


Machine Translation Hype

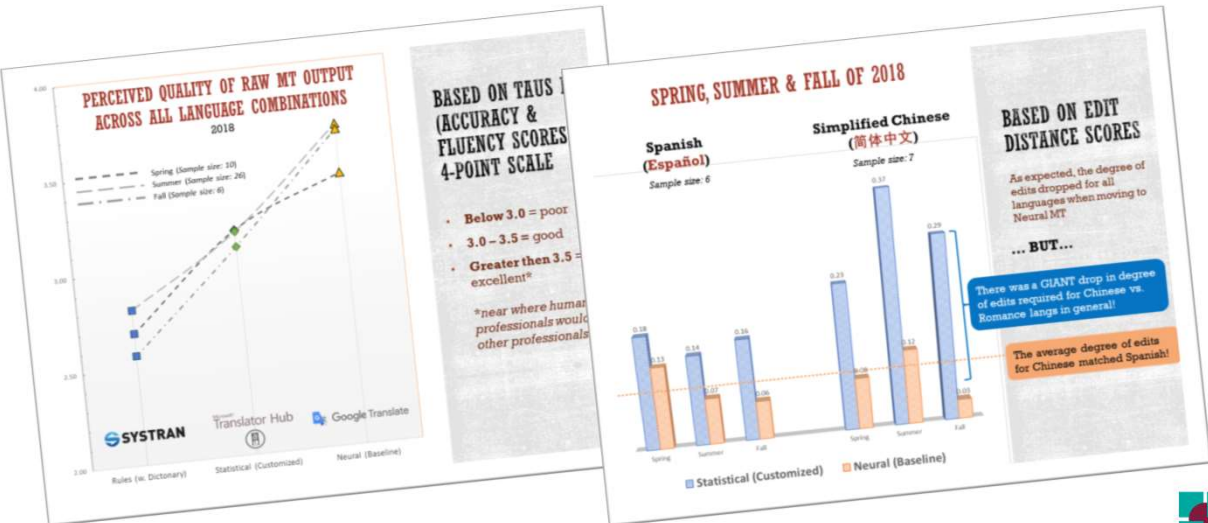
Crash-Tested by Translation Students



www.rws.com/moravia

1

Since 2016 I have been treating the last four weeks of my *Translation Technology* course as a testing laboratory...



PERCEIVED QUALITY OF RAW MT OUTPUT ACROSS ALL LANGUAGE COMBINATIONS 2018

System	Spring (Sample size: 50)	Summer (Sample size: 26)	Fall (Sample size: 6)
SYSTRAN	~3.4	~3.4	~3.4
Translator Hub	~3.4	~3.4	~3.4
Google Translate	~3.4	~3.4	~3.4
Neural (Baseline)	~3.4	~3.4	~3.4

BASED ON TAUS (ACCURACY & FLUENCY SCORES 4-POINT SCALE)

- Below 3.0 = poor
- 3.0 – 3.5 = good
- Greater than 3.5 = excellent*

*near where human professionals would other professionals

SPRING, SUMMER & FALL OF 2018

Spanish (Español) Sample size: 6

System	Spring	Summer	Fall
Statistical (Customized)	0.29	0.14	0.16
Neural (Baseline)	0.13	0.07	0.08

Simplified Chinese (简体中文) Sample size: 7

System	Spring	Summer	Fall
Statistical (Customized)	0.29	0.12	0.29
Neural (Baseline)	0.08	0.07	0.08

BASED ON EDIT DISTANCE SCORES

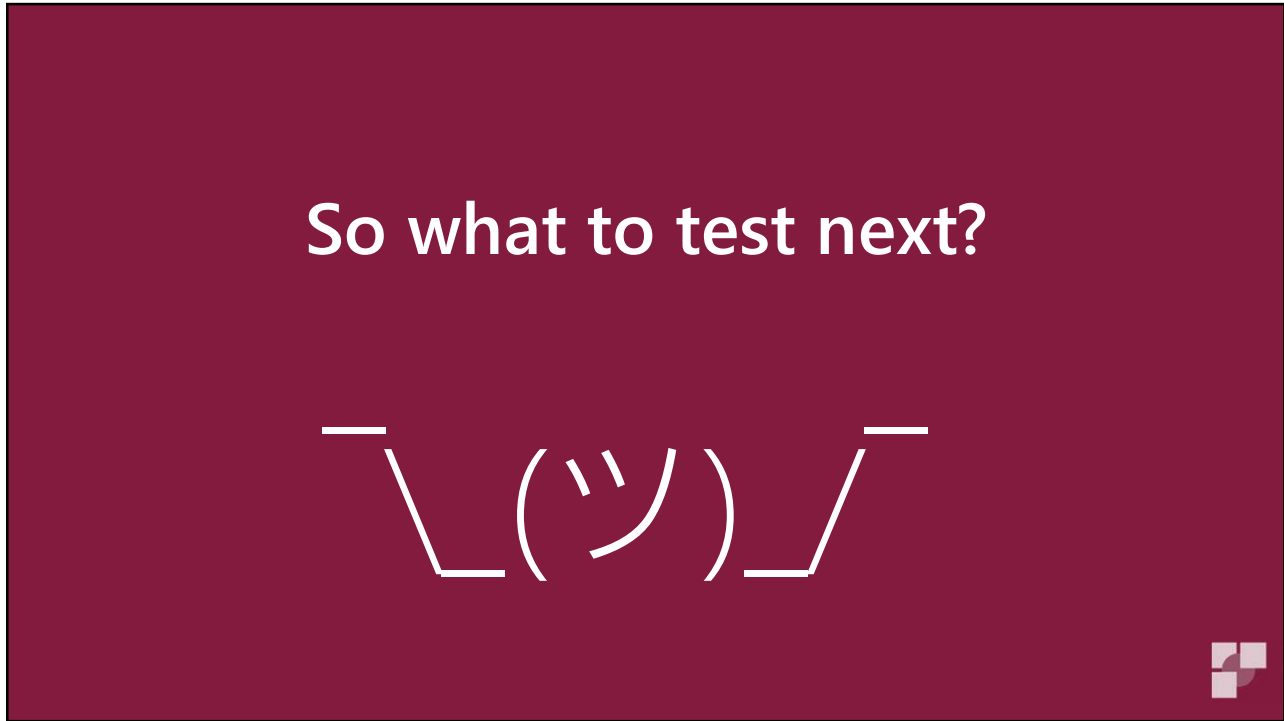
As expected, the degree of edits dropped for all languages when moving to Neural MT

... BUT ...

There was a GIANT drop in degree of edits required for Chinese vs. Romance langs in general!

The average degree of edits for Chinese matched Spanish!

2



3



4

“

What is the sweet spot in the hybrid process
 what is the ideal
 memory is no
 transla

Differences Between the Languages Are Substantial


The trends in MT productivity have a quite stable pattern over time, as the trend reports show. But that does not mean that machine translation is equally productive across different languages. There are considerable differences between languages when it comes to the average time that is needed to edit a machine translation for every 100 characters of the source text. Again we took the same sample, and filtered on a few of the bigger target languages that used MT and had English as the source language.

As shown in figure 10, it appears that the MT productivity in the Western-European languages is twice or almost three times as high as in the Asian languages.

Brazilian Portuguese and Spanish being the MT champions in DQF, how do MT and TM compare in these languages? MT here is on par with fuzzy matches between 75 and 85%, and it shows that the sweet spot for switching to machine translation might move up to the higher TM match rates.

- TAUS DQF BI Bulletin - Q1 2019; March

5



Classroom Laboratory

A When should we ditch TM and take an MT generated segment instead?

1. At what point does MTPE require the same amount of editing as a "fuzzy" TM match on average?
2. Which requires more editing overall:
 a full text of MT generated segments OR
 a full text of "fuzzy" TM matches in the 89%-69% range ("low fuzzies")?

B What's the relation between edit-distance and actual time spent in a TM vs. MT scenario?

1. When editing the content, which takes less time:
 "fuzzy" TM edits OR MTPE?
2. Is there a matching correlation between edit-distance and time spent fixing segments for "low fuzzies" and MTPE?

6

Experiment A

When should we ditch TM and take an MT generated segment instead?

www.rws.com/moravia

7

Methodology

- 8 students; en-US > 3 target language groups (es-419 (es-LA), zh-CN, zh-TW)
- There are 22 source segments in total split equally into "Set A" and "Set B"
- All are *full* sentences from older (2015 or earlier) Apple iOS documentation (pulled from PDFs) with an average length of 15 wds
- For **TM**, pre-populate with fuzzy matches in the 89%-69% range (avg. match across Set A = 79%; Set B = 78%)
- For **MT**, pre-populate with Google Translate generated sentences
- One student *per language* will complete Set A with a "fuzzy match TM" and Set B as MTPE; the other student will do the reverse. Then they will switch.
- Edit distance will be measured for every segment
- NOTE:** Students can be unpredictable on occasion, so the official Apple translations were added as a "control group" to minimize this risk

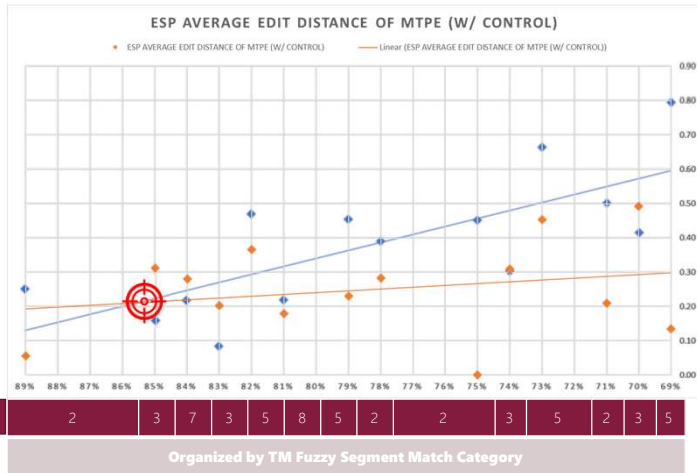
For every language...

 MT	Set B - 161 wds - 11 sentences; avg. 15 wds long	Set A
 Fuzzies TM	Set A - 165 wds - 11 sentences; avg. 15 wds long	Set B

9

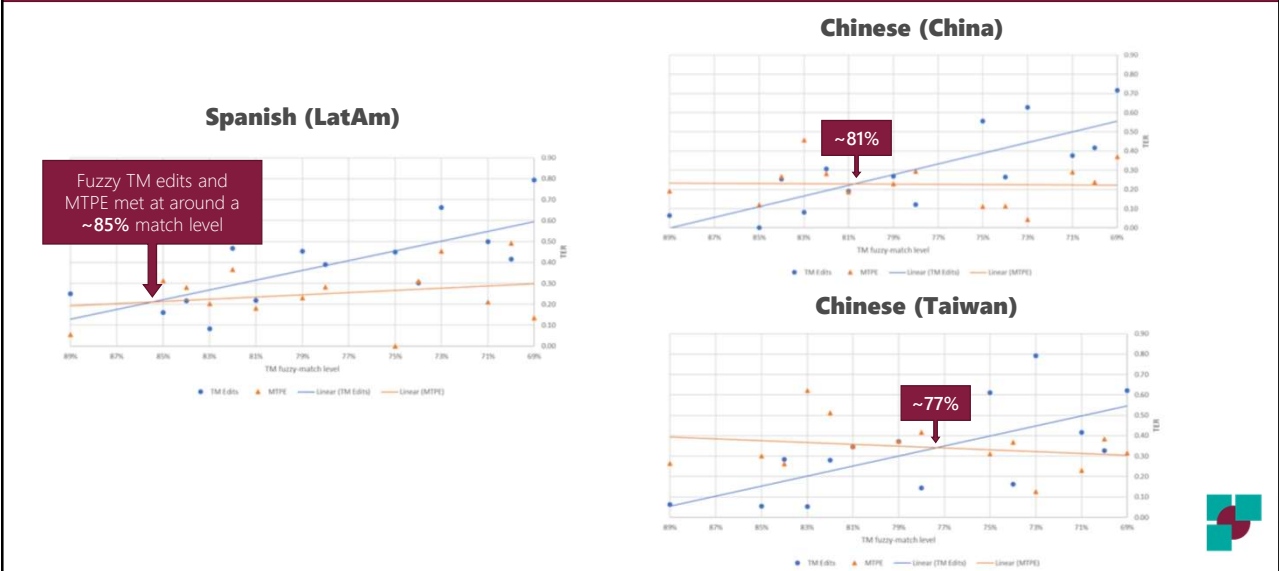
1 At what point does MTPE require the same amount of editing as a fuzzy TM match on average?

- > Organize segments by original TM match categories
- > Calculate location of point based on average edit-distance for total sampled segments by language
- > Do this both for fuzzy TM edited & MTPE segments
- > Plot the trendlines for both
- > Find intersection

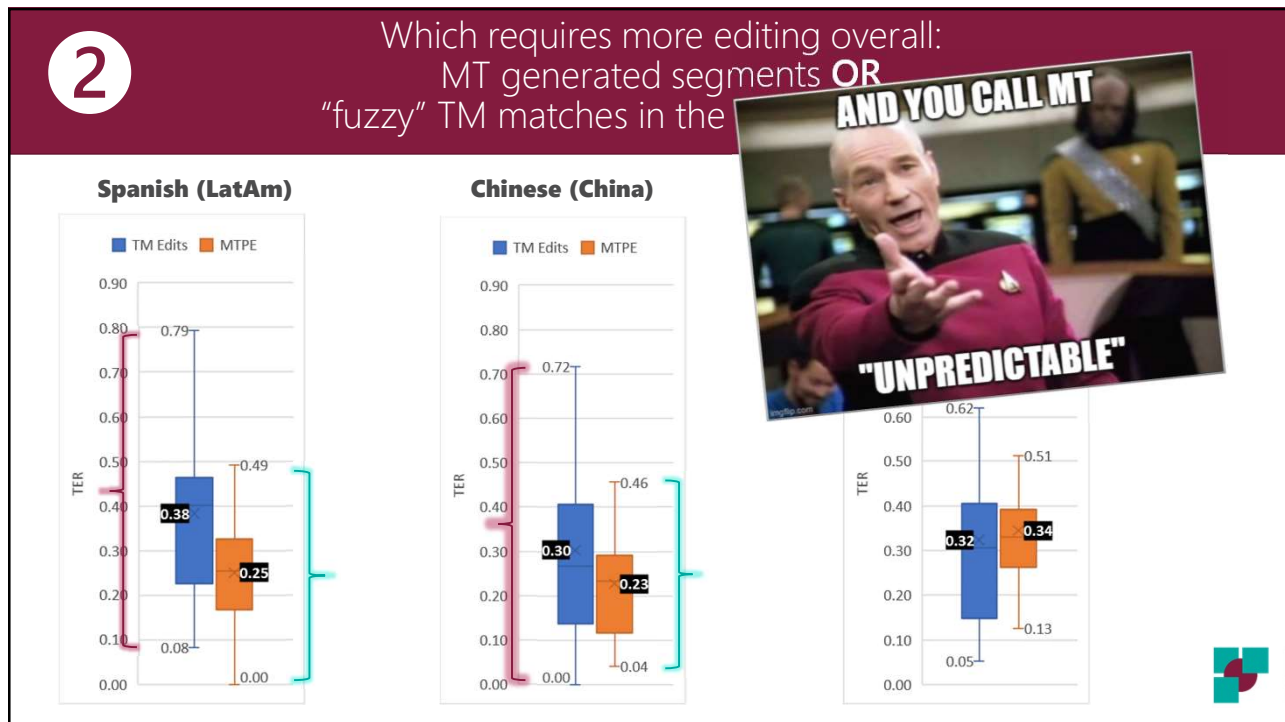


10

At what point does MTPE require the same amount of editing as a fuzzy TM match on average?



12



13

Experiment B

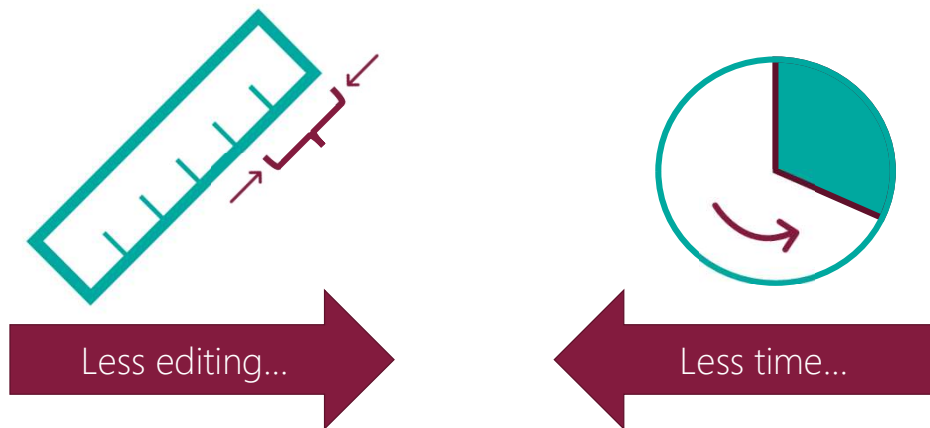
What's the relation between edit-distance and actual time spent in a TM vs. MT scenario?

www.rws.com/moravia

16

Why this test?

Our “knee jerk” reaction when seeing less edit distance is...



17

SDL* Blog
Topics ▾ About ▾
SDL Trados Blog 🔍

[Machine Translation](#)

Edit Distance: Not a Miracle Cure

by [Izabella Lizuka](#)
March 13, 2019 - read time: 10 min

With the current rapid developments in Neural Machine Translation (NMT), discussions on its market impact are gathering pace particularly around post-editing. While some suggest that paying by the hour – rather than the traditional per-word rate – is the way to go, others feel that translators should only be paid for actual changes to the MT output. This second option is usually operationalized by using what’s known as an edit distance metric.

Edit distance metrics have also been suggested as a means to replace or support BLEU (bilingual evaluation understudy), which is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another and is used for assessing MT quality - thus support MT engine development. These discussions are not necessarily new; however, edit distances have not become mainstream as of yet. Why? The apparent ease of this solution does in fact hide a lot of complications. Let’s take a closer look.

What is edit distance?

There are various ways to measure edit distance, with [Levenshtein](#) distance and TER(p) among the best-known. In essence, all edit distance metrics work on the same principle: they measure the minimal number of edits necessary to change one string (in this case, the MT output) into another string (in this case, the final translation). An edit can

-
-
-
-
-

18

“

t

The reasoning is that
on the post-ed
payment should
wo

Can edit distance be used to measure post-editing effort?

Another way to utilize edit distance is as an indication of post-edit effort. The reasoning is that fewer edits indicate less effort on the post-editor's part – and hence the payment should be less, so as to only pay for work completed. This is not entirely true though, which becomes clear when the translation and post-editing tasks are broken down into their parts:

```

graph TD
    A[Job intake (download, admin)] --> B[Read source of one segment]
    B --> C[Read target and compare to source]
    C --> D[Read segments before and after for context]
    D --> E[Check term list and style guide]
    E --> F[Do research and/or ask query]
    F --> G[Think about best approach]
    G --> H[Typing: translate or edit]
    H --> I[Run QA on full job]
    I --> J[Return job + admin]
    
```

- Edit Distance: Not a Miracle Cure; March 2014

19

My one advantage in measuring for time...

Classroom

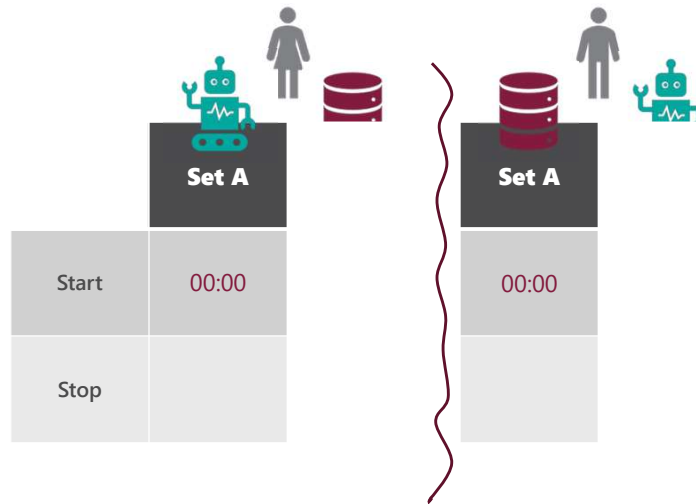
Home

20

Methodology

- > Begin stopwatch at start of translation with each set
- > Pause timer and call instructor over when...
 - > All segments confirmed
 - > All terminology checks against the term database were cleared out (or personally verified as a "non-error")
 - > All other automated quality checks that could be cleared out were completed (capitalization, punctuation, number mismatches, etc).
- > If issue was spotted, unpause the timer and call instructor over again when fixed.
- > Instructor writes down time of completion
- > **NOTE:** there is no "control group" for this experiment

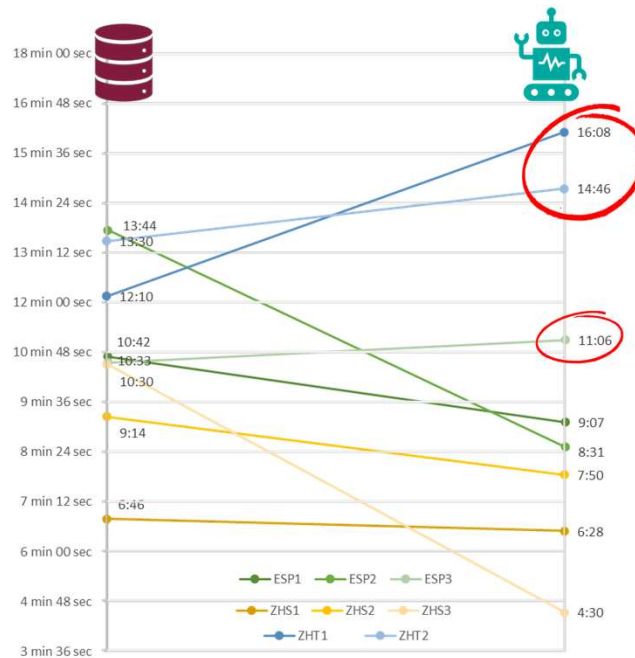
Every student has a stopwatch (i.e., phone) so easy to accommodate



22

1

When editing ~160 wds / 11 sentences of content, which takes less time:
 "fuzzy" TM edits (89%-69%)
 or MTPE?



23

↔

More time...

More editing?

When asked "why" they spent more time on MTPE than TM edits...

ESP3

"TM was easier because the sections to change are clearly marked so I only needed to cut out minor sections."

TM	MT
.20	.22

ZHT1 & ZHT2

"spent a lot of time looking up the proper terms for Taiwan since the MT kept giving Chinese for mainland China, but in traditional characters."

TM	MT
.26	.23
.26	.25

24

2

Is there a matching correlation between edit-distance and time spent fixing segments for both "low fuzzies" and MTPE?

TM Edits

Nearly flat


MTPE

Pitched

No matter the level of editing, the time spent is fairly constant. Conclusion: It's faster to figure out what to edit with TM fuzzies.

...whereas with MT, time spent does go up with increasing edit-distance.


26



Subjective Experiments

What's the "experience" when working with MT?


www.rws.com/moravia





27

What did you prefer?
PE of wholly generated raw MT output OR
a trained "adaptive" MT system where you had control over the output

Nearly 50/50 in a show of hands every year since 2016



 ...those who prefer "adaptive MT" are nearly always Romance language speakers



28

I've mixed the reference translations in with raw MT from 4 major engines. Which do you think is human?

Students are given homework the week before to provide multiple sentences of human-made content (bilingual) that they believe will be "tough" for MT.

Poor Tom Fool, yonder behind the wagon, mumbling his bone with the honest family which lives by his tumbling.	可憐的湯姆 帕爾，馬車後面的遠處，他可憐的湯姆 傻瓜，就在馬車後面，噙理的骨頭與他的翻滾生活的誠實家庭喃喃自語。	可憐的湯姆 傻瓜 (Tom Fool) 身處馬車後面，與誠實的家庭糊塗骨頭，誠實的家庭靠他的翻滾生活。	可憐的湯姆 愚人，在馬車後面，喃喃自語與誠實的家庭，這是由他的翻滾。	再過去是可憐的小丑湯姆躺在馬車後頭帶著一袋老小時骨頭，這些老實人就靠他翻筋斗賺來的錢過活。	
The young lady's countenance, which had before worn an almost livid look of hatred, assumed a smile that perhaps was scarcely more agreeable.	只是這笑容比起方才惡狠狠嫩青的臉色來，也好看不了多少。	這位年輕女士的臉容，曾經穿著一種幾乎鮮豔的仇恨外表，假裝一個微笑，也許幾乎沒有比較愉快。	這位年輕女士的臉容，以前帶著近乎憤怒的神情，露出了一絲可能更討人喜歡的微笑。	這位年輕女士的臉上以前帶著幾分仇恨的嫩青色，現在露出了一種也許再好不過討人喜歡的微笑。	這位年輕女士的容顏曾帶過幾乎幾乎是充滿生氣的仇恨表情，但露出了微笑，這也許簡直讓人難以接受。
The world is a looking-glass, and gives back to every man the reflection of his own face.	這世界是一面鏡子，每個人都可以在裡面看見自己的影子。	這個世界是一個扁薄的玻璃，並將自己臉上的反射回轉給每個人。	世界是一面鏡子，把自己臉上的倒影還給每個人。	這個世界是一個窺視鏡，並且將每個人的面孔反射給每個人。	世界是一個看起來玻璃，並回轉每個人自己的臉的反射。
A very stout, puffy man, in buckskins, and Hessian boots, with several immense neckcloths that rose almost to his nose.	他穿著鹿皮褲子，筒上有流蘇的靴子，圍著好幾條寬大的領巾，幾乎直達到鼻子。	一個非常健壯，浮腫的男人，穿著鹿皮，黑森靴，幾條幾乎高到鼻子的大領巾。	一個非常粗壯，浮腫的人，在鹿鹿，和黑森靴子，與幾個巨大的頸布，幾乎上升到他的鼻子。	一個非常粗壯，浮腫的男人，在牛皮和黑森靴子，幾乎上升到他的鼻子幾乎巨大的領口。	一個非常矮胖，矮胖的人，穿著鹿皮和黑森州的靴子，幾條巨大的圍巾圍在他的鼻子上。



Fall 2019 = 1st time students in a language group (Spanish) chose MT generated output over the human reference text



29

I'm looking for your ideas...



30



Q&A

31