Translator Brief

In this project we wish to translate data from several domains for use in evaluation of Machine Translation (MT). The translations produced by you will be compared against the translations produced by a variety of different MT systems. They will be released to the research community to provide a benchmark, or "gold-standard" measure for translation quality. The translation therefore needs to be a high-quality rendering of the source text into the target language, as if it was originally written directly in the target language. However, there are some constraints imposed by the intended usage:

- All translations must be "from scratch", without post-editing from machine translation or usage of CAT tools. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing and will reject translations that are post-edited.
- Translation should preserve the paragraph boundaries but may change number of sentences per paragraph. The source texts contain one paragraph per line and the translations should be the same.
- Translators should avoid inserting parenthetical explanations into the translated text and obviously avoid losing any pieces of information from the source text. We will check the translations for quality and will reject translations that contain errors.
- If the original data contain errors, typos, or other problems, do not change the source sentences, instead try to prepare correct translation as if the error wouldn't be in the source.
- The data contain several domains, each folder containing one domain source.

The source files will be delivered as text files (sometimes known as "notepad" files), with one paragraph per line. We need the translations to be returned in the same format. The translation file needs to have the same name as the original file.

Speech Domain

The texts are the transcriptions of audio, edited by native speakers. Each file represents one segment of audio (you are also provided with correspondent audio in WAW format). Phrases said by different speakers are located on different lines. Audios correspond to different domains, they differ in formality, style, topics and number of speakers. The idea is to translate using the most similar language in the target language, matching as best as possible the characteristics of the source text.

Social domain

The texts are from the social network Mastodon (similar to Twitter). Each file represents a thread or part of a thread from one or several users. Different posts within a thread are presented on different lines in the file, although individual posts can also span several lines. The sentences have been selected so that they do not contain offensive or sensitive content (hate speech, taking-drugs, suicide, politically sensitive topics, 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 the whole line blank and let us know that you have not translated it.

The texts are particular in that they may contain spelling errors, slang, acronyms, marks of expressivity, etc. The idea is to translate using the most natural language in the target language, matching as best as possible the style and familiarity of the source text.

- Spelling mistakes should not be preserved in their translations, i.e. the translation should be spelt correctly
- Marks of expressivity (e.g. asterisks *wow*, capitals letters WOW) should be conserved as best as possible. However, we recommend not to attempt to reproduce repeated characters (e.g. woooow) in translation, as the choice as to which character to repeat is often arbitrary.
- There will be abbreviations and acronyms (e.g. btw -> by the way, fwiw -> for what it’s worse). These do not need to be translated using abbreviation or acronyms unless an abbreviation/acronym is the best translation choice in the target language.
- Users have been pseudo-anonymised (e.g. @user1, @user2). These should be left as they are, i.e. not translated.
- Platform-specific elements such as hashtags should be translated as hashtags, but the content should be translated as appropriate into the target language.
- Punctuation can be added if it necessary to avoid comprehension difficulties. Otherwise we recommend following the punctuation of the source text.

A file entitled README-social-domain-translation-notes.pdf has been distributed with the texts to translate. This file should not be translated. It contains some notes to provide additional context on the topic and terms used in some of the texts.