



Asking MOSES to help with translation verification

Yuri Pettinicchi

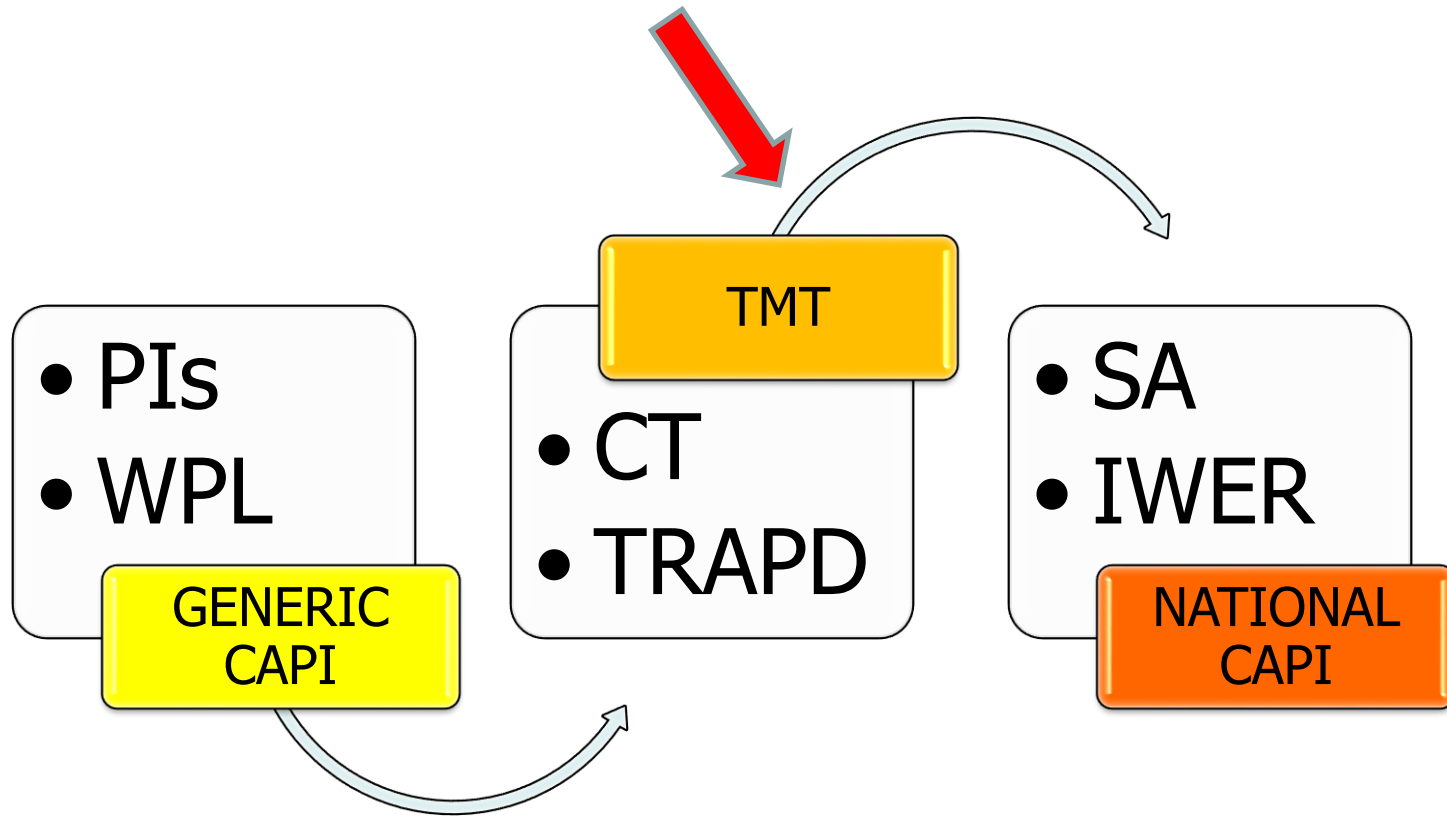
CSDI

16-18 March 2017 – Mannheim



- ▶ **SHARE - cross-national survey**
 - ▶ 28 national CAPI
 - ▶ 40 languages
- ▶ **Translation verification:**
 - ▶ National CAPI tools display properly translated questionnaire
 - ▶ Green flag to go on-field
- ▶ **Aim**
 - ▶ Avoiding avoidable mistakes
 - ▶ Improving data quality

Where are we?



- ▶ How SHARE handles translations
- ▶ Possible issues
- ▶ Current approach
- ▶ New approach
- ▶ Next steps

▶ SHARE does follow TRAPD procedure

From generic

- ▶ “Do you have another child that was not already mentioned? Again, please think of all natural children, fostered, adopted and stepchildren including those of Lorenza.”

To national

- ▶ “Haben Sie ein Kind, das noch nicht genannt wurde? Denken Sie bitte wieder an alle leiblichen Kinder, Pflegekinder, Adoptivkinder und Stiefkinder einschließlich die von Lorenza.“

Multiple fills in the translation

Dynamic fill with hidden text

Gender specific fills

Numbered answer options

Wave	7
Items	1099
Text	6837

Q1text

Do you have ***FL_CH001a_1**? Again, please think of all natural children, fostered, adopted and stepchildren ***FL_CH001a_2** ***FL_CH001a_3** ***FL_CH001a_13**

Haben Sie ***FL_CH001a_1**? Denken Sie bitte wieder an alle leiblichen Kinder, Pflegekinder, Adoptivkinder und Stiefkinder ***FL_CH001a_2** ***FL_CH001a_3** ***FL_CH001a_13**

Answer type:

a1.	1. Yes	1. Ja
a2.	*FL_CH001a_7	*FL_CH001a_7
a3.	*FL_CH001a_8	*FL_CH001a_8
a4.	*FL_CH001a_9	*FL_CH001a_9
a5.	*FL_CH001a_10	*FL_CH001a_10
a6.	*FL_CH001a_11	*FL_CH001a_11
a97.	*FL_CH001a_12	*FL_CH001a_12

Translate fills for this question:

- *FL_CH001a_1** → (dynamic constructed text based on how the child was loaded) → **FL_CH001a_1**
- *FL_CH001a_2** → , including those of/(empty) → einschließlich die von
- *FL_CH001a_4** → (child/children loaded from FLDefault 71-73) → Ihres Ehemannes
- *FL_CH001a_5** → (name of child if available else empty) → **FL_CH001a_5**
- *FL_CH001a_6** → (further information like age and gender if available else empty) → **FL_CH001a_6**
- *FL_CH001a_7** → 2. Yes, but child's name, gender or year of birth is incorrect/(empty) → 2. Ja, aber der Name, das Geschlecht oder das Geburtsjahr des Kindes sind falsch/
- *FL_CH001a_8** → 3. No, child of partner from whom R separated/(empty) → 3. Nein, Kind des Partners von dem RP getrennt lebt
- *FL_CH001a_9** → 4. No, child died/(empty) → 4. Nein, Kind verstorben
- *FL_CH001a_10** → 5. No, child unknown/5. No → 5. Nein, Kind unbekannt/5. Nein
- *FL_CH001a_11** → (empty)/6. Yes, but already mentioned earlier → 6. Ja, aber bereits früher erwähnt
- *FL_CH001a_12** → 97. No, other reason/(empty) → 97. Nein, anderer Grund
- *FL_CH001a_13** → (empty)/@/@/@IWER:@/If a child is listed twice, delete the second one with category "6. Yes, but already mentioned earlier", and keep the first@! → @/@/@IWER:@/Wenn ein Kind zweimal in der Liste vorkommt, behalten Sie das erste Kind und löschen Sie das zweite Kind mit der Kategorie 6. Ja, aber bereits früher erwähnt@!
- *FL_CH001a_14** → (empty) → **FL_CH001a_14**
- *FL_CH001a_3** → your husband/your wife/your partner/(empty) → Ihrem Mann/Ihrer Frau/Ihrem Partner/Ihrer Partnerin
- *FL_CH001a_15** → your husband/your wife/your partner/(empty) → **FL_CH001a_15**
- *FL_CH001a_16** → (empty) → **FL_CH001a_16**

Possible issues

- ▶ Misspelling a word
- ▶ Misspelling a command
- ▶ Empty fields / missing sentence
 - ▶ (full sentence vs part of a sentence)
- ▶ *Flipped* translations
 - ▶ Negative effects on qnn routing

Wave	7
Items	43960
Text	273480

GENERIC	NATIONAL
1. Employed	1. Arbeitslos
2. Unemployed	2. Abhängig

Current checking

- ▶ Visual inspections of the TMT
- ▶ Testing the CAPI – (Generic vis-a-vis national)
- ▶ CAPI remarks from two small scale field runs
- ▶ Analysis of the data

PROS	CONS
Flexible	Not systematic
Decentralized	Effort and time demanding

Automated checks

▶ Ingredients:

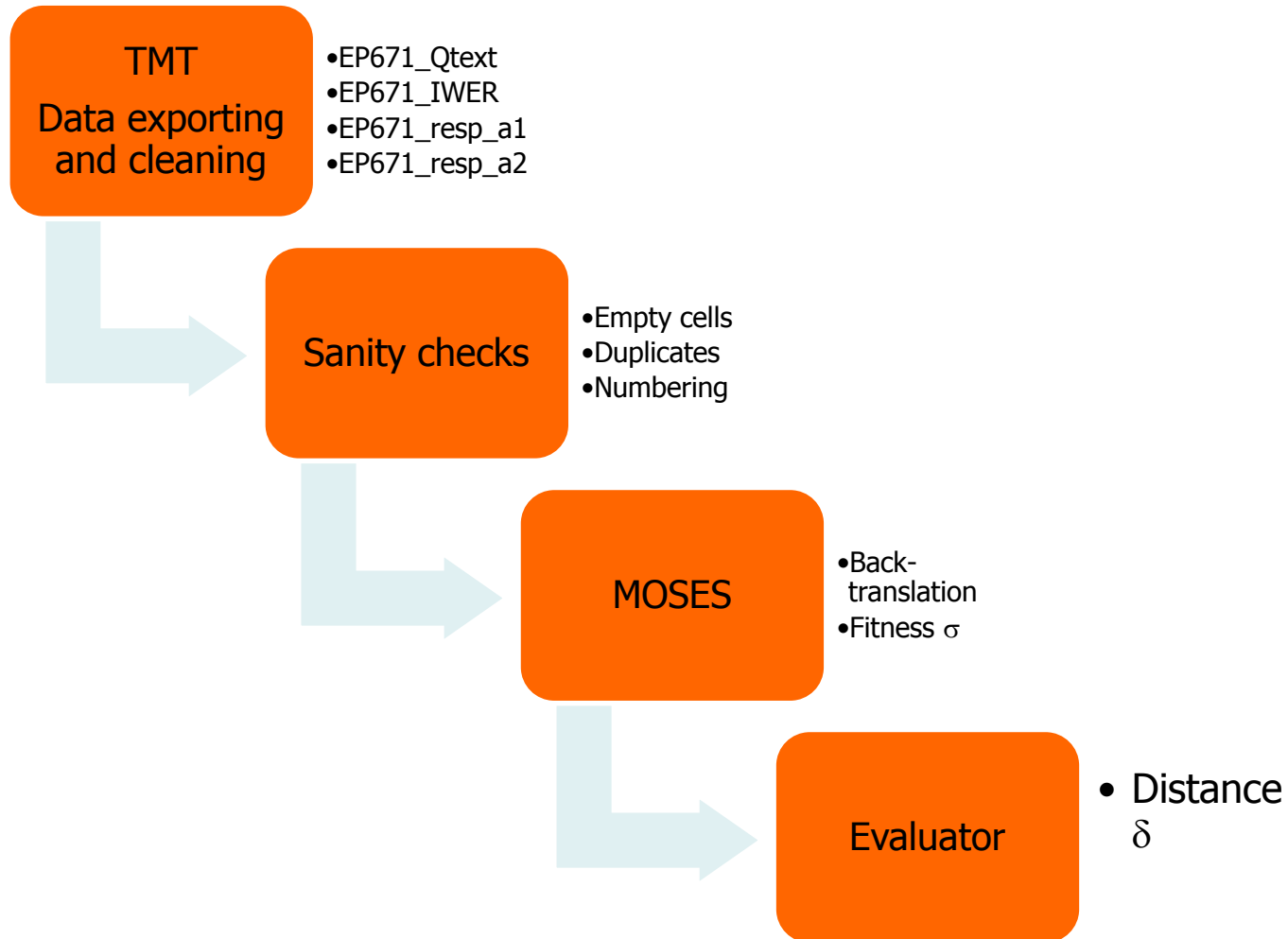
- ▶ Text data
- ▶ Sanity checks
 - ▶ Empty cells
 - ▶ Duplicates
 - ▶ Numbering
 - ▶ Translation

▶ Outcome:

- ▶ Flagged items
- ▶ Feedback to/from country teams

PROS	CONS
Systematic	Prone to false positive
Centralized	

Workflow



- ▶ **Statistical machine translation system (SMT)**
 - ▶ **Training pipeline**
 - ▶ from raw data (parallel corpora) to a machine translation model
 - 1. Prepare data
 - 2. Align words
 - 3. Lexical translation
 - 4. Extract and score phrases
 - 5. Reordering model and Generation model
 - 6. Configuration file
 - ▶ **Decoder**
 - ▶ find the highest scoring sentence in the target language (according to the translation model) corresponding to a given source sentence.

- ▶ **Open source software:** <http://www.statmt.org/moses/>

- ▶ Measure the distance between
 - ▶ the back translation (EN)
 - ▶ the source sentence (EN)

- ▶ Different metrics:
 - ▶ Counting how many words are in common

▶ Report for the project manager

ITEM	Mod	Lang	Source Text	Sanity Check1	Moses Check1	Back translation	Fitness σ	Distance δ	Flag
EP671_Qtext									
EP671_IWER									
EP671_resp_a1									
EP671_resp_a2									

- ▶ List of items to be checked to the CTO
- ▶ Feedback implemented in the procedure

- ▶ Pilot to test
 - ▶ Sanity checks
 - ▶ Moses trained on UN corpora
 - ▶ One language (French)
 - French – France
 - French – Swiss
 - French – Belgium
 - French – Luxemburg

- ▶ Statistics
 - ▶ Percentage of flagged items
 - ▶ Percentage of false positive

- ▶ Comments? Questions?
 - ▶ Thank you