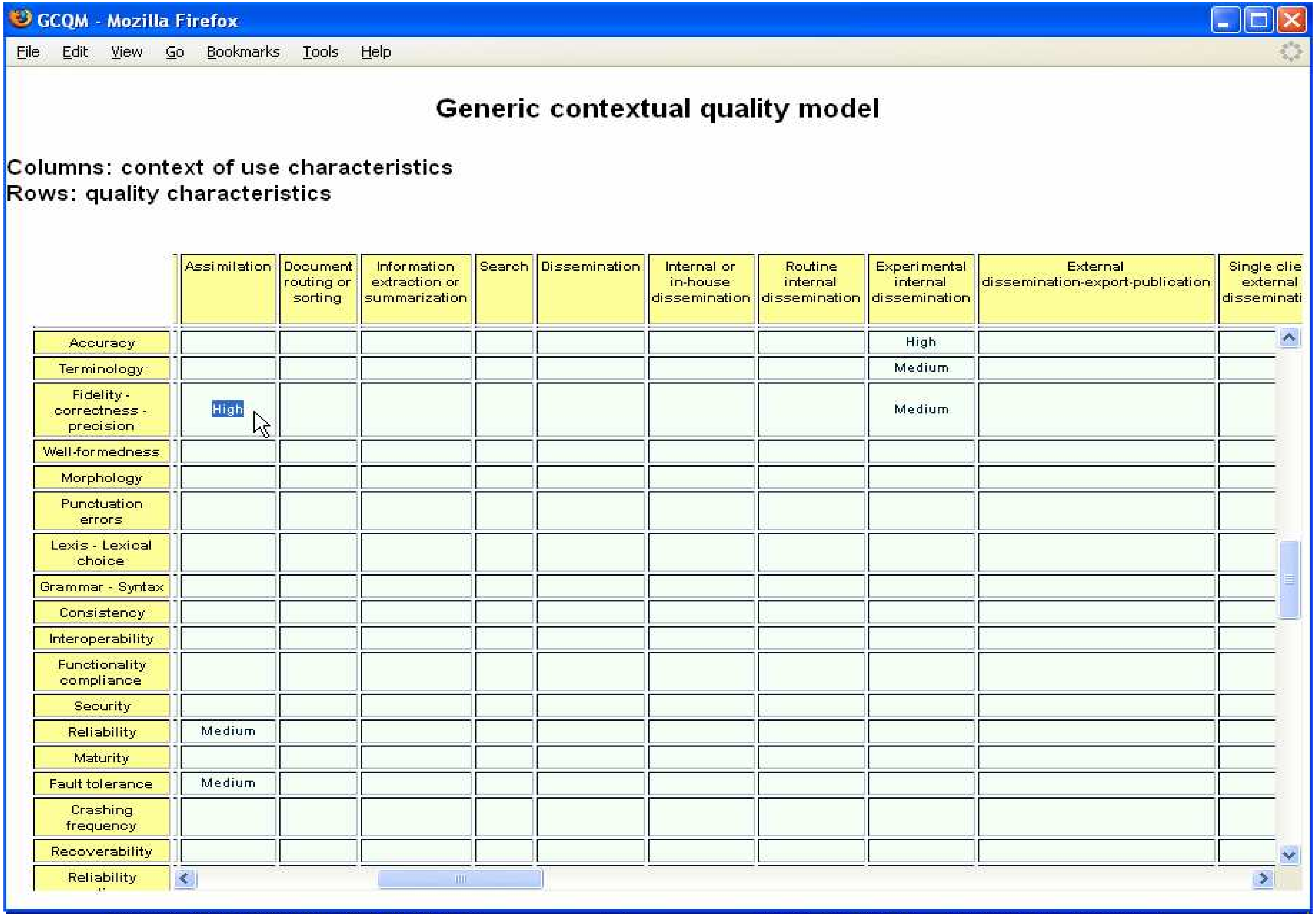
Using FEMTI as expert

- Similar interface as before
  - Experts work on 1 CC at time

- First select 1 context characteristic to work on

- Create links to QC
  - by selecting them in Part II
  - If possible, indicate relevance of the link

- Save/view GCQM



Generic contextual quality model

Columns: context of use characteristics
Rows: quality characteristics

	Assimilation	Document routing or sorting	Information extraction or summarization	Search	Dissemination	Internal or in-house dissemination	Routine internal dissemination	Experimental internal dissemination	External dissemination-export-publication	Single clie external disseminati
Accuracy								High		
Terminology								Medium		
Fidelity - correctness - precision	High							Medium		
Well-formedness										
Morphology										
Punctuation errors										
Lexis - Lexical choice										
Grammar - Syntax										
Consistency										
Interoperability										
Functionality compliance										
Security										
Reliability	Medium									
Maturity										
Fault tolerance	Medium									
Crashing frequency										
Recoverability										
Reliability										

Future work on FEMTI

- Keep refining both taxonomies
  - refine contexts of use
  - refine qualities and metrics

- Better management of weights for a quality model
  - change the selection mode of characteristics
    - current: binary
    - future: ‘essential’, ‘important’, ‘not important’

- Poll experts for the two taxonomies
- Poll experts for individual GCQMs
- Best method to integrate several GCQMs

Overview of the tutorial

1. Principles
   - 1.1. Role of the context of use in MT evaluation
   - 1.2. ISO standards for software evaluation: terminology and role of context of use
   - 1.3. FEMTI guidelines: theoretical model
2. Tools
   - 2.1. Implementation of FEMTI interfaces
   - 2.2. Use of FEMTI by evaluators & evaluation experts
3. Practical application

Outline of the exercise (1/3)

- Objective
  - define a contextualized evaluation plan for an MT system
  - compare the plans defined by various groups
  - improve the current Generic Contextual Quality Model

1. Select one of the two scenarios of use outlined below for the MT system under evaluation
   - options: (a) focus the entire group on only one scenario; (b) enrich the selected scenario with additional specifications of the intended use; (c) propose your own scenarios of use

Outline of the exercise (2/3)

2. What context characteristics are relevant? Which are the most *vs*. least important?
   - select from the list of characteristics of the context of use (FEMTI Part I) the ones that best describe the intended context of use of the MT system under evaluation

3. What quality characteristics correspond to each of the system characteristics you have picked out? What is their relative importance?
   - Based on the context characteristics, on your own experience of MT systems, and on the indications available in FEMTI for these characteristics, proceed to select (from FEMTI Part II) a list of relevant quality characteristics that the MT system under evaluation should possess.

Outline of the exercise (3/3)

- Indicate for each characteristic of the context, which qualities from the list (FEMTI Part II) are important for an MT system that will be used in that context; you can also quantify the importance on a 3-point scale (3: very important; 2: important, 1: nice to have)
  - Use the numbers of the characteristics (rather than names) to refer to them on the form.
- The final list of quality characteristics constitutes the contextualized quality model to be used for evaluation

4. When you have finished defining your contextualized quality model, please hand your form to the presenters, who will synthesize the results in preparation for a general discussion.

Conclusion

- FEMTI in its current state is useful, but …
  - ❑ Content still needs work
- Feedback is needed
  - ❑ To improve FEMTI's content in three directions
    - Improve taxonomies and GCQMs
    - Diversify contexts of use based on MT case studies
      - ❑ Questionnaire at http://www.issco.unige.ch/mt-use/
    - Integrate general/specific suggestions
      - ❑ Using FEMTI's integrated comments mechanism
- http://www.issco.unige.ch/femti/

Acknowledgments

- Houcine Benantar
- Francine Braun-Chen
- Olivier Hamon
- Tony Hartley
- Eduard Hovy
- Sandra Manzi
- Keith J. Miller
- Florence Reeder
- Nancy Underwood
- Michelle Vanni
- Sandrine Zufferey
- John S. White

SNSF projects
200021-103318
200020-113604

FNS SNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

EU project
IST-1999-10647

ist
information
society
technologies

References

Church Kenneth W. & Hovy Eduard H. (1993): "Good Applications for Crummy MT", *Machine Translation*, vol. 8, n°, pp. 239-258.

Estrella Paula, Popescu-Belis Andrei & Underwood Nancy (2005): "Finding the System that Suits you Best: Towards the Normalization of MT Evaluation", *Proc. 27th ASLIB International Conference on Translating and the Computer*, London, UK, pp. 23-34.

ISO/IEC (1999): *ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC (2001): *ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

King Margaret (2005): "Accuracy and Suitability: New Challenges for Evaluation", *Language Resources and Evaluation*, vol. 39, n° 1, pp. 45-64.

King Margaret & Underwood, Nancy (2006): "Evaluating Symbiotic Systems: the challenge", *Proc. LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy.

Nomura Hirosato & Isahara Hitoshi (1992): "JEIDA's Criteria on Machine Translation Evaluation", in *IPSJ SIGNotes Natural Language*, Tokyo, Japan, Information Processing Society of Japan, pp. 107-114.

Popescu-Belis Andrei, Estrella Paula, King Maghi & Underwood Nancy (2005): "Towards Automatic Generation of Evaluation Plans for Context-based MT Evaluation", *Working Paper* n. 64, ETI/TIM/ISSCO, 18 p.

Popescu-Belis Andrei (2003): "An experiment in comparative evaluation: humans vs. computers", *Proc. Machine Translation Summit IX*, New Orleans, LA, USA, pp. 307-314.

Popescu-Belis Andrei, Estrella Paula, King Margaret & Underwood Nancy (2006): "A model for context-based evaluation of language processing systems and its application to machine translation evaluation", *Proc. LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 691-696.

Van Slype Georges (1979): Critical Study of Methods for Evaluating the Quality of Machine Translation, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142.

White John S. & Taylor Kathryn B. (1998): "A Task-Oriented Evaluation Metric for Machine Translation", *Proc. LREC 1998 (1st International Conference on Language Resources and Evaluation)*, Granada, Spain, vol. 1/2, pp. 21-25.