8th Live Subtitling and Accessibility Symposium
Barcelona, 19 April 2023

Measuring the quality of interlingual live subtitles via respeaking: insights from the SMART project

Elena Davitti, University of Surrey, UK

Annalisa Sandrelli, Università degli Studi Internazionali di Roma(Italy)

Economic and Social Research Council

UKRI Economic and Social Research Council

SMART

Shaping Multilingual Access through Respeaking Technology

ESRC UK, 2020-2023, ES/T002530/1

[redacted]

Research team

Elena Davitti, PI (University of Surrey, CTS)
Simon Evans, CI (University of Surrey, School of Psychology)
Lucile Desblache, CI (University of Roehampton)
Pablo Romero-Fresco, CI International (University of Vigo, Spain)
Annalisa Sandrelli, CI International (UNINT Rome, Italy)
Tomasz Korybski, Research Fellow (University of Surrey, CTS)
Zoe Moores, Research Fellow (University of Surrey, CTS)
Anna-Stiina Wallinheimo, Research Fellow (University of Surrey, CTS)

Advisory Board

Academic members

Jan-Louis Kruger (Macquarie University)
Franz Pöchhacker (University of Vienna)
Aline Remael (University of Antwerp)

Industry members

Ai-Media
Sky
SUB-TI and FRED FILM RADIO

Website: **https://smartproject.surrey.ac.uk/**
Twitter: **@SMARTatSurrey**

[redacted]

Quality in SMART

Quality as a **multidimensional**, **elusive** and **relative** concept

Our focus is on **ACCURACY** in **interlingual respeaking**

To refine our understanding of what contributes to output accuracy

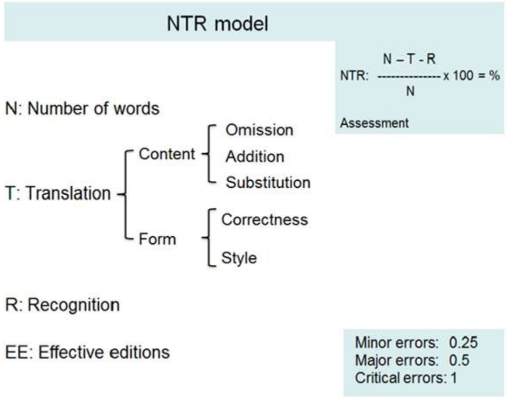
- what accuracy benchmark can language professionals achieve after 25h of upskilling
- which variables are predictors of accuracy
- how do different conditions impact on performance

[redacted]

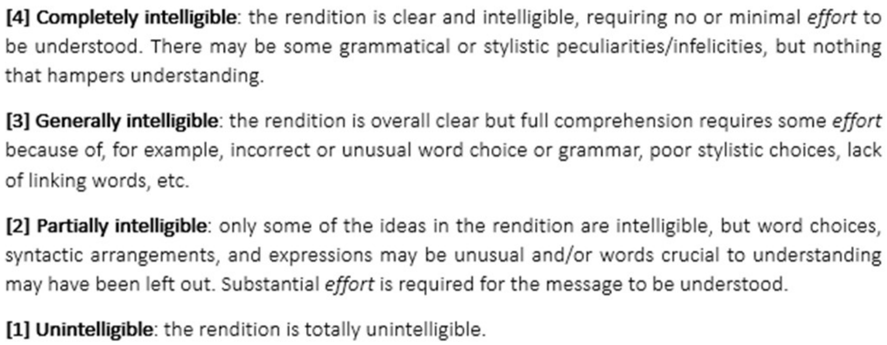
Approach to measuring accuracy

Accuracy operationalised as **informativeness + intelligibility**

Accuracy measured via NTR model (Romero-Fresco and Pöchhacker 2017) applied to 153 performances under different scenarios.



Intelligibility scale (based on Tiselius 2009) to determine high and low performers, which was validated in the results obtained.



[4] Completely intelligible: the rendition is clear and intelligible, requiring no or minimal *effort* to be understood. There may be some grammatical or stylistic peculiarities/infelicities, but nothing that hampers understanding.

[3] Generally intelligible: the rendition is overall clear but full comprehension requires some *effort* because of, for example, incorrect or unusual word choice or grammar, poor stylistic choices, lack of linking words, etc.

[2] Partially intelligible: only some of the ideas in the rendition are intelligible, but word choices, syntactic arrangements, and expressions may be unusual and/or words crucial to understanding may have been left out. Substantial *effort* is required for the message to be understood.

[1] Unintelligible: the rendition is totally unintelligible.

Participants

51 language professionals selected out of 250+ applicants

Professional backgrounds: 2,000h+ work experience in translation, interpreting and/or pre-recorded/live subtitling; majority with 3+ professions (composite profiles)

Languages: 17 participants between EN and each romance language (French/Italian/Spanish);
32 EN>Romance; 19 Romance>EN

Demographics: 8 males, 43 females (*Mage* = 40.12 years, *SD* = 10.97 years); from 11 countries (UK, Spain, Italy, France, Germany, Belgium, Australia, Argentina, New Zealand, USA, Peru)

Materials

- Intra and interlingual tests – INTERLINGUAL results analysed
- 12 speeches
  - 4 languages: English, Spanish, French, Italian
  - 3 different source input conditions

SPEED	PLANNED/UNPLANNED	MULTIPLE SPEAKERS
M duration 15'+	*M* duration 12'	*M* duration 12'
140 wpm	110 wpm	120 wpm

- Controlled variables: topic (respeaking-themes), vocabulary (brief), numbers
- Randomisation of testing (ABC-CBA)

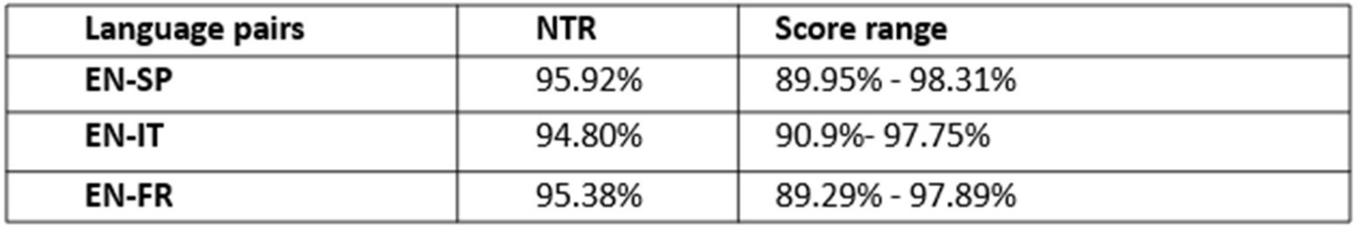
> I really loved that, it [...] enabled demonstration of practical skills with as little interference from an unfamiliar topic as possible.

> Testing materials [...] could correspond to the difficulty level to everyday demanding tasks

Accuracy after 25h of upskilling

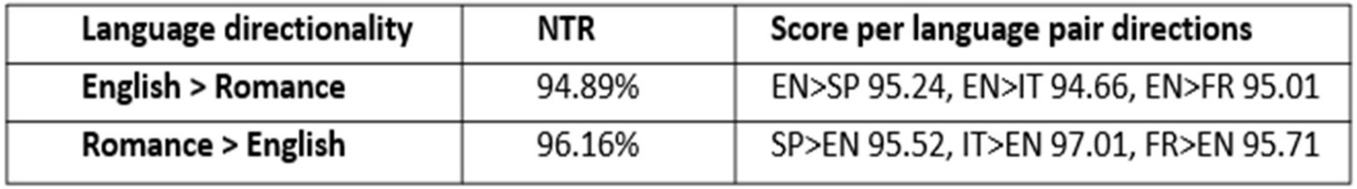
Average NTR score across all participants and conditions: 95.37%

Average NTR scores per language pair



Language pairs	NTR	Score range
EN-SP	95.92%	89.95% - 98.31%
EN-IT	94.80%	90.9%- 97.75%
EN-FR	95.38%	89.29% - 97.89%

Average NTR scores per language directionality



Language directionality	NTR	Score per language pair directions
English > Romance	94.89%	EN>SP 95.24, EN>IT 94.66, EN>FR 95.01
Romance > English	96.16%	SP>EN 95.52, IT>EN 97.01, FR>EN 95.71

Language professionals

SUBGROUPS

HIGH/LOW PERFORMERS

- High performers: 27/51
- Low performers 24/51

** Informativeness threshold: 96%*
** Intelligibility threshold: 16*
TOT: 45/153 performances

PROFESSIONAL CLUSTERS

- Spoken-to-Spoken: 17/51
- Spoken-to-written: 16/51
- Mixed: 16/51

** 2 outliers*

Accuracy after 25h of upskilling

HIGH vs LOW performers

Significant difference in accuracy performance across all scenarios, *p* < .001

***M* = 96.3% (high) and *M* = 94.4% (low)**

***M* = 97.1% (top 12) and *M* = 94.8% (other 39)**

PROFESSIONAL CLUSTERS

No statistical differences between clusters, *p* > .05

- Spoken to spoken: *M* = 95.4%
- Spoken to written: *M* = 95.3%
- Mixed: *M* = 95.3%

Accuracy predictors: errors

Across all participants and all source inputs

Errors as **negative** predictors

- **Omissions (OM)** (β = -1.12, p < .001) [MAJ β = -.071; MIN β = -.19]
- **Recognition (R)** (β = -.34, p < .001)
- **Substitutions (SUB)** (β = -.17, p < .001)

Effective editions (EE) as **positive** predictor (β = .31, p = .03)

High vs low performers

MajOM (HM = 24.84; LM = 36.59), p = .004.

MinOM (HM = 34.84; LM = 40.10), p = .09.

MajSUB (HM = 4.58; LM = 6.44), p = .03.

MajR (HM = 4.43; LM = 7.29), p = .07.

MinADD (HM = 1.52; LM = 2.78), p = .02.

MinCORR (HM = 12.02; LM = 18.69), p = .02.

EE Ave: F(1, 49) = 4.71, p = .04. H used EEs more (M = 43.19) than L (M = 37.17).

High performers

MinOM (β = -.35), p = .027.

MajOM (β = -.58), p < .001.

Low performers

MajOM (β = -.72), p < .001.

[redacted]

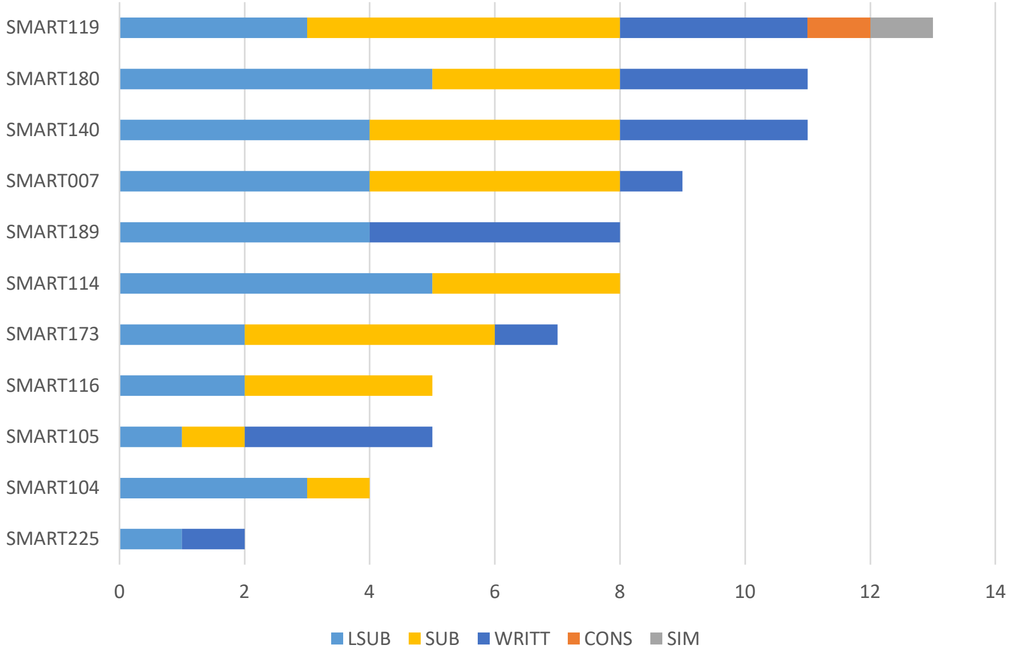
Accuracy predictors: professional background

Professional background – ALL

No statistical differences (*p* > .05) between professional clusters (spoken-to-spoken; spoken-to-written; mixed) pointing to no cluster providing an advantage over another, but…

Linear Regression

- Live subtitling as a **positive predictor** *F*(1, 49) = 2.38, *p* = .02, β = .32.



Impact of source input on performance

Average NTR scores **source input condition**

Significant difference as *p* = .008

SPEED: 94.8%

- **Spanish 95.3%**
- **French 95.1%**
- **Italian 95.9%**

PLANNED/UNPLANNED: 95.8%

- **Spanish 95.6%**
- **French 96%**
- **Italian 94.9%**

MULTIPLE SPEAKERS: 95.5%

- **Spanish 95.9%**
- **French 95.13%**
- **Italian 95.5%**

Impact of source input on performance

HIGH (27) vs LOW (24) performers:
significant difference in accuracy performance across all scenarios, *p* < .001

- Speed: *M* = 95.5% (high) and *M* = 93.9% (low)
- PU: *M* = 96.8% (high) and *M* = 94.8% (low)
- MS: *M* = 96.5% (high) and *M* = 94.4% (low)

TOP (12) vs OTHERS (39) performers:
significant difference in accuracy performance across all scenarios, *p* < .001

- Speed: *M* = 96.8% (top) and *M* = 94.1% (others)
- PU: *M* = 97.2% (high) and *M* = 95.4% (low)
- MS: *M* = 97.2% (high) and *M* = 95.0% (low)

A qualitative approach: TAP data analysis

TAP comments produced by the 27 **HIGH performers**

- **Speed:** 8 subjects
- **Multiple speakers:** 15 subjects
- **Planned/unplanned:** 22 subjects

The TAP comments were analysed and grouped by **thematic category** to identify the root cause of the reported problem and the strategy adopted to tackle it (if any)

- **Source-input related**
- **Technique-related**
- **Technology-related**
- **Person-related**

Key findings from TAP data analysis

- Most TAP comments focused on **TECHNIQUE** rather than on the characteristics of the source materials.
- Most frequently mentioned challenges:
  - **décalage** (keeping up the pace)
  - **live error correction**
  - (**audiovisual**) **monitoring**
  - **software-adapted delivery (SAD):** clear pronunciation (dictionary form) + neutral intonation + clear articulation + strategic pausing behaviour for chunking
- Comprehension issues mentioned in some TAPs, but often related to other challenges (i.e., missed part of a sentence because of time lag, voice overlap, typing a correction, etc.)
- Low number of comments on technology *per se*. Some comments on **human-machine interaction** (i.e., insufficient vocabulary training, inefficient macros, etc.)

SPEED task

On average, more TAP comments on the SPEED task > the longest and hardest test (lowest NTR scores)

- Most of the comments focus on
  - comprehension problems
  - the effect of speed on the respeaker's SAD
  - output monitoring
  - performing live corrections at speed
- All the challenges encountered in the other tasks are magnified by speed
- Suggested coping strategies:
  - increasing décalage to gain more context and then compressing
  - anticipating potential recognition problems and avoiding certain words or typing them
  - strategic omissions of secondary information

MULTIPLE SPEAKERS task

- The majority of TAP comments are focused on technique:
  - comprehension problems often related to décalage
  - SAD often mentioned in conjunction with output monitoring or translation difficulties
  - issues with sound and volume management
  - overlapping talk/cross over between speakers (question-answer)

- Coping strategies:
  - omission of less important items (e.g. hesitations, interjections, conversation markers...)
  - pausing to improve recognition (better chunking)
  - live correction: pause, wait for the text to be displayed, correction

PLANNED/UNPLANNED task

- Again, the most common comments are on technique:
  - SAD issues
  - Output monitoring (multiple visual input, in relation to the questions that were displayed in a written form)
- Technology: software preparation and working set-up
- A higher number of TAPS on the source material, i.e., audio quality, technical topic and complex structures > comprehension problems

Coping strategies:

- longer décalage for better comprehension and better TL reformulation
- omitting items that have not been understood
- prioritising meaning over error correction
- anticipating recognition problems and using macros or typing

Implications of TAP analysis

- When reporting **problems**, subjects were often able to indicate **solutions**
- Given the short duration of the course, the fact that subjects have been able to automate some behaviours and develop coping strategies is encouraging

- Examples:
  - dictating has become second nature;
  - SAD still poses challenges but overall has become more of a habit
  - being able to anticipate recognition problems and using either synonym, macro or typing;
  - pausing frequently to enable *Dragon* to display the output faster;
  - chunking to avoid using too much punctuation;
  - strategic omissions (of less important items or items that have not been fully understood)

Conclusions/I

- Large-scale validation of NTR model (**intertextual dimension**)
- Significance of NTR data enhanced by integration of an intelligibility scale (**intratextual dimension**)
- Other aspects of the live subtitling service (such as delay) to be added for a more holistic view (**instrumental dimension**)
- Need to **review and validate the accuracy benchmark** for interlingual respeaking?

Conclusions/II

- Integration of statistical methods allowed for focus from **macro** (all participants) to **micro** (specific subgroups) to build **evidence-base** – requires expertise
- Implications for upskilling:

  - **Evidence** that experience in live (intralingual) subtitling provides a clear advantage: automated processes (interaction with technology) that make it easier to add language transfer component

  - **Evidence** that other profiles (spoken-to-spoken, mixed...) may also acquire interlingual respeaking skills, but may need to focus more on the human-machine interaction component

- Modular approach to upskilling ("pick and choose")

8th Live Subtitling and Accessibility Symposium
Barcelona, 19 April 2023

Thank you for your attention!

e.davitti@surrey.ac.uk
annalisa.sandrelli@unint.eu

[redacted]

UKRI Economic and Social Research Council

SMART

UNIVERSITY OF SURREY

Centre for Translation Studies