texts from NEA and SBTC will be used in future extension of the corpus when resources allow for manual data entry (e.g. mechanical turks) or proprietary OCR software for Asian languages that performs reasonably well. Also, we are constantly requesting for parallel public informational text from other governmental authorities.

Although we have exploited prior knowledge put into the design of the POS tag sets and token segmentations using different (ad-hoc) tools, the philological perspective on segmentations and POS varies within each individual language and across languages. To fill these philological and cross-lingual gaps in the monolingual annotations, we are working to provide syntactic annotation with the Deep Linguistic Processing with HPSG Initiative (DELPH-IN)[6] and semantic annotation with the Global WordNet Association (GWA).[7] From the parses of the individual languages, the multi-layered annotation will allow extraction of the syntactic annotations (e.g. POS from HPSG word classes, word boundary from HPSG lexicon) and semantic annotations (e.g. semantic constraints from HPSG lexicon and its corresponding word senses mapped to WordNet). Wordnet sense annotation of the Indonesian and Japanese data from the `yoursingapore` subcorpus is ongoing.

For cross-lingual annotation, sentence-level, word-level and concept-level alignment will be carried out as resources permit. These word alignments from the hitherto under-represented language pairs should provide rich data for language technologies like MT and IR.

The NTU-MC is being used as a teaching tool, both in courses on corpus linguistics and semantics and as material for student projects. In the semantics class, students annotated short tourism pages (three students to a page) then looked at their inter-annotator agreement and reported on words where they had disagreed as to the correct sense as well as on words missing from the sense inventory (Princeton Wordnet). Students said that they found the concrete task interesting and that it really made the issues involved in defining word meanings clear. A similar task was done on the Chinese portion for a class in Chinese lexicography. When the corpus has been checked once more we intend to submit it as a sense tagged corpus multi-text to the Natural Language Tool Kit.

5 Conclusion

This project has produced a text collection, the NTU Multilingual Corpus, small in size but rich in language diversity. The NTU-MC contains a layer of monolingual annotation (POS tags and some sense tags) as well as a layer of cross-lingual annotation (sentence-level alignments) valuable for cross-lingual NLP tasks. The texts and annotation are released under an open license (CC by). In a cosmopolitan city like Singapore, there is a wealth of parallel text. This project urges future research to continue to draw diverse data through readily available yet untapped resources for corpus compilation. By progressively extending the NTU-MC with a larger dataset and multiple layers of annotation, it expands the scope of the usage and becomes a better corpus for general or computational linguistics researches. By building corpora of more diverse cross-lingual nature, it provides information on the unique sociolinguistic situation in linguistically diverse societies (e.g. translatability researches, language choice and language domain researches); also it pushes

[6] http://www.delph-in.net/

[7] http://www.globalwordnet.org/