WIKIMEDIA
FOUNDATION

Machine translation, human editors

Language quality assessment trends

Language Team | Design Strategy
Last updated January 2023

An opportunity to take a look at quality

- The primary focus is to analyze human edits made to MT outputs
- But -- we had a chance to take a look at quality of the paragraphs randomly sampled

The wikis

ISO	Language	Size	Articles	Active users	Boost	CX pubs	CX[4] out of beta	Better with CX?[5]
sq	Albanian	sm	84k	212	Core boost	4900 (88% source=en)	yes	yes (+14.39%)
id	Indonesian	med	593k	3300	no	12k (96% source=en)	no	no (-1.39%)
zh	Standard written Chinese	lg	1.2m	9500	no	22k (86% source=en)	no	yes (+8.31%)

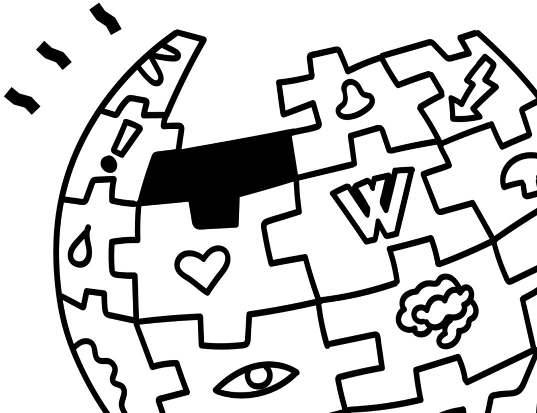
维基百科

海納百川，有容乃大

人人可編輯的自由百科全書

Selamat datang di Wikipedia

Mirësevini në Wikipedia



The process and dataset

3-step Sampling Process

1. **Randomly sample** the pool of Content Translation (CX) published articles within a set of defined parameters
2. **Automate a data pull to retrieve data (text) from these articles.** For each of the articles, there are three types of language data corresponding to each article section (plus a bit of metadata):
   a. Text of the source article
   b. Machine translation output generated from this source text
   c. Final published version (post-edited machine translation output)
3. **Clean the data set** by removing article paragraphs for which machine translation outputs were not used, or for which machine translation outputs were generated but the corresponding section was not published. Also, cap the number of paragraphs for each article at 10 to ensure a more balanced representation from the articles sampled.

Parameters used in sampling

A. **Source language** - Only articles with English as a source language were included. English is the most frequent source language for all CX publications, with rates as high as 80-90% depending on the target wiki.

B. **Translator diversity and experience** - For each of the wikis, to establish a minimal amount of individual translator variation (i.e., we didn't want to inadvertently retrieve translations from a single editor), the 50 articles represented work of 10 or more individual editors, with no individual editor contributing more than 8 to the sampled data. In addition, 50% of the articles were published by a 'newer' editor, defined as an account created no longer than 2 years prior. The other half of articles were published by editors with CX publications beginning at least 3 years prior.

C. **Machine translation engine** - Being one of the most common services used by CX users (overall, across all language pairs), we tried to narrow to articles/paragraphs produced exclusively using initial machine translation outputs provided by Google Translate.

D. **Topic-category** - All articles belonged to the 'nature/natural phenomena' or 'biography' category. This was done to limited the amount of language variation we recognize is present across different genres.

E. **Article length** - All articles contained a minimum of 7+ paragraphs. These paragraphs could be contained in a single article section, or across multiple sections of an article. In other words, there was no minimum number of article sections specified.

F. **Percent machine translation modified** - The current CX quality algorithm calculates an approximate 'percentage the MT is modified.' We sampled articles from three categories defined as (1) less than 10% modified, (2) between 11 and 50% modified, and (3) more than 51% modified.

3-step Sampling Process

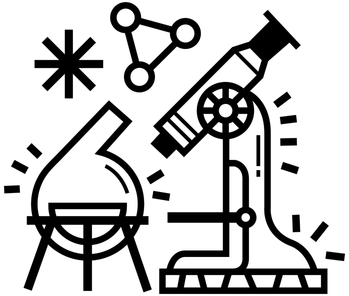
1. **Randomly sample** the pool of Content Translation (CX) published articles within a set of defined parameters
2. **Automate a data pull to retrieve data (text) from these articles**. For each of the articles, there are three types of language data corresponding to each article section (plus a bit of metadata):
   a. Text of the *source article*
   b. *Machine translation output* generated from this source text
   c. Final *published version* (post-edited machine translation output)
3. **Clean the data set** by removing article paragraphs for which machine translation outputs were not used, or for which machine translation outputs were generated but the corresponding section was not published. Also, cap the number of paragraphs for each article at 10 to ensure a more balanced representation from the articles sampled.

Results of this process: What we analyzed

Indonesian *(id)*: **294 paragraphs**[1] across **38 articles**[2]

Albanian *(sq)*: **381 paragraphs** across **47 articles**

Standard Written Chinese *(zh)*: **407 paragraphs** across **47 articles**



[1] Paragraph length is variable and ranges from one to around five sentences of medium length on average, but occasionally also includes some section titles, which may be best characterized as long phrases.
[2] Currently coordinating to sample additional data to increase article count to 47-50; this is also the goal for zhwiki

What we analyzed

Detailed breakdown (after cleaning process)

	Indonesian (id)	**Albanian** (sq)	**Chinese** (zh)	**Overall/combined**
Articles	38	47	47	**132 articles**
Paragraphs	294	381	407	**1082 paragraphs**
Paragraphs per article analyzed (median)	8[1]	8	10	**9 paragraphs per article**
Translator experience (on wiki)[2]	*junior* = 138 (47%) *senior* = 158 (53%)	*junior* = 203 (53%) *senior* = 178 (47%)	*junior* = 206 (51%) *senior* = 201 (49%)	***junior* = 547 (51%)** ***senior* = 535 (49%)**
Percent of machine translation outputs modified[3]	*<10%* = 3 (1%) *11-50%* = 159 (54%) *50%+* = 132 (45%)	*<10%* = 200 (53%) *11-50%* = 161 (42%) *50%+* = 20 (5%)	*<10%* = 145 (36%) *11-50%* = 148 (36%) *50%+* = 114 (28%)	***<10%* = 348 (32%)** ***11-50%* = 468 (43%)** ***50%+* = 266 (25%)**
Machine translation engine[4]	*Google* = 155 (53%) *Yandex* = 139 (47%)	*Google* = 369 (67%) *Yandex* = 11 (3%)	*Google* = 195 (48%) *Yandex* = 212 (52%)	***Google* = 748 (72%)** ***Yandex* = 150 (28%)**

[1] Included in dataset and analyzed; not median paragraphs present in actual article
[2] Via 'time since first translation'; junior=less than 2 years; number represents number of sections produced by junior or senior, not unique number of individuals
[3] Measured at level of the overall article; section-level measurements may vary
[4] To constrain independent variables, we attempted to narrow only to Google Translate, but this was not possible for Indonesian

Quality assessment

- Assessment was provided by professional translators/language experts/linguists*
- **We assessed each article at the level of the paragraph, according to 2 measures:**

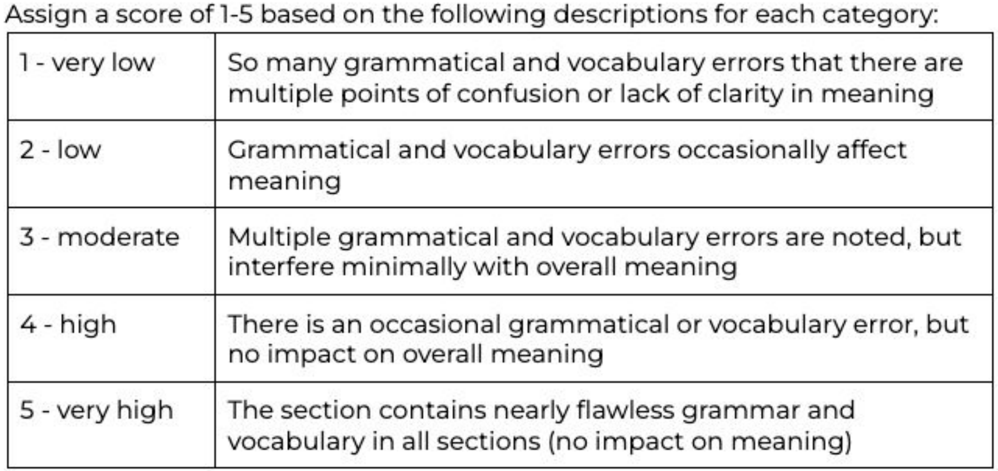
  ① **Language quality** - to what degree does the text contain grammatical errors, and to what degree do these impact meaning and understanding?

  ② **Machine translation markedness -** Compared to human-written texts, to what degree are there language choices in the text indicating that machine translation outputs were used to produce it?
- *Why?* We wanted to take a look at quality trends/patterns, and potentially identify sample pools of articles (of 'high' and 'low' quality) that we could use to test any future iterations/changes to the quality algorithm and checks system.

* To avoid potential bias, no access to metadata was provided to reviewers at this time; for example, information about translator experience, degree of MT output modification, etc… Reviewers were simply provided with the language samples they were asked to rate.

Language quality score assignment

Grammar + Meaning



Assign a score of 1-5 based on the following descriptions for each category:

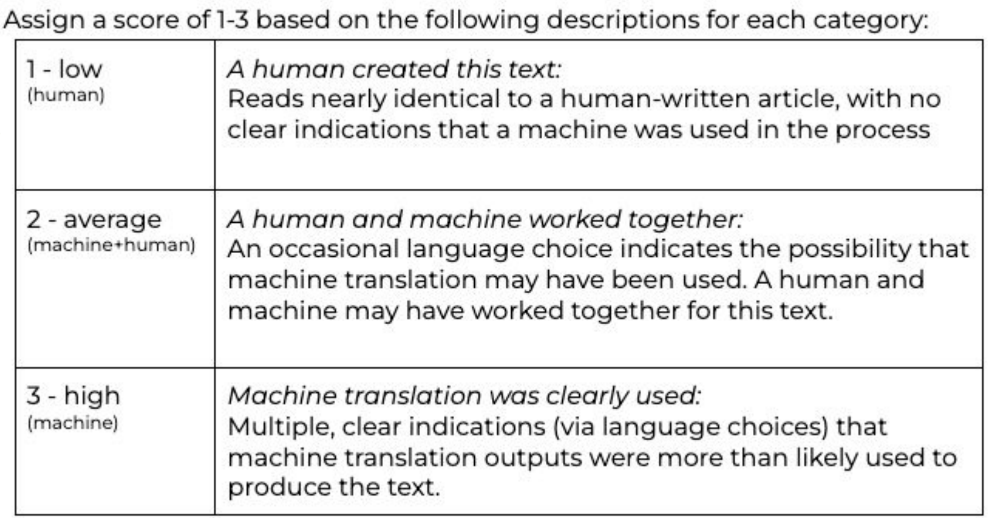
1 - very low	So many grammatical and vocabulary errors that there are multiple points of confusion or lack of clarity in meaning
2 - low	Grammatical and vocabulary errors occasionally affect meaning
3 - moderate	Multiple grammatical and vocabulary errors are noted, but interfere minimally with overall meaning
4 - high	There is an occasional grammatical or vocabulary error, but no impact on overall meaning
5 - very high	The section contains nearly flawless grammar and vocabulary in all sections (no impact on meaning)

Machine translation markedness

Signals/signs that MT outputs were used to create the text

✍️

✍️➕



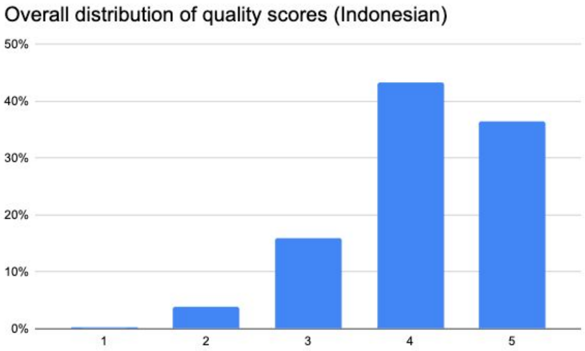
Assign a score of 1-3 based on the following descriptions for each category:

1 - low (human)	*A human created this text:* Reads nearly identical to a human-written article, with no clear indications that a machine was used in the process
2 - average (machine+human)	*A human and machine worked together:* An occasional language choice indicates the possibility that machine translation may have been used. A human and machine may have worked together for this text.
3 - high (machine)	*Machine translation was clearly used:* Multiple, clear indications (via language choices) that machine translation outputs were more than likely used to produce the text.

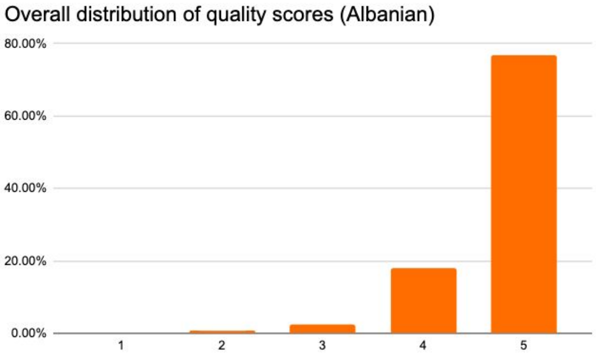
Quality trends and results

Overall distribution of quality scores

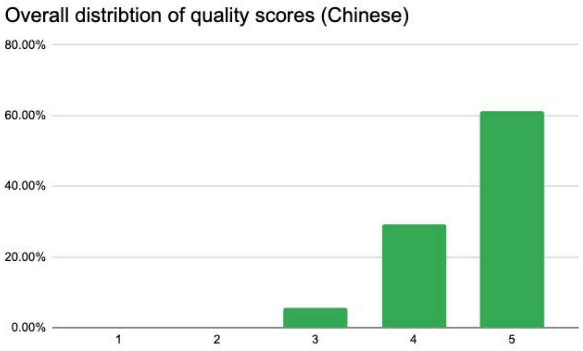
Overall scores were high, but highest for Albanian and Chinese



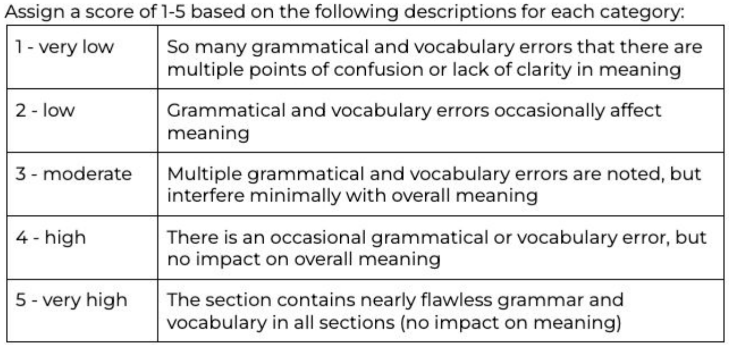
Paragraph quality score



Paragraph quality score



Paragraph quality score

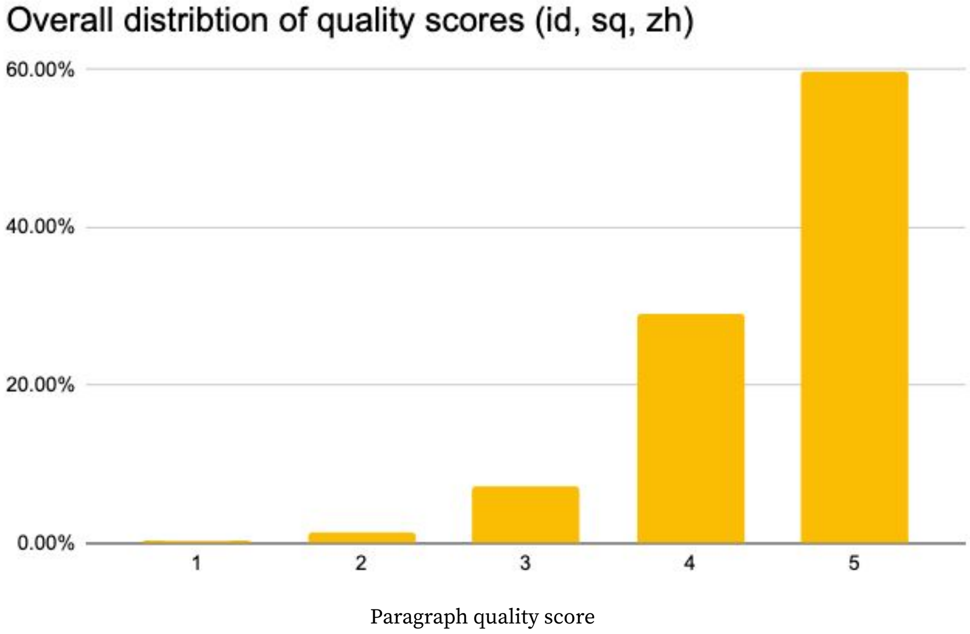


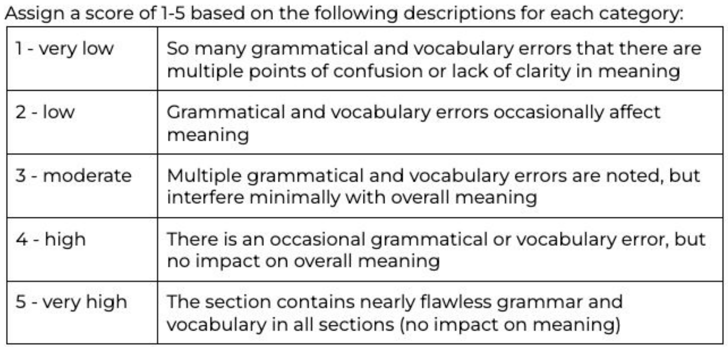
Assign a score of 1-5 based on the following descriptions for each category:

1 - very low	So many grammatical and vocabulary errors that there are multiple points of confusion or lack of clarity in meaning
2 - low	Grammatical and vocabulary errors occasionally affect meaning
3 - moderate	Multiple grammatical and vocabulary errors are noted, but interfere minimally with overall meaning
4 - high	There is an occasional grammatical or vocabulary error, but no impact on overall meaning
5 - very high	The section contains nearly flawless grammar and vocabulary in all sections (no impact on meaning)

Overall distribution of quality scores

Combined languages, overall skewed high for the sample set





Assign a score of 1-5 based on the following descriptions for each category:

1 - very low	So many grammatical and vocabulary errors that there are multiple points of confusion or lack of clarity in meaning
2 - low	Grammatical and vocabulary errors occasionally affect meaning
3 - moderate	Multiple grammatical and vocabulary errors are noted, but interfere minimally with overall meaning
4 - high	There is an occasional grammatical or vocabulary error, but no impact on overall meaning
5 - very high	The section contains nearly flawless grammar and vocabulary in all sections (no impact on meaning)

Quality scores results

General observations

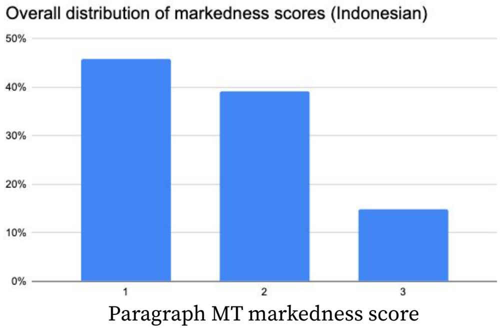
- Overall, quality scores were high for sampled paragraphs from the randomly sampled articles
- The Indonesian sample contained more instances of ‘high’ quality ratings, whereas the majority of Albanian and Chinese samples were rated as ‘very high’
- While very few observations were made of ‘poor’ or ‘very poor’ quality, a meaningful number of observations were observed for the ‘moderate’ and ‘high’ quality categories

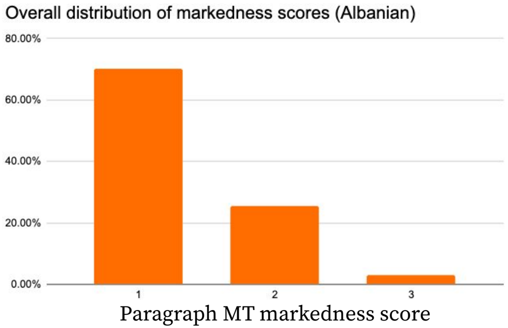
Survivor bias and critiques of deletion ratio

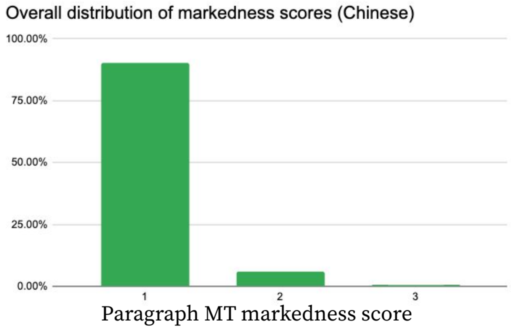
- We may be observing a survivor bias as presumably any articles with paragraphs rated ‘very low’ quality would have been deleted shortly after creation, and therefore not represented in our sample
- At the same time, a critique of the deletion ratio as a proxy for article creation quality is sometimes that admins cannot keep up with reviewing and deleting articles. For the wikis sampled, we do not observe evidence that this is the case. (We would expect more observations of low quality)

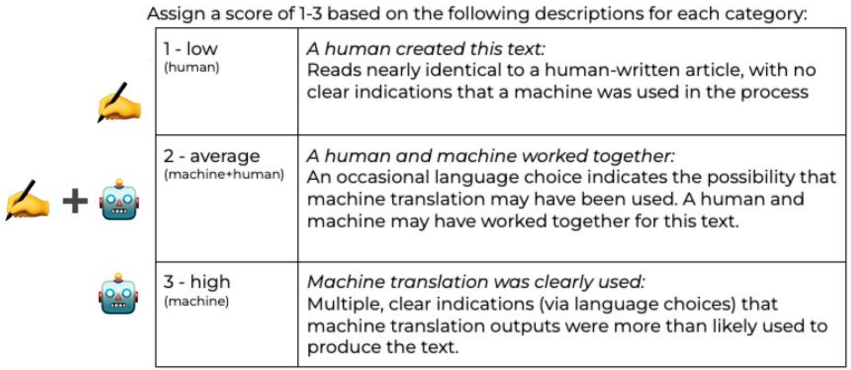
Overall distribution of MT markedness scores

MT markedness was highest for Indonesian, followed by Albanian and Chinese, respectively







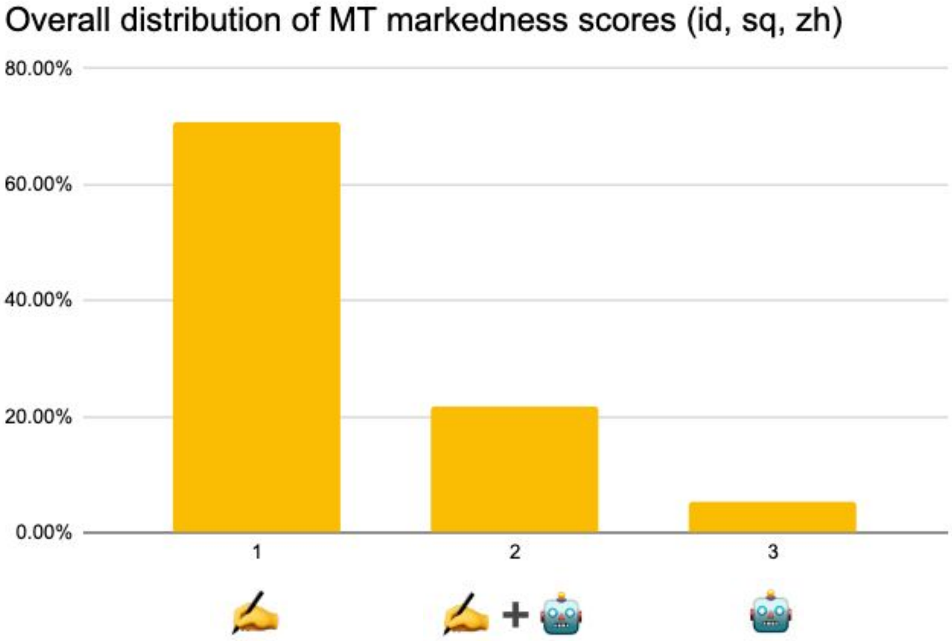


Assign a score of 1-3 based on the following descriptions for each category:

✍️	1 - low (human)	*A human created this text:* Reads nearly identical to a human-written article, with no clear indications that a machine was used in the process
✍️ + 🤖	2 - average (machine+human)	*A human and machine worked together:* An occasional language choice indicates the possibility that machine translation may have been used. A human and machine may have worked together for this text.
🤖	3 - high (machine)	*Machine translation was clearly used:* Multiple, clear indications (via language choices) that machine translation outputs were more than likely used to produce the text.

Overall distribution of MT markedness scores

Combined languages



Paragraph MT markedness score

MT markedness scores results

General observations

- **Overall, the majority of samples reviewed were rated as ‘created by a human’.**
- However, a smaller but meaningful number were noted as containing cues that MT outputs were used. That is, the presence of MT outputs in CX publications is notable by readers.
- Unsurprisingly, markedness scores were negatively associated with language quality (i.e., higher MT markedness was associated with lower language quality)

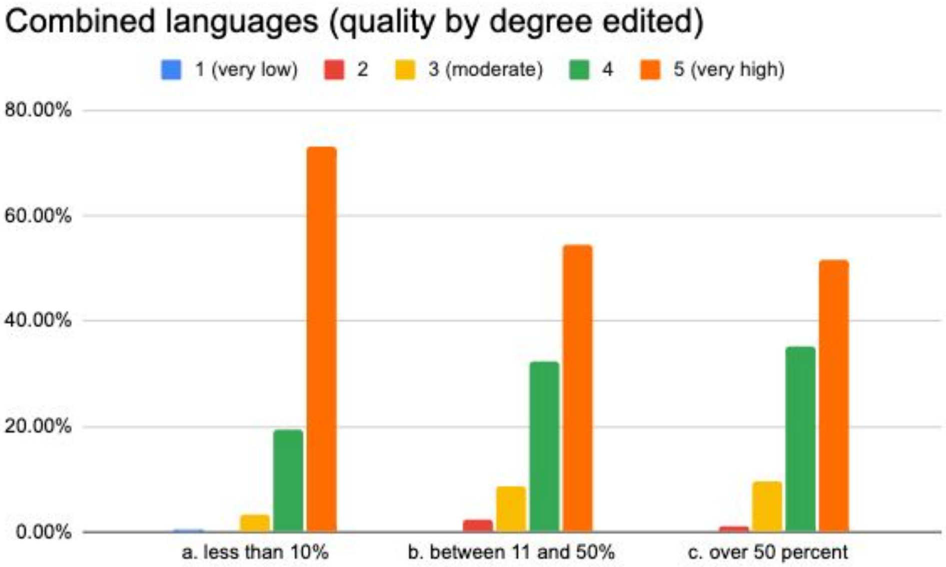
Wiki/language variation

- The sampled Indonesian articles notably contained more instances of language indicative of MT outputs (~40%), especially compared to Albanian (~20%) and Chinese (~5%).

For full details of exploratory analyses, please reference this sheet. Thanks to collaborator Kai Zhu for his time and assistance.

MT output degree of editing and quality

Do articles with increased editing of MT outputs correlate with paragraphs more likely to be rated high quality?



Percent of MT outputs edited (at article level)

High quality paragraphs are observed across all ‘degree edited’ categories, including the lowest ‘less than 10%’ category.

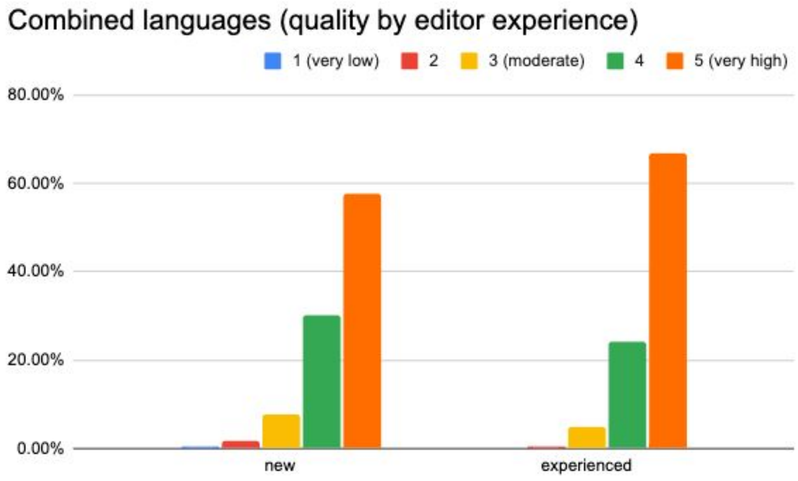
Distribution of quality scores is nearly identical for the ‘11-50%’ and ‘over 50%’ categories.

Yet, overall we find a significant association between degree of human intervention/editing and higher quality scores.

- Note that quality was assessed at the level of the paragraph, but ‘degree edited’ is calculated at the article level. Thus, the specific question we can address is, ‘Is a paragraph sampled from an article that is overall more edited by a human more likely to be rated as high quality?’ This note about metadata coming from the article level applies to the following slides.
- Blue boxes indicate the comparisons we may have enough data to make. Anything outside of them represents figures for which there’s more limited data.
- For full details of exploratory analyses, please reference this sheet. Thanks to collaborator Kai Zhu for his time and assistance.

Editor experience and quality

Do articles published by users who have edited Wikipedia for a longer time correlate with paragraphs more likely to be rated high quality?

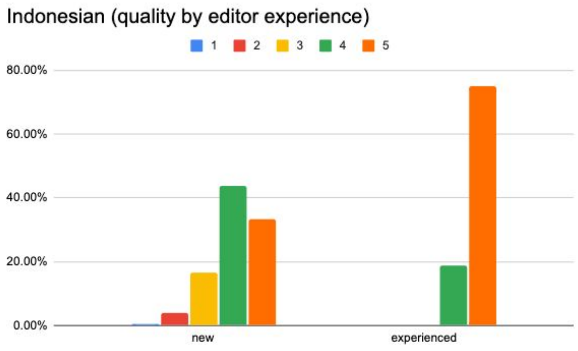


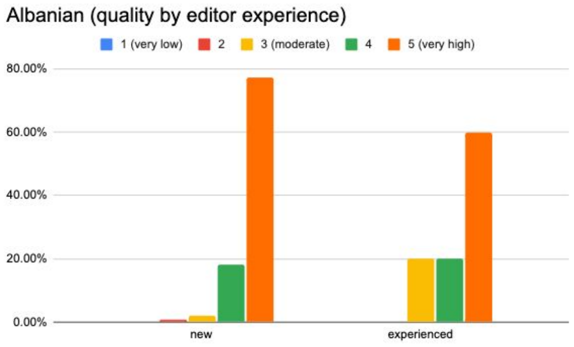
Both new and experienced editors are observed producing generally 'high' quality language articles with Content Translation.

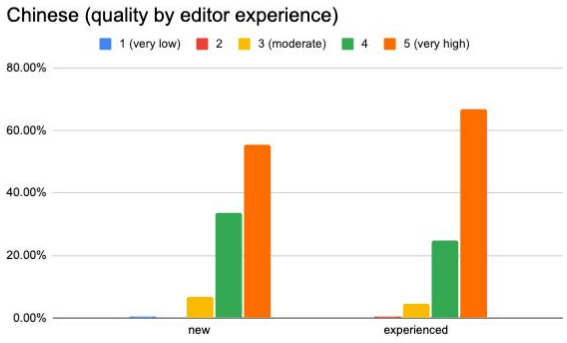
- 'New' is defined as a global registration date of less than 2 years. 'Experienced' includes editors whose global registration date is 2+ years past.
- At least for combined languages, running this same analysis, but looking instead at time since first translation (with a 2 year cutoff for determining junior vs. senior shows a very similar pattern to what is presented here (which is based on time since global registration).

Editor experience and quality - *by wiki*

When looking wiki-by-wiki, we observe more high quality language among experienced Indonesian editors. Chinese is a similar, but less pronounced pattern. Albanian data shows paragraphs produced by new editors to more commonly be rated as high quality.







Editor experience and quality

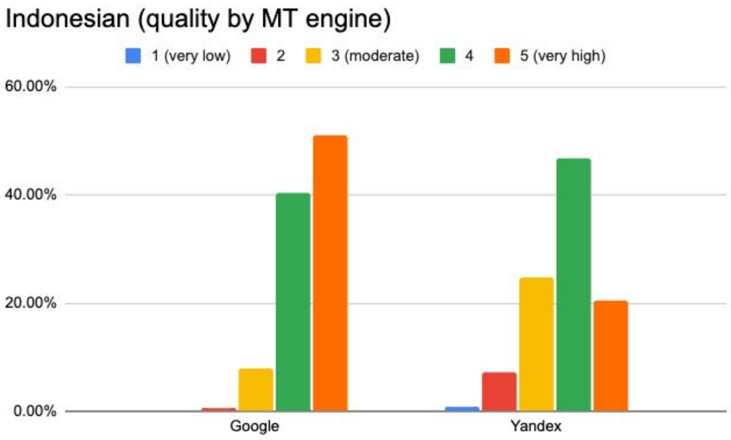
Time since registration vs. Time since first translation

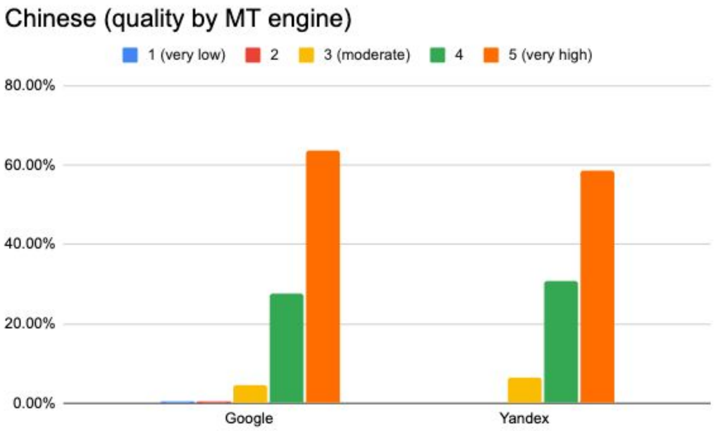
- **There is a significant association between editor experience with CX (“time since first translation”) and higher language quality**
- There is **not** a significant association between higher language quality and overall experience on Wikipedia (“time since registration”)
- **In other words, more experience translating articles is a good predictor of quality; overall editing experience (non-translation) is not.**

For full details of exploratory analyses, please reference this sheet. Thanks to collaborator Kai Zhu for his time and assistance.

MT engine and quality

Does the use of certain MT engines correlate with paragraphs more likely to be rated high quality?





Indonesian paragraphs produced with Google Translate are more likely to be high quality than those produced with Yandex. The difference is less extreme for Chinese.

All Albanian data sampled was produced exclusively with Google Translate so comparisons are not available.

MT engine and quality

Use of Google Translate is associated with higher quality language

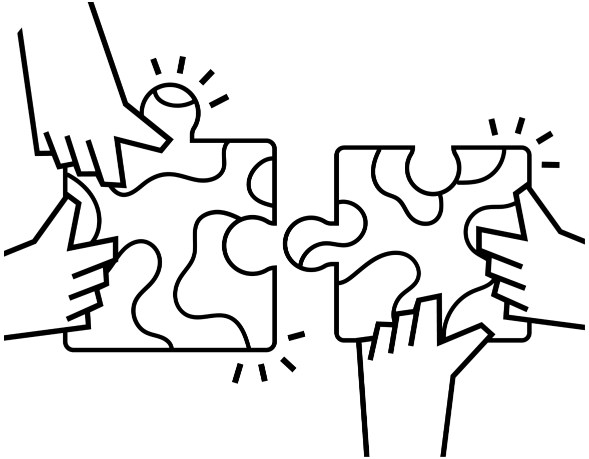
- On average, language quality of sections created with Google Translate is higher than with Yandex
- *Note:* When we say 'language quality of sections' here, we refer to the final published (human-edited) version (*not* the unedited MT output)

Section length and quality

Does section length correlate with quality?

- *Yes* - The longer the section, the lower the language quality score

Recommendations



WIKIMEDIA
FOUNDATION

1. A review support system

Supporting new and/or struggling translators, while building trust of other editors/reviewers

Context:

- Based on this work (which included careful sampling techniques), we don't observe much evidence for very low quality translations that aren't getting appropriately deleted.
- We know that a certain number of translations will be deleted by reviewers. This is variable by wiki, and not unlike how a certain number of new articles will get deleted (a naturally occurring survival rate).
- We've also validated that relative experience with Content Translation correlates with higher quality article language quality.

1. A review support system

Supporting new and/or struggling translators, while building trust of other editors/reviewers

***Recommendation:* Introduce a *translation review support system* that can automatically and/or allow reviewers to manually flag translations for further review and improvement.**

- The goal of such a system is to support and increase confidence for new translators, and provide assistance for good faith translators who are not yet meeting wiki quality standards. It would also provide an alternative to more rigid indiscriminate blocks to the tools, while reassuring the community that translation quality is taken seriously
- *Option A:* Publish into some kind of draft/review space the potentially problematic translations (detected automatically or by the author)
- *Option B:* Publish into main namespace (flagging the potentially problematic subset to reviewers), and making it easier for reviewers to move to a draft/review space (easier decision than keep vs. delete). Translators would then have the ability to further review and improve, helping that investment of time and effort translating survive and reach readers.

2. Support for existing limits override system

Validation that this mechanism is needed

Context: Although we've validated that human intervention for machine translation outputs has a positive association with language quality, we also know that not all edits have an equal impact on quality. The current CX limits system treats all edits as equal in the pursuit of achieving high quality.

Recommendation: Continue to ensure that editors are able to override any limits system alerts, particularly on the level of the paragraph. Consider interactions between a review support system and excessive limits system overrides. For example, auto flagging could take into account multiple points of data; for example, new translators overriding more than a certain number of intervention limits could trigger a peer review.

3. Build around translator, not editor, experience

Experience translating is what matters for quality

Context: We observe a clear association between relative level of experience using CX and language quality, but we don’t observe the same correlation between general editing experience and high language quality. It appears that experience translating may be a better predictor of quality than general experience editing Wikipedia.

3. Build around translator, not editor, experience

Experience translating is what matters for quality

Recommendations: Avoid restrictions on CX that are based solely on general editing experience. Promote alternative mechanisms to avoid low quality translations. Some options:

- Peer review support system (previously mentioned)
- Provide user facing estimations of how likely a translation is to be deleted based on features such as length, degree modified, etc.
- Tailor recommendations based on translator experience (setting them up for success by prioritizing suggestions with characteristics making the translation task easier)
    - Encourage the translation of sections over articles for newcomers
    - Limit the number of articles that can be started if previous translations were deleted - as a stronger motivation to expand sections as a starting point
- Warn users selecting a long piece of content to translate that a short high quality translation is better than a longer one that is more likely to get deleted

4. An additional input for calibration of default MT engines for wikis

Context: We have preliminary evidence that certain MT engines can be associated more or less with higher or lower quality language in published translations. We also know that certain MT engines may also have positive or negative associations with higher/lower deletion ratios (which have been used as a way of approximating quality on large scale). [1]

Recommendation: In the absence of a clear way of easily and automatically measuring language quality in published translations, we can leverage the deletion ratios associated with certain MT engines to provide another input into calibrating MT engine defaults.

- For example, there was an overall positive correlation between language quality and Google Translate for Indonesian and Chinese (vs. Yandex). This should be a factor in determining MT engine defaults for the wikis (in combination with other signals, such as community feedback, etc…

5. Tailoring the translation experience around parameters known to associate with quality

Context: We're beginning to learn of new associations between certain article/section characteristics and language quality. *For example, we find that longer passages/paragraphs are less likely to be associated with high quality language than shorter ones.*

- *Note:* A related open-ended question is to what degree length affects a paragraphs chances of making it into a translation (since we know that translators can choose to omit parts of source articles - we just don't fully know on what basis they're making those decisions)

Recommendation: Increasingly tailor the overall experience on factors, such as length, which we learn to have associations with higher or lower quality outputs. Some examples may include:

- Prioritize the suggestion of shorter sections and/or articles with fewer sections for newer translators
- Potentially weight longer sections differently in the MT abuse calculation (needs further input)
- Break down longer passages in the CX workflow to promote more manageable tasks with higher quality results (the system design could serve up shorter passages, thereby increasing the chances of inclusion and high quality)

Limitations

- Quality analysis was an opportunity that presented itself in this project
- A more comprehensive quality evaluation of Content Translation publications should include:
  - A greater number of raters
  - A quality rating for both the MT output and published text version
  - If possible, a sample that includes deleted articles
- While we were able to use this sample and process to generate a potential pool of articles we can consider vetted for high quality standards, the process did not result in a similar pool for low quality articles

Sample articles for future testing?

(Indonesian & Albanian as test case)

High quality article pool

Examples of articles we definitely wouldn’t want to block publication of via any quality control system (n=50)

⭐⭐⭐ ‘Very high quality, created by humans’ (n=8)

(Articles with all paragraphs rated as 5 (very high quality) and 1 (human-created))

639553 (SQ), 662912 (SQ), 925602 (SQ), 949375 (SQ), 963826 (SQ), 970322 (SQ), 1010240 (SQ), 1054967 (SQ)

⭐⭐ ‘High quality, created by humans’ (n=5)

(Articles with all paragraphs rated as 4-5 (high-very high quality) and 1 (human created))

91237 (ID), 971962 (ID), 983436 (ID), 812091 (SQ), 971343 (SQ)

⭐ ‘High quality, created by humans working with machines’ (n=37)

(Articles with all paragraphs rated as 4-5 (high-very high quality) and 1-2 (human created or human-machine mix))

222962 (ID), 225463 (ID), 352407 (ID), 405825 (ID), 636908 (ID), 729932 (ID), 774781 (ID), 843893 (ID), 940809 (ID), 969460 (ID), 1046534 (ID), 1148471 (ID), 1374520 (ID), 636781 (SQ), 849051 (SQ), 875223 (SQ), 876316 (SQ), 922698 (SQ), 963888 (SQ), 964495 (SQ), 966932 (SQ), 1019396 (SQ), 1027714 (SQ), 1139900 (SQ), 1186662 (SQ), 1197098 (SQ), 1211691 (SQ), 1295984 (SQ), 1307533 (SQ), 1308059 (SQ), 1309013 (SQ), 1311779 (SQ), 1387812 (SQ), 1387942 (SQ), 1396806 (SQ), 1440221 (SQ), 1441194 (SQ)

Low quality article pool

Examples of articles we'd possibly want a quality control system to flag for further editing (n=0)

🚩🚩🚩 Articles with all paragraphs rated as 1 (very low quality) and 3 (machine-created)

None observed to fall in this category

🚩🚩 Articles with all paragraphs rated as 1-2 (low-very low quality) and 3 (machine-created)

None observed to fall in this category

🚩 Articles with all paragraphs rated as 1-2 (low-very low quality) and 2-3 (machine-created or human-machine mix)

None observed to fall in this category

- No articles we sampled and reviewed met these criteria.
- However, individual paragraphs were observed that met these criteria.
- **In other words, low quality language indicative of machine translation outputs used to produce it, when observed, is most commonly spread across various articles.**

Thank you!

Questions & comments?

Thanks to…
Megan Neisler and Mikhail Popov for analytics assistance
Kai Zhu for assistance with exploratory statistical analysis
Our language experts for careful evaluation
As always, the Language Team for their input

WIKIMEDIA
FOUNDATION

Contact Eli
eli@wikimedia.org