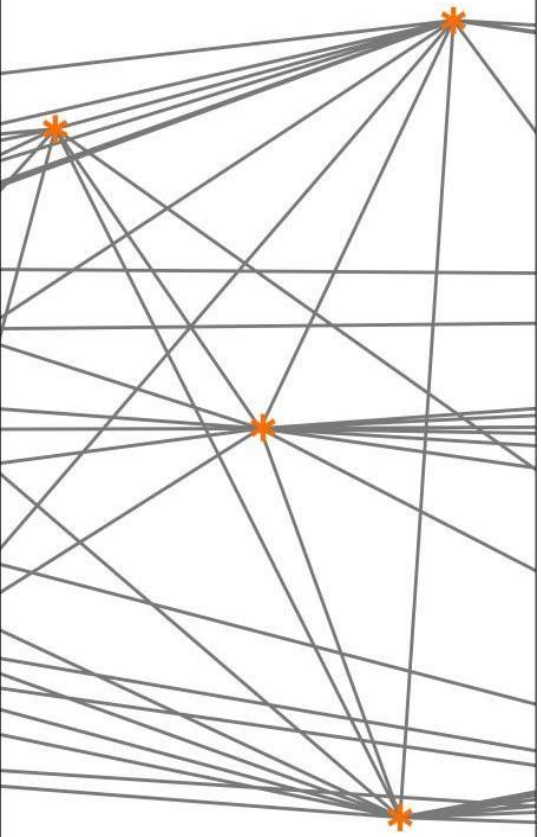
SHARE

Survey of Health, Ageing and Retirement in Europe



Asking MOSES to help with translation verification

Yuri Pettinicchi

CSDI

16-18 March 2017 – Mannheim

HORIZON 2020

Federal Ministry of Education and Research

NIH National Institute on Aging

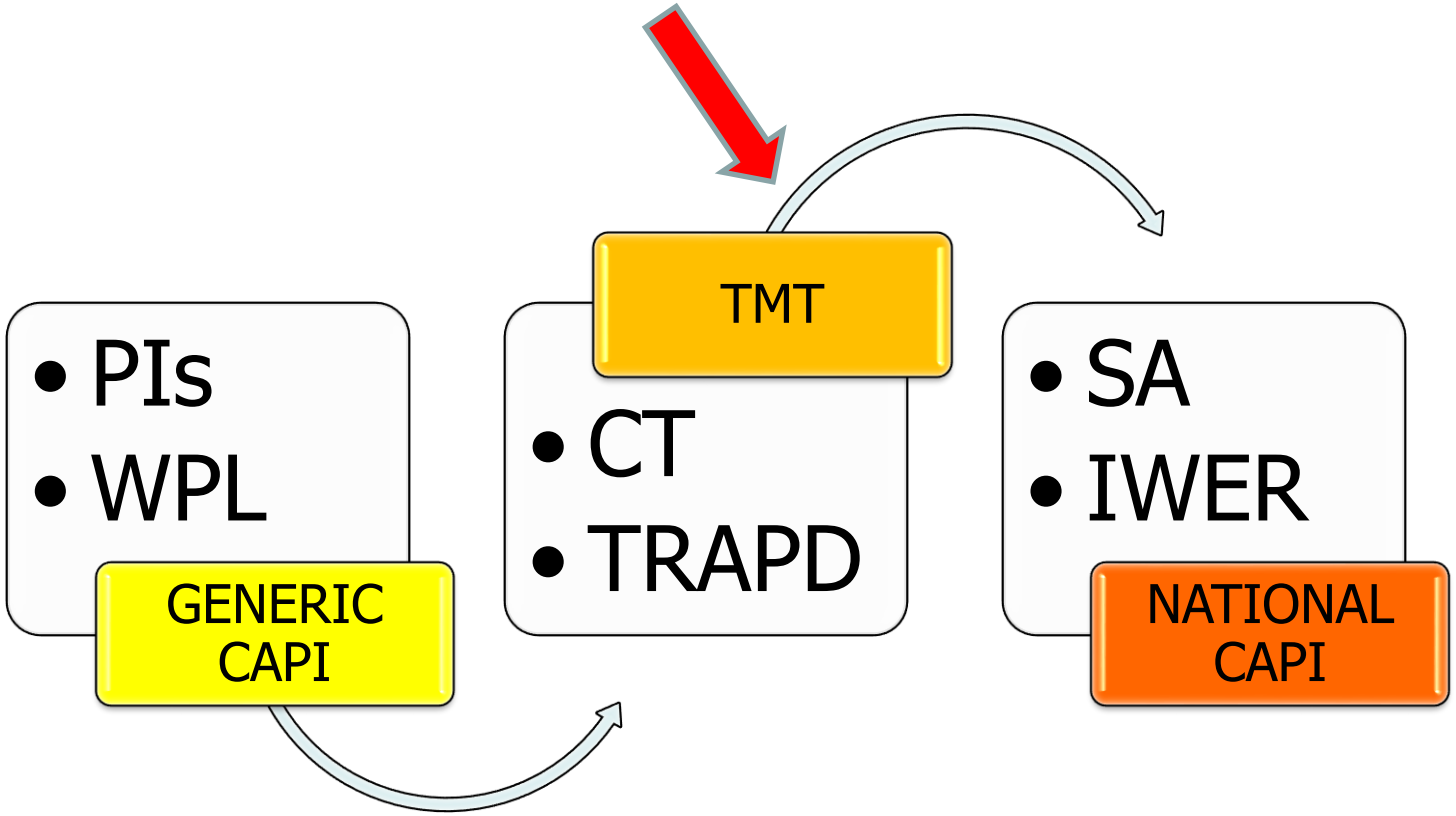
DFG

mea MAX PLANCK INSTITUTE FOR SOCIAL LAW AND SOCIAL POLICY
Munich Center For The Economics Of Aging

Motivation

- 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?



Outline

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

SHARE

- 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."

SHARE - TMT

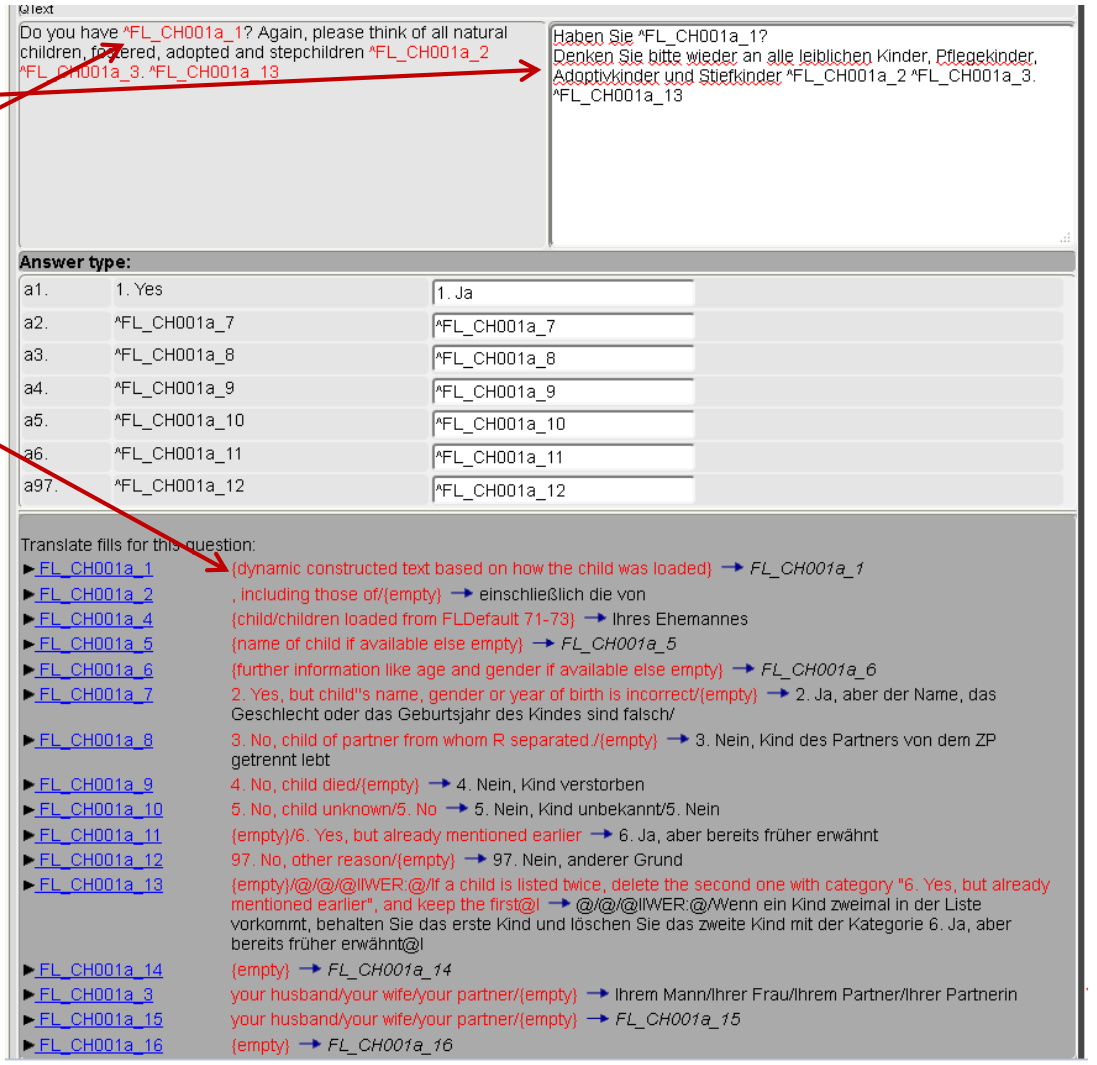
Multiple fills in the translation

Dynamic fill with hidden text

Gender specific fills

Numbered answer options

Wave	7
Items	1099
Text	6837



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

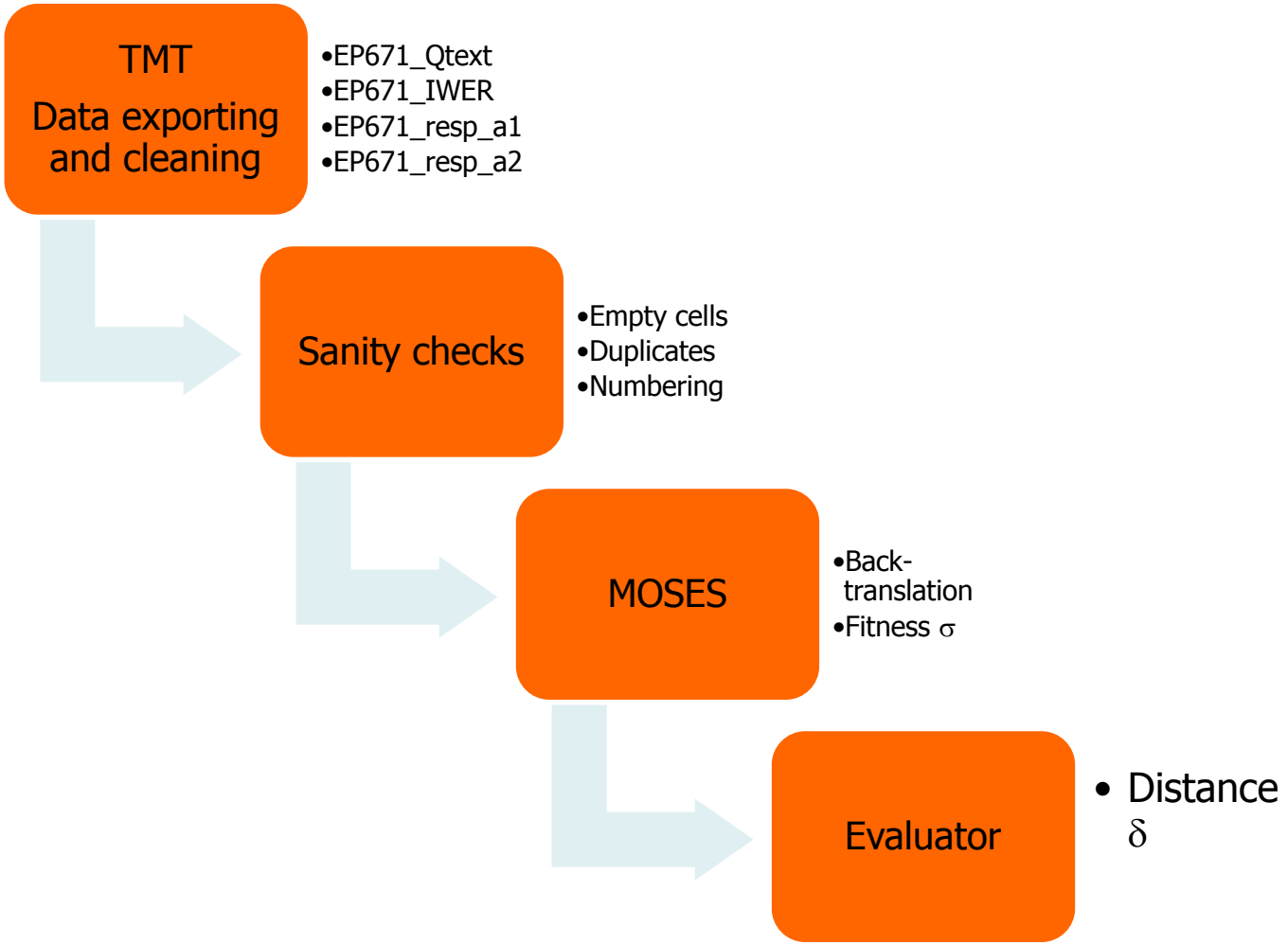
PROS	CONS
Flexible	Not systematic
Decentralized	Effort and time demanding

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


TMT
Data exporting and cleaning
•EP671_Qtext
•EP671_IWER
•EP671_resp_a1
•EP671_resp_a2
Sanity checks
•Empty cells
•Duplicates
•Numbering
MOSES
•Back-translation
•Fitness σ
Evaluator
• Distance δ

MOSES

- 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/

Evaluator

- Measure the distance between
  - the back translation (EN)
  - the source sentence (EN)
- Different metrics:
  - Counting how many words are in common

Outcome

- 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

- 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