23 December 2019

DQF BI Bulletin - Q4 2019



When it comes to metrics in translation, there are several levels to approach the subject. The first is the high-level approach that in the end there is just a body of text that needs to be translated within a certain time frame. Productivity rate is an essential metric, but it alone does not tell much about the actual job. To analyze the workload for the translator or the post-editor, metrics like edit density give a lot more insight.

But the work of translation deserves a much richer analysis, one that reaches beyond the number of words or edits, and specifies whether a translation accurately conveys the message of the source, in the right terminology and style, and according to the rules of the language. This is how we can understand the quality of a translation, and how resources such as MT and TM can be improved. The way to get there is to familiarize oneself with the quality review results, and the next step is to understand how TM and MT impact the quality of the translation.

Review Data Gaining Importance

DQF as a framework has many different aspects. Over the last year, this bulletin covered many productivity-related statistics. However, DQF is also well known for lending its name to the DQF-MQM error typology that reviewers can use for the quality evaluation of translations. Quality review has been a part of the DQF API since the start. Interestingly, though each DQF project can be measured in terms of productivity and edit density, the pool of reviewed projects is much smaller, as only a small portion of the translation projects actually gets reviewed using error annotation.

However, things changed in 2019. Whereas in 2018 only a very small portion of the projects were actually reviewed using the DQF-MQM scoring method, in 2019, the quantity of data has grown to proportions that justifies a presentation in this BI Bulletin. The growth shows in the number of reviewed words, but foremost in the number of projects, in which the growth is much more stable.

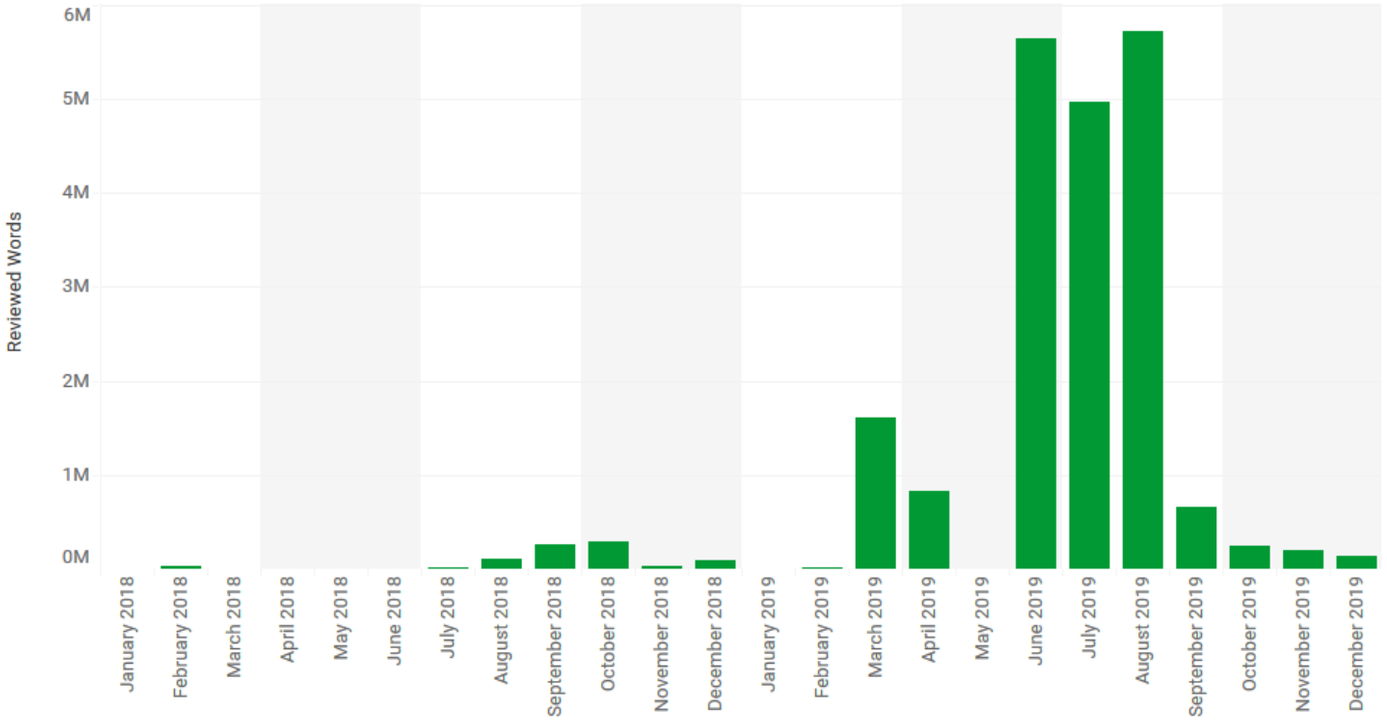


Figure 1: Number of words reviewed in DQF

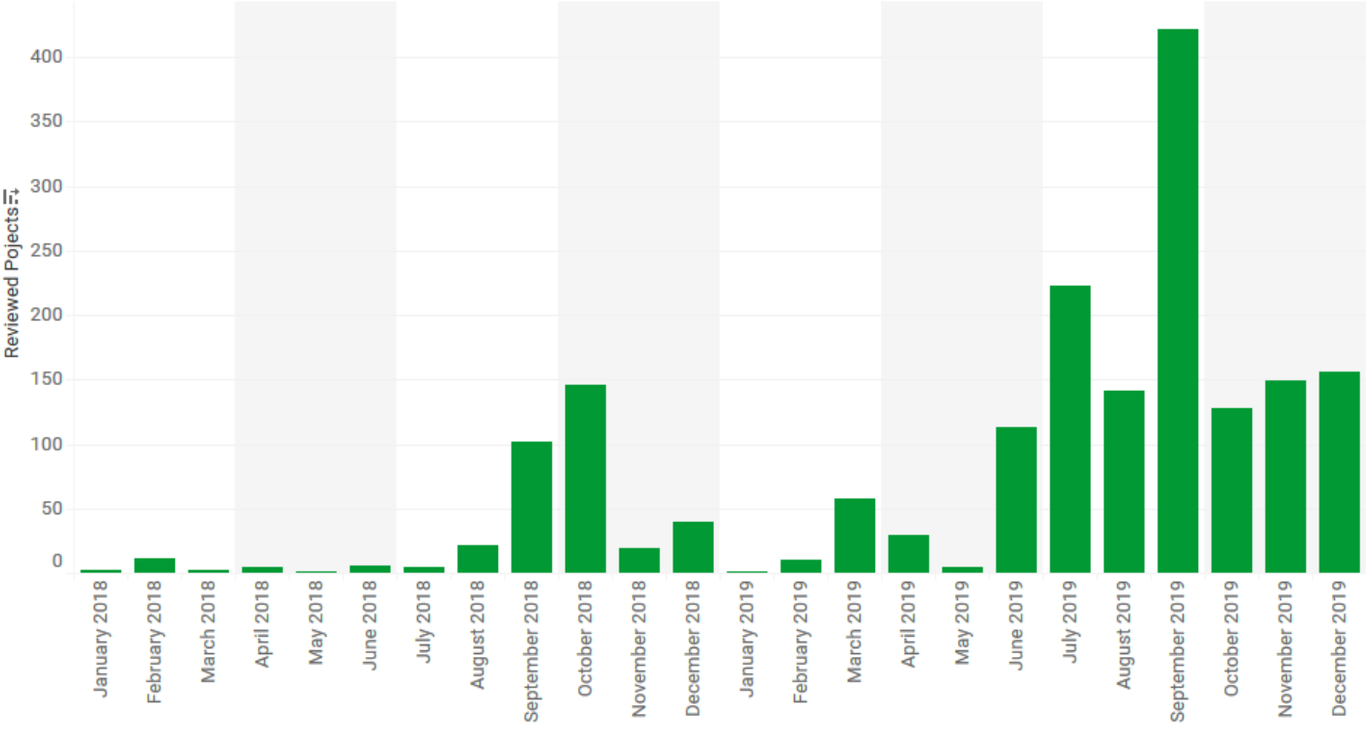


Figure 2: Number of reviewed projects in DQF

The number of target languages for which there is review data is large: around 80 different languages. In most of these languages, several thousand words have been reviewed. But one language stands out. As we reported in the previous edition of the BI Bulletin, the target language with the largest volume in DQF is Hindi. Hindi is also by far the most prominent target language for reviews.

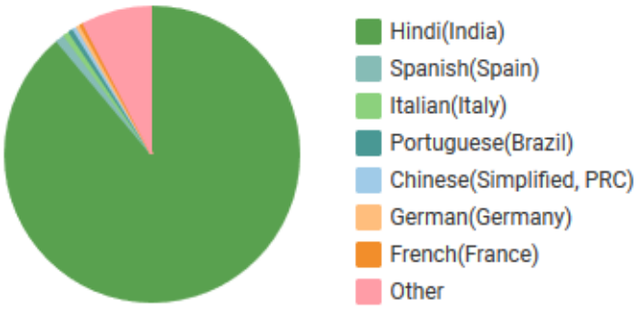


Figure 3: Target languages with largest amounts of reviewed words

Looking for the New Normal

Quality evaluation in DQF means that a project for which a translator has already provided a translation, is reviewed. A reviewer marks errors in the translation by determining the error type and the severity of the error, which can be either neutral, minor, major or critical.

As this is the first statistical analysis of DQF review projects on a big scale, let's first have a look at some of the basic numbers. For this bulletin, we took the reviews over the last two years, starting in January 2018. In this period over 30,000 translation errors have been reported. The majority of these errors were assessed as minor errors. Only 900 critical errors were found, that totals up to 3% of the errors.

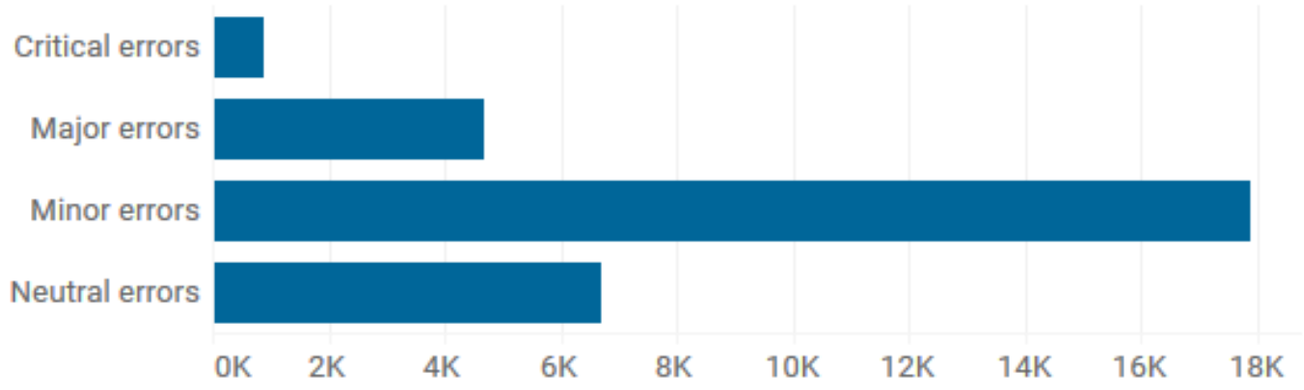


Figure 4: Distribution of errors per severity level

For the analysis of the error type, we only looked at the eight main error dimensions, and rolled up all the errors from the subcategories into these main categories. Most errors are identified as accuracy errors or fluency errors.

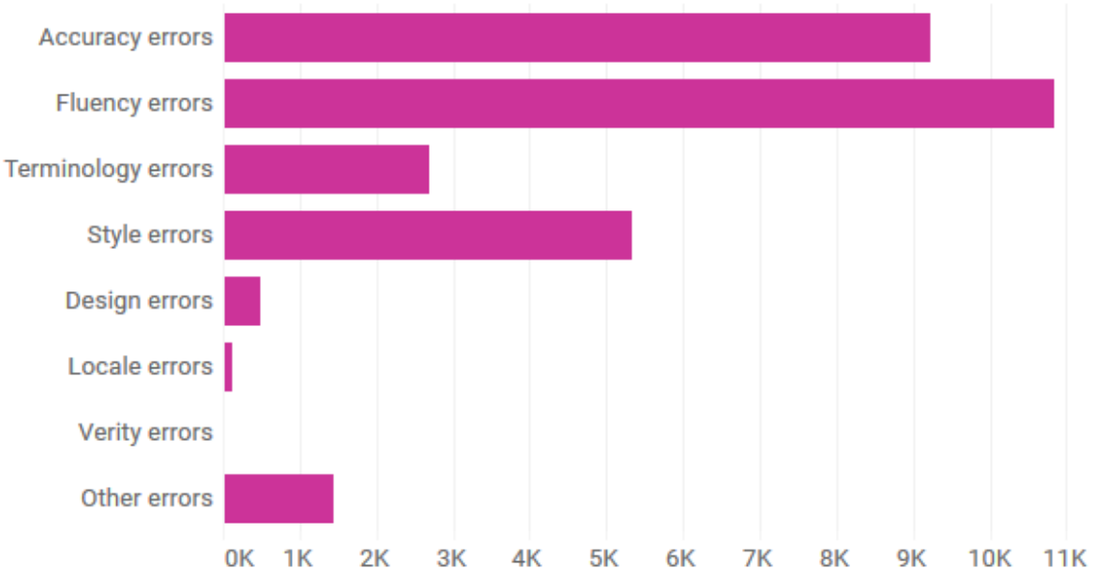


Figure 5: Distribution of errors per dimension

With the growth of review data, it is to be expected that certain common patterns will be established over a large number of individual reviews. Even though the estimation of the severity of an error is based on a common definition, a single estimation in itself is a subjective judgment. But on a large scale over many translations and many reviewers, a pattern emerges, and a 'new normal' settles. TAUS DQF Dashboard uses error density, and weighted error density to estimate the relative quality of a reviewed translation. For error density, the number of errors, regardless of severity, are counted, together with the number of reviewed source words. From these numbers the average number of errors per 1000 words can be calculated.

The weighted error density is different as it takes the severity of the error into account. Each severity is assigned a certain value, so that, for example, a major error puts more weight in the calculation than a minor error. After multiplying the errors by the value assigned to their severity, the total sum of these weighted errors is divided by the total number of words that were reviewed. This is multiplied by 1000 to produce the number of penalty points per 1000 words.DQF-MQM does not impose these multipliers and leaves it up to the user to choose their own severity values. For this analysis the 0-1-5-25 model is used, that is default for DQF-MQM: we don't count neutral errors, we count minor errors just once, major errors are multiplied by 5 and critical errors are multiplied by 25.

With all this in place, what does that mean for the average reviewed project? What is a good score, and when is it bad? It is possible to calculate the error density for each reviewed project, and group these projects according to the error density score? This shows which scores occur the most. The same process is applied to the weighted error density. It turned out that projects with an error density score close to 0 occur the most: almost 30% of the projects fall under this group. These projects have less than 1 error per 2000 words. The error density for projects is a typical long-tail distribution, in which a small percentage of projects have an extremely high error density.

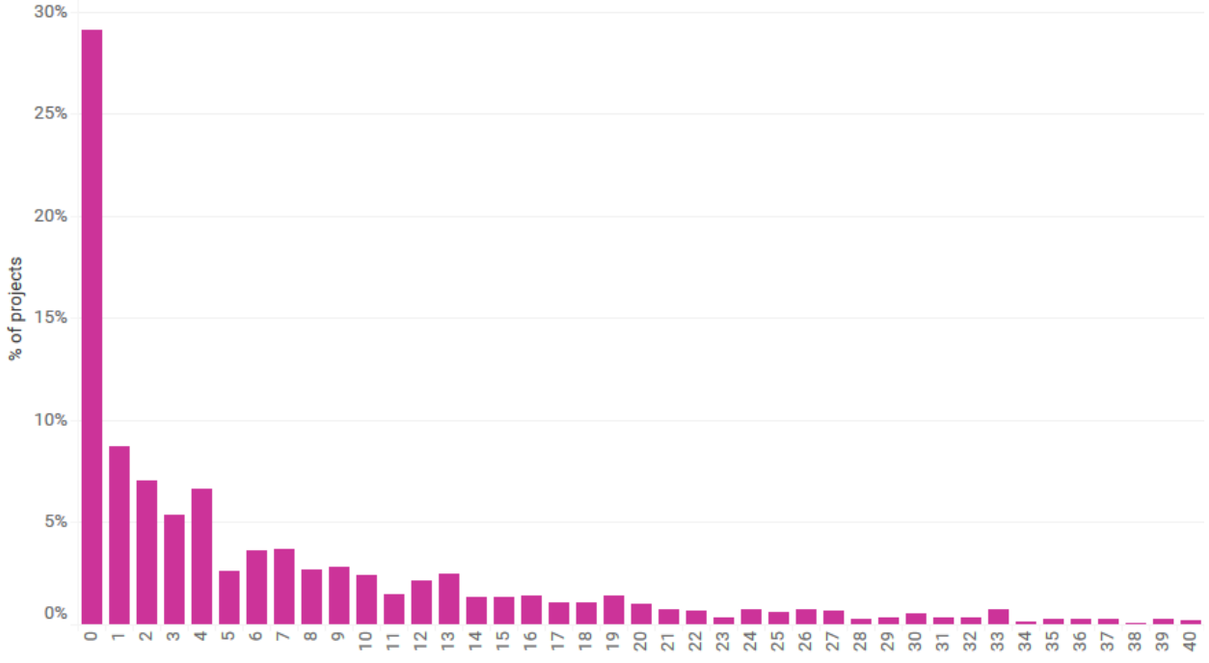


Figure 6: Error density scores per project (cut off for all error density values above 40)

When using the weighted error density, it is possible to set a threshold that separates failed projects from projects that passed. Similar to the severity multipliers, there is no fixed value for this threshold, but for the sake of this Bulletin, let's set the pass/fail threshold for all projects at 25, which means that one critical error per 1000 words is enough to make the project fail. Here we also see a long-tail distribution. But counting all projects that have a weighted error density above the pass/fail threshold, we see that 9% of the reviewed projects did not pass.

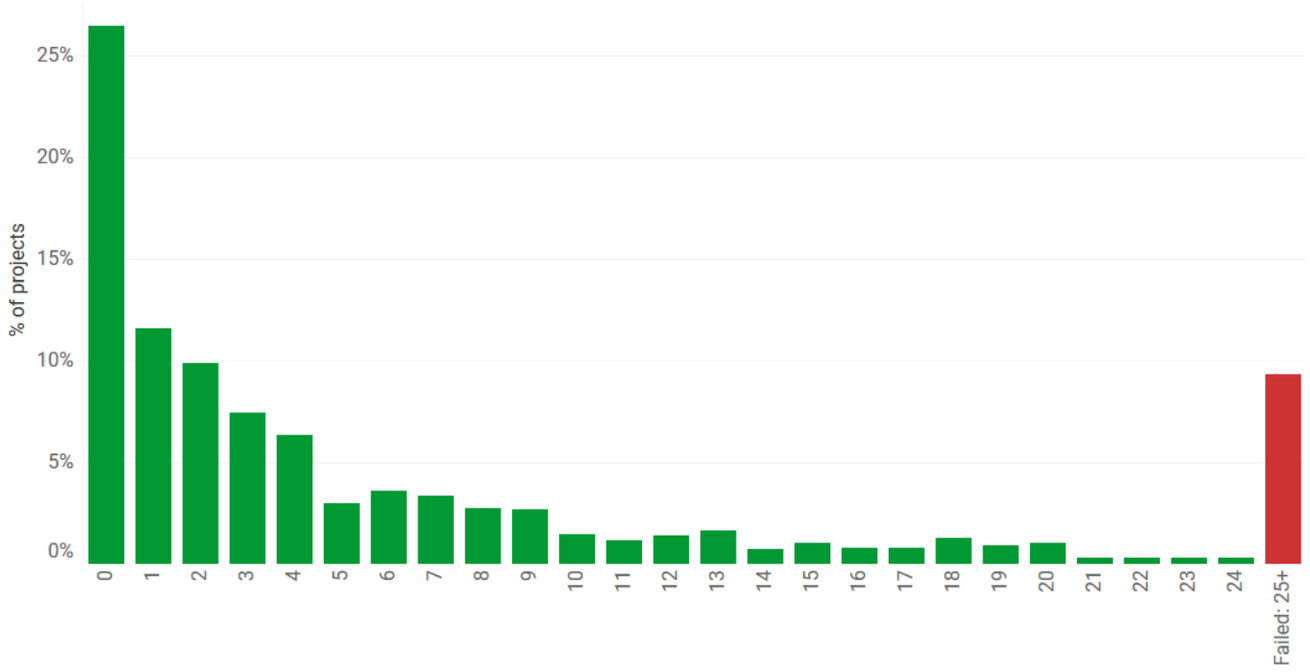


Figure 7: Weighted error density scores per project. Project with a weighted error density higher or equal to 25 count as failed

The Origin of the Errors

Previous bulletins reported that, with regards to productivity and edit density, the origin of the pretranslation makes quite a difference. Whether a translator receives a suggestion from MT or from TM impacts the time that is needed per edit. What influence does segment origin have on the errors in translations? The weighted error density can be calculated for each segment origin separately. Translations that have TM as segment origin have a quite low weighted error density on average, whereas translations with suggestions from MT have a high weighted error density, and therefore are lower in quality. Human translated segments that were translated without direct assistance from either TM or MT have an average weighted error density that is closer to that of segments with TM origin, but still have a relatively higher error density. It appears that TM pretranslations improve quality of the translation.

In this short analysis, it is difficult to explain the difference between MT and TM. It is of course possible that segments with MT origin have followed a different workflow. It might be the case that for MT reviews are more often being performed on raw output, making it entirely different from translated segments originating from TM.

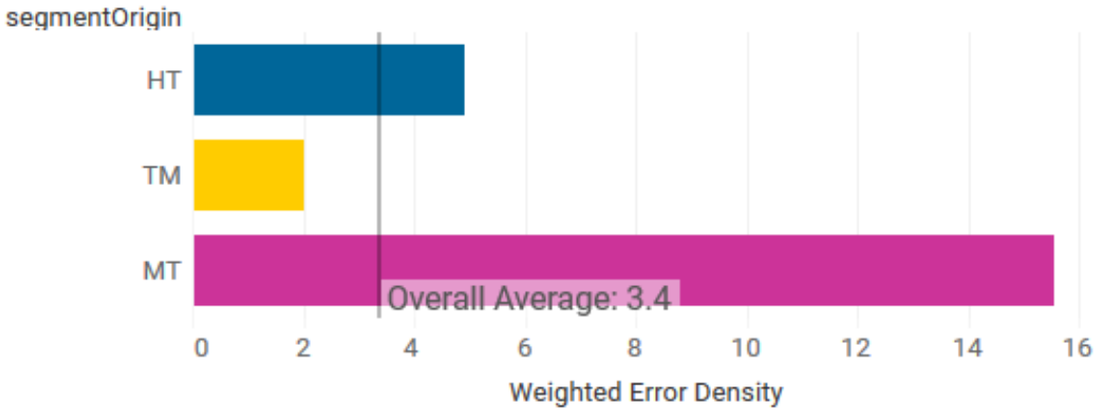


Figure 8: Average weighted error density per segment origin

Remarkable are the correlations between error type and segment origin. In figure 5, we have seen the overall distribution of the error types. However, this distribution is very different for each of the segment origins.

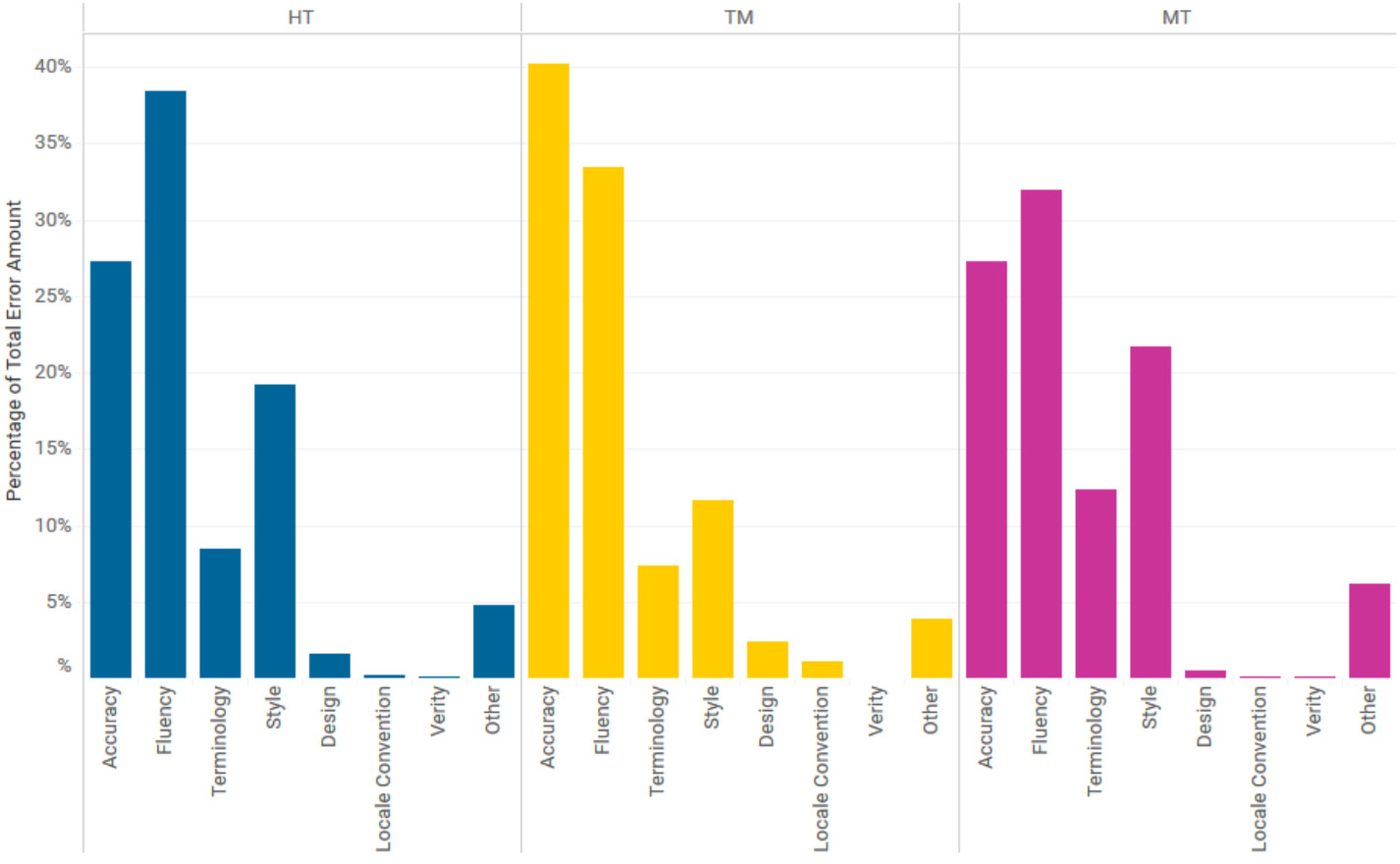


Figure 9: Distribution of error types in each segment origin

Compared to errors in segments with MT or HT origin, errors in segments with TM origin are more often about accuracy, and less often about style and terminology. Errors in segments with MT origin are more often about terminology and style of the translation. This seems to confirm the intuition that TM in general has better local knowledge. As TM is built and maintained within a specific organization, it includes more information about the organization's style and terminology. As such, it even outperforms human translation without any pretranslation. MT, unless it is trained, lacks the local knowledge and this might be the reason for a relatively high error density in the terminology and style error dimensions.

A relatively high error density in TM accuracy also can be explained: pretranslations from TM with a high match rate usually have a good fluency, terminology and style, but their source segment differs slightly, and therefore they aren't an accurate translation. If working with suggestions from TM implies changing the pretranslation to reflect the meaning of the new translation, it is no surprise that relatively more accuracy errors (additions, omissions, mistranslations, over- and under-translations) might occur.

Learn from Errors

Error review gives you a lot of insight in the quality of your translation process. DQF-MQM gives the tools for scoring and categorizing the issues found in the translated content. While it provides comprehensive guidelines of how you should annotate errors, it also gives a lot of freedom for the final score interpretation. This means that, once you have a score calculation, interpreting the results is a practice that has to develop. This BI Bulletin aims towards raising this score awareness, by presenting the results collected over the past two years.

Learning to recognize quality is a process that takes time. The first step is a regular review of your translations, and next, the analysis of the results by comparison. This will enable you to analyze the strong and weak points of your resources. It can help you decide whether or not to define your terminology and style more clearly. It informs you about the quality of your TM or your MT engine. But most of all, review gives you the security where you stand when it comes to quality. In the end, localization is about trust that you translate a message with the least possible distortion, accurate and fluent. Using a standard review framework gives you the tools to compare and communicate your quality.

