MT Quality and its effects on post-editors and end-users

Maria Stasimioti and Vilelmini Sosoni

Department of Foreign Languages, Translation and Interpreting,
Ionian University, Corfu, Greece

43rd Translating and the Computer Conference,TC43
16-18 November 2021

Introduction

- ✓ Machine translation (MT) quality has improved considerably in recent years, not least because of the rise of neural machine translation (NMT) models.
- ✓ MT is thus increasingly used in industrial settings to produce raw translations to be further post-edited by translators (Lommel and DePalma, 2016; Koponen, 2016).

Post-editing (PE) is influenced by the MT system used, as well as the number and the type of errors included therein, given that some errors have been found to be more demanding than others (Koponen, 2012). Numerous studies have shown that the improved quality of the NMT system output, especially at the level of fluency, requires the correction of fewer segments, mainly due to the lower number of morphological errors (Castilho et al., 2017a, 2017b). However, this does not always result in lower PE effort mainly due to NMT errors at the level of accuracy being more difficult to identify and correct, compared to the obvious word-order errors and disfluencies occurring in phrase-based machine translation (PBMT) and statistical machine translation (SMT) outputs (Castilho et al, 2017b), requiring, thus, longer post-editing times (Carl and Báez, 2019).

PE effort

✓ According to Krings (2001), there are three categories of post-editing effort:

(1) the temporal effort, which refers to the time taken to post-edit a sentence to a particular level of quality, and which "is undoubtedly the most important aspect of post-editing from an economic perspective", yet "only the obvious external form of post-editing effort" (Krings 2001: 54),

(2) the technical effort, which refers to keystroke and mouse activities such as deletions, insertions, and text re-ordering and

(3) the cognitive effort, which refers to the "type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation" (Krings 2001: 179).

✓ Interestingly, temporal, technical, and cognitive effort do not necessarily correlate.

Reception studies

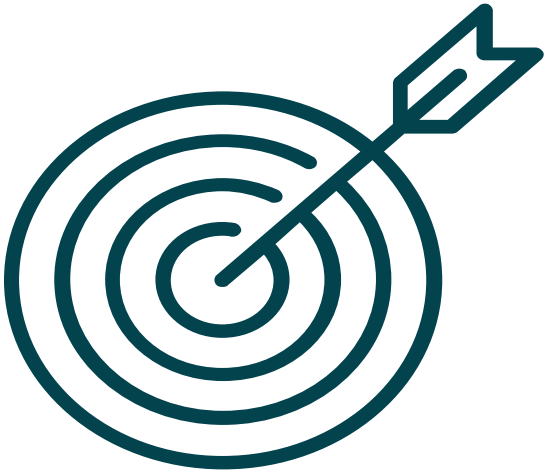
PE gives insights into the nature of the translators' job and can be used to also evaluate the quality of the final product on the basis of metrics such as TER

But

It does not provide any indication about the acceptability of the final product by its end users, i.e. its readers-recipients.

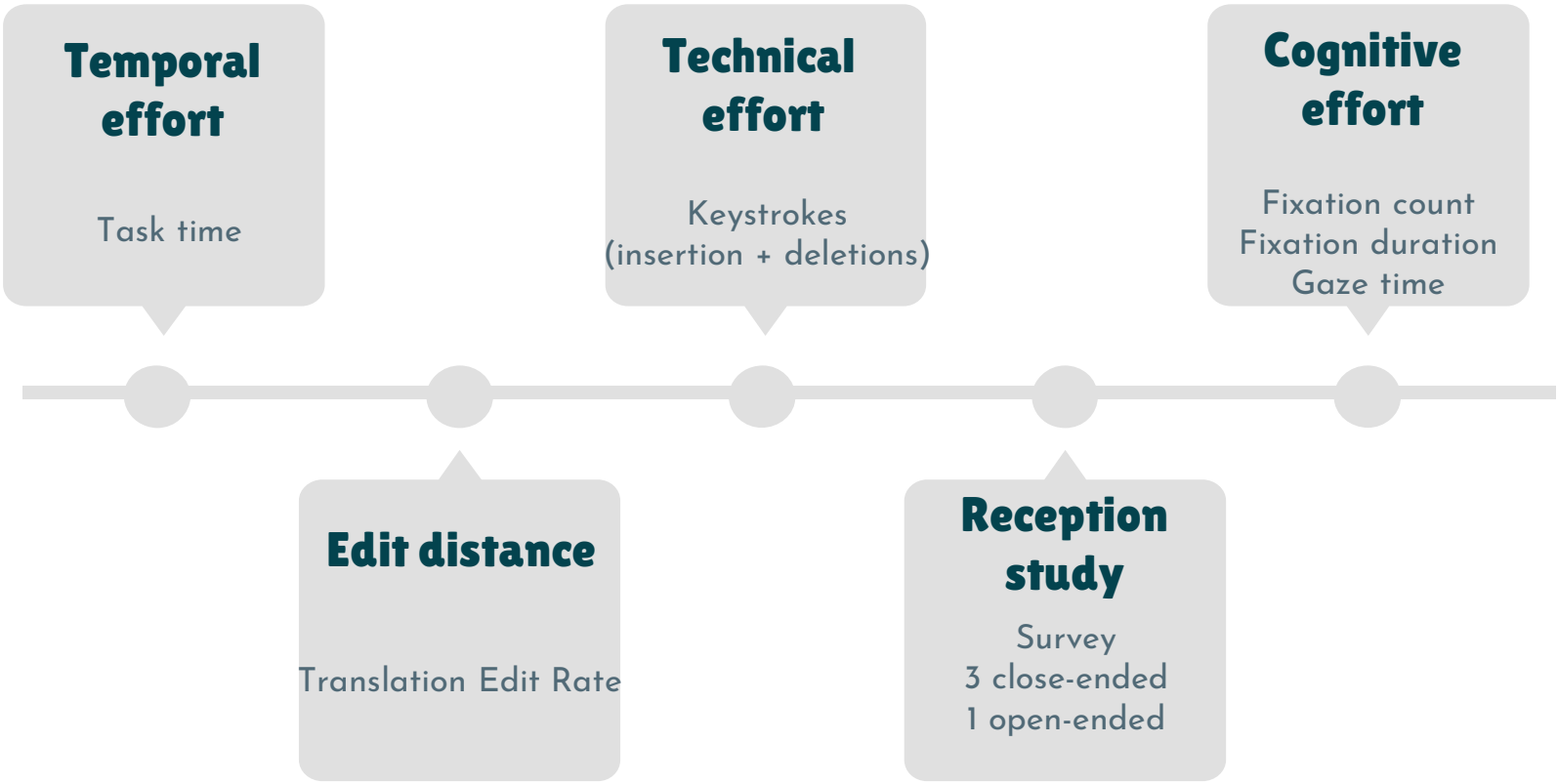
For that, reception studies are used which offer a uniquely thorough picture of the life and afterlife of texts and give a central role to readers as the final target text readers-recipients.

Aim of the study

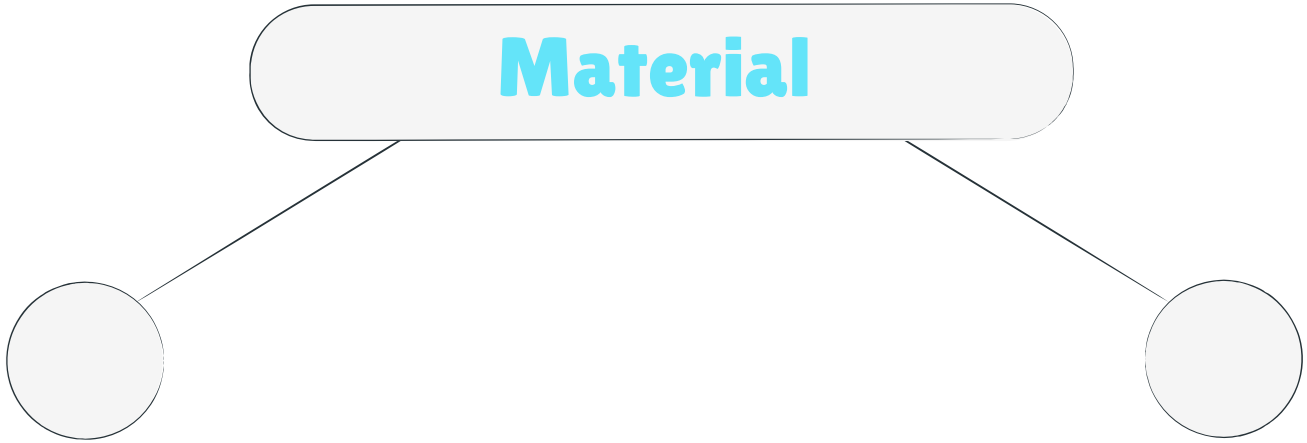


- Compare the **temporal, technical and cognitive effort** required for full PE of NMT output with the effort required for full PE of SMT output in the English-Greek language pair
- Compare the **extrinsic quality** of final texts, i.e. its acceptability by the end-users in terms of readability and comprehensibility

Methodology



Material



Source Texts

2 short (~140 words) semi-specialized authentic texts about the 2019 EU elections with comparable readability scores (between 1200L and 1300L)

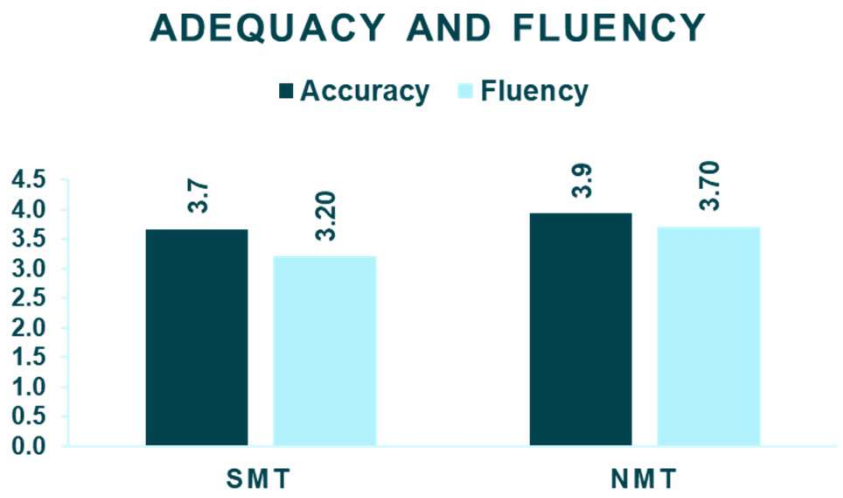
MT outputs

SMT system developed by Google
NMT system developed by Google

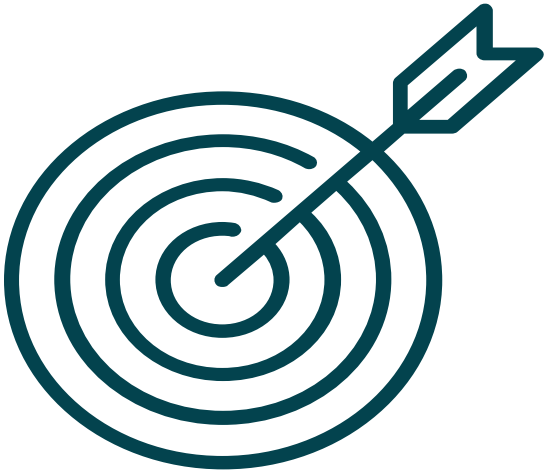
SMT & NMT outputs: AEMs and accuarcy/fluency rating

	SMT	NMT
BLEU	0.38	0.37
METEOR	0.51	0.51
WER	0.50	0.51
TER	0.52	0.48

Table 3. Average of AEMs per system



Experimental design



- Participants were asked to fully post-edit the NMT and the SMT outputs
- The temporal effort (total task time), the technical effort (keystrokes: insertions and deletions) and the cognitive effort (number of fixations, mean fixation duration and total gaze time) they expended during the PE tasks were registered using a Tobii X2-60 remote eye-tracker and the Translog-II software
- A meeting was organized before and a warm-up session preceded the actual execution of the PE tasks

Participants

Gender	Female	10
	Male	0
Age distribution	18-24	5
	25-34	4
	45-54	1
Education level	Undergraduate degree holder	7
	Postgraduate degree holder	2
	PhD holder	1
Degree type	Translation	5
	Other	5
Experience in Translation	Yes	3
	No	7
Experience in PE	Yes	0
	No	10

PE training

"Translation tools" compulsory module

- theory and history of MT and PE,
- basic principles of MT technology,
- analysis of the dominant systems in the market,
- importance of controlled language and pre-editing for MT,
- quality metrics and evaluation of MT output,
- PE levels of quality,
- PE effort and productivity,
- MT output error identification,
- MT engine implementation in the translation workflow,
- post-editor profile and associated skills.

Reception study

Question 1: The comprehensibility of the text was:

1. Very easy 2. Easy 3. Fairly easy 4. Not very easy 5. Not at all easy

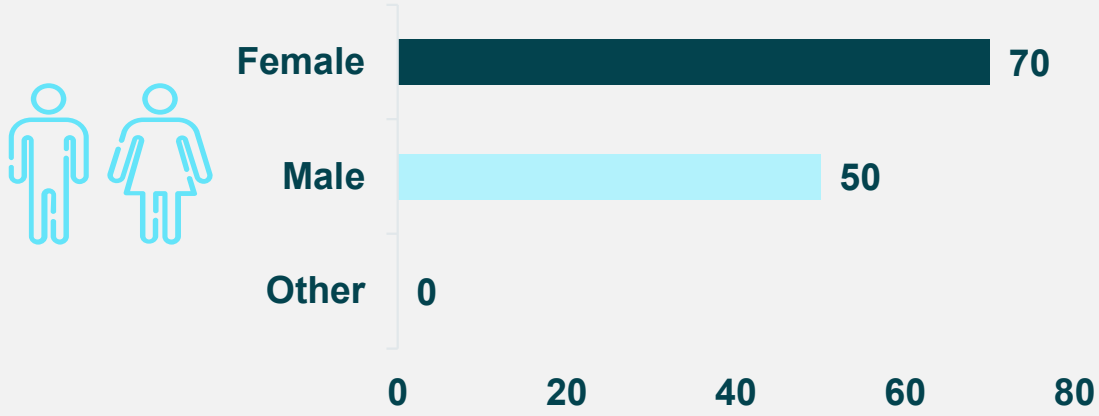
Question 2: The flow of the text was:

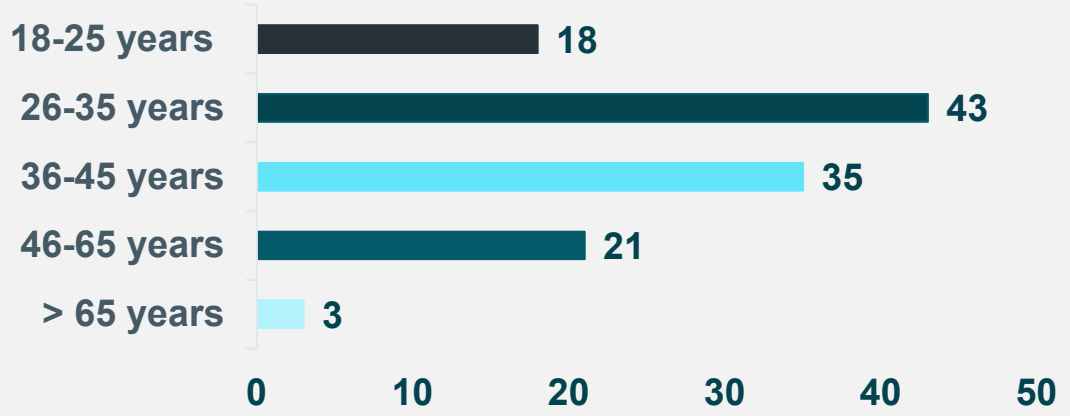
1. Very good 2. Good 3. Fairly good 4. Not very good 5. Not good at all

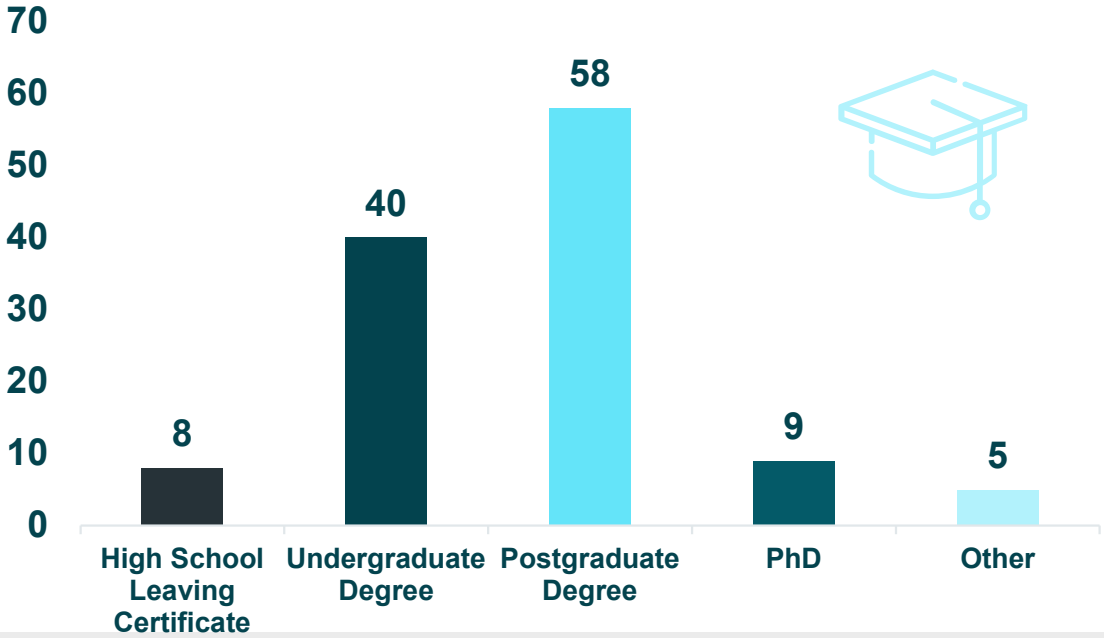
Question 3: Do you think the text could be improved?

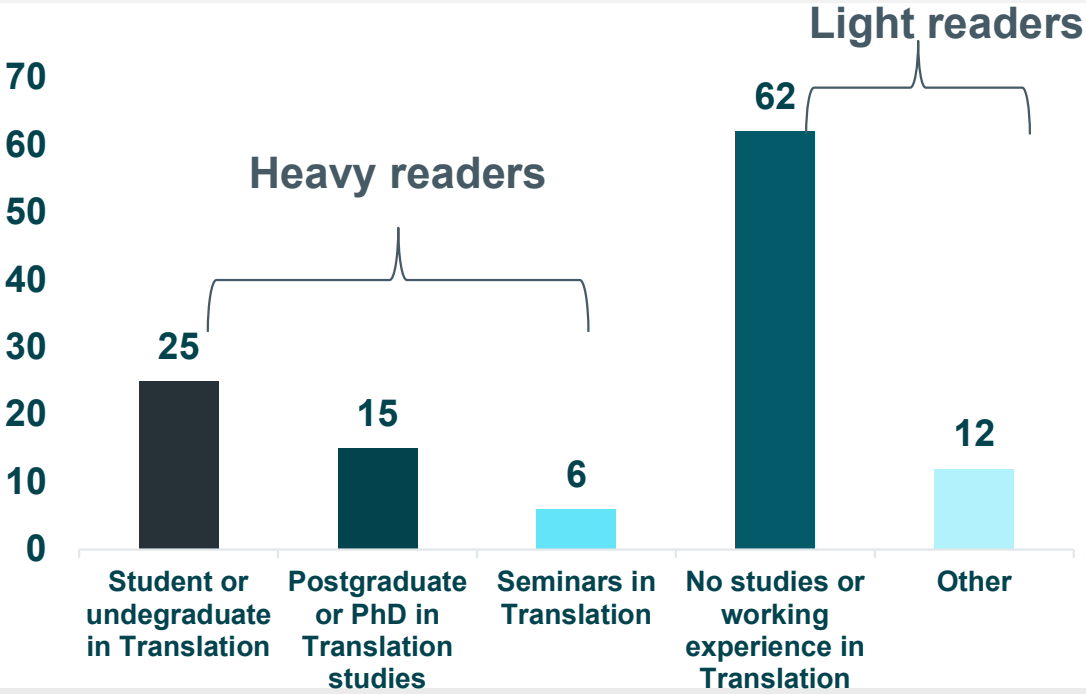
1. Yes 2. No

Question 4: If yes, how?

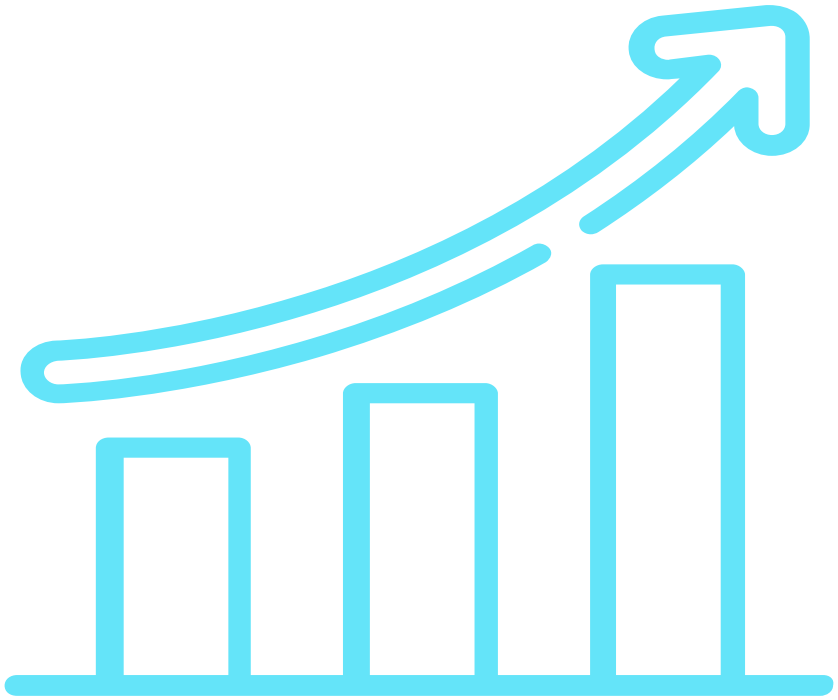

Female
70
Male
50
Other
0
0
20
40
60
80


18-25 years
18
26-35 years
43
36-45 years
35
46-65 years
21
> 65 years
3
0
10
20
30
40
50

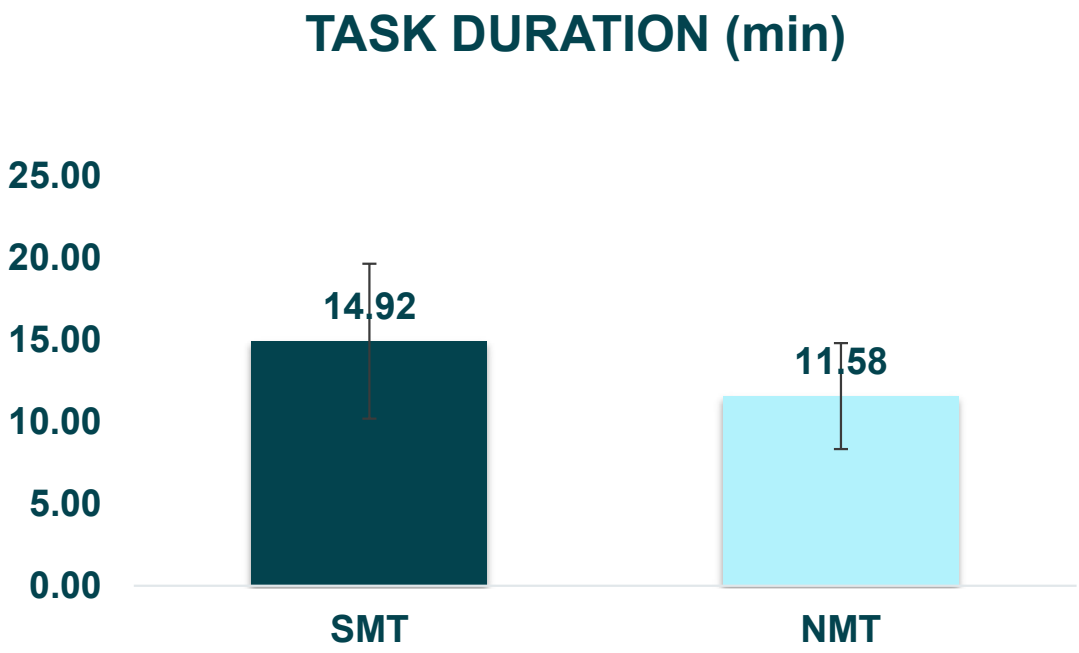

70
60
50
40
30
20
10
0
8
40
58
9
5
High School Leaving Certificate
Undergraduate Degree
Postgraduate Degree
PhD
Other


Light readers
Heavy readers
70
60
50
40
30
20
10
0
25
15
6
62
12
Student or undegraduate in Translation
Postgraduate or PhD in Translation studies
Seminars in Translation
No studies or working experience in Translation
Other

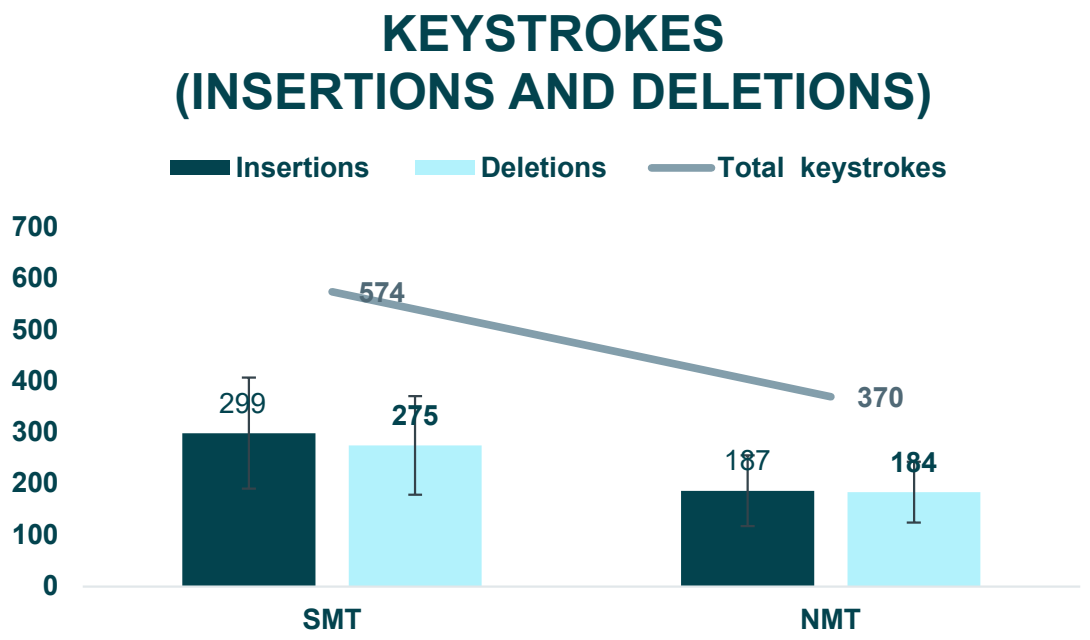
Findings and analysis



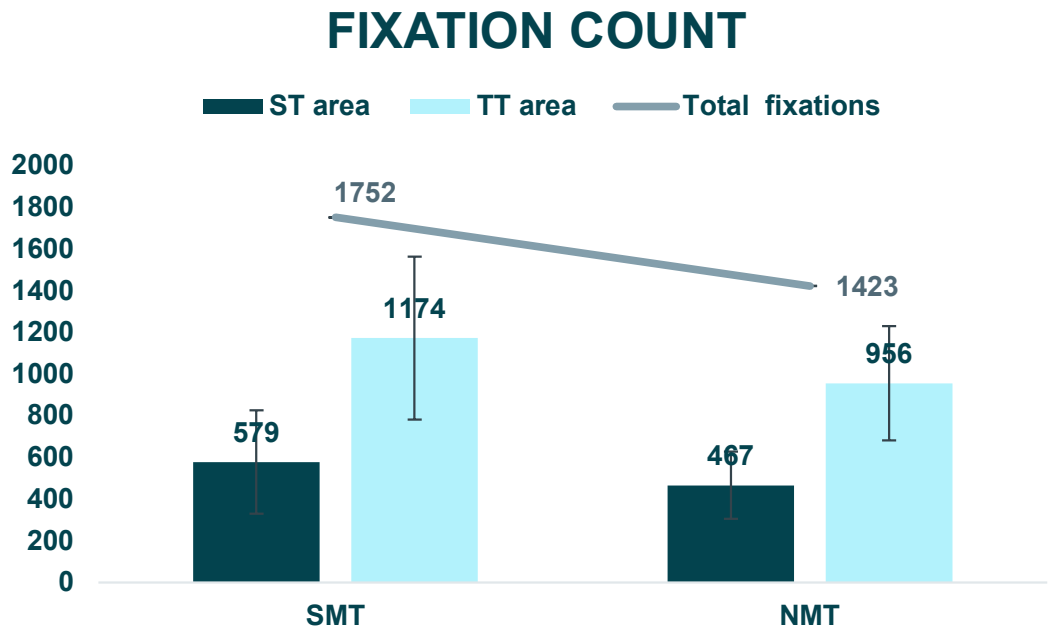
Temporal effort: task duration



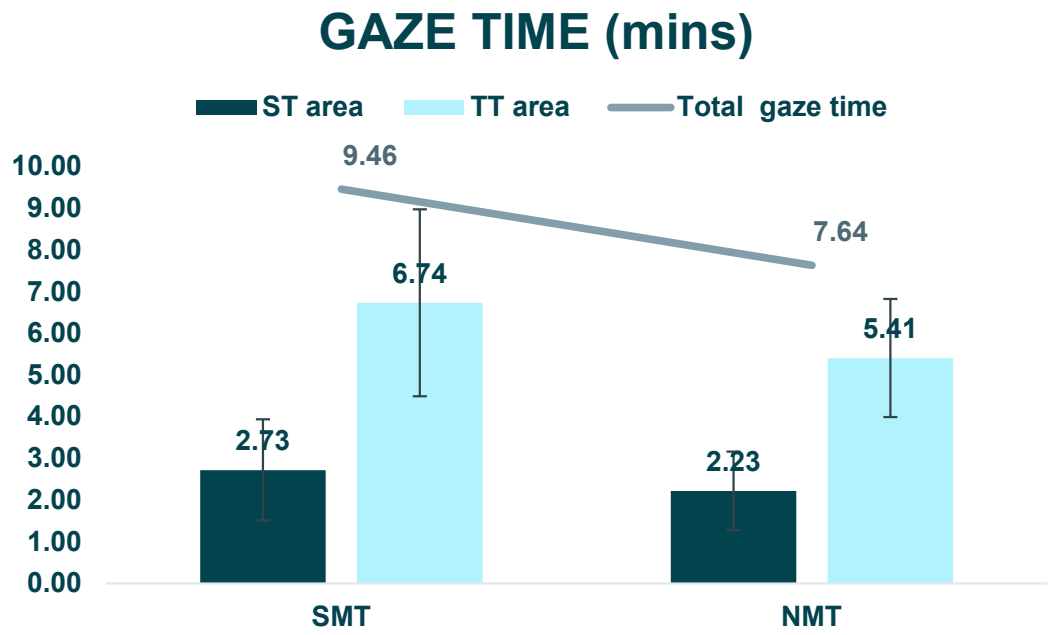
Technical effort: keystrokes



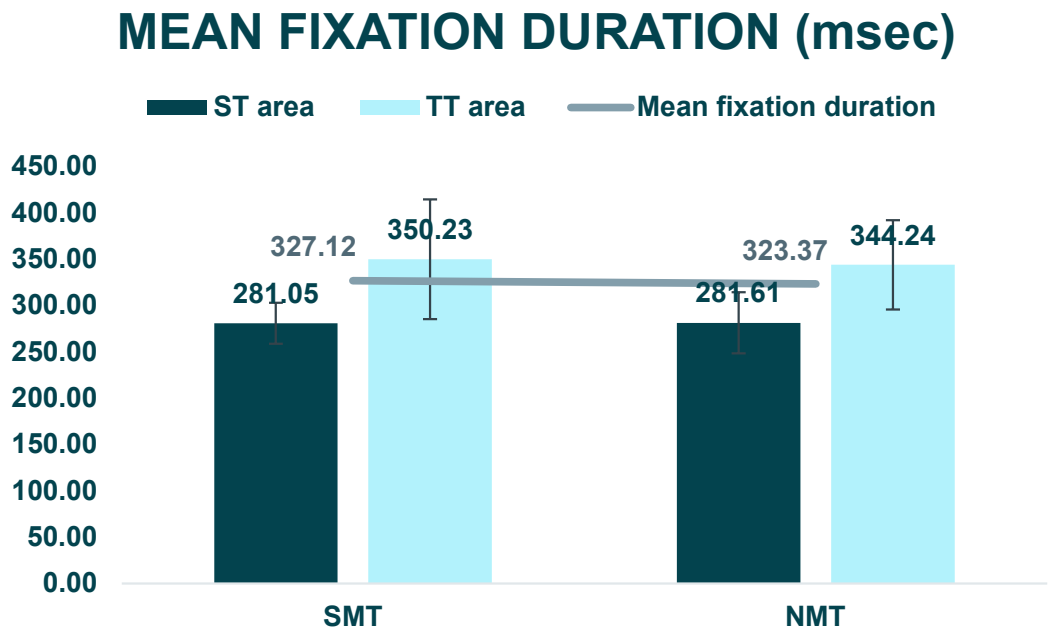
Cognitive effort: fixation count



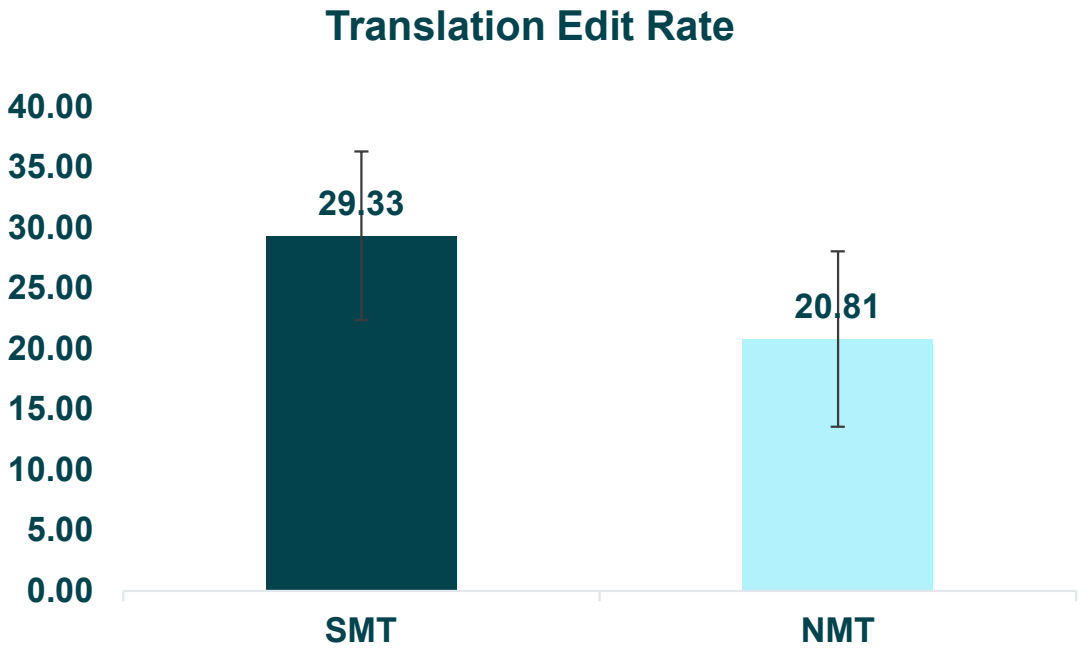
Cognitive effort: gaze time



Cognitive effort: mean fixation duration

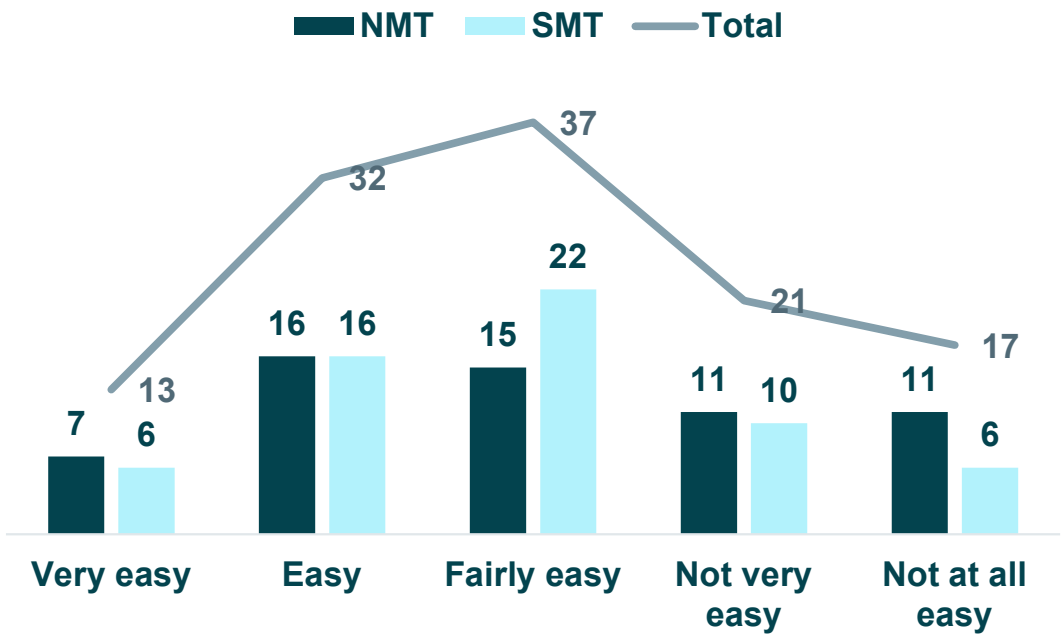


Edit Distance
TER

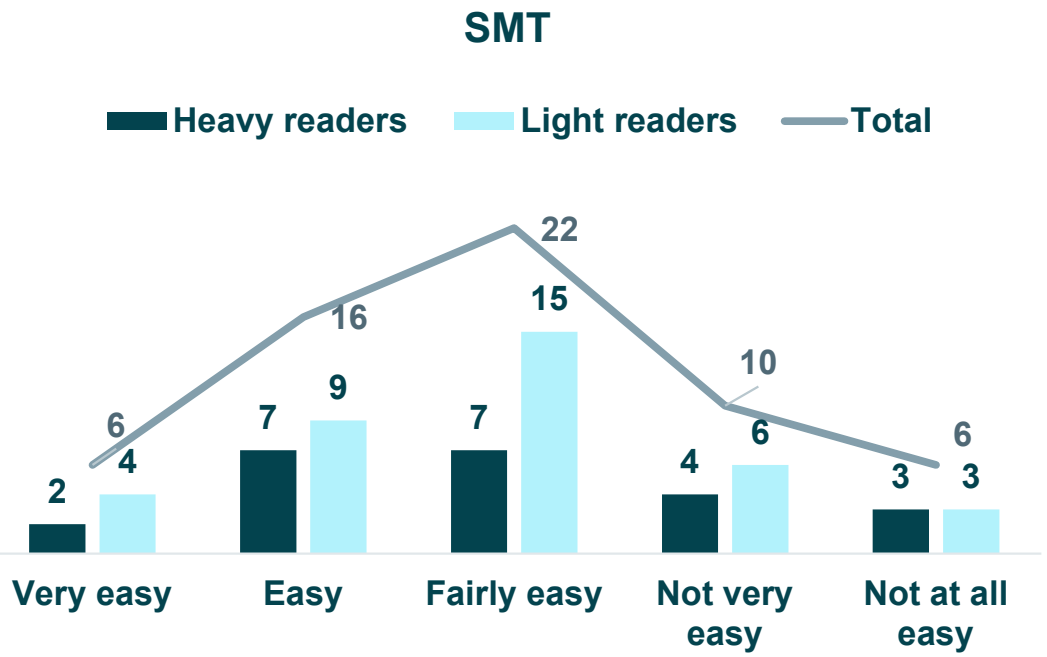


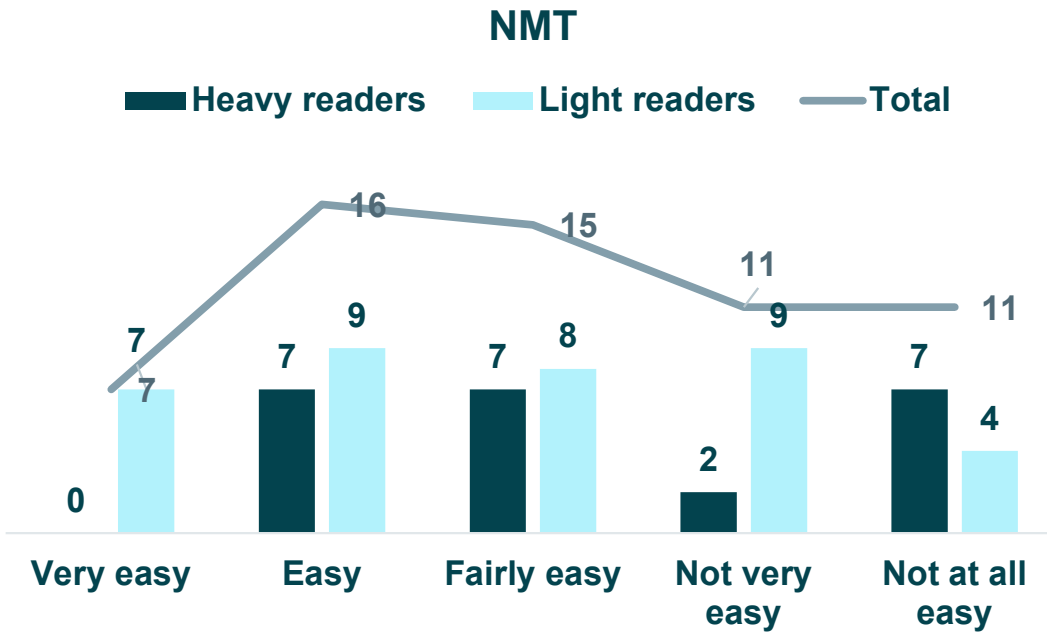
Reception study:

The comprehensibility of the text was...



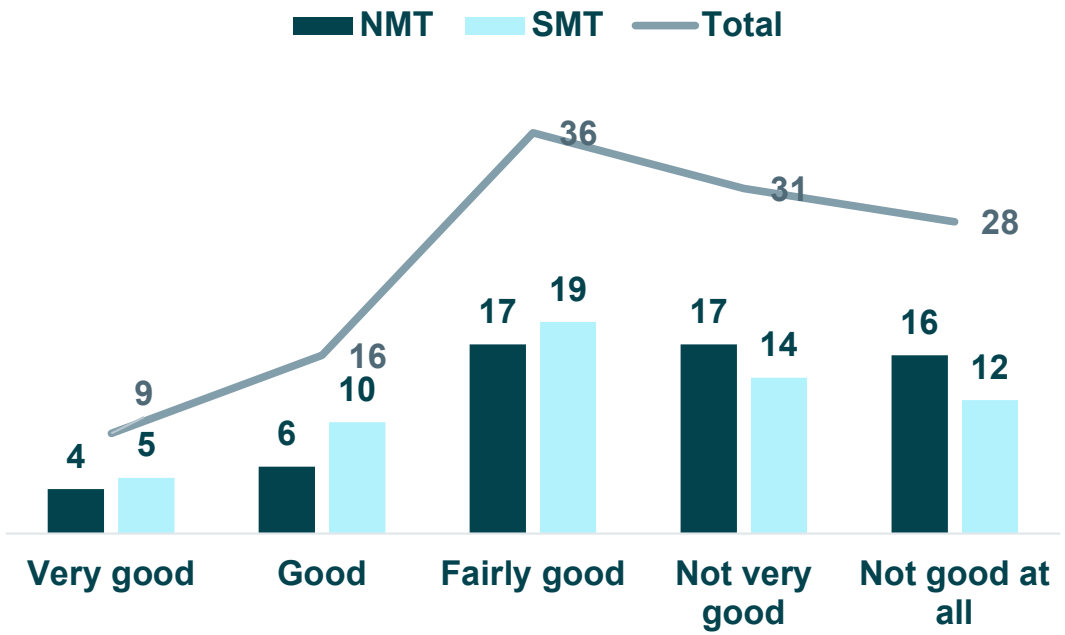
Reception study: The comprehensibility of the text was...



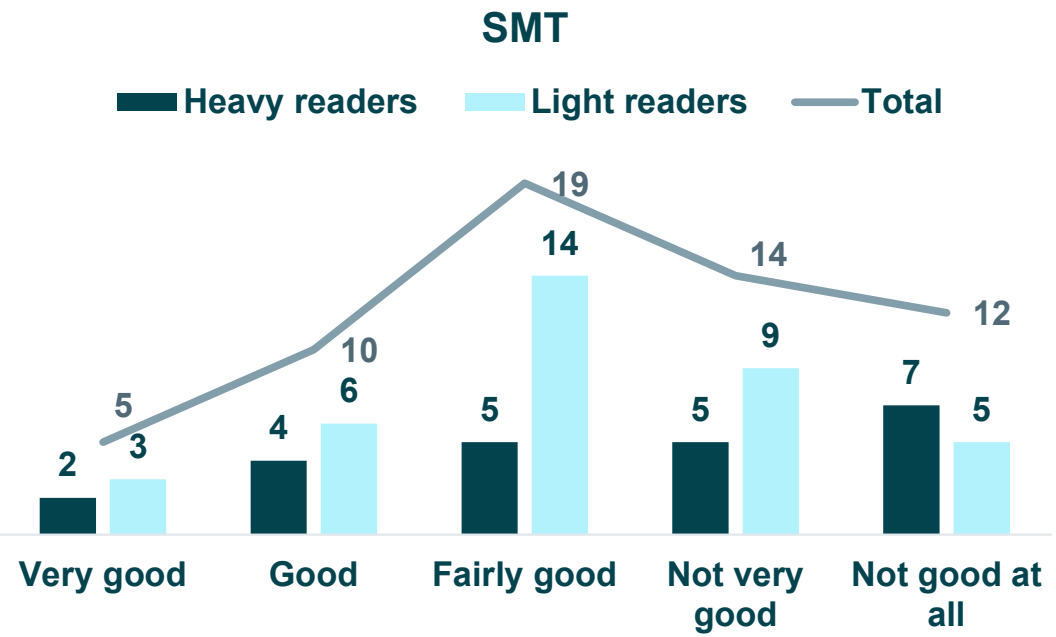


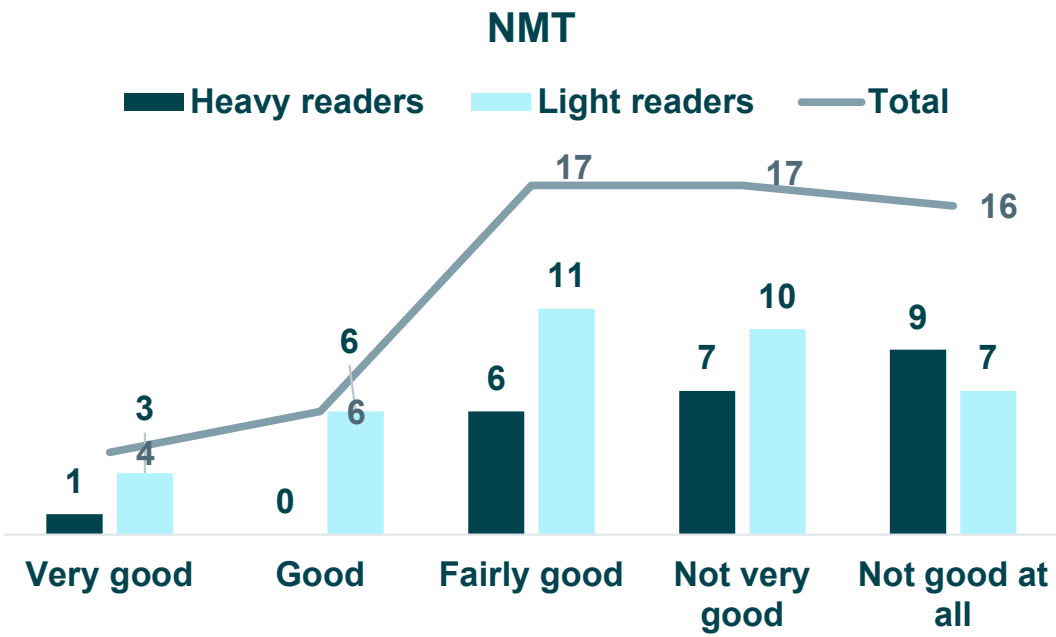
Reception study:

The flow of the text was...



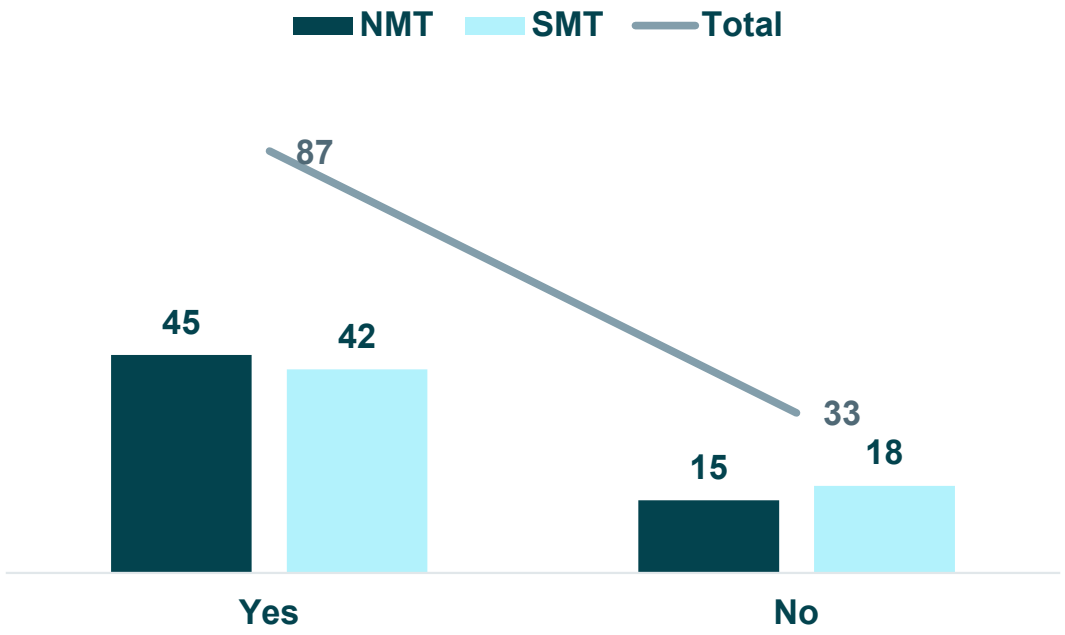
Reception study: The flow of the text was...



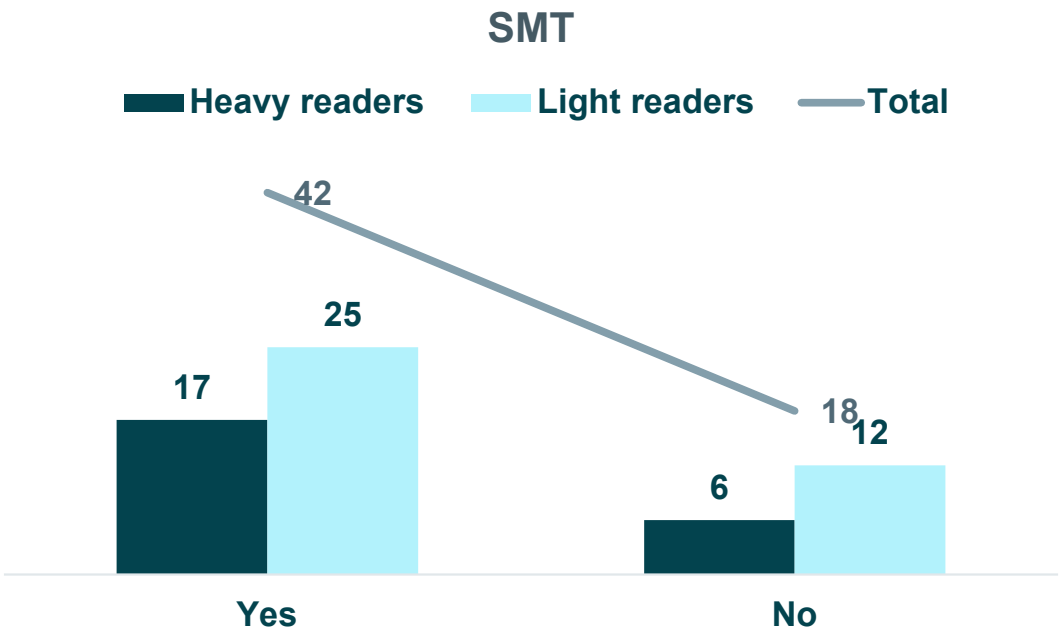


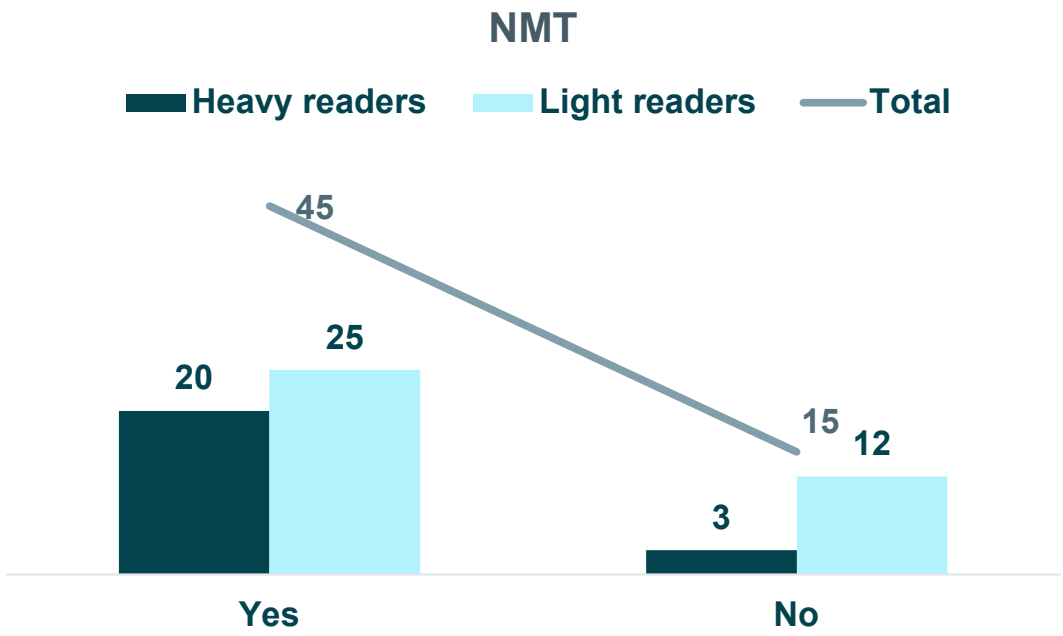
Reception study

Do you think the text could be improved?



Reception study: Do you think the text could be improved?





Conclusion

The analysis reveals that:

- The PE the SMT was more demanding in terms of temporal, technical and cognitive effort.
- The PE of the SMT output required more edits.
- The comprehensibility of both the SMT and NMT post-edited texts was fairly easy for both the heavy and light readers.
- The flow of the SMT post-edited texts was fairly good for both the heavy and light readers, while flow of the NMT post-edited texts was fairly good for the light readers and not very good for the heavy readers.
- Both the SMT and NMT post-edited texts could be improved.

Thank you

Do you have any questions?

Maria Stasimioti
stasimioti@ionio.gr

Vilelmini Sosoni
sosoni@ionio.gr

Acknowledgement We would like to thank the HUBIC Lab for providing the Tobii X2-60 remote eye-tracker for the purposes of this study.