MT Summit 2021

James Phillips
Director

PCT Translation Division, WIPO

August 2021

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Outline: Main Topics

- (A)MTQE

- Neural Machine Translation Evaluation

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

PCT Translation Division

Translation of patent abstracts and patentability reports

From: AR, DE, EN, ES, FR, JA, KO, PT, RU, ZH (10 languages)

Into: EN, FR

680,000 translations / 183 million translated words in 2020

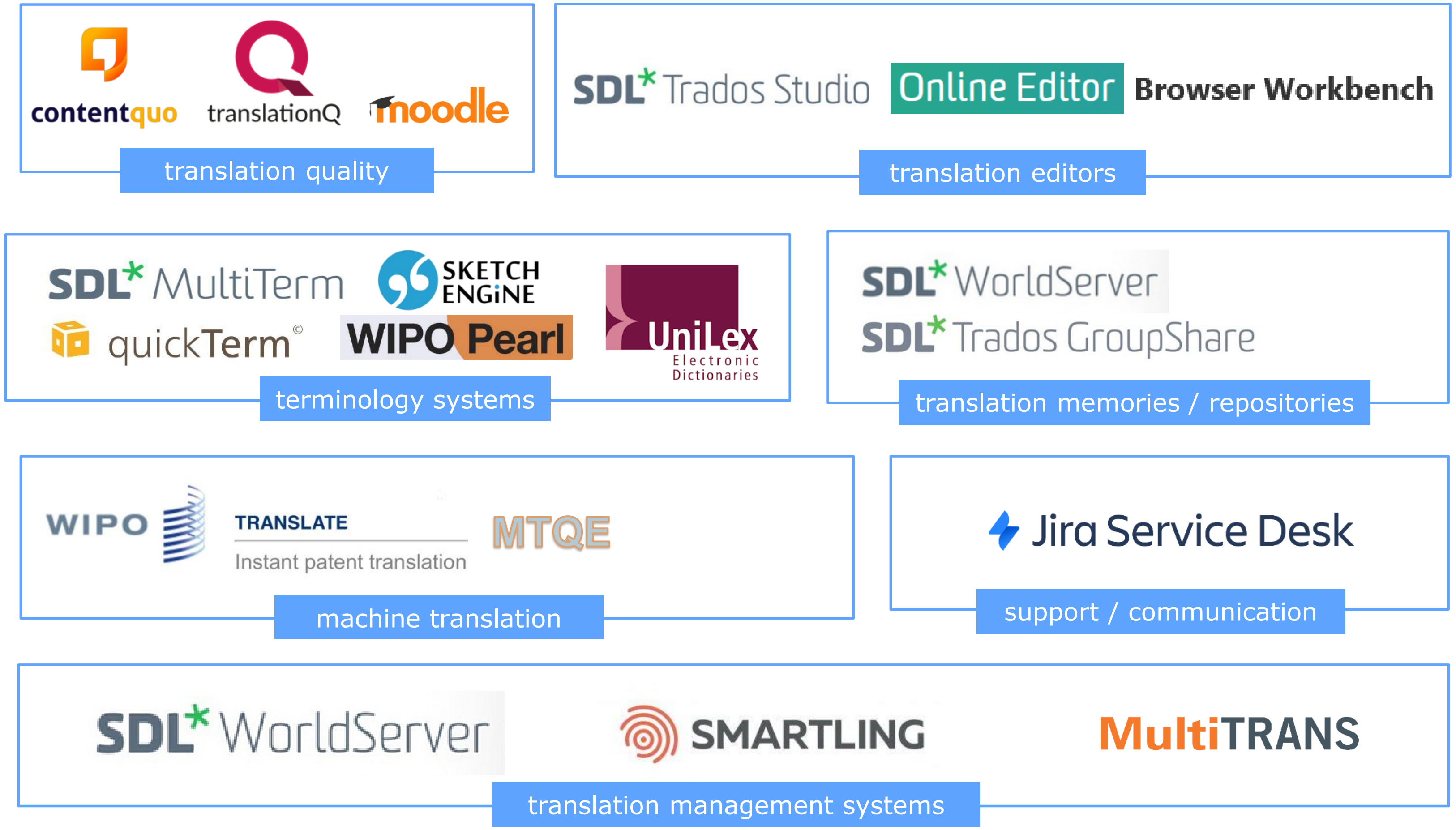
In-house translators + outsourcing (91%)

CAT tools

WIPO FOR OFFICIAL USE ONLY

WIPO

WIPO translation technology stack



contentquo translationQ moodle

translation quality

SDL Trados Studio Online Editor Browser Workbench

translation editors

SDL MultiTerm SKETCH ENGINE quickTerm© WIPO Pearl UniLex Electronic Dictionaries

terminology systems

SDL WorldServer
SDL Trados GroupShare

translation memories / repositories

WIPO TRANSLATE Instant patent translation MTQE

machine translation

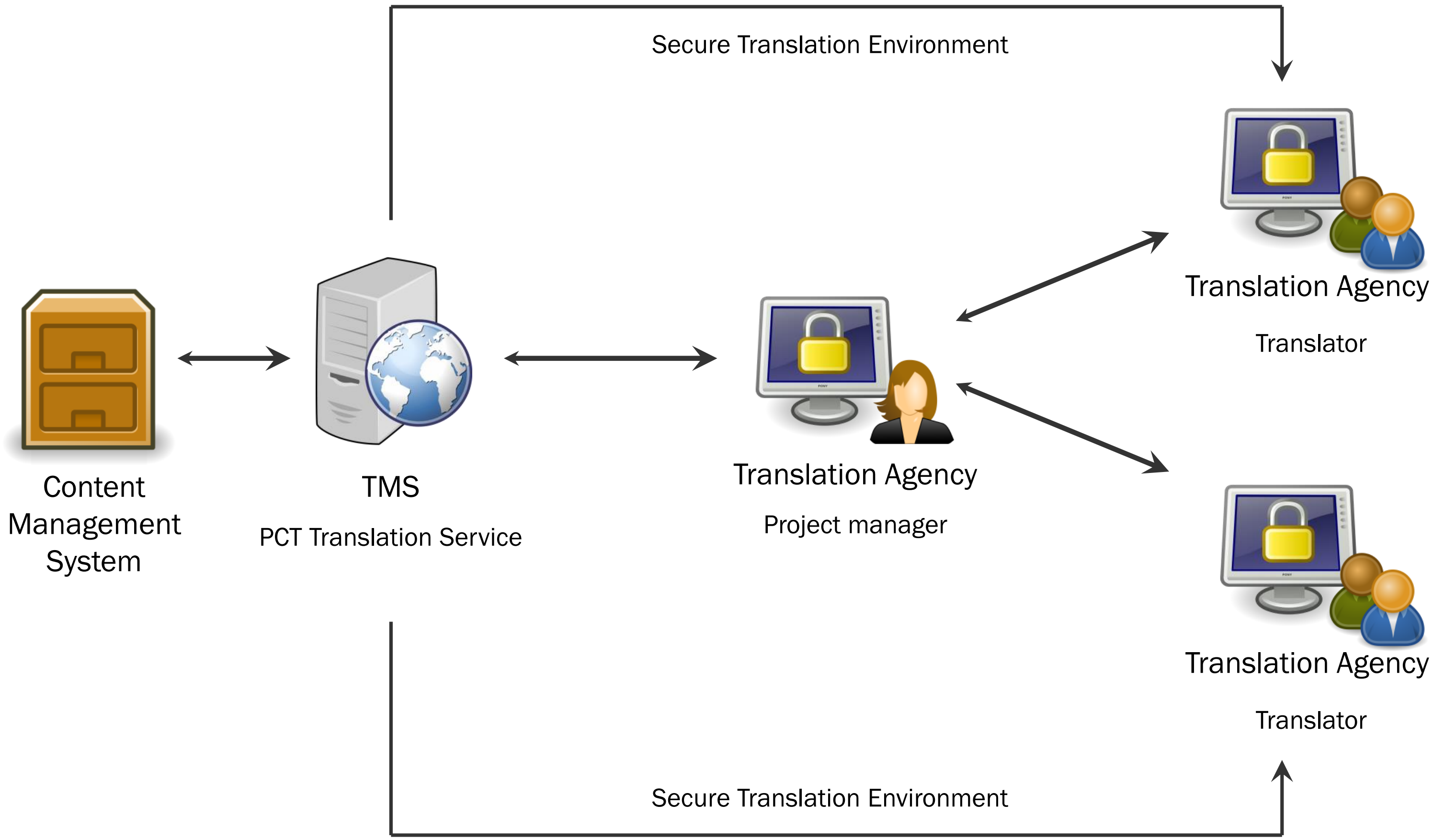
Jira Service Desk

support / communication

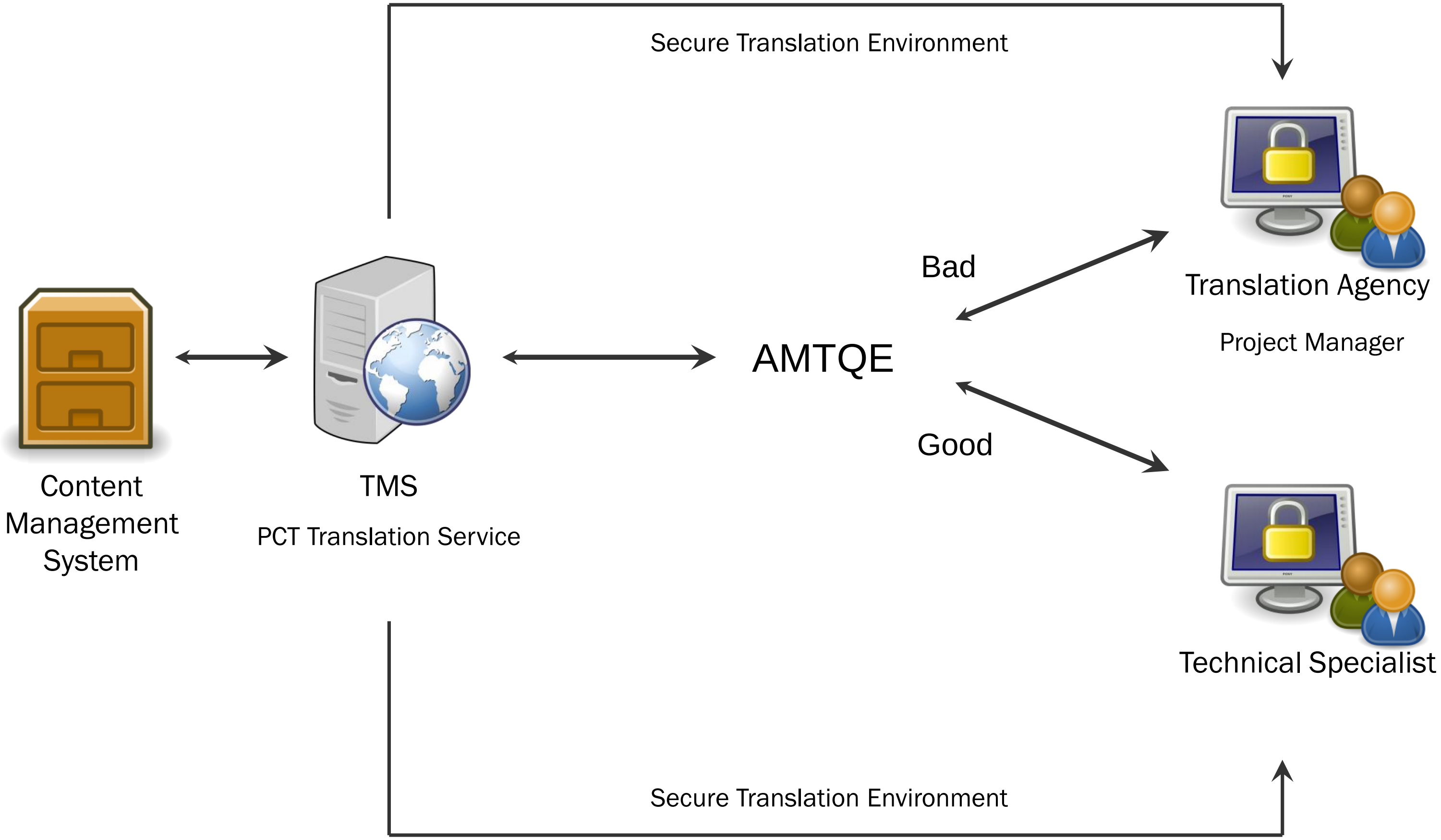
SDL WorldServer SMARTLING MultiTRANS

translation management systems

WorldServer high-level architeture



WorldServer high-level architeture



Secure Translation Environment

Content Management System

TMS

PCT Translation Service

AMTQE

Bad

Good

Translation Agency

Project Manager

Technical Specialist

Secure Translation Environment

Post-editing at WIPO-PCT

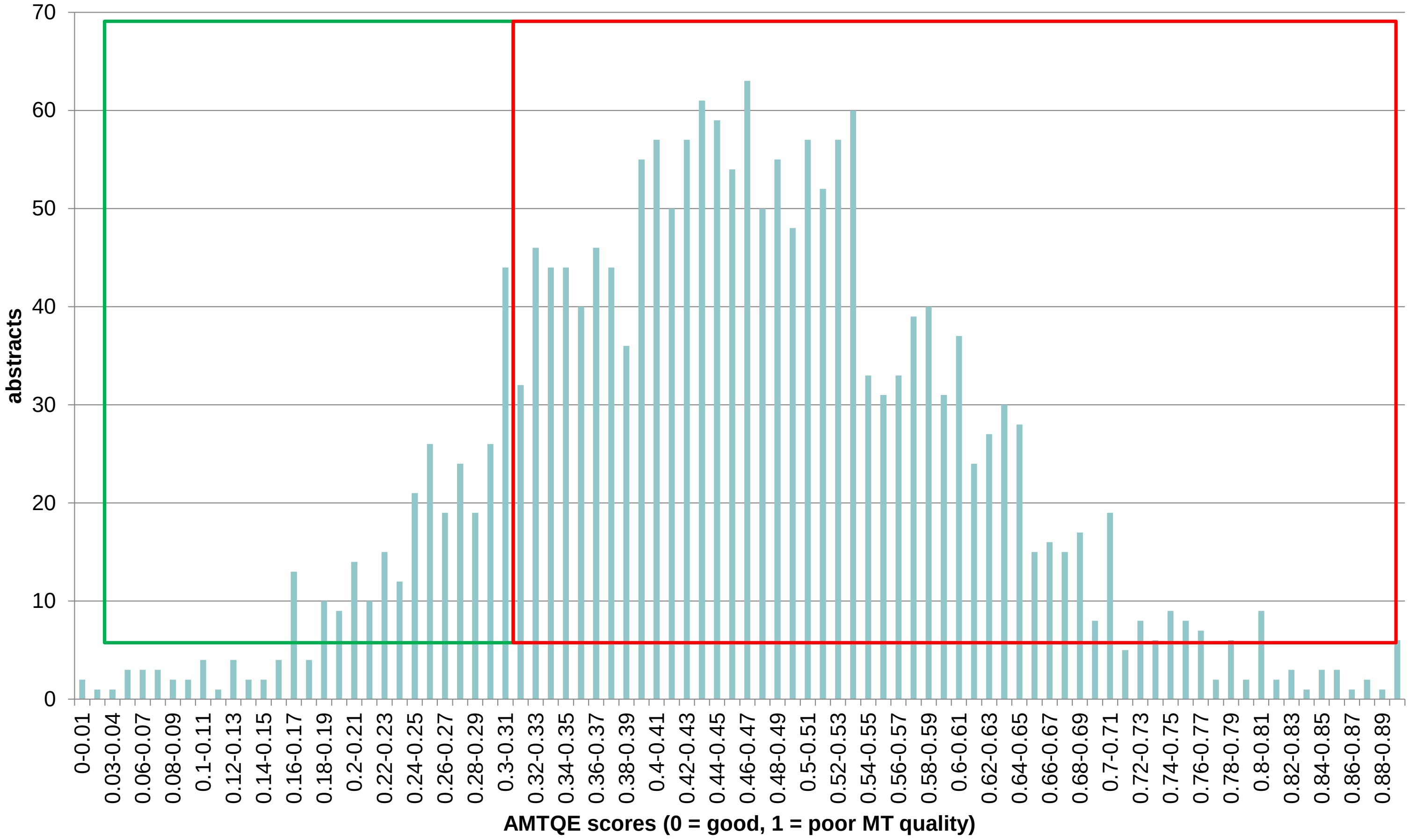
Could visually observe that some of the machine translations were good and decided to try to identify them.

Started collecting triplets (source, NMT output, final agency translation) in 2016. Took six months to build-up sufficient triplets.

Initially difficult to confirm quality threshold at which post-editing becomes feasible. This evaluation process has now been refined.

Decided to attempt AMTQE (Automatic Machine Translation Quality Estimation) using the QuEst framework by Lucia Specia.

AMTQE score distribution & human evaluation



AMTQE

- 3 human evaluation rounds conducted to determine reliability of AMTQE score.
- Evaluators asked to think in number of necessary post-edits.
- Threshold of 0.3 identified
  - AMTQE scores of < 0.3 effectively correlate well with translators' perception of good MT quality for documents of up to 50 words in length.
  - Strong correlation between document length and good AMTQE score.

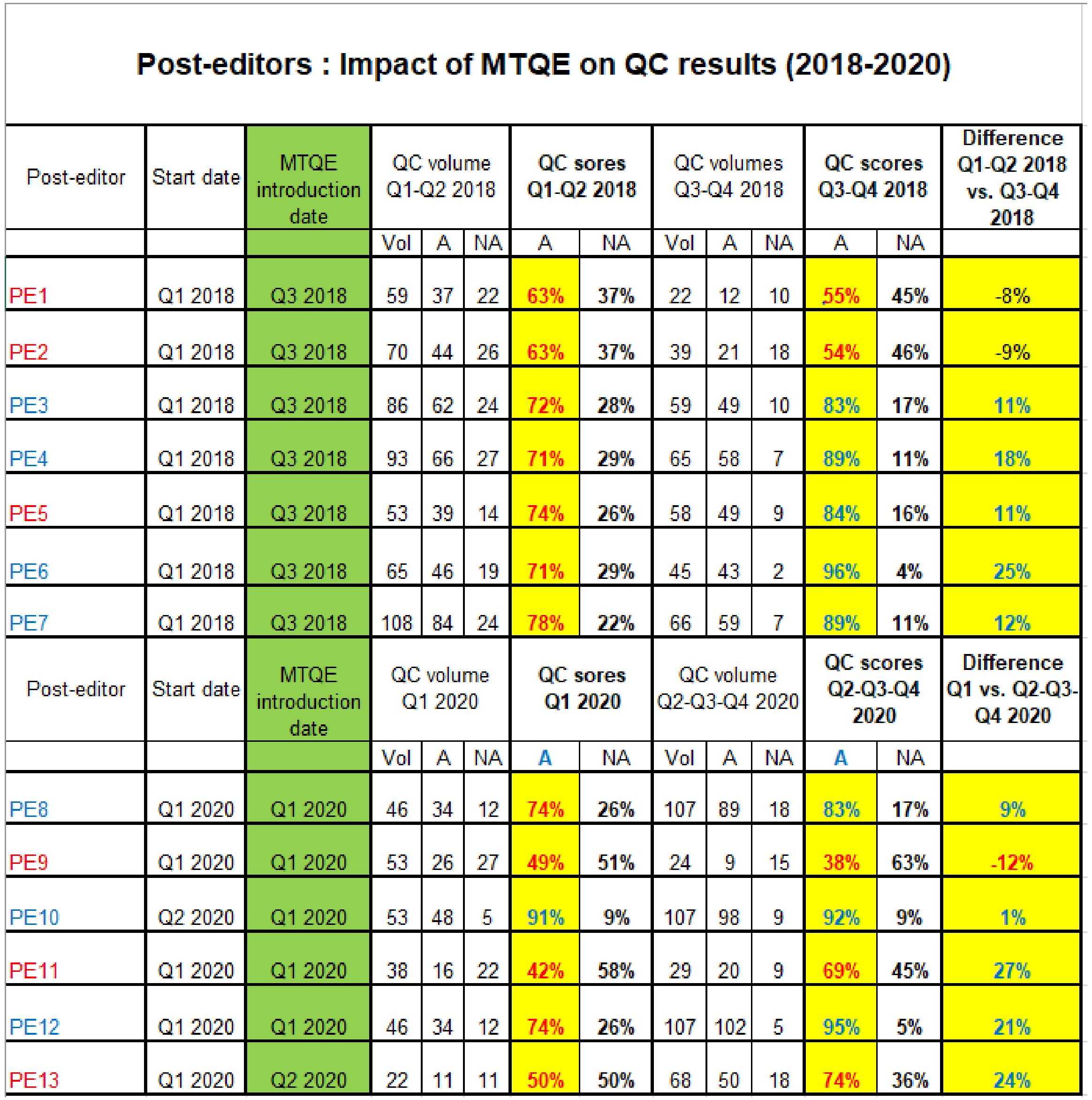
Post-editing at WIPO-PCT

Project 1: Post-editing by Technical Specialists

- Technical specialists (not translators) only given documents in their field.
- Combination of Automatic Machine Translation Quality Estimation score and IPC routing could potentially mean we could adopt post-editing without a dip in quality.

Recruiting challenges

- Recruitment and testing procedures were gradually refined.
- Providing training without imparting bias critical.
- Incorporating translation guidelines into WorldServer glossary extremely helpful.



Post-editors : Impact of MTQE on QC results (2018-2020)

Post-editor	Start date	MTQE introduction date	QC volume Q1-Q2 2018			QC sores Q1-Q2 2018		QC volumes Q3-Q4 2018			QC scores Q3-Q4 2018		Difference Q1-Q2 2018 vs. Q3-Q4 2018
			Vol	A	NA	A	NA	Vol	A	NA	A	NA	
PE1	Q1 2018	Q3 2018	59	37	22	63%	37%	22	12	10	55%	45%	-8%
PE2	Q1 2018	Q3 2018	70	44	26	63%	37%	39	21	18	54%	46%	-9%
PE3	Q1 2018	Q3 2018	86	62	24	72%	28%	59	49	10	83%	17%	11%
PE4	Q1 2018	Q3 2018	93	66	27	71%	29%	65	58	7	89%	11%	18%
PE5	Q1 2018	Q3 2018	53	39	14	74%	26%	58	49	9	84%	16%	11%
PE6	Q1 2018	Q3 2018	65	46	19	71%	29%	45	43	2	96%	4%	25%
PE7	Q1 2018	Q3 2018	108	84	24	78%	22%	66	59	7	89%	11%	12%

Post-editor	Start date	MTQE introduction date	QC volume Q1 2020			QC sores Q1 2020		QC volume Q2-Q3-Q4 2020			QC scores Q2-Q3-Q4 2020		Difference Q1 vs. Q2-Q3-Q4 2020
			Vol	A	NA	A	NA	Vol	A	NA	A	NA	
PE8	Q1 2020	Q1 2020	46	34	12	74%	26%	107	89	18	83%	17%	9%
PE9	Q1 2020	Q1 2020	53	26	27	49%	51%	24	9	15	38%	63%	-12%
PE10	Q2 2020	Q1 2020	53	48	5	91%	9%	107	98	9	92%	9%	1%
PE11	Q1 2020	Q1 2020	38	16	22	42%	58%	29	20	9	69%	45%	27%
PE12	Q1 2020	Q1 2020	46	34	12	74%	26%	107	102	5	95%	5%	21%
PE13	Q1 2020	Q2 2020	22	11	11	50%	50%	68	50	18	74%	36%	24%

Post-editing at WIPO-PCT

Project 2 : Light post-editing

Instigated from the bottom-up as a result of observations by the translators

Use internal resources

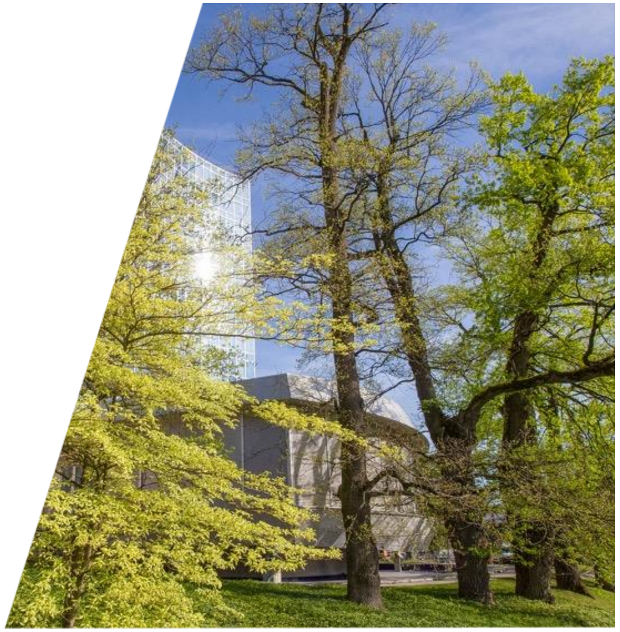
5 to 6 days of work/week

Preselection of abstracts

Only abstracts with good MT are (lightly) post-edited

500 abstracts translations per week under project 1

NMT Evaluation



Evaluate multiple engines and translator profiles

Minimum Team: Senior Translator, Junior translator, external translator, multiple engines, minimum two revisers (must be different people)

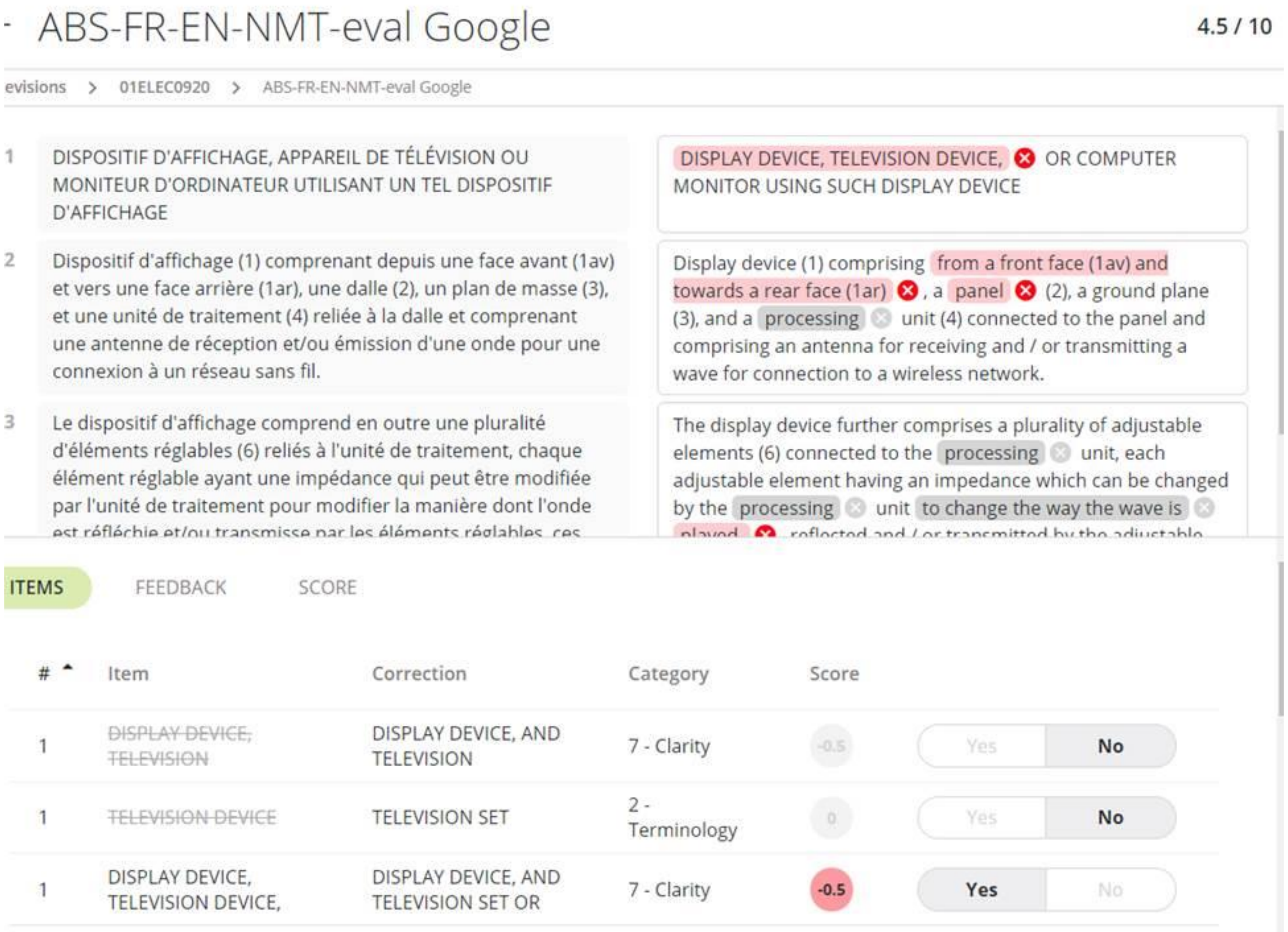
Penalty scoring system: 4 point deduction for major error, 0.5 points for minor error

Recently published documents only (two weeks)

Ten documents minimum (same field)

Error Categories

Error Categories (Major/Minor Errors Applied)			Document or sentence level?	
1	Meaning	Over-translation: more specific. Under-translation: less specific. Verity: contradictions that are not pivotal language. Mistranslations	Sent.	
2	Terminology		Sent.	Doc.
3	English usage	Poor/incorrect English usage	Sent.	
4	Omission/Addition	Addition Omission	Sent.	
5	Consistency		Sent.	Doc.
6	Proof-reading/Spelling	Numbers, citations, reference signs, spelling errors, currency, dates, names, etc.	Sent.	Doc.
7	Clarity	Penalty if difficult to understand, misleading, or ambiguous.	Sent.	
8	Fluency	Penalty if not fluent. How smoothly does it read? To be restricted to being a minor error only when the sentence does not read smoothly at all. It could, for example, be grammatically correct, accurate, and clear, but quite painful to read, which would incur a fluency penalty.	Sent.	
9	Pivotal Language (Reports Only)	Contradictions that are pivotal to the document. i.e. calling something novel when the document says not novel. To be classed as a critical error.		



ABS-FR-EN-NMT-eval Google

4.5 / 10

evisions > 01ELEC0920 > ABS-FR-EN-NMT-eval Google

1	DISPOSITIF D'AFFICHAGE, APPAREIL DE TÉLÉVISION OU MONITEUR D'ORDINATEUR UTILISANT UN TEL DISPOSITIF D'AFFICHAGE	DISPLAY DEVICE, TELEVISION DEVICE, OR COMPUTER MONITOR USING SUCH DISPLAY DEVICE
2	Dispositif d'affichage (1) comprenant depuis une face avant (1av) et vers une face arrière (1ar), une dalle (2), un plan de masse (3), et une unité de traitement (4) reliée à la dalle et comprenant une antenne de réception et/ou émission d'une onde pour une connexion à un réseau sans fil.	Display device (1) comprising from a front face (1av) and towards a rear face (1ar), a panel (2), a ground plane (3), and a processing unit (4) connected to the panel and comprising an antenna for receiving and / or transmitting a wave for connection to a wireless network.
3	Le dispositif d'affichage comprend en outre une pluralité d'éléments réglables (6) reliés à l'unité de traitement, chaque élément réglable ayant une impédance qui peut être modifiée par l'unité de traitement pour modifier la manière dont l'onde est réfléchie et/ou transmisse par les éléments réglables, ces	The display device further comprises a plurality of adjustable elements (6) connected to the processing unit, each adjustable element having an impedance which can be changed by the processing unit to change the way the wave is played, reflected and / or transmitted by the adjustable

ITEMS FEEDBACK SCORE

#	Item	Correction	Category	Score	
1	~~DISPLAY DEVICE, TELEVISION~~	DISPLAY DEVICE, AND TELEVISION	7 - Clarity	-0.5	Yes [✓ No]
1	~~TELEVISION DEVICE~~	TELEVISION SET	2 - Terminology	0	Yes [✓ No]
1	DISPLAY DEVICE, TELEVISION DEVICE,	DISPLAY DEVICE, AND TELEVISION SET OR	7 - Clarity	-0.5	[✓ Yes] No

	Average	Abs. 1	Abs. 2	Abs. 3	Abs. 4	Abs. 5	Abs. 6	Abs. 7	Abs. 8	Abs. 9	Abs.10
Difficulty		E	D	E	E	M	D	M	M	M	D
Senior Translator	9.75	10	10	10	10	9.5	8.5	10	10	9.5	10
Junior Translator	9.45	10	8	9	10	10	8	10	10	9.5	10
Agency Translator	8.55	8.5	8.5	8.5	9.5	8.5	7	8	9	9.5	8.5
Engine 1	-2.85	4.5	-13	-1.5	9	1	-20	4.5	-0.5	0	-12.5
Engine 2	-3.9	8	-7.5	-2	7.5	-1.5	-23	8	-10	3	-21.5
Engine 3	-5.55	-3.5	-8	2	-2	5	-16.5	0.5	-3.5	0	-29.5
Engine 4	-6.5	0	-8.5	-2.5	-6	-3.5	-20.5	3.5	-17	-1.5	-9
Engine 5	-15.85	3	-11.5	-11	-8.5	-5.5	-39.5	-14.5	-18	-17	-36

Lessons Learned

- It is time-consuming to configure an AMTQE algorithm.
- We would prefer off-the-shelf.
- Need reliable human evaluation that can preferably be carried out quickly and give clear indication of whether post-editing will be cost-effective.

WIPO

Thank you!

james.phillips@wipo.int

laurent.gottardo@wipo.int

Photo credits: