Translation guidelines

These translation guidelines must be acknlowedged by all translators who will be contributing data.

Note! The latest version of these guidelines is available at https://oldi.org/translation-guidelines.pdf.

Important note

Your translations will be used to help train or evaluate machine translation engines. For this reason, this project requires **human translation**.

- If you are translating data to be used for evaluation purposes, such as for FLORES+, using or even referencing machine translation output is not allowed (this includes post-editing).
- If you are translating data to be used for training purposes, such as Seed, the use of post-edited machine translated content is allowed, provided all data is manually verified and edited where necessary.

Caution! Some commercial machine translation or LLM services – including DeepL, Google Translate, ChatGPT, Claude and Gemini – prohibit the use of their output for training other translation or AI models. Their use is not permitted.

General guidelines

1. You will be translating sentences coming from different sources. Please refer to the source document if available.
2. Do not convert any units of measurement. Translate them exactly as noted in the source content.
3. When translating, please maintain the same tone used in the source document. For example, encyclopedic content coming from sources like Wikipedia should be translated using a formal tone.
4. Provide fluent translations without deviating too much from the source structure. Only allow necessary changes.
5. Do not expand or replace information compared to what is present in the source documents. Do not add any explanatory or parenthetical information, definitions, etc.
6. Do not ignore any meaningful text that was present in the source.

Named entities

Named entities are people, places, organisations, etc., that are commonly referred to using a proper noun. This section provides guidance on how to handle named entities. Please review the following guidelines carefully:

1. If there is a commonly used term in the target language for the Named Entity:
   a. If the most commonly used term is the same as in the source language, then keep it as it is.
   b. If the most commonly used term is a translation or a transliteration, then use that.
2. If there is no commonly used term:
   a. If possible, a transliteration of the original term should be used.
   b. If a transliteration would not be commonly understood in the context, and the source term would be more acceptable, you may retain the original term.