

# Survey translation according to the team approach: On the impact of post-edited translations on final review output

---

gesis Leibniz Institute  
for the Social Sciences

upf. Universitat  
Pompeu Fabra  
Barcelona

**RECSM**  
Research and Expertise Centre  
for Survey Methodology

Dorothee Behr, Brita Dorer, Veronika  
Keck, Diana Zavala-Rojas, & Danielly  
Sorato

ESRA, 9 July, 2021  
Virtual

European  
Social  
Survey

SSHOC   
social sciences & humanities open cloud



# Overview

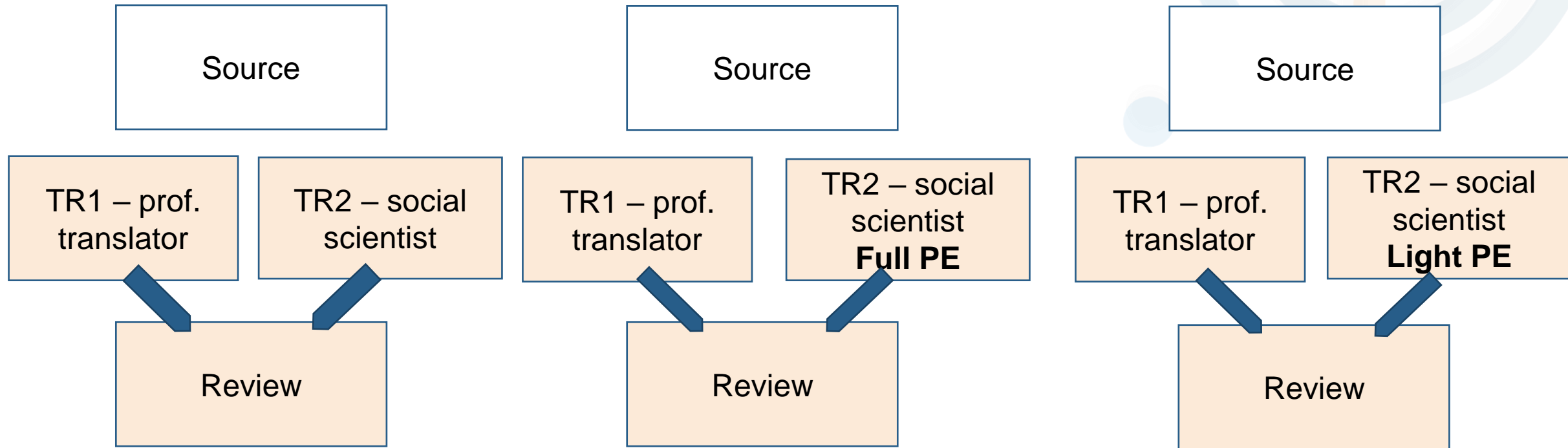
- 🌀 Research question
- 🌀 Operationalization
- 🌀 Preliminary results
- 🌀 Conclusions
- 🌀 Outlook



# Research question

- 🌀 How useful are the post-edited versions in the production of the final review output compared to the human translations?
- 🌀 And how are these findings linked to the final quality of the review output?
- 🌀 The more useful the post-edited versions are and the more they contribute to high-quality review outputs, the more MT and post-editing can have a place in survey translation.

# Focus on relationship TR1/TR2 vs. Review



Baseline  
all human

Full PE

Light PE

# Operationalization: “rich points” (anchor terms)

- Source text items had been selected because of anticipated MT or general translation problems (“rich points”), e.g.,
  - (potentially) challenging terminology/wording → “race or ethnic group”, “environment”, “political action”
  - (potentially) challenging questionnaire design characteristics → response scale labels, typical survey wording such as “generally speaking”, “around”
  - Known MT issues → gender issues (“partner”), missing grammatical information (“might do”)
- For these aspects (n = 189), we coded where the final review translation came from.

# Operationalization: “rich points” – example

🌀 **Source:** In the last 12 months , that is since [MONTH, YEAR], were you ever unable to get a medical consultation or the treatment you needed for any of the reasons listed on this card?

🌀 **Review output:** In den letzten 12 Monaten, d. h. seit [MONAT, JAHR]: Ist es jemals vorgekommen, dass es Ihnen aus einem oder mehreren der auf dieser Liste aufgeführten Gründe nicht möglich war, eine ärztliche Beratung oder Behandlung in Anspruch zu nehmen?

🌀 **Issue:** Translation was in both TR1 and TR2


🌀 **Issue:** Translation was in TR1

🌀 **Issue:** Translation was in both TR1 and TR2

🌀 **Issue:** Translation was newly produced during the review discussion

# Operationalization: “rich points” – background

## Be aware:

-  The fact that one translation (e.g., TR1) was chosen over another one (TR2) does not automatically mean that TR2 was erroneous; sometimes, decisions need to be made between two equally working solutions and/or preferential choices are made.

# Operationalization: full identity of segments

EN	TRANSLATION_1	TRANSLATION_2	REV	
The waiting list was too long	Die Warteliste war zu lang	Die Warteliste war zu lang	Die Warteliste war zu lang	TR1 = TR2 = REV
There were no appointments available	Es gab keine freien Termine	Es waren keine Termine mehr verfügbar	Es gab keine freien Termine	TR1 = REV ( $\neq$ TR2)
Other reason	Andere Gründe	andere Gründe	Anderer Grund	REV ( $\neq$ TR1 or TR2)

- Checking for full identity in segment (= row) content for TR1 vs. REV and TR2 vs. REV
- Calculation for 268 segments in Excel
- As before, does not necessarily mean that the lack of identity is due to errors.



# Operationalization: Levenshtein distance

- 🌀 Automated metric, measuring the similarity between two strings (edit distance)
- 🌀 Minimum number of edits needed to change one string into another, with the allowable edits being insertion, deletion, or substitution of a single character\*
  - 🌀 *Stimme zu vs. Stimme zu (0) – Stimme nicht zu vs. Lehne ab (12)*
- 🌀 Here:
  - Comparing TR1 vs. REV and TR2 vs. REV
  - Calculation for 268 segments in Excel, mean across all segments
  - The higher the number, the more different the two strings are

\*[https://rosettacode.org/wiki/Levenshtein\\_distance](https://rosettacode.org/wiki/Levenshtein_distance)

# Operationalization: Post-review questionnaire for participants in PE conditions

## 🌀 Item “impact of PE on review” (closed question)

🌀 To what extent did the post-edited version shape the final Review version?

- 🌀 In (almost) all segments
- 🌀 In many segments
- 🌀 About half with the other translation
- 🌀 In only a few segments
- 🌀 In (almost) no segments

## 🌀 Item “ease of use” (open-ended question)

🌀 During the Review discussion, which translation version – the human or the post-edited one – was easier to work with? Please explain in more detail why one of the texts (if at all) was easier to work with.

# Preliminary results – German

Method	Rich points (%)	Identical segments (%)	Levensthein distance (mean)	Error count German (w/o punctuation/spelling)
<b>Baseline</b>	<b>Translation 1 (TR1): 29.10</b> Translation 2 (TR2): 17.45 Both: 30.69 New in Review (REV): 22.75	<b>TR1: 41.04</b> TR2: 4.10 Both: 16.42 New in Rev: 38.43	<b>TR1: 10.07</b> TR2: 15.73	n = 31 (least errors)
<b>Full PE</b>	<b>TR1: 32.28</b> TR2: 10.58 Both: 33.33 New in REV: 23.81	<b>TR1: 30.6</b> TR2: 6.34 Both: 26.12 New in Rev: 36.94	<b>TR1: 9.17</b> TR2: 16.14	n = 36
<b>Light PE</b>	<b>TR1: 31.75</b> TR2: 6.88 Both: 42.33 NEW in REV: 19.05	<b>TR1: 36.57</b> TR2: 5.97 Both: 25.0 New in Rev: 32.46	<b>TR1: 5.28</b> TR2: 14.73	<b>n = 74 (most errors)</b>

- **Across all settings**, TR1 (professional translation/human) had a much stronger impact/was closer to the final review version than TR2 (social scientist, human or PE).
- In **Baseline** and **Full PE**, this lead to good/OK quality.
- In **Light PE**, relying on TR1 seems to have caused mistakes, which apparently could not be mitigated by the light PE version.

# Preliminary results – Russian

Method	Rich points (%)	Identical segments (%)	Levenshtein distance (mean)	Error count Russian (w/o punctuation/spelling)
Baseline	<b>TR1: 25.13</b> TR2: 21.03 Both: 32.3 New in REV: 21.5	<b>TR1: 13.06</b> TR2: 7.84 Both: 8.59 New in REV: <b>70.52</b>	<b>TR1: 14.97</b> TR2: 17.09	n = 44 (most errors)
Full PE	<b>TR1: 26.1</b> TR2: 25.53 Both: 38.3 New in REV: 10.1	TR1: 19.78 <b>TR2: 24.25</b> Both: 20.15 New in REV: 35.82	<b>TR1: 9.96</b> TR2: 12.74	n = 36 (least errors)
Light PE	TR1: 17.5 <b>TR2: 24.5</b> Both: 37.5 New in REV: 20.5	TR1: 7.46 <b>TR2: 22.76</b> Both: 19.40 New in REV: <b>50.37</b>	TR1: 17.71 <b>TR2: 10.86</b>	n = 41

- In the **Baseline** version (most errors), TR1 - the human translation by the professional translator - influenced the review version a bit more, but there were also extremely high percentage of newly discussed elements.
- The influence of the human translation by the professional translators (TR1) vs. the post-edited version by the social scientist (TR2) seems even in the **Full PE** setting (least errors).
- In the **Light PE** setting, the influence of the post-edited version by social scientist was stronger.

# Results – Post-review questionnaire

Questions	Measurement	German Full PE (n=3)	German Light PE (n=3)	Russian Full PE (n=3)	Russian Light PE (n=3)
Impact of PE on review*	2 = in many segments, 3 = About half with the other translation, 4 = In only a few segments	3.67	3.3	3.3	2.67
Ease of use of PE vs. human translation	1 = equal 2 = human better 3 = PE better 4 = uncertain	1: 2 2: 1	1: 2 2: 1	1: 1 2: 2	1: 1 3: 1 4: 1

- According to the participants, TR1 (human translation by professional translator) shaped the review output in a (slightly) stronger way; the exception though is the Russian Light PE team.
- Overall, there was a tendency to judge both versions (MT/human) or the human translation better to work with; the exception, though, is the Russian Light PE team. Here, there was a tendency to prefer the PE version.

\*To what extent did the post-edited version shape the final Review version?

# Conclusions

- 🌀 **German:** TR1 (i.e., the human translation by professional translator) did contribute much more to the final review version than TR2 (i.e., version by social scientist, human or post-edited).
  - 🌀 In two out of three teams, this impact turned out to be successful.
  - 🌀 In the Light PE setting, this impact was less successful; additionally, the light PE version by the social scientist was apparently not considered as suitable/could not mitigate mistakes from TR1.
- 🌀 **Russian:** The picture is mixed; the influence of the post-edited version (TR2, full and light PE) by the social scientist was quite strong and the resulting review output was of good/OK quality.
- 🌀 Including post-editing at the translation stage can make sense; employing light PE may be less successful, though, in particular if the second translation is (also) weak.

# Outlook: Further analyses/research

- 🌀 Statistical analyses beyond pure descriptive results, linking quality of individual segments to results presented here.
- 🌀 Linking results to other papers or research, e.g., how “usable” is PE (Keck et al., ESRA), how are the PE versions perceived during the review discussion (Dorer et al., future work), or what kind of errors need to be corrected in MT/how good are the initial translations/PE versions before they go into the review (*coding ongoing*).
- 🌀 Comparability notions of each participant, their individual skills, and discussion routines do certainly influence our results.
  - 🌀 We encourage to replicate the study/conduct similar studies and/or analyse the data once the data is available online in a research repository.

Thank you!

[dorothee.behr@geis.org](mailto:dorothee.behr@geis.org)

