AMTA PRESENTATION

MT Quality Evaluations: From Test Environment to Production

ELAINE OCURRAN
Welocalize

October 2015

welocalize
doing things differently

AGENDA

- Our MT evaluation methodologies
- Correlations between automatic scores and human evaluations
- Differences between system autoscores and PE autoscores
- MT evaluations in a production setting
- MT evaluations of post-edited files: a case study

welocalize
doing things differently

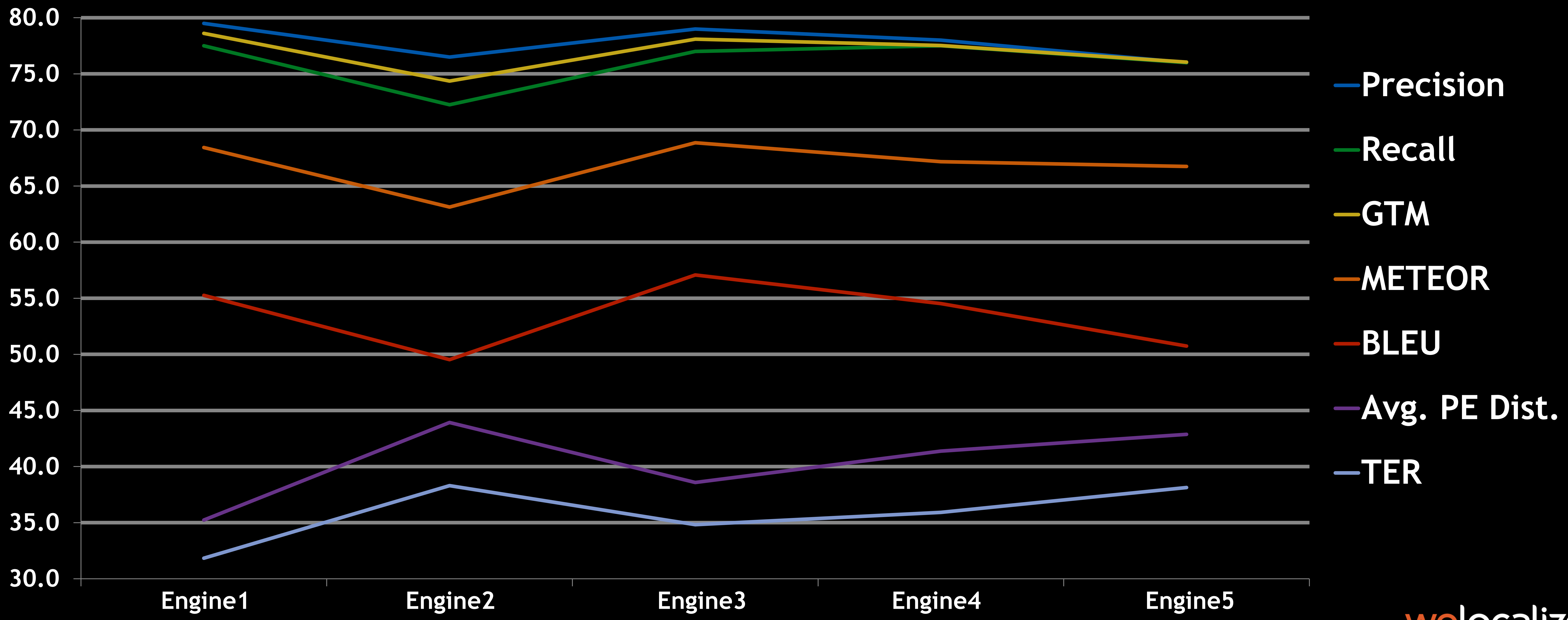
OUR EVALUATION METHODS

A TYPICAL EVALUATION PROCESS PER LOCALE AND PER ENGINE

1. Engine Training
2. Autoscoring Test Sets (2500 TUs)
3. Engine Ranking (100 TUs)
4. Adequacy & Fluency Scoring (100 TUs)
5. Error Typology Scoring (25 TUs)
6. Analysis & Report

OUR EVALUATION METHODS

AUTOMATIC SCORES GENERATED BY WESCORE



welocalize
doing things differently

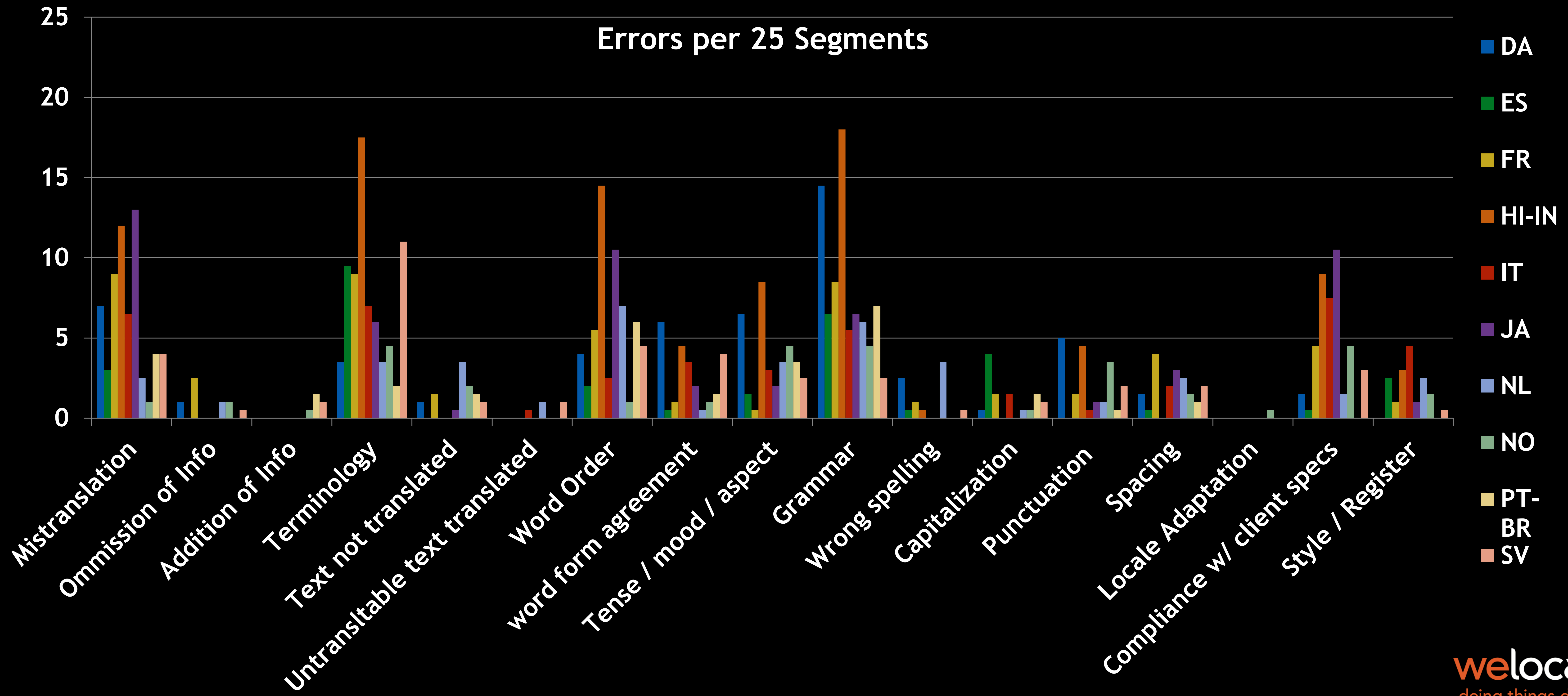
OUR EVALUATION METHODS

HUMAN EVALUATIONS:
ADEQUACY AND FLUENCY SCORING

SCORE	ACCURACY	FLUENCY
5	All meaning expressed in the source fragment appears in the translation fragment.	Native language fluency. No grammar errors, good word choice and syntactic structure. No PE required.
4	Most of the source fragment meaning is expressed in the translation fragment.	Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little PE required.
3	Much of the source fragment meaning is expressed in the translation fragment.	Not very fluent. About half of translation contains errors and requires PE.
2	Little of the source fragment meaning is expressed in the translation fragment.	Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of PE required.
1	None of the meaning expressed in the source fragment is expressed in the translation fragment.	No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.

OUR EVALUATION METHODS

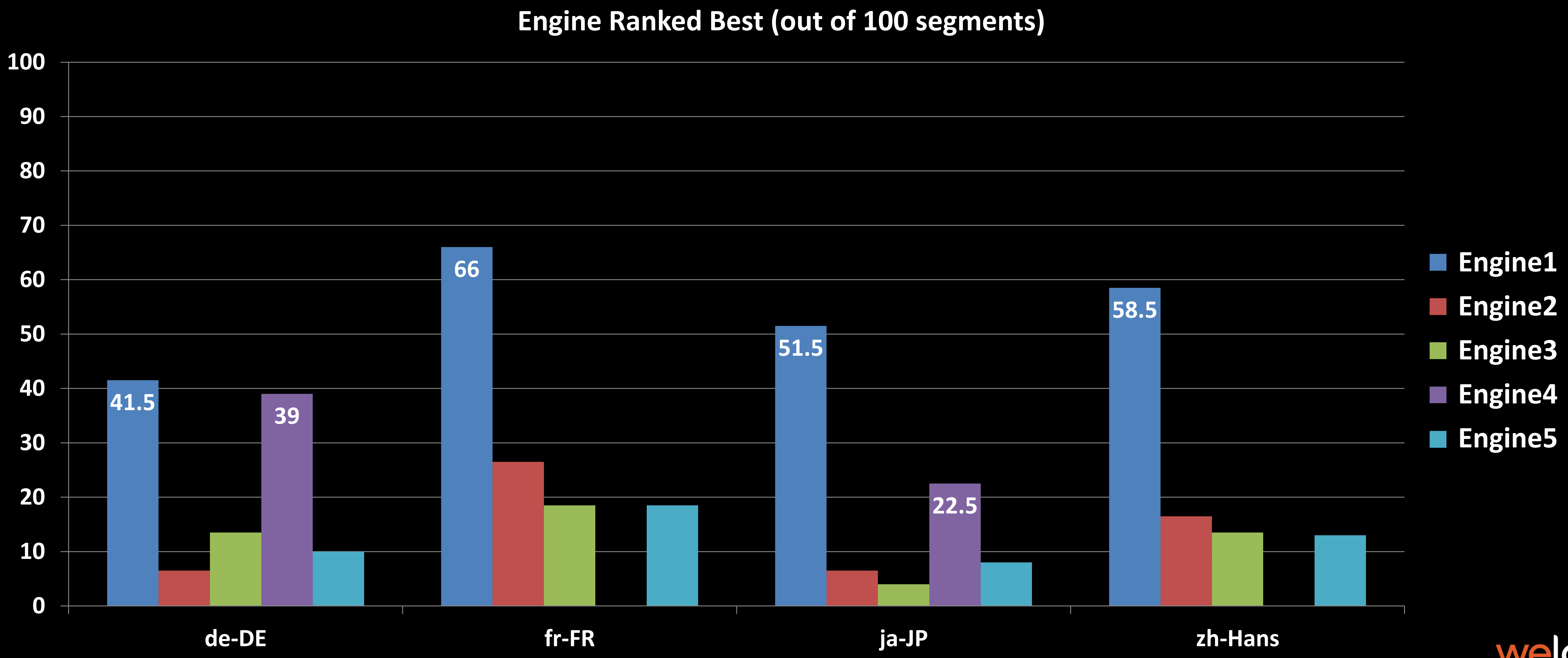
HUMAN EVALUATION: ERROR TYPOLOGY



welocalize
doing things differently

OUR EVALUATION METHODS

HUMAN EVALUATION: ENGINE RANKING



welocalize
doing things differently

LESSONS LEARNED

- We always perform autoscoring PLUS human scoring for all our MT evaluations. We have internal thresholds that qualify an engine ready for deployment and it's level of maturity.
- For bake-offs between several engines, we always include engine ranking in addition to our standard scores.
- Productivity tests are valuable during the initial phase of an MT program to build up productivity data for future reference across languages, domains and MT systems.
- Our MT program is now mature and we are able to perform most of our evaluations based on autoscoring PLUS human scoring, and by referencing the productivity data we have collected over a number of years.

NEXT

Correlations between automatic scores and human evaluations

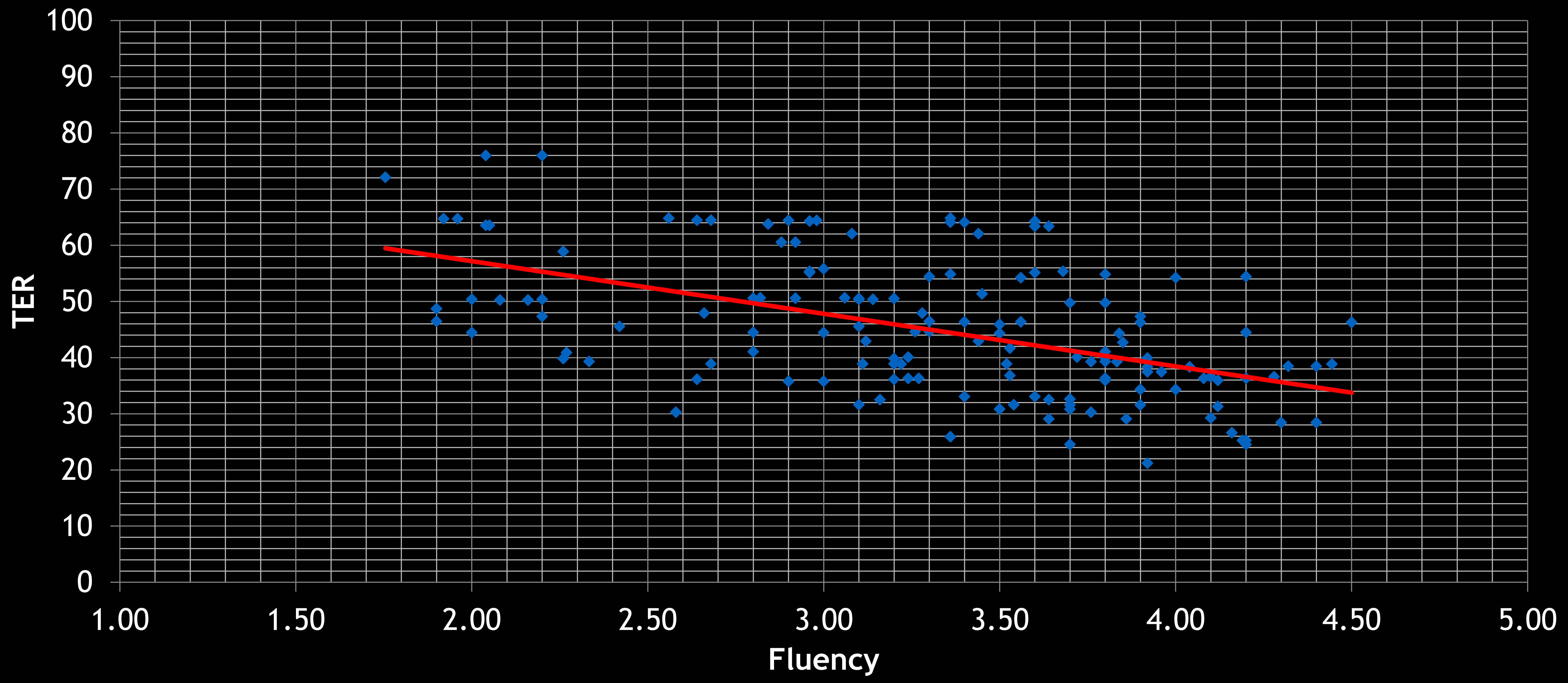
CORRELATIONS

CORRELATIONS BETWEEN AUTOMATIC SCORES AND HUMAN EVALUATIONS

Pearson's r	Variables	Strength of Correlation	Tests (N)	Locales
0.50576955	Fluency & METEOR	Strong positive relationship	150	11
0.50070425	Fluency & BLEU	Strong positive relationship	150	11
0.49816365	Fluency & Recall	Strong positive relationship	150	11
0.49724893	Fluency & NIST	Strong positive relationship	150	11
0.49195687	Fluency & GTM	Strong positive relationship	150	11
0.47064566	Fluency & Precision	Strong negative relationship	150	11
0.38293518	Adequacy & NIST	Moderate negative relationship	150	11
0.31354314	Adequacy & METEOR	Moderate negative relationship	150	11
0.2940756	Adequacy & Recall	Weak positive relationship	150	11
0.28586852	Adequacy & GTM	Weak positive relationship	150	11
0.28386332	Adequacy & BLEU	Weak positive relationship	150	11
0.26685854	Adequacy & Precision	Weak positive relationship	150	11
-0.40270902	Adequacy & TER	Strong negative relationship	150	11
-0.4788575	Fluency & PE Distance	Strong negative relationship	150	11
-0.5385275	Adequacy & PE Distance	Strong negative relationship	150	11
-0.5421933	Fluency & TER	Strong negative relationship	150	11

CORRELATIONS

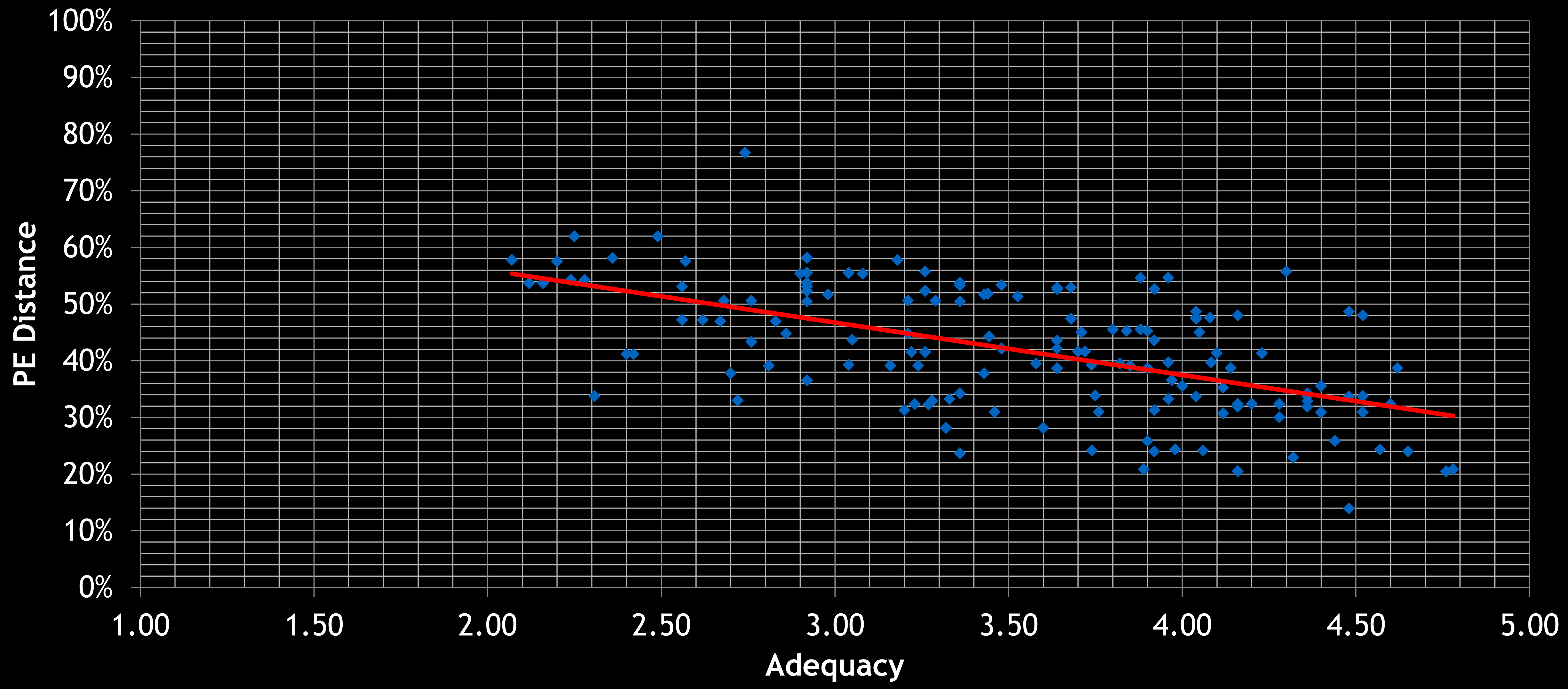
THE STRONGEST CORRELATION WAS FOUND BETWEEN FLUENCY AND TER



welocalize
doing things differently

CORRELATIONS

THE 2ND STRONGEST CORRELATION WAS FOUND BETWEEN ADEQUACY AND PE DISTANCE



welocalize
doing things differently

LESSONS LEARNED

- It seems that we cannot rely solely on autoscores as long as the correlation with human judgment is not stronger than the data suggests
- TER and PE Distance show the strongest correlation to both Fluency and Adequacy, and therefor seem closer to human judgment than the other scores.
- Fluency correlates stronger with system autoscores than Adequacy overall.
- PE Distance is the only metric that correlates stronger with Adequacy than Fluency. PE Distance is also the only character-based metric.

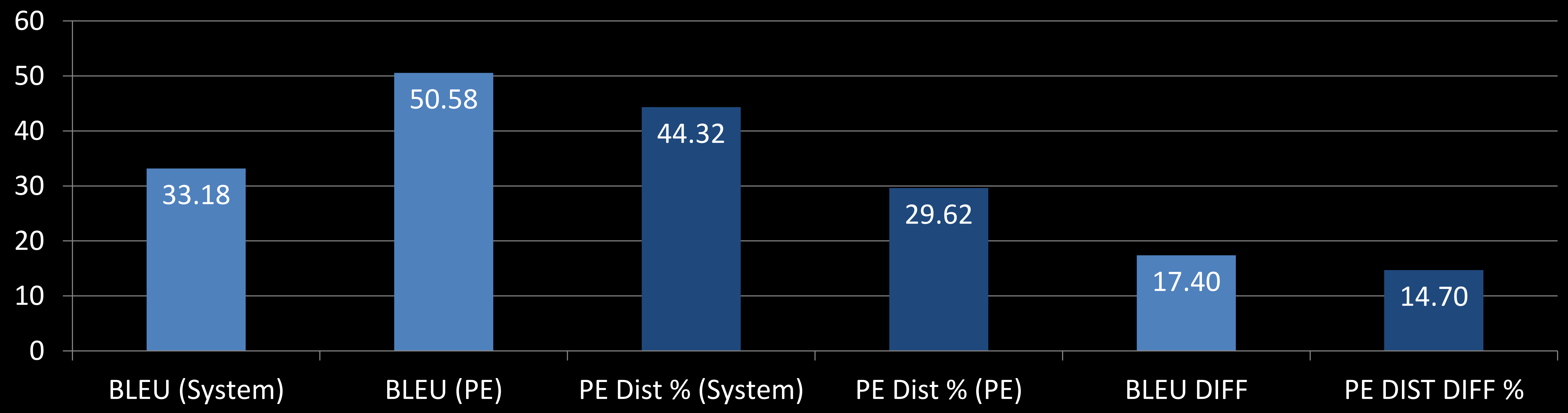
NEXT

Differences between system autoscores and post-editing autoscores

SYSTEM VS PE AUTOSCORES

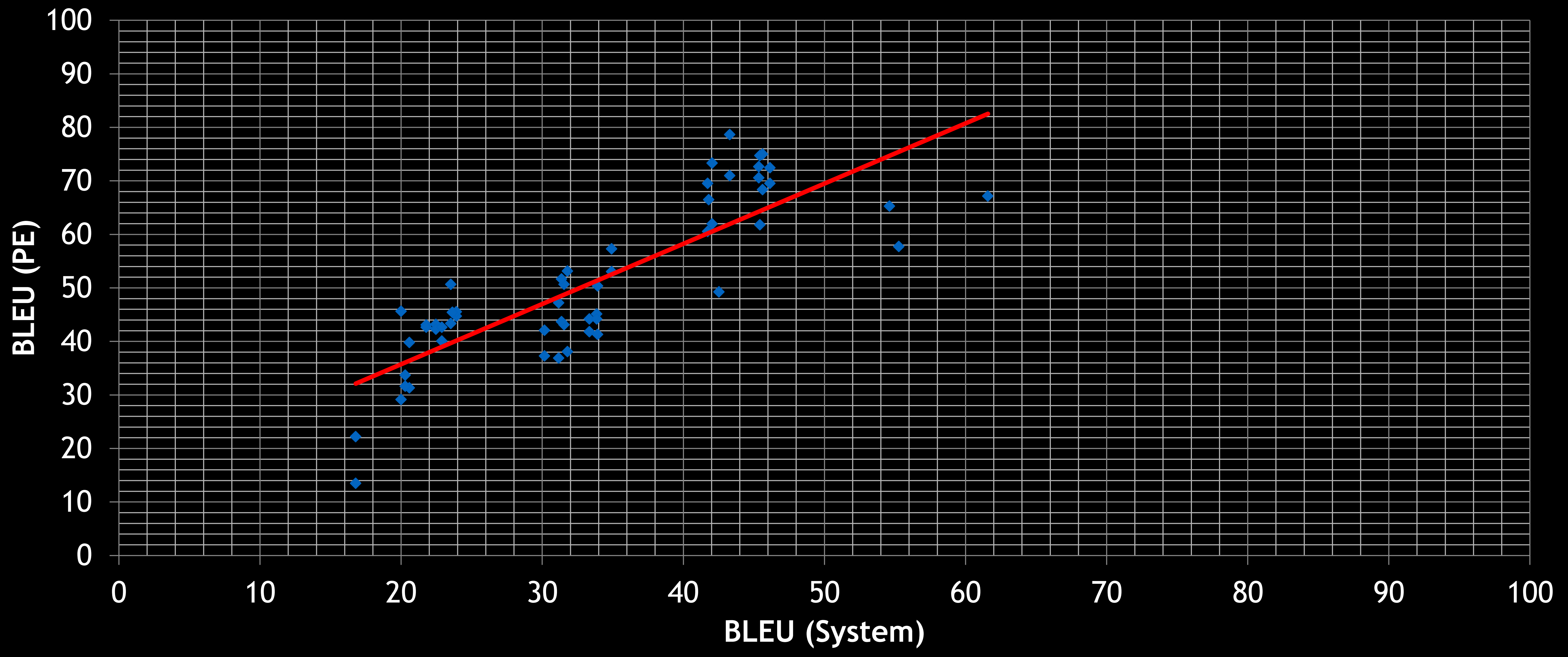
ON AVERAGE, THE POST-EDITING SCORE IS 15 AND 17 POINT HIGHER FOR PE DISTANCE AND BLEU RESPECTIVELY

Pearson's r	Variables	Strength of Correlation	Tests (N)	Locales
0.832226688	BLEU (System) & BLEU (PE)	Very strong positive relationship	57	9
0.832218909	PE Distance (System) & PE Distance (PE)	Very strong positive relationship	57	9



SYSTEM VS PE AUTOSCORES

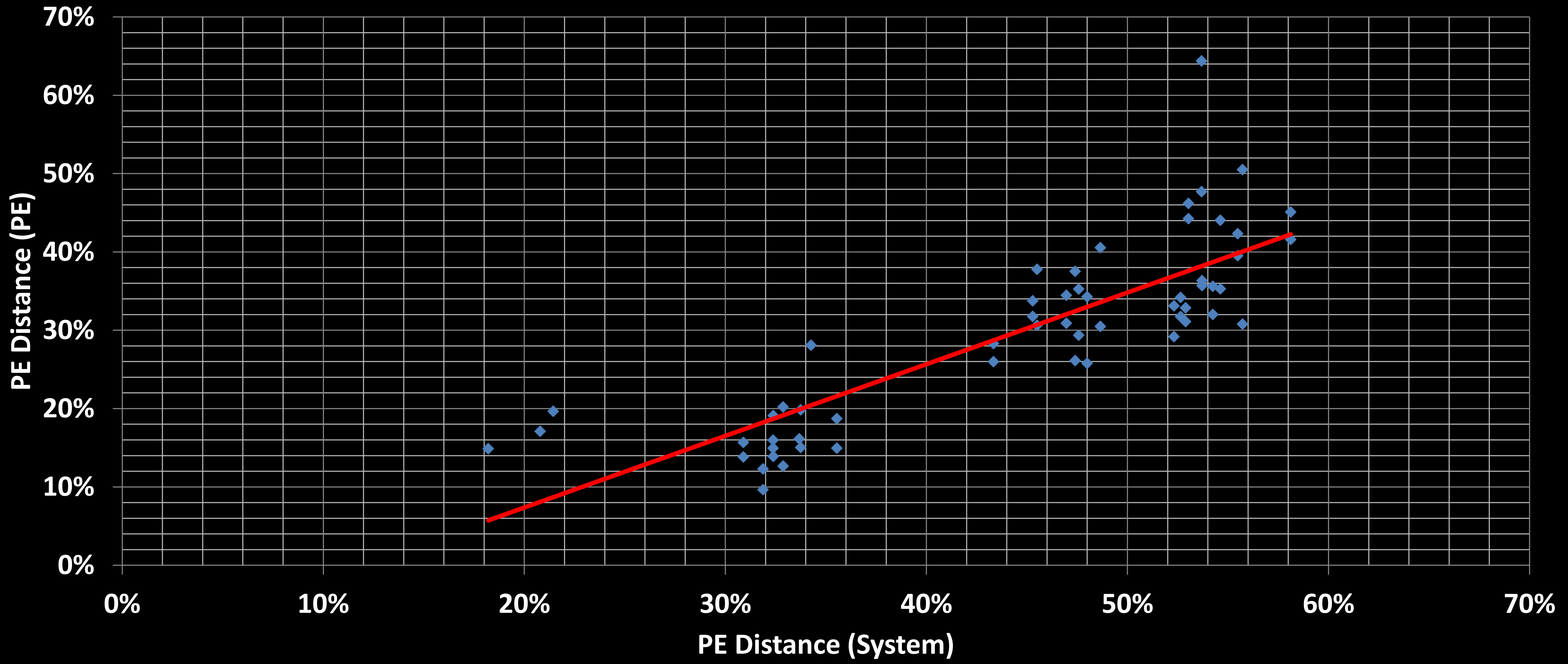
CORRELATIONS BETWEEN SYSTEM BLEU AND POST-EDITING BLEU



welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

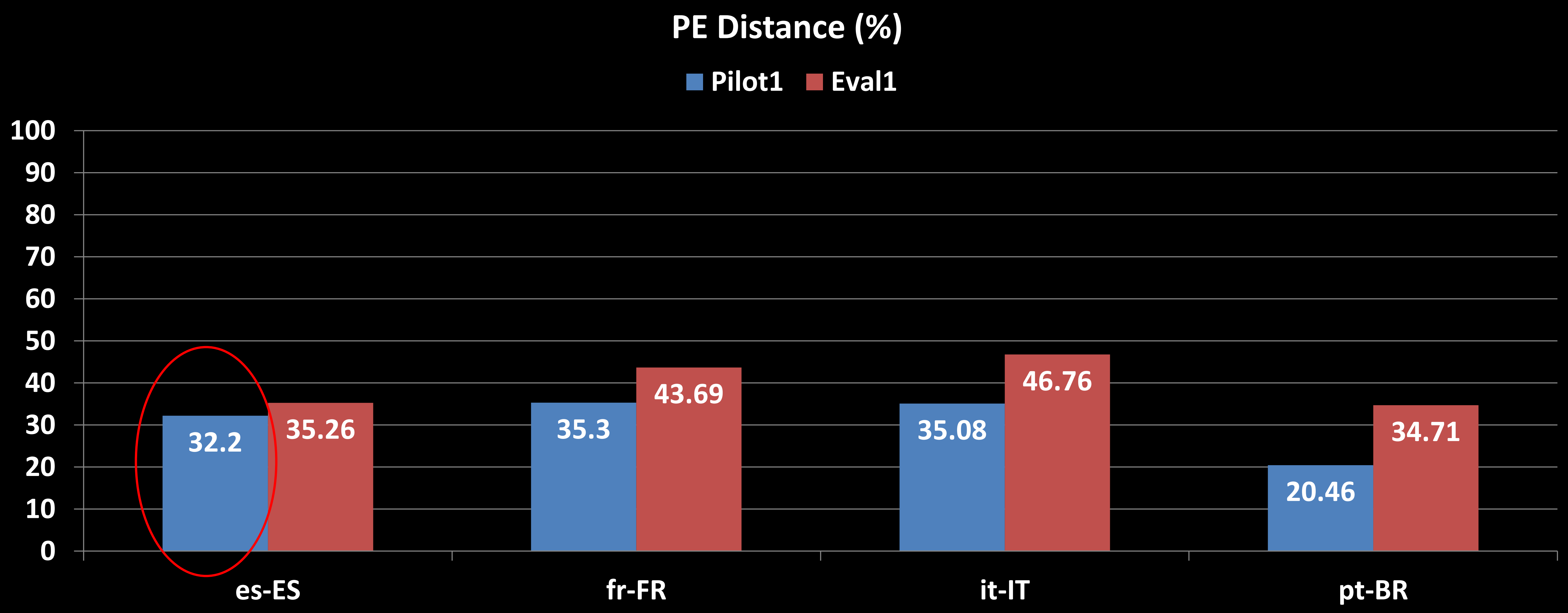
CORRELATIONS BETWEEN SYSTEM PE DISTANCE AND POST-EDITING PE DISTANCE



welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

REAL DATA WHERE WE COMPARE EVALUATION SCORES WITH SCORES FROM A 3-MONTH PILOT



LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS

welocalize
doing things differently

LESSONS LEARNED

- There is a very high correlation between the MT system autoscores generated during the evaluation phase and the autoscores generated from production using the same engines.
- However, the post-editing autoscores are considerably better than the MT system autoscores by around15%.
- We now differentiate the autoscores in our database as ‘System’ and ‘PE’.

welocalize
doing things differently

NEXT

MT evaluations in a production setting

welocalize
doing things differently

PRODUCTION SETTING

HOW TO MEASURE POST-EDITING EFFORT

- It is important to monitor the performance of MT and post-editors, especially during the initial launch of a new program
- The use of autoscoring to analyze post-project files is a valuable and cost-effective method to measure the post-editing effort
- They support rate negotiations and can help us to identify over- or under-editing by post-editors
- TER and PE Distance are useful metrics, with different underlying algorithms

PRODUCTION SETTING

HOW TO MEASURE POST-EDITING EFFORT

PE Distance - lower is better!

- Measures the number of insertions, deletions, substitutions required to transform MT output to the required quality level
- PE Distance values are derived by comparing the post-edited segments with the corresponding machine translation segments
- In our analysis the PE distance applies the Levenshtein algorithm and is **character-based**. This captures morphological post-edits, such as fixing word forms.

PRODUCTION SETTING

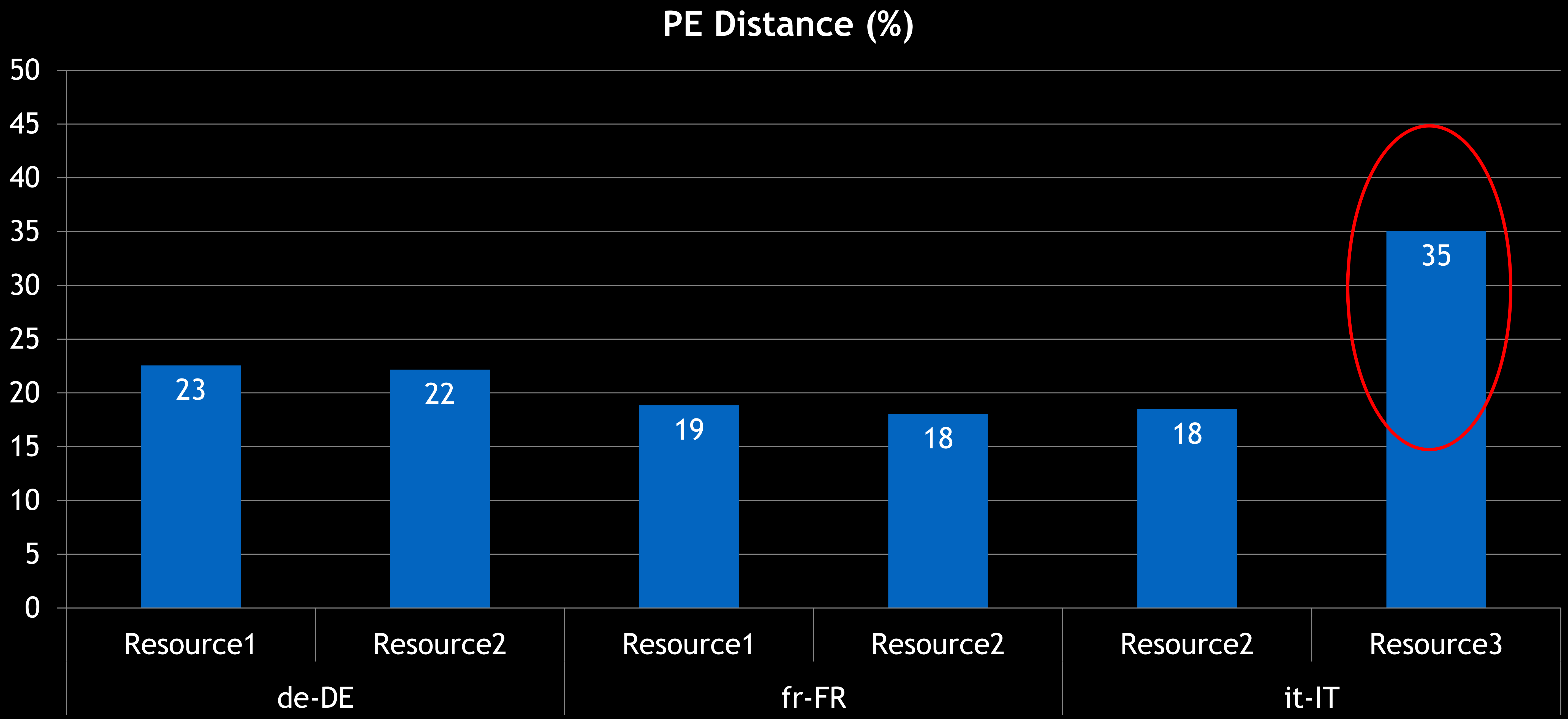
HOW TO MEASURE POST-EDITING EFFORT

TER - lower is better!

- TER stands for Translation Edit Rate
- It is an error metric for machine translation that measures the number of edits required to change a system output into the post-edited version
- Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences.
- Unlike PE Distance, TER is a **word-based** error metric and therefor does not capture morphological changes during post-editing.

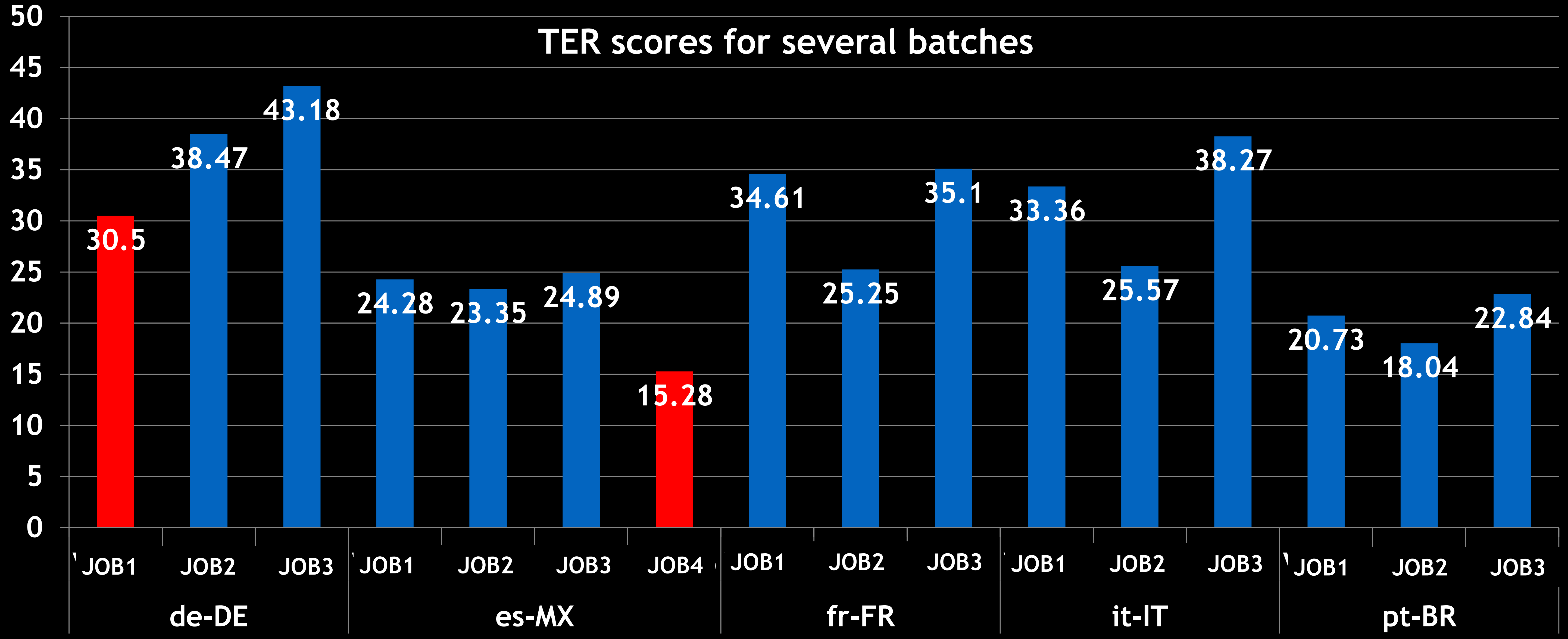
PRODUCTION SETTING

LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS



PRODUCTION SETTING

LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS: POST-PROJECT AUTOSCORES INDICATE UNDEREDITING



welocalize
doing things differently

PRODUCTION SETTING

TOOLS TO MEASURE POST-EDITING EFFORT

TOOL	INPUT FILES	OUTPUT REPORT	PROS	CONS
iOmegaT	xliff & more	xml	Includes productivity data	Generated in the CAT tool during translation, requires post-editor buy-in
MateCat	xliff	Excel	Includes productivity data as a built in feature	Generated in the CAT tool during translation, requires post-editor buy-in
Okapi	xliff	html	Allows us to measure PE distance post-project	Requires access to pre- and post-edited file sets
Post-Edit Compare	sdlxliff	html	Allows us to measure PE distance post-project	Requires access to pre- and post-edited file sets
Qualitivity	sdlxliff	Excel	Includes productivity data	Generated in the CAT tool during translation, requires post-editor buy-in
wescore	tmx	Excel	Allows us to measure PE distance post-project	Proprietary tool, Requires access to pre- and post-edited file sets

welocalize
doing things differently

PRODUCTION SETTING

MATECAT IS A FREE ONLINE CAT TOOL WITH EDITING LOG

www.matecat.com/support/translation-toolbox/editing-log/

The Editing Log contains **statistical information about the translation.**

matecat 11601337 (43563) > en-US > fr-FR

< Back to Translation | Export All Data in CSV

Job 43563 - Editing Log

Slowest 5.000 segments by time-to-edit

Summary

Words	Avg Secs per Word	% of MT	% of TM	Total Time-to-edit	Avg PEE %	% of words in too SLOW edits	% of words in too FAST edits
877	6.1s	100%	0%	01h:25m:24s	38%	4%	0%

help

Editing Details

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	PE Effort
254.4	43563	21799870	18.00	Machine Translation	85%	16m:18s	24%

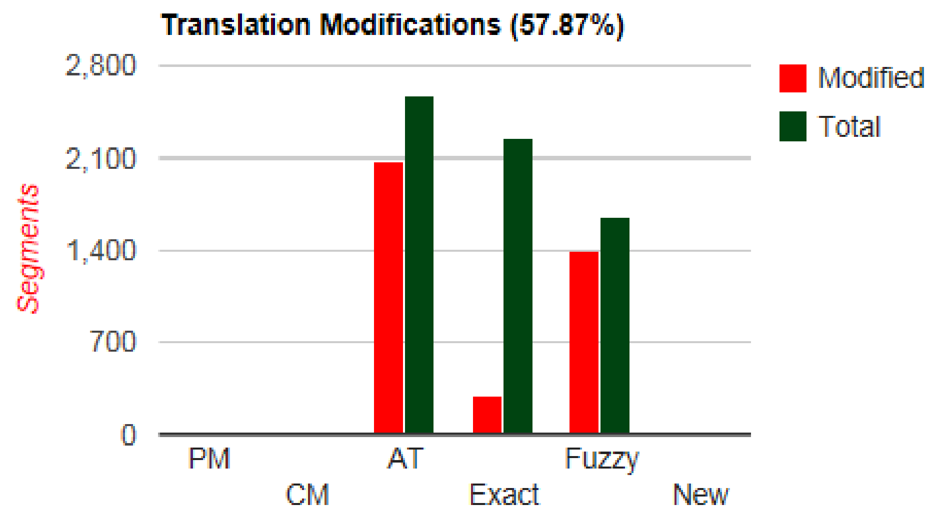
Segment	To view an article in 3D, select it and press the <g id="185">3D View</g> button below the preview pane.
Suggestion	Pour voir un article en 3D, sélectionnez-le et appuyez sur la <g id="185"> Vue 3D </g> bouton ci-dessous le panneau de prévisualisation.
Translation	Pour voir un article en 3D, sélectionnez-le et appuyez sur le bouton <g id="185">3D View</g> sous le panneau de prévisualisation.
Diff View	Pour voir un article en 3D, sélectionnez-le et appuyez sur ~~la <g id="185"> Vue 3D </g> bouton ci-dessous~~ le bouton <g id="185">3D View</g> sous le panneau de prévisualisation.

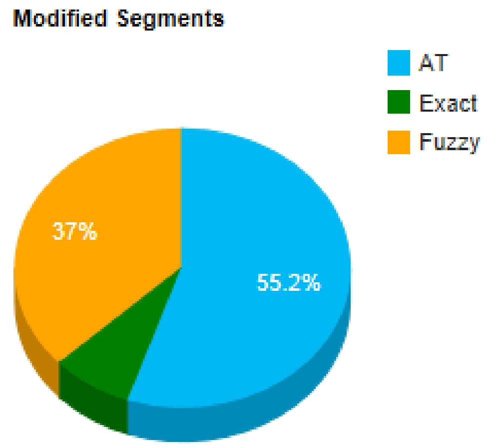
http://www.matecat.com/support/translation-toolbox/editing-log/

PRODUCTION SETTING

USE POST-EDIT COMPARE TO ANALYSE SDLXLIFF FILES

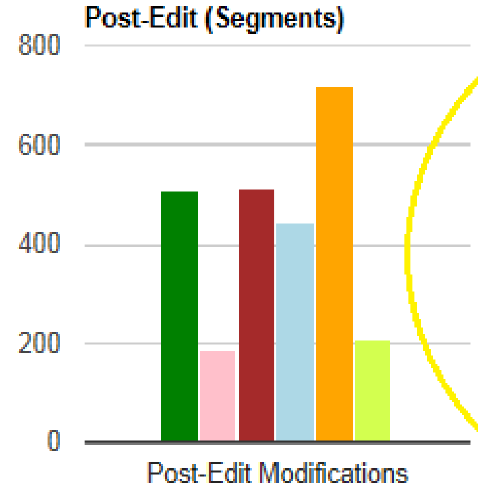
Translation Modifications	Segments			Words		Characters		Tags/Placeables	
	Total	*Modified*	*%*	*Added*	*Removed*	*Added*	*Removed*	*Added*	*Removed*
Perfect Match	0	0	*0%*	0	0	0	0	0	0
Context Match	0	0	*0%*	0	0	0	0	0	0
Exact Match	2251	294	*13%*	339	322	1963	2282	47	47
Automated Translation	2582	2074	*80%*	9128	9057	53240	50679	1337	1337
Fuzzy Match	1661	1391	*84%*	3049	2709	18279	15052	671	30
New	1	0	*0%*	0	0	0	0	0	0
Sub-Total				**12516**	**12088**	**73482**	**68013**	**2055**	**1414**
Total	**6496**	**3759**	***58%***	**24604**		**141495**		**3469**	

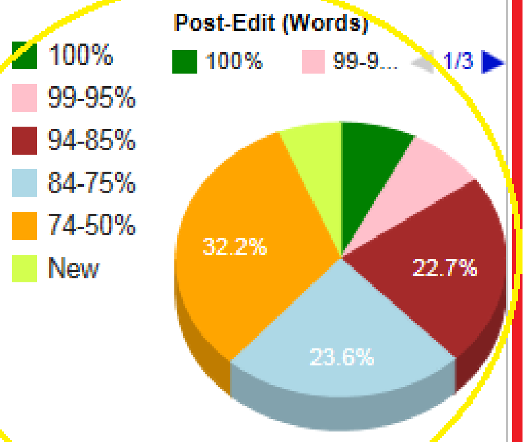




Post-Edit Modifications Analysis

Type	Segments	Words	Characters	Percent	Tags	Total
100%	509	2296	22581	7.50%	0	0 ()
95% - 99%	188	2361	12352	7.71%	54	0 ()
85% - 94%	514	6953	36004	22.71%	276	0 ()
75% - 84%	443	7217	38309	23.57%	428	0 ()
50% - 74%	721	9868	52569	32.23%	1286	0 ()
New	208	1926	11008	6.29%	582	0 ()
-						
Total	**2583**	**30621**	**172823**	**100%**	**2626**	**0 ()**





http://www.translationzone.com/openexchange/app/post-editcompare-495.html

welocalize
doing things differently

PRODUCTION SETTING

OKAPI FRAMEWORK TRANSLATION COMPARISON STEP

Summary

Repartition for Trans1 to Trans2:

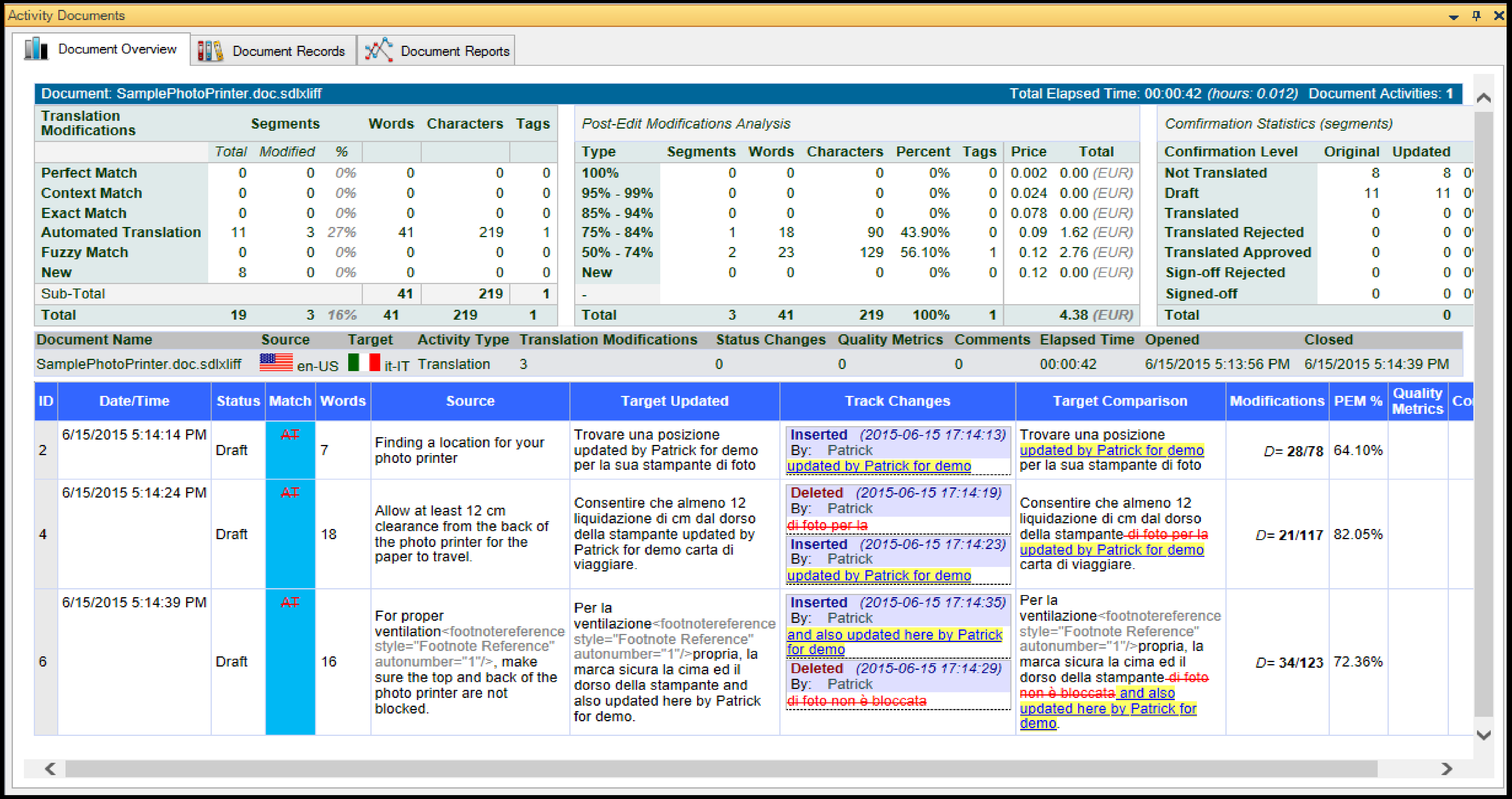
Scores	ED-Scores				FM-Scores			
	Segments	%	Words	%	Segments	%	Words	%
100	139	3	1414	3	176	4	1802	4
90 - 99	350	8	3954	8	346	8	3864	8
80 - 89	862	20	9850	20	674	16	7659	15
70 - 79	971	22	11137	23	804	19	9191	19
60 - 69	1078	25	12423	25	805	19	9332	19
50 - 69	598	14	6794	14	655	15	7500	15
40 - 59	197	5	2215	4	392	9	4479	9
30 - 39	33	1	359	1	240	6	2775	6
20 - 29	2	0	22	0	102	2	1159	2
10 - 19	1	0	4	0	36	1	398	1
0 - 9	104	2	1258	3	105	2	1271	3
Total	4335	100%	49430	100%	4335	100%	49430	100%

```
Total Number of Segments:      4335
Total Number of Words:         49430
Average word count per segment: 11.40
Average ED-Score (by segment): Trans1 to Trans2 = 69.95
Average FM-Score (by segment): Trans1 to Trans2 = 65.48
Average ED-Score (by word):    Trans1 to Trans2 = 69.76
Average FM-Score (by word):    Trans1 to Trans2 = 65.18
Edit Effort Score:             32.53
```

http://www.opentag.com/okapi/wiki/index.php?title=Translation_Comparison_Step

PRODUCTION SETTING

QUALITIVITY PLUGIN FOR SDL TRADOS STUDIO


Activity Documents

Document Overview | Document Records | Document Reports

Document: SamplePhotoPrinter.doc.sdlxliff — Total Elapsed Time: 00:00:42 *(hours: 0.012)* Document Activities: 1

Translation Modifications	Segments Total	Segments Modified	%	Words	Characters	Tags
Perfect Match	0	0	0%	0	0	0
Context Match	0	0	0%	0	0	0
Exact Match	0	0	0%	0	0	0
Automated Translation	11	3	27%	41	219	1
Fuzzy Match	0	0	0%	0	0	0
New	8	0	0%	0	0	0
Sub-Total				41	219	1
Total	19	3	16%	41	219	1

Post-Edit Modifications Analysis

Type	Segments	Words	Characters	Percent	Tags	Price	Total
100%	0	0	0	0%	0	0.002	0.00 (EUR)
95% - 99%	0	0	0	0%	0	0.024	0.00 (EUR)
85% - 94%	0	0	0	0%	0	0.078	0.00 (EUR)
75% - 84%	1	18	90	43.90%	0	0.09	1.62 (EUR)
50% - 74%	2	23	129	56.10%	1	0.12	2.76 (EUR)
New	0	0	0	0%	0	0.12	0.00 (EUR)
-							
Total	3	41	219	100%	1		4.38 (EUR)

Confirmation Statistics (segments)

Confirmation Level	Original	Updated
Not Translated	8	8
Draft	11	11
Translated	0	0
Translated Rejected	0	0
Translated Approved	0	0
Sign-off Rejected	0	0
Signed-off	0	0
Total		0

Document Name	Source	Target	Activity Type	Translation Modifications	Status Changes	Quality Metrics	Comments	Elapsed Time	Opened	Closed
SamplePhotoPrinter.doc.sdlxliff	en-US	it-IT	Translation	3	0	0	0	00:00:42	6/15/2015 5:13:56 PM	6/15/2015 5:14:39 PM

ID	Date/Time	Status	Match	Words	Source	Target Updated	Track Changes	Target Comparison	Modifications	PEM %	Quality Metrics	Co
2	6/15/2015 5:14:14 PM	Draft	AT	7	Finding a location for your photo printer	Trovare una posizione updated by Patrick for demo per la sua stampante di foto	Inserted (2015-06-15 17:14:13) By: Patrick updated by Patrick for demo	Trovare una posizione updated by Patrick for demo per la sua stampante di foto	D= 28/78	64.10%		
4	6/15/2015 5:14:24 PM	Draft	AT	18	Allow at least 12 cm clearance from the back of the photo printer for the paper to travel.	Consentire che almeno 12 liquidazione di cm dal dorso della stampante updated by Patrick for demo carta di viaggiare.	Deleted (2015-06-15 17:14:19) By: Patrick ~~di foto per la~~ Inserted (2015-06-15 17:14:23) By: Patrick updated by Patrick for demo	Consentire che almeno 12 liquidazione di cm dal dorso della stampante ~~di foto per la~~ updated by Patrick for demo carta di viaggiare.	D= 21/117	82.05%		
6	6/15/2015 5:14:39 PM	Draft	AT	16	For proper ventilation<footnotereference style="Footnote Reference" autonumber="1"/>, make sure the top and back of the photo printer are not blocked.	Per la ventilazione<footnotereference style="Footnote Reference" autonumber="1"/>propria, la marca sicura la cima ed il dorso della stampante and also updated here by Patrick for demo.	Inserted (2015-06-15 17:14:35) By: Patrick and also updated here by Patrick for demo Deleted (2015-06-15 17:14:29) By: Patrick ~~di foto non è bloccata~~	Per la ventilazione<footnotereference style="Footnote Reference" autonumber="1"/>propria, la marca sicura la cima ed il dorso della stampante ~~di foto non è bloccata~~ and also updated here by Patrick for demo.	D= 34/123	72.36%		

http://www.translationzone.com/openexchange/app/qualitivity-788.html

welocalize
doing things differently

LESSONS LEARNED

- The use of autoscoring to analyze post-project files is a valuable and cost-effective method to measure the post-editing effort.
- A productivity test requires upfront organization and buy-in from translators.
- It is important to find a tool that works with the given file format and workflow.
- Access to pre- and post-edit versions of projects is required. This is a challenge on some accounts.
- Identification and separation of MT segments from fuzzy segments may be required for some tools.
- Look for consistency across languages and resources. Unusually high or low scores can be a sign of over-editing or under-editing.

welocalize
doing things differently

NEXT

MT evaluations of post-edited files: a case study

welocalize
doing things differently

CASE STUDY

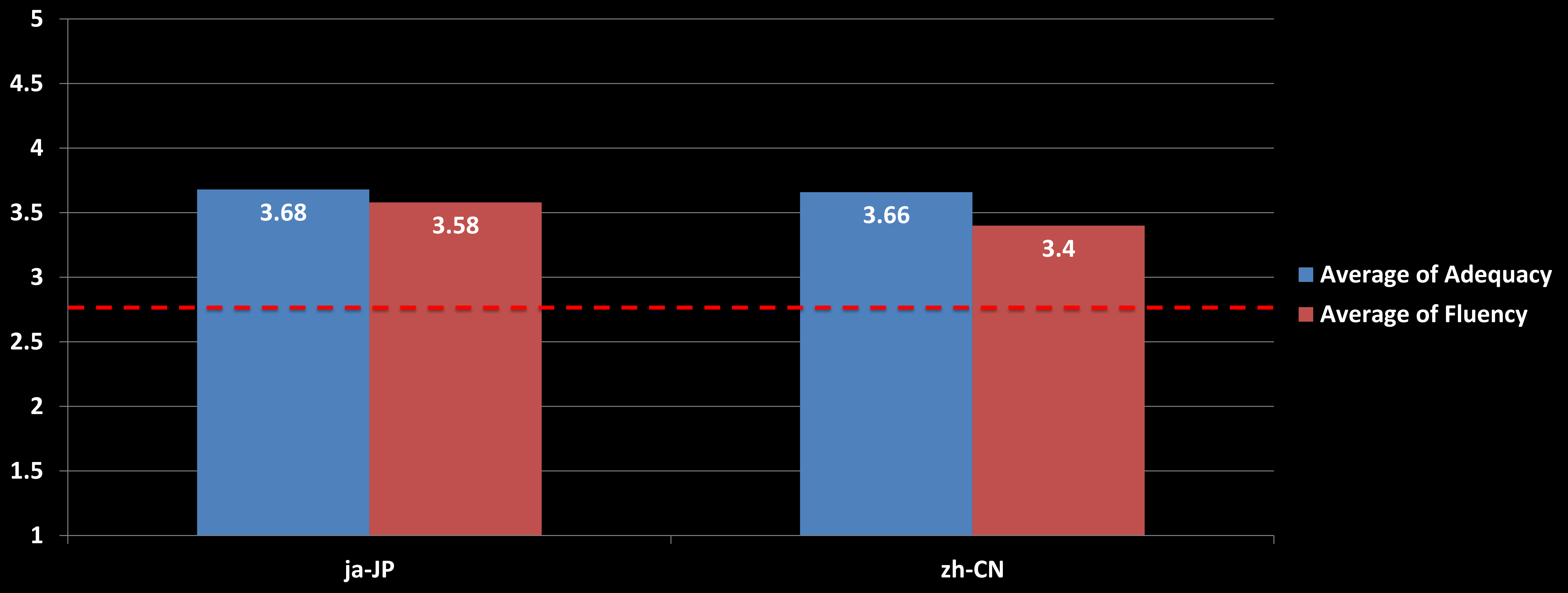
TEST PILOT FOR LIGHT AND FULL POST-EDITING

- Languages: Chinese (Simplified) and Japanese
- The resources are regular translators for this client
- In order to have comparable data, the same resource performed both light and full post-editing tasks of 438 segments

Full Post-editing → Light Post-editing → LQA of PE Kits → Autoscoring PE Kits → Adequacy & Fluency Scoring → Error Typology Scoring → Analysis & Report

CASE STUDY: HUMAN EVALS

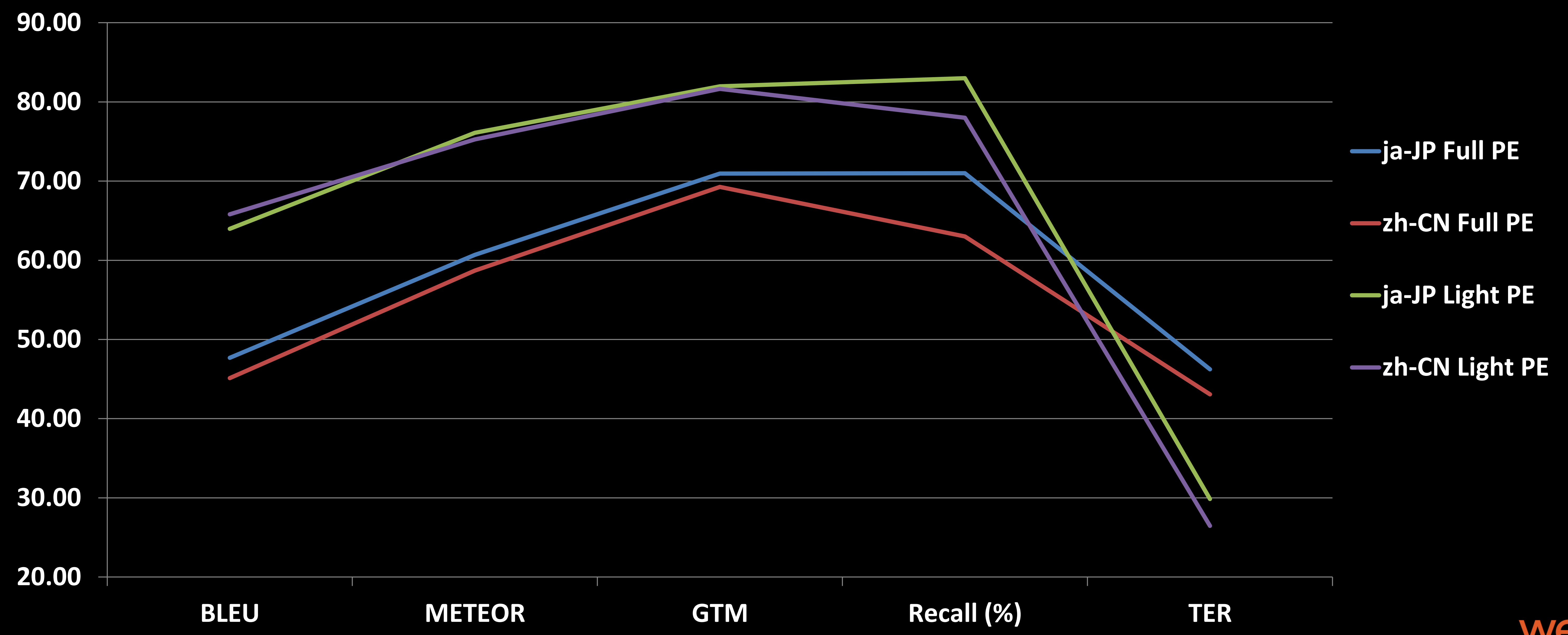
ADEQUACY AND FLUENCY SCORES



welocalize
doing things differently

CASE STUDY: AUTOSCORES

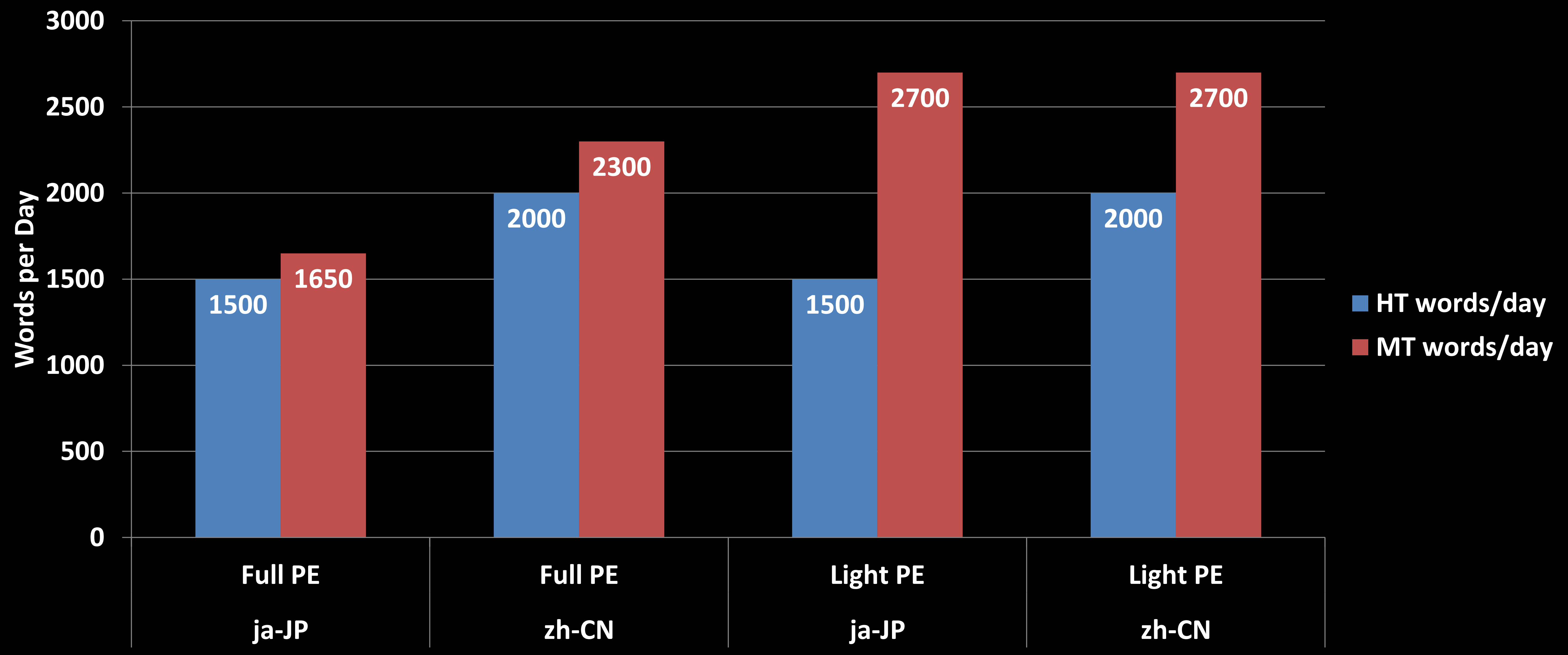
AUTOSCORES FOR LIGHT AND FULL POST-EDITING



welocalize
doing things differently

CASE STUDY: PRODUCTIVITY

PRODUCTIVITY FOR LIGHT AND FULL POST-EDITING



welocalize
doing things differently

CASE STUDY: LESSONS

LESSONS LEARNED

- Using autoscores on post-edited translations can indicate the level of post-editing effort involved for a specific content and MT engine
- The autoscores also illustrate the difference in effort between Light and Full Post-editing, approximately 20 point delta for BLEU and 15 point delta for TER
- The autoscores confirm that the resources have indeed managed to perform two distinct post-editing levels