Executive summary

Deliverable D5.1.1 reports on the work carried out in **WP5: Translation equivalent selection.** WP5 "*relates to the second phase of the machine translation process. To this end, a translation equivalent selection module will be designed and implemented, which will retrieve for each phrase of the source sentence the best-matching translational equivalent included within the monolingual corpus.*"

WP5 is divided into two tasks, *T5.1: Design and implementation of the translation equivalent selection module* that "*focuses on designing and implementing the second part of the main machine translation process, which consists in defining the best-matching translation patterns within the target language monolingual corpus with the help of semantic relevance of words defined via novel computational intelligence methods.*" and *T5.2: Optimisation of module-specific parameters,* which "*involves a series of similarity weights, whose values need to be optimally set via the use of learning algorithms. Based on the experience of the project partners, metaheuristic optimisation techniques (such as genetic algorithms) are expected to be used, the appropriate one[s] to be defined at the start of the task via a comparative evaluation process*".

In the PRESEMT architecture, the translation process is split into two phases. The first phase, **Structure selection**, determines the overall structure of the target language (TL) sentence that is the translation of the input source language (SL) sentence with the help of the syntactic information contained in a small bilingual corpus. The second phase, **Translation equivalent selection**, handles more fine-grained properties of the target language. In particular it aims at performing the following tasks:

1. **Word translation disambiguation:** Out of the translation alternatives of the SL words, the best translation in the given context has to be chosen.
2. Resolving micro-level **word order** issues
3. Insertion or deletion of **functional words** (such as articles or prepositions)
4. Specifying **additional morphological information** such as case or number that may or may not be present in the source language
5. Generation of **TL tokens** for the lemmas based on the morphological information provided

The Translation equivalent selection module uses for these tasks the information contained in a huge monolingual target language corpus. No bilingual information is used at this stage, apart from the output of the first phase of the translation process. The output of the Translation equivalent selection module is the final translation of the source language sentence.

The search by the module in the monolingual corpus data is guided by several parameters. At first, these parameters will be set manually to approximate values. Later, the parameters will be optimised in an offline optimisation process, which is provided for. The optimised parameters are then used online by the Translation equivalent selection module.

The deliverable has the following structure: First, the role of the Translation Equivalent module in the overall PRESEMT architecture is explained (Section 2). Then the aforementioned tasks of the module are covered: the translation disambiguation (Section 3), the adjustment of word order and insertion or deletion of functional words (Section 4), the incorporation of morphological information (Section 5) and the generation of tokens (Section 6). Finally, there are some remarks as to the status of the implementation and an outlook regarding the work that is to be done in the second year of the project.