Steps of extracting zh-en MWEs from parallel corpus:

Input: tokenised en and word-segmentation finished zh, parallel corpus

1 tag zh and en corpus with treetagger for POS and lemma information.

2 convert tagged corpus into xml format with MWEtoolkit.

3 extract en MWEs and zh MWEs respectively with MWEtoolkit using the POS-based patterns

3.1 prepare the zh pos patterns, a mapping I did from en (Matiss used, MWE-Tools) to zh.

3.2 map the en patterns from the one that Matiss used to the other kind of pos-tagset that was used in this experiments.

4 annotate the extracted MWE-files and covert to MPAligner format (with the tool from MWE-Tools, from Matiss)

5 align the extracted en and zh MWEs with MPAligner tool.

5.1 run Moses/Giza++ tools with zh-en parallel corpus to get the word translation table of zh-2-en and en-2-zh

5.2 prepare the zh stop-words list

5.3 modify the xxxxx file of MPAligner by adding the zh-en word translation probability files and zh stop words list into the correct directory.

5.4 run the MPAligner.

Details of running of process: