RoCS-MT Translation guidelines

These guidelines are included because there are some specific constraints as to how the translations are to be carried out, and some particularities of the dataset to explain. The sentences to be translated are found in the excel spreadsheet in the column "Normalised segment". However, we also provide additional information that can help translation (see below for more information).

Origin of the text
The texts to be translated are from the Reddit online forum (extracted using the API), taken from a range of different subreddits (so of different genres of text, e.g. relationship advice, advice about pets, video gaming strategy, etc.). They were selected due to their non-standard nature (spelling mistakes, abbreviations, lack of punctuation etc.).

Preprocessing of the text
The texts have been manually pseudo-anonymised (usernames and names other than those representing celebrities and other well-known public figures are replaced with new names), split into "sentences" and normalised. It is the normalised versions of the sentences that are to be translated.

The sentences have been filtered to remove offensive or sensitive content (hate speech, taking drugs, suicide, etc.). However, profanities were kept as they were taken to be illustrative of the sociolect of online language. If however, you do not feel comfortable with translating something, please leave it blank and write a comment indicating that you have not translated it.

Additional context provided to help translation
The text is split into short documents with one or several sentences per document. In the excel document, a sentence's document is indicated by the value in the column "Post number", and the cells are also coloured such that it is visually easier to see which sentences belong to the same document (alternating grey and white). A Reddit post is associated with a title and a text with the main content of the post. The documents can contain either the title or a subset of the text or even both. The type of text associated with each sentence is indicated in the column "Text type". Titles are marked in bold to make them visually easier to see. Although the normalised text may be sufficient to carry out the translation, we also give access to the additional information just in case:

- the title of the post
- the entire body of text associated with the post
- the raw version of the sentence (after pseudo-anonymisation and segmentation into sentences)
- some translation notes have been added to provide some context about the posts (e.g. to give an idea of what is the subject of conversation, the meaning of some expressions and abbreviations, etc. in order to make translation easier). Very occasionally there are indications about how to translate (for instance for meta-linguistic questions where people discuss particular words, it is best to keep the English words, e.g. *One word I simply can't say properly is water…* -> *water* should be kept in English in the translation.

Constraints (important)

The dataset will be used to evaluate machine translation systems on their ability to handle non-standard texts. This crucially means that:

- the sentence boundaries that have been defined must not be modified. It is possible to translate a sentence using several sentences if that is what is natural. However, it is not possible to merge several source sentences to produce a single translation of both (i.e. one translation per row).
- translators should not use machine translation systems or other computational systems to aid translation as this could bias the translations to look like translations produced by GoogleTranslate, DeepL, ChatGPT, etc.

More specific guidelines

- There are multiple posts that use slang terms (e.g. gaming or general online slang such as *lol*) and it possible that the correct translation will be an English borrowing. It is fine to use an English borrowing in this case, if this is what is generally used online.
- The punctuation choices should be kept as much as possible, as appropriate for the target language of translation (e.g. conserving full stops, exclamation marks, quotes, etc.).
- As described above, there are some instances of people talking about English words, and in this case, the English words should be kept as is. Another example: *One says "Let's eat granny" making it seem like someone's going to eat their nan. However, the other example says "Lets eat, granny", implying a different meaning to the sentence*. The phrases "*Let's eat granny*" and "*Let's eat, granny*" should be kept in English. These are indicated in the translation notes.
- Use of "non-standard" language:
  - Any spelling mistakes that were in the raw sentence should not be reproduced in the translation (i.e. the normalised version should be used as the source sentence to translate).
  - Formatting, including things like capitalisation, should (for the same reasons) follow the conventions of the normalised translation.
  - Abbreviations, acronyms and simplifications (e.g. in English *wdym* = what do you mean, *bc* = because, *rly* = really, etc.) should be expanded, unless the result would not make a natural sentence that could realistically be found. An example of a non-natural expansion would be *lol* = *laughing out loud*, since this is not practically used.
  - However, abbreviations linked to the names of places (e.g. *USA*, *UK*, *UCL* (=University College London) should be kept as they are if the acronym is also commonly used in the target language. In other cases, the most frequent equivalent translation should be used. (e.g. English *UN* = French *ONU*, English *NATO* = French *OTAN*).

- The overall idea is that the translations should be natural and not contain the types of non standard language that were normalised in the English versions, although they should match as best possible the style and familiarity.

Additional questions

If you have any doubts or questions about the meaning of the sentences, please contact me at rachel.bawden@inria.fr to discuss things further.