Executive summary

The present deliverable falls within Task *T3.3: Development of parallel corpus – Phrase aligner module* (**WP3: Corpus extraction & processing algorithms**), which *"aims at defining the algorithm required to define phrases in sentences in both the source and target languages of a given language pair, these phrases being aligned in the two sentences"*. The work described relates to processing a bilingual corpus with a twofold purpose: (a) corpus alignment and (b) elicitation of a phrasing model compatible for both members of a given language pair.

In an attempt to overcome compatibility issues when processing a bilingual corpus, the underlying idea implemented here is that one need not supply two parsers (one per language of a given language pair) and then try to make the two different parsing outputs converge. On the contrary, only one parser for either side (source or target language) suffices for deriving a phrasing model appropriate for the other side of a bilingual corpus.

More specifically, a bilingual corpus is needed, whose source language (SL) side is annotated with Part of Speech and lemma information and whose target language (TL) side additionally bears phrasing information, i.e. it is parsed into phrases. Then the corpus is aligned at word level. The next step involves mapping the TL parser-provided model onto SL, which entails grouping the SL words into phrases in accordance with the phrases of the TL corpus.

It is noteworthy that the mapping method described above is not unidirectional; an SL parser could be employed and its phrasing model used as a guide for phrasal segmentation of the TL. In the current implementation, for reasons of simplicity the target language corpus has served as a basis for generating a phrasing model for the source language.

For the aforementioned processing of a bilingual corpus, two distinct modules are being developed, the **Phrase aligner module** (**PAM**) and the **Phrasing model generator** (**PMG**), both of which are envisaged to function as language-independent methods. The Phrase aligner module is responsible for aligning a bilingual corpus at word level and mapping the TL phrasing model onto the source language, whereas the Phrasing model generator, being dependent on the first module, is assigned with the task of (a) eliciting the SL phrasing model, which is implicit in the aligned corpus, and (b) applying it to any new SL text.

The deliverable has the following structure: Section 2 provides a concise summary of the machine translation approach being followed in the PRESEMT project. Section 3 introduces the two modules and Section 4 reviews the related literature. Sections 5 and 6 respectively present the basic aspects, design and implementation of the Phrase aligner and the Phrasing model generator. Future related research is exemplified in Section 7. References are listed in Section 8, while an excerpt of an early internal report on the two modules (Appendix I) concludes the deliverable.