The Impact of MT Quality Estimation on Post-Editing Effort

Carlos Teixeira **Sharon O'Brien**
carlos.teixeira@dcu.ie sharon.obrien@dcu.ie

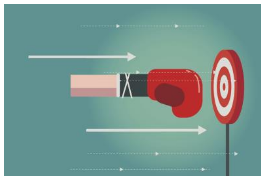
Centre for Translation and Textual Studies (CTTS)
ADAPT Centre for Digital Content Technology
Dublin City University (Ireland)

MOTIVATION

- Professional translators edit suggestions coming from translation memories (TM) and machine translation (MT)
- Handling those two **types of linguistic support** requires different strategies
- TM suggestions incorporate metadata to increase efficiency and quality (e.g. Fuzzy Match scores)
- QE scores are an attempt to provide **relevant metadata** for MT suggestions

Novelty: Despite recent advances in QE research, little is known about the **real impact of QE scores** on the translation process.

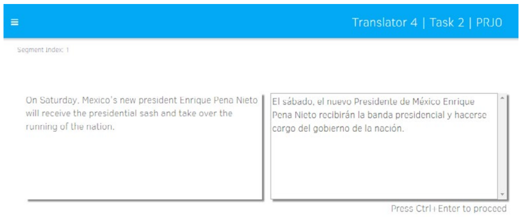
POTENTIAL IMPACT



- Improve translators’ efficiency when working with MT
- Reduce cognitive strain on translators
- Increase translators’ trust in MT output by offering accurate QE
- Reduce their need to search for validation from additional, external resources (cf. Bundgaard 2017, Daems et al. 2016)

EXPERIMENT DESIGN

- Online post-editing tool (HandyCAT)



Translator 4 | Task 2 | PRJ0

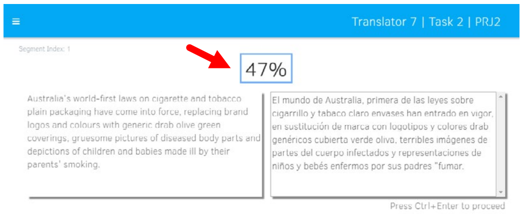
Segment Index: 1

On Saturday, Mexico's new president Enrique Pena Nieto will receive the presidential sash and take over the running of the nation.

El sábado, el nuevo Presidente de México Enrique Pena Nieto recibirán la banda presidencial y hacerse cargo del gobierno de la nación.

Press Ctrl+Enter to proceed

Only source text and MT displayed



Translator 7 | Task 2 | PRJ2

Segment Index: 1

47%

Australia's world-first laws on cigarette and tobacco plain packaging have come into force, replacing brand logos and colours with generic drab olive green coverings, gruesome pictures of diseased body parts and depictions of children and babies made ill by their parents' smoking.

El mundo de Australia, primera de las leyes sobre cigarrillo y tabaco claro envases han entrado en vigor, en sustitución de marca con logotipos y colores drab genéricos cubierta verde oliva, terribles imágenes de partes del cuerpo infectados y representaciones de niños y bebés enfermos por sus padres "fumar.

Press Ctrl+Enter to proceed

Quality estimation (QE) scores displayed

- Participants: 20 professional translators
- Materials: 4 texts (WMT13 news material)
- Languages: English → Spanish
- Four different QE modes (more details below)

EXPERIMENT DESIGN (cont.'d)

QE mode consists of two parts:

- **Score Type**:
  - No QE: the QE box is hidden in HandyCAT
  - Accurate QE: scores obtained from the automatic scoring system that ranked best in the WMT13 shared task (automatic, accurate)
  - Inaccurate QE: 'random' scores (automatic, inaccurate)
  - Human QE: scores obtained using a human evaluation method (human, accurate) (Graham et al. 2015)
- **Score Level**: Percentage (between~20% and 99 %)

EXPERIMENT DESIGN (cont.'d)

Research question:

- What is the impact of the different modes of QE scores on:
  - temporal effort (time spent)
  - physical effort (number of keystrokes)
  - cognitive effort (gaze behaviour)

Data collection:

- activity logging
- screen recording
- eye tracking

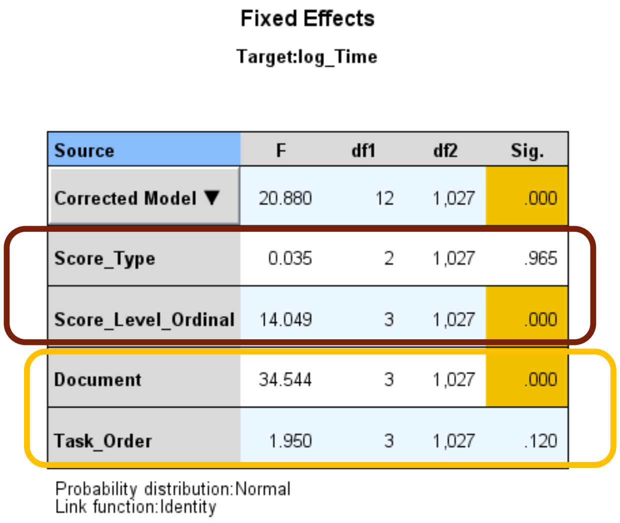
EXPERIMENT DESIGN (cont.'d)

Full range of variables being considered:

Role		Name	Type	Measurement / Levels
Dependent	Temporal –	Translation time	numeric	seconds per word
	Physical –	Amount of typing		keys per word
	Cognitive –	Fixation count		n per word
		Fixation duration		seconds per word
		Pupil dilation		mm (variance)
Independent (Fixed effects)	Primary	QE score type	categorical	No_QE, Acc_QE, Inacc_QE, Human_QE
		QE score level		N/A (No_QE condition) L0: 0.1 to 19.9% L2: 20 to 39.0% L4: 40 to 59.9% L6: 60 to 79.9% L8: 80% to 100%
	Secondary	Document		SRC1, SRC2, SRC5, SRC7
		Task order		T01, T02, T03, T04

RESULTS - Temporal effort

(time spent per word)



Fixed Effects

Target:log_Time

Source	F	df1	df2	Sig.
Corrected Model ▼	20.880	12	1,027	.000
Score_Type	0.035	2	1,027	.965
Score_Level_Ordinal	14.049	3	1,027	.000
Document	34.544	3	1,027	.000
Task_Order	1.950	3	1,027	.120

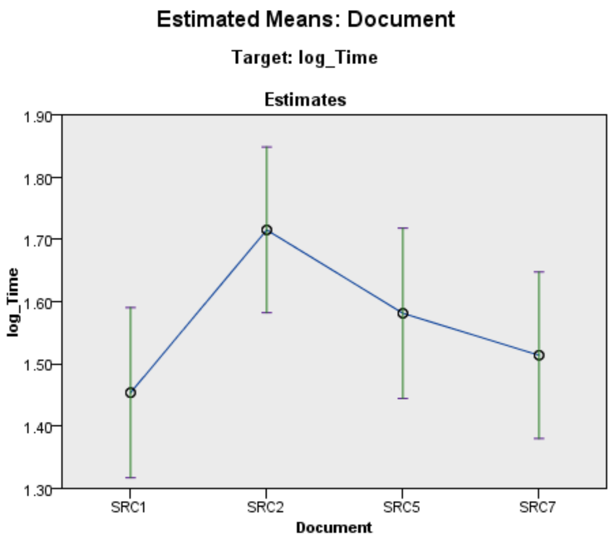
Probability distribution:Normal
Link function:Identity

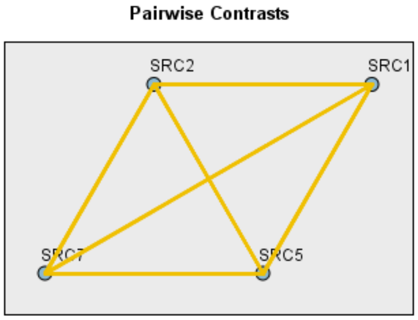
Primary variables

Secondary variables

RESULTS - Temporal effort

Effects found for **Document**

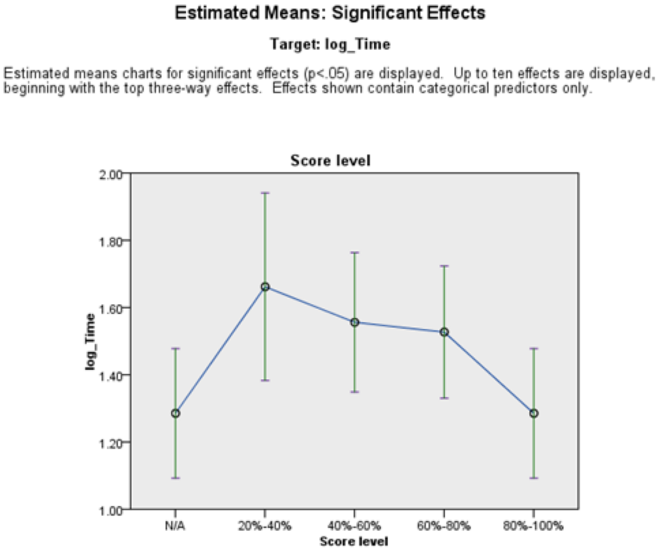




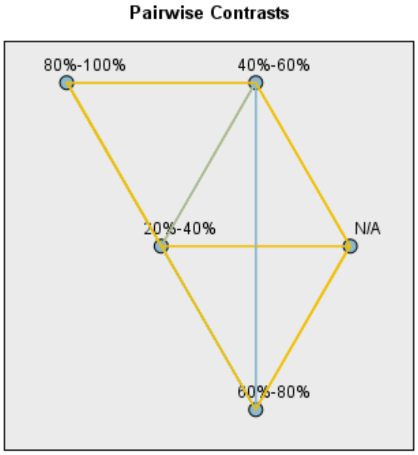
Significant contrasts are shaded gold. The sequential Bonferroni adjusted significance level is .05.

RESULTS - Temporal effort

Effects found for **Score Level**



Estimated means charts for significant effects (p<.05) are displayed. Up to ten effects are displayed, beginning with the top three-way effects. Effects shown contain categorical predictors only.



Significant contrasts are shaded gold. The sequential Bonferroni adjusted significance level is .05.

RESULTS – Physical effort

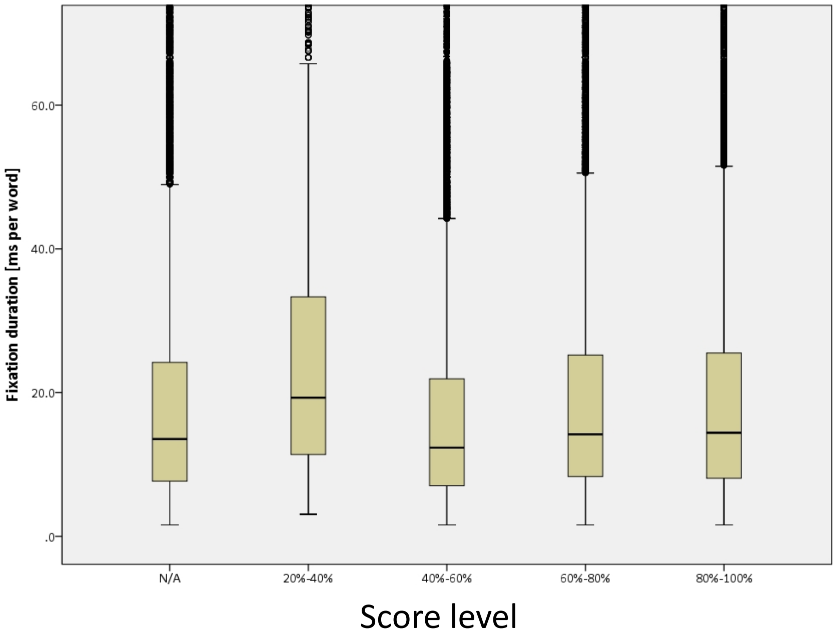
(# of keys typed per word)

Results are similar to the ones found for Temporal effort:

- **No significant** differences in average # of keys according to **Score Type**
- **Significant** differences in average # of keys according to **Score Level**

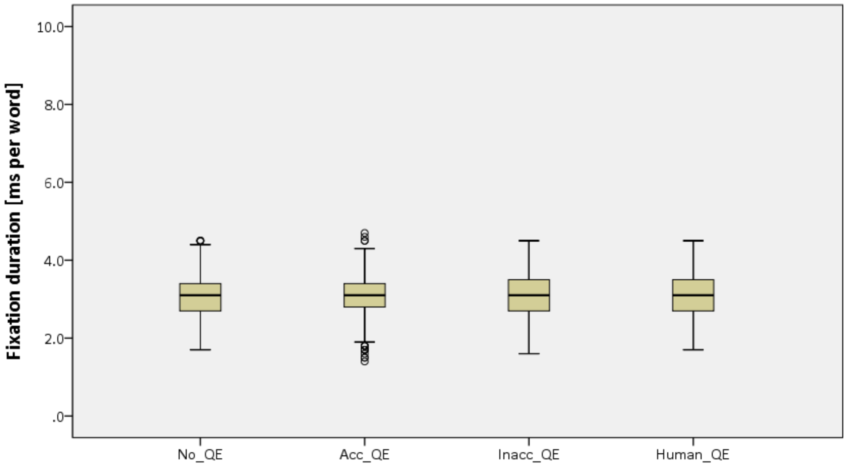
RESULTS – Cognitive effort

Fixation duration per Score Level – No significant effects found



RESULTS – Cognitive effort

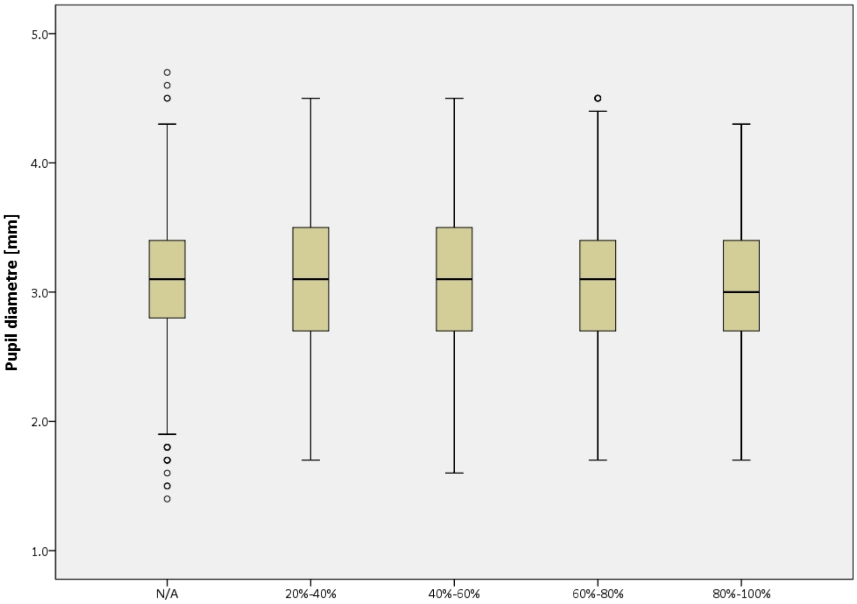
Fixation duration per Score Type – No significant effects found



Score type

RESULTS – Cognitive effort

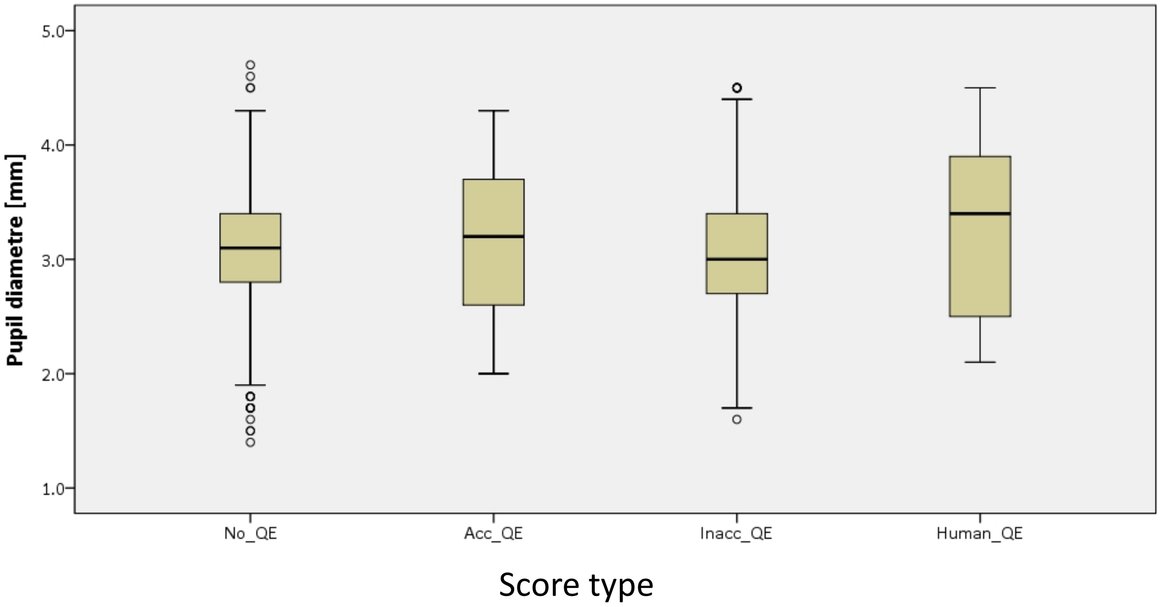
Pupil diameter per Score Level – No significant effects found

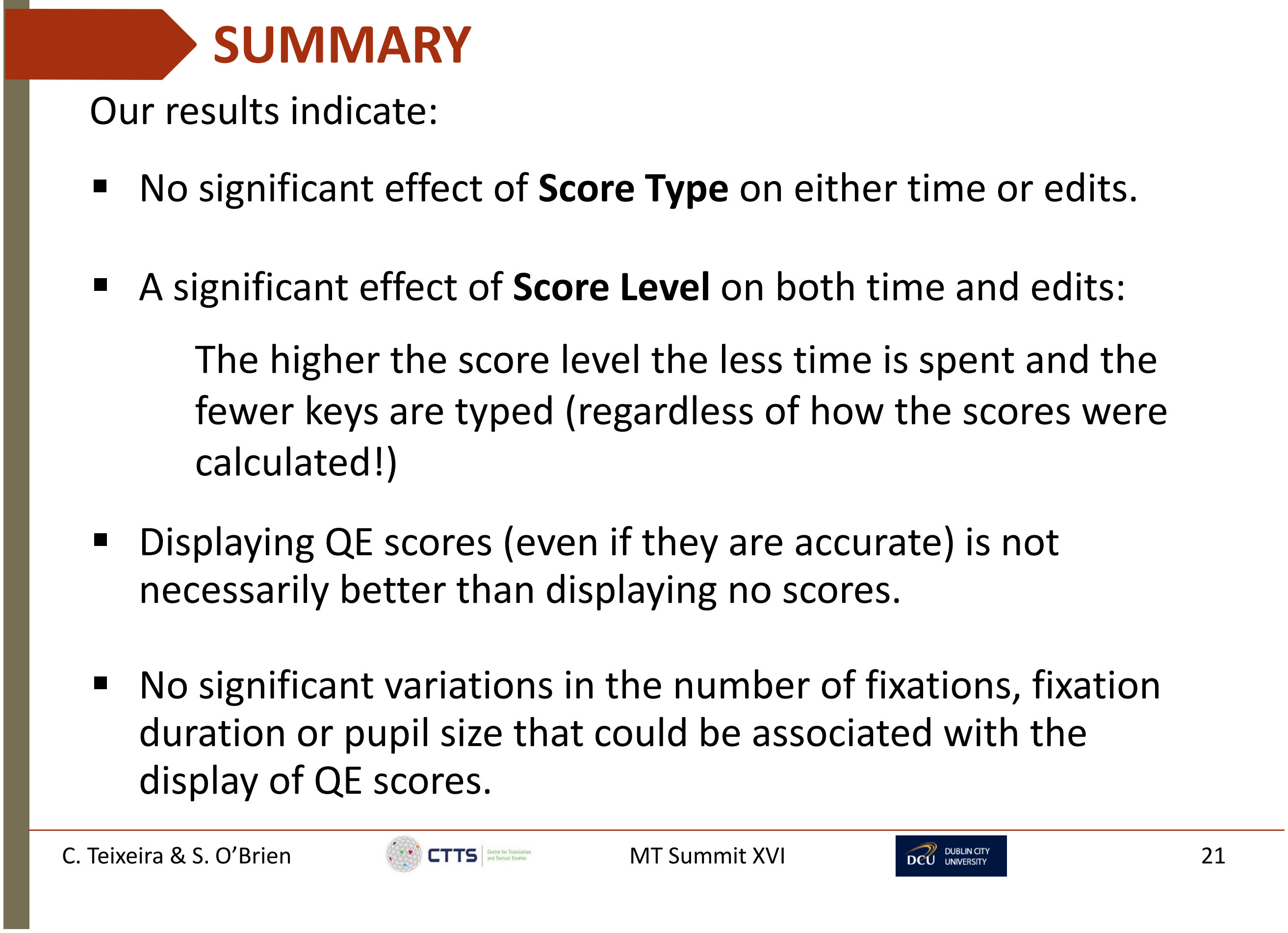


Score level

RESULTS – Cognitive effort

Pupil diameter per Score Type – No significant effects found





SUMMARY

Our results indicate:

- No significant effect of **Score Type** on either time or edits.
- A significant effect of **Score Level** on both time and edits:

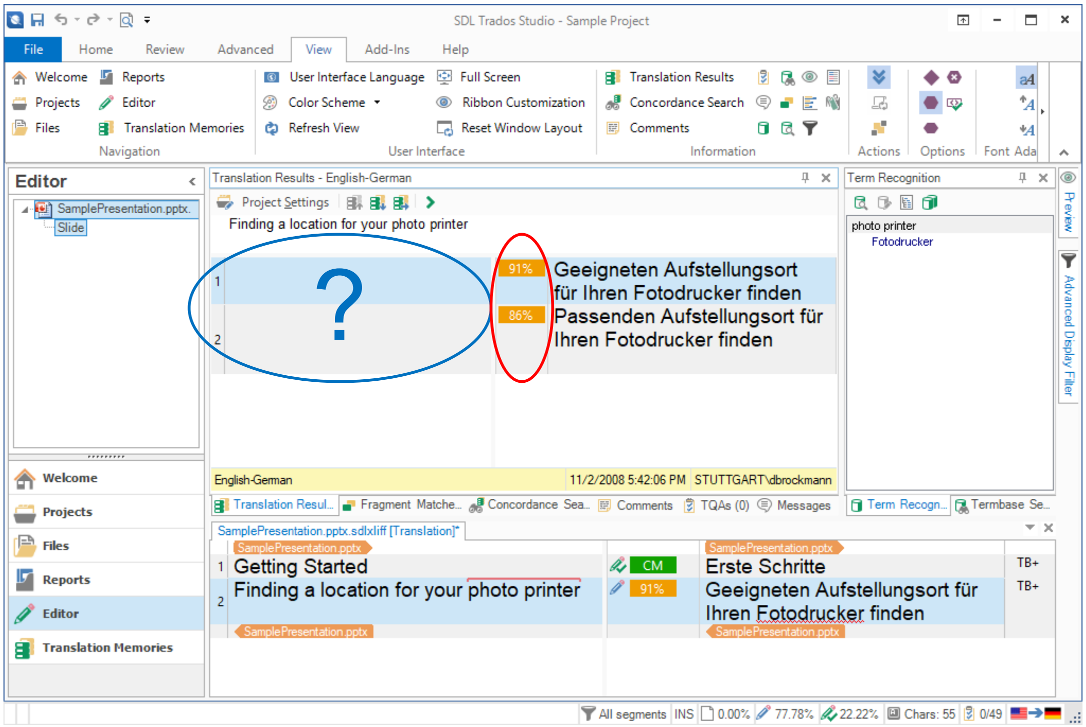
  The higher the score level the less time is spent and the fewer keys are typed (regardless of how the scores were calculated!)

- Displaying QE scores (even if they are accurate) is not necessarily better than displaying no scores.
- No significant variations in the number of fixations, fixation duration or pupil size that could be associated with the display of QE scores.

DISCUSSION

- In our experiment, only a QE percentage was displayed.
- Perhaps the same results would have been found for TM if we had removed the *diff* indication and just left the Match percentages?

DISCUSSION (cont.'d)


SDL Trados Studio - Sample Project

Translation Results - English-German

Finding a location for your photo printer

1	91%	Geeigneten Aufstellungsort für Ihren Fotodrucker finden
2	86%	Passenden Aufstellungsort für Ihren Fotodrucker finden

?

English-German 11/2/2008 5:42:06 PM STUTTGART\dbrockmann

Term Recognition

photo printer
Fotodrucker

SamplePresentation.pptx.sdlxliff [Translation]*

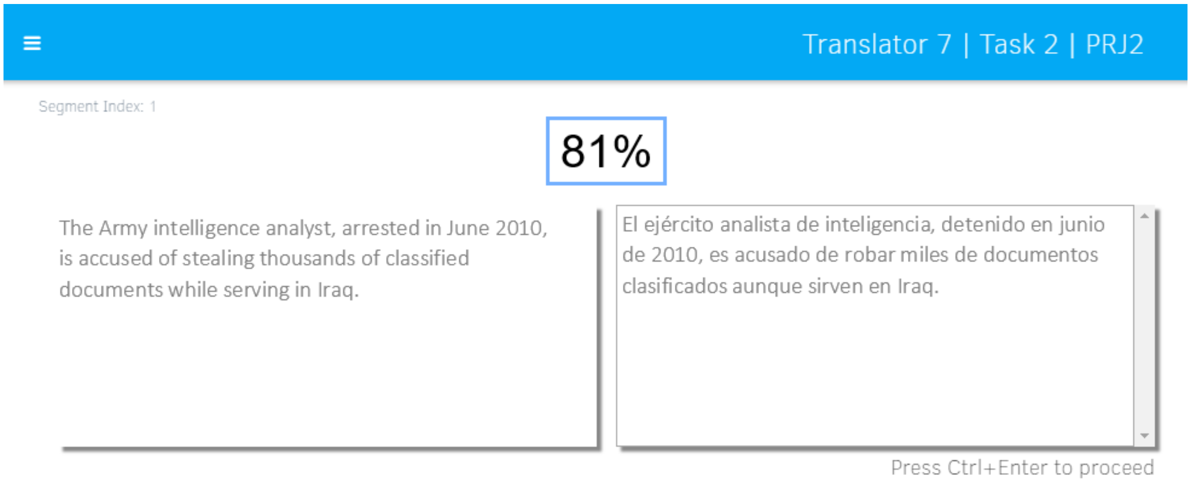
1	Getting Started	CM	Erste Schritte	TB+
2	Finding a location for your photo printer	91%	Geeigneten Aufstellungsort für Ihren Fotodrucker finden	TB+

All segments INS 0.00% 77.78% 22.22% Chars: 55 0/49

DISCUSSION (cont.’d)

- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This is what we displayed:



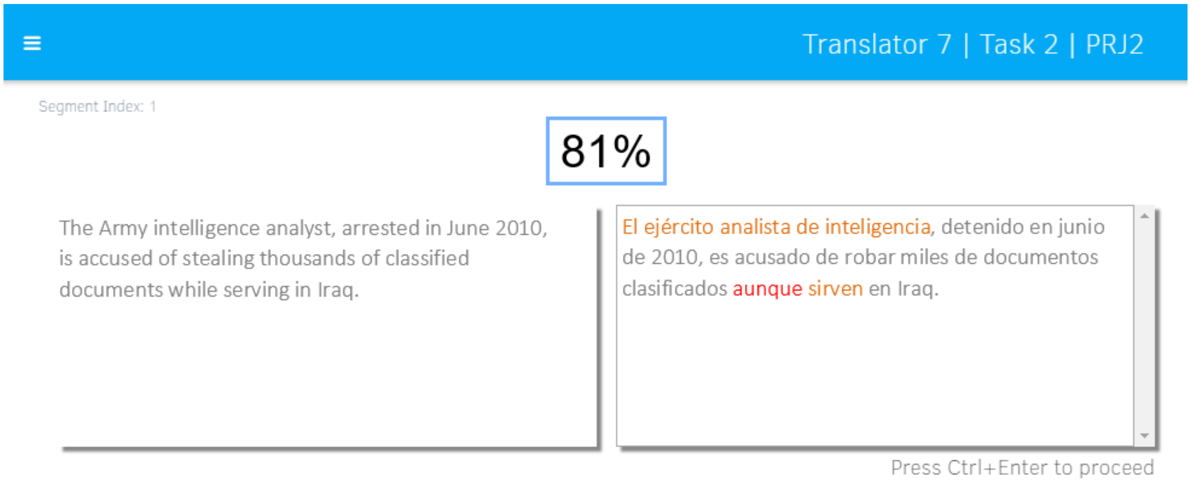
The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq.

El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq.

DISCUSSION (cont.’d)

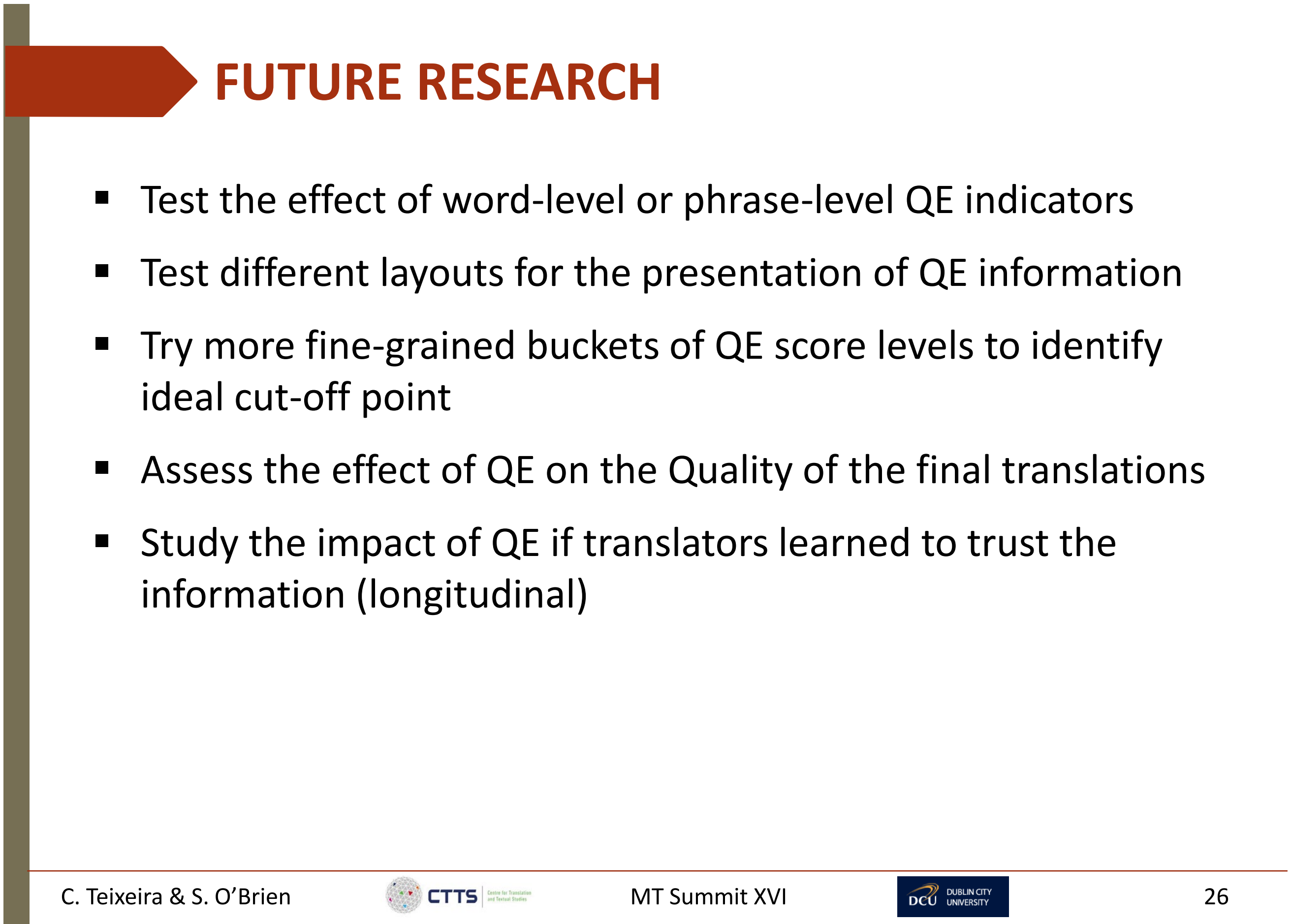
- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This might be the way forward to make QE more effective:



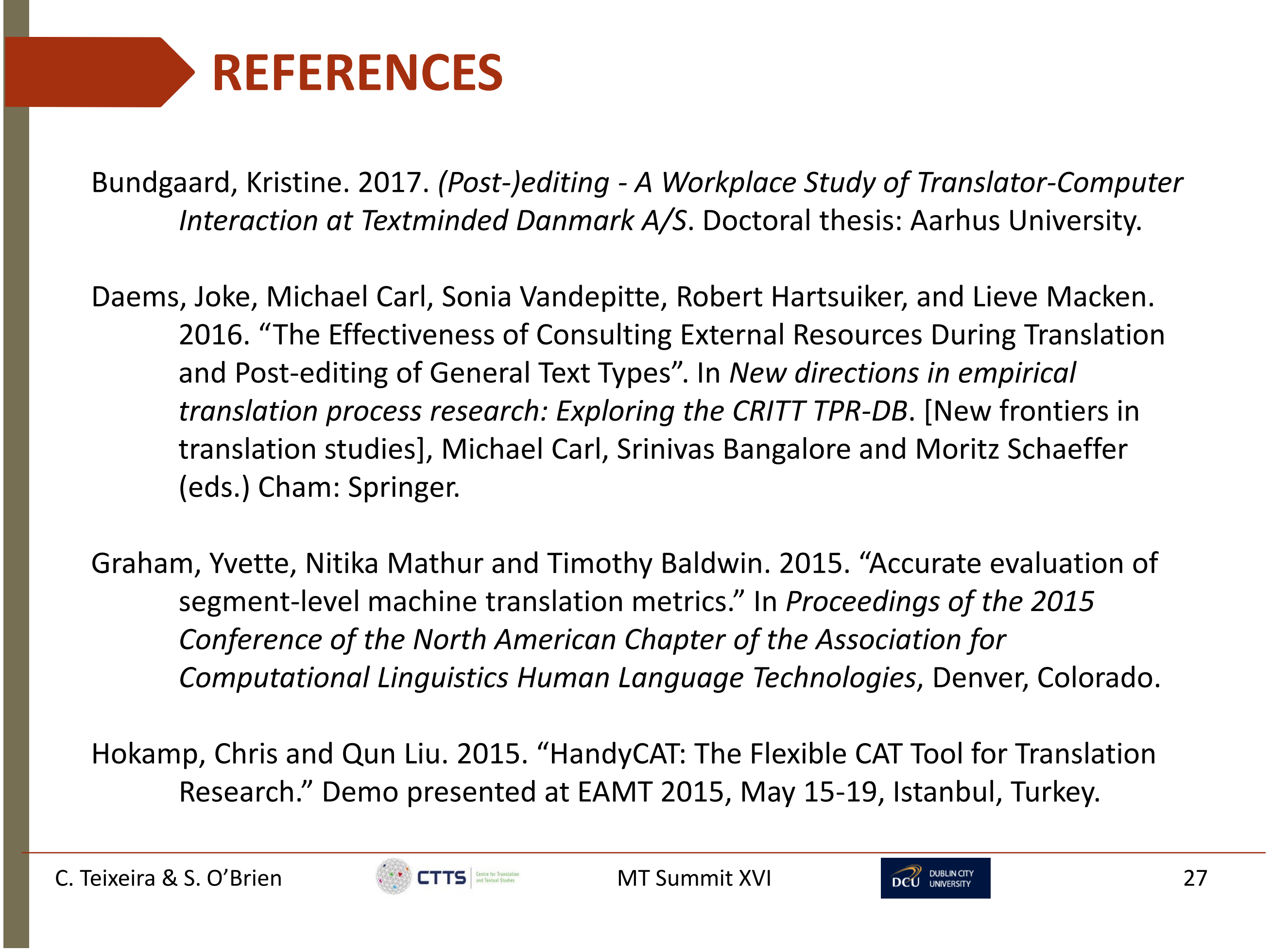
The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq.

El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq.



FUTURE RESEARCH

- Test the effect of word-level or phrase-level QE indicators
- Test different layouts for the presentation of QE information
- Try more fine-grained buckets of QE score levels to identify ideal cut-off point
- Assess the effect of QE on the Quality of the final translations
- Study the impact of QE if translators learned to trust the information (longitudinal)



REFERENCES

Bundgaard, Kristine. 2017. *(Post-)editing - A Workplace Study of Translator-Computer Interaction at Textminded Danmark A/S*. Doctoral thesis: Aarhus University.

Daems, Joke, Michael Carl, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2016. "The Effectiveness of Consulting External Resources During Translation and Post-editing of General Text Types". In *New directions in empirical translation process research: Exploring the CRITT TPR-DB*. [New frontiers in translation studies], Michael Carl, Srinivas Bangalore and Moritz Schaeffer (eds.) Cham: Springer.

Graham, Yvette, Nitika Mathur and Timothy Baldwin. 2015. "Accurate evaluation of segment-level machine translation metrics." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, Denver, Colorado.

Hokamp, Chris and Qun Liu. 2015. "HandyCAT: The Flexible CAT Tool for Translation Research." Demo presented at EAMT 2015, May 15-19, Istanbul, Turkey.

MT Summit XVI

Nagoya, 20 September 2017

Thank you!

ありがとうございました

Carlos Teixeira
carlos.teixeira@dcu.ie

Sharon O'Brien
sharon.obrien@dcu.ie

ADAPT
CTTS Centre for Translation and Textual Studies
DCU DUBLIN CITY UNIVERSITY
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn