

Sampling issues: 11 reasons for non-uniformity of error distribution

Created in ASTM WK46396 working group by Serge Gladkoff, Merle Tenney, Sue Ellen Wright and Arle Lommel. Version 5/8/2020.

For any statistical approach to be applicable, it is critical to know what statistical distribution is valid before choosing the right distribution for further statistical analysis. The most widely used option is the normal distribution. Application of the normal distribution would be based on the notion that (a) all errors are independent and (b) the probability of errors in the text is uniform.

Both assumptions are in fact not true with respect to texts (and translation products) in general.

Practitioners in the language and translation industry know that translation errors are not, in general, uniformly distributed in content and, what is more, over time. Furthermore, their significance and “weights” are also different in various parts of the material and/or types of material, and also vary according to other sometimes unpredictable factors. Also, different types of errors may depend on each other. Indeed, some errors only occur when triggered by other fundamental errors. The following factors can influence why error distribution in text is not, in fact, uniform:

1. Errors exhibit different probability of appearing in various parts of the material that were translated by different people.
2. Errors exhibit different probability of appearing in parts of different kinds of texts, depending on degrees of abstraction, register, domain-specific style, etc., which can make some texts more or less difficult to translate. For these reasons any **representative sample**¹ needs to include all parts of the text, or at least the most important parts of the text, such as headers and titles, the abstract, and concluding paragraphs, etc. The challenge in these cases is that **biased sampling**² may be the best way to choose the smallest representative sample.
3. Errors exhibit different probability of appearing in different materials created or processed by even the same translator or reviewer working at different times (due to the human factor of instability of work, or varying qualifications depending on the topic).
4. Errors exhibit different probability of appearing in different materials even for the output of NMT engines, because the quality of machine output is also not stable and depends on the current state of the system. Because developers are always working on the engines, it is not unusual that the Google Translate engine, for example (as quoted by some MT deployment partners), may unexpectedly produce unacceptable

¹ Representative sampling – a method of selecting smaller parts of text for quality evaluation and using the evaluation of the partial segment to extrapolate an assumed evaluation projection for the whole text or translation product. [We need to talk about this – We need to achieve consensus on the ideal wording.]

² Biased sampling – a method of non-random sampling where the subset for quality evaluation is taken not at random, but from specific parts of the material in order to select a suitably smaller sample while attempting to ensure a representative evaluation of the whole translation product. [We need to work on this one as well.]

- product, or at least very different product from previous output for even the same identical text.
5. Errors should be assigned different weights depending on the part of the material they appear in, because text is not a uniform media; it has structure and some parts of the whole are more important than others.
 6. Errors themselves may depend on the context, and the same text can be an error in one context and not an error in another. The meaning of words, in fact, depends on the context. For instance, depending on context or the purpose of the text, a flowering plant could be referred to as a weed or a flower. Conceptual semantics is very context-dependent, and the semantics of a given section may exceed the scope of a given sample.
 7. Errors may not be uniformly distributed, because when translation updates insert new content into already translated content. inconsistencies can arise.
 8. Errors may not be uniformly distributed in translation product because the source text is not uniform in quality. (It may have been authored by different individuals or harmonization of disparate sections may have been inadequate.)
 9. Errors may not be uniformly distributed because different types of errors are interdependent, and consequently tend to cluster. Generally speaking, the frequent tacit assumption that errors of various types are independent³ is not true.
 10. Due to the progress of technology, errors can exhibit different probability of appearing in parts of content of different age. For example, if the Help System was created for version 1.0 and version 2.0 of a computer application, and it is now being updated for product version 3.0, new content may have matches with both 1.0 and 2.0 translation units, but “100% matches” from version 1.0 may have become obsolete due to the changes in technology or changes in terminology, and consequently, updates may tend to have more errors in units from version 1.0 than is the case for later content. These kinds of changes can be interlaced in the text in such a way that it can be very difficult for translators and reviewers to recognize the differences.
 11. Errors may not be uniformly distributed in the translation product because parts of the source text were not actually originally authored in the source language itself and represent some sort of inadequate translation from a third language.

Consequently, for all these reasons, representative sampling is a highly unreliable undertaking, and there are fundamental reasons why it is frequently desirable to check the whole text rather than do “spot checks”, because although extrapolating “spot-check” results to the entire material may sometimes be valid and may seem temptingly expedient, in many cases it may not be. In any case, it is only possible to apply statistical models/distributions when certain assumptions can be reliably viewed as valid.

³ <https://www.statisticshowto.com/probability-and-statistics/dependent-events-independent/>