Building and Annotating the Linguistically Diverse NTU-MC (NTU– Multilingual Corpus)*

Liling Tan and Francis Bond

Division of Linguistics and Multilingual Studies, Nanyang Technological University,
14 Nanyang Drive, Singapore, 637332, Singapore
alvations@gmail.com, bond@ieee.org

Abstract. The NTU-MC compilation taps on the linguistic diversity of multilingual texts available within Singapore. The current version of NTU-MC contains 375,000 words (15,000 sentences) in 6 languages (English, Chinese, Japanese, Korean, Indonesian and Vietnamese) from 6 language families (Indo-European, Sino-Tibetan, Japonic, Korean as a language isolate, Austronesian and Austro-Asiatic). The NTU-MC is annotated with a layer of monolingual annotation (POS tags) and cross-lingual annotation (sentence-level alignments). The diverse language data and cross-lingual annotations provide valuable information on linguistic diversity for traditional linguistic research as well as natural language processing tasks. This paper describes the corpus compilation process with the evaluation of the monolingual and cross-lingual annotations of the corpus data. The corpus is available under the Creative Commons – Attribute 3.0 Unported license (CC by).

Keywords: Multilingual, Corpus, Parallel text

1 Introduction

"The rapidly growing gap between the demand for high-quality multilingual content and the lag in the supply of language professionals is driving the requirement for technology that can dramatically improve translation turnaround time while maintaining exceptionally high output quality" (McCallum, 2011). Cross-lingual training using parallel corpora has been gaining popularity in NLP application tasks such as word sense disambiguation (e.g. Sarrafzadeh et al. 2011; Saravanan et al. 2010; Mitamura et al. 2007), information retrieval and question-answering. In addition, parallel corpora are valuable resources for advancing linguistic annotations morphologically, syntactically and semantically (e.g. Snyder and Barzilay; 2008, Hwa et al. 2005; Resnik, 2004).

The essential knowledge resource in building these language technologies are grounded on parallel corpora. The present pool of resources holds a sizable amount of European parallel corpora (e.g. Ralf et al. 2006; Erjavec, 2004), an increasing interest in building Asian languages-English bitexts (e.g. Xiao, 2004) but only a handful of parallel Asian language corpora (e.g. Zhang, 2005).

To fill the lack of parallel corpora of Asian languages, the NTU–Multilingual Corpus (NTU-MC) taps on the array of multilingual texts available in Singapore; ranging from the multilingual sign boards with official languages of Singapore (English, Chinese, Malay, Tamil) to posters, signs and guides targeted towards migrants, expats and tourists (in Indonesian, Japanese, Korean, Vietnamese, Thai, Tagalog, etc.). Singapore's multicultural and multilingual

* The authors of this paper thank Ms Joan Lee, New Media Manager of the Singapore Tourist Board, for granting permission to use and redistribute text from their multilingual websites (*www.yoursingapore.com.sg*). This research was partially funded by a joint JSPS/NTU grant on *Revealing Meaning through Multiple Languages.*

society has necessitated the use of parallel text for signboards, public announcements and information dissemination. The NTU-MC presents multilingual data from a modern cosmopolitan city where people interact in different languages. Empirically, the NTU-MC represents unique societal linguistic diversity; computationally, the NTU-MC provides diverse parallel text for NLP tasks. The NTU-MC presents a wealth of data to inform our analysis of language in a modern multicultural city from the traditional and computational linguistic point of view. This paper discusses the compilation of the NTU-MC from data collection to the present state of POS tagged and sentence-aligned parallel texts.

The rest of the paper is structured as follows: Section 2 describes the sub-tasks in the corpus compilation, the monolingual annotation and cross-lingual annotation process; Section 3 present the NTU-MC outputs and evaluates the layers of annotations; Section 4 presents future work on the NTU-MC and Section 5 concludes.

2 Corpus Construction

The NTU Multilingual Corpus adopts an opportunistic data collection approach and it is representative of the linguistically diverse data available within the Singapore language habitus. Currently, the corpus data contains the tourism domain of Singapore where multiple foreign languages are used on Singapore Tourism Board's website to entreat the tourism from the countries that speaks the respective languages. The NTU-MC is built on a Linux operating system with UTF-8 as the standardized encoding for its outputs.

2.1 Data Collection

The corpus project was granted the permission to use the websites[1] that are published by Singapore Tourism Board (STB). In the initial phase we have built a corpus totaling 375,000 words (15,000 sentences) in 6 languages (English, Chinese, Japanese, Korean, Indonesian and Vietnamese), from 6 language family trees[2] (Indo-European, Sino-Tibetan, Japonic, Korean as a language isolate, Austronesian and Austro-Asiatic) based on texts from the Singapore Tourism Board's *www.yoursingapore.com* website.

2.2 Crawling and Cleaning

`Httrack` (Roche, 2007) was used for data-collection and it was completed with a single command `httrack http://www.yoursingapore.com -o +*.yoursingapore.com/content/traveller/*/*.html -p1` .The raw HyperText Markup Language (HTML) files were downloaded without the embedded media files (e.g. images, flash files, embedded videos, etc.) from the webpages. As the markup language used to construct the websites were consistent, a custom-made perl script was created to extract the main body paragraph instead of using the commonly used Condition Random Field (CRF) algorithm (Marek et al. 2007). The markup cleaning extracted the text bounded by `<p>...</p>` within the `<div class = paragraph section>...</div>` attributes. The perl script successfully extracted the main body text from each webpage and ignored the subtexts that were headers to other pages.

During the annotation tasks, non-break spaces (`0xa0`) were found in the extracted text. These caused errors in POS tagging and sentence alignment. They appear before the start of the sentence and after the full stops in the sentence. A second round of cleaning was carried out to

[1] STB hosts a series of tourism related websites in particular websites with parallel texts, viz. *http://www.singaporemedicine.com/index.asp* , *http://app.singaporeedu.gov.sg/asp/index.asp* and *www.yoursingapore.com*

[2] Language family in this paper refers to the highest level of language classification from the Ethnologue (Lewis, 2009)

remove non-break spaces and the texts were then re-tokenized and re-annotated. All the textfiles were saved in UTF-8 encoding.

2.3 Sentence Segmentation

The English, Korean and Indonesian Texts use the same punctuation and the Natural Language Tool Kit (NTLK) `sent_tokenize` module (Bird et al, 2009) was sufficient to segment the English, Korean and Indonesian text. The `sent_tokenize` program uses stop punctuations (i.e. ! ? .) to identify the end of the sentence and it also correctly segmented sentence with websites by differentiating the sentence end full stop and full stops within a website.

The multi-byte Chinese and Japanese sentences were separated by the same sets of ！？。 punctuations. Thus the `nltk.RegexpTokenizer(u'[^！ ？ 。 ]*[!?。]')` was used to segment the Chinese and Japanese sentences. The Japanese regex has a minor tweak from the common `nltk.RegexpTokenizer(u'[^「 」 ！ ？ 。 ]*[ ！ ？ 。 ]')` , as recommended by the Hagiwara's Japanese chapter of the 「入門 自然言語処理」*nyumon shizen gengo shori* "Japanese Natural Language Processing with python" (Bird et al, 2010). The tweak was necessary to include non-sentence phrases bounded by「...」 brackets. Normally the Japanese「」brackets would have an individual sentence within the bracket, the text from www.yoursingapore.com used the 「」differently by embedding not only sentence but also proper names (e.g. 「マリーナ貯水池」*mari-na chosuichi* "Marina Reservoir"; 「スターバックス」*suta-bakkusu* "Starbucks") or loan phrases(e.g.「三歩一拝」*san ho ichi hai* "three step a bow" - a Chinese Buddhism term; 「ハラール」 *hara-u* "halal"; 「カルーセル」*karu-seru* "carousal") .

2.4 Tokenization

The tokenization (i.e. word level segmentation) tasks splits sentences up into individual "meaningful units" and these meaningful units are dependent on the philological stance of different word segmenter programs. In this paper, the term word and token will be used interchangeably to refer to the individual tokens output by the POS taggers and tokenizers.

For English and Indonesian data, whitespaces are the delimiter for the tokens. Although Vietnamese words are separated by whitespaces in the orthography, sometimes two "words" separated by whitespace are supposed to mean a single thing. For example, the Vietnamese word 'quốc tế' mean international but the individual "word" separated by the space does have its meaning ('quốc' means country and 'tế' means to run). Thus the JVnSegmenter module within JVnTextPro (Nguyen and Phan, 2007) was used to tokenize the Vietnamese data.

For the Japanese and Korean word level segmentation, the segmenter is incorporated into the POS-taggers that this corpus project is using. The Stanford Chinese word segmenter was used to segment the Chinese sentences in this corpus (Tseng et al, 2005).

Mis-segments generated from Stanford segmenter were local street names that were transliterated from English to Chinese. For example, the Stanford Chinese word segmenter wrongly tokenized 乌节路 *wujielu* "Orchard road" as 乌 节路 *wu jielu* "black joint-road". These topological terms were re-segmented with a manually crafted dictionary built using Wikipedia's Chinese translations of English names of Singapore places and streets.

2.5 Monolingual Annotation – Part Of Speech (POS) Tagging

Different programs were used to tag the individual languages with their respective POS tag sets. Due to the lack of an open source POS-tagger for Bahasa Indonesian, the Indonesian texts were not POS-tagged. All the tagged output was formatted into the Corpus Work Bench (CWB) verticalized text format with eXtensible Markup Language (XML) tags to encode the start and end of a sentence (i.e. <s>…</s>). Table 1 presents a brief summary of the sentence segmentation and POS-tagging task for the corpus compilation.

Table 1: Summary of Tokenization and Monolingual Annotation (POS tagging) Task

Language	Sentence Segmenter	Word Segmenter	POS-tagger (Tagger Encoding)	Tagset
English (en)	NLTK sent_tokenize	Whitespaces	HunPos (ISO-8859-1)	Penn Treebank II
Japanese (ja)	NLTK RegexpTokenizer	MeCab	MeCab (UTF-8)	IPAdic
Korean (ko)	NLTK sent_tokenize	POSTAG/ Sejong	POSTAG/Sejong (EUC-KR)	Sejong
Vietnamese (vi)	NLTK sent_tokenize	JVnSegmenter	JVnTagger (UTF-8)	VSLP
Chinese (zh)	NLTK RegexpTokenizer	Stanford Segmenter	Stanford POS tagger (UTF-8)	Penn Chinese Treebank
Indonesian (in)	NLTK sent_tokenize	Whitespaces	–	–

The HunPos tagger applied the Penn Treebank II POS annotations to the English texts (Halacsy et al, 2007). The pre-trained Wall Street Journal English (`en_wsj.model`) model was used with the HunPos tagger to tag the English data.

The Japanese data was tagged by the MeCab tagger (Kudo et al, 2004). The MeCab tagger was used with the `-0chasen` model, which was trained by the ChaSen tagger (Matsumoto et al. 1999). Different from the other POS-tagger used in this project, the MeCab morphological analyser provided more than a layer of POS annotations; MeCab output adheres to the IPADIC 2.7.0 standards (Matsumoto and Asahara, 2004).

The POSTech TAGger –Korean (POSTAG/Sejong) was used to tag the Korean text. As an agglutinative language, POSTAG/Sejong tagged the tokens at a morpheme level rather than a word level. A custom tagset with 41 tags was used by POSTAG/Sejong to suit the Korean morphemes. The POSTAG/Sejong tagger is only available on Microsoft Windows OS but we managed to run it under the WINE emulator (scripts for this are available with the corpus).

The JVnTagger (part of the JVnTextPro tool) with the `MaxEnt` model was used to annotate the Vietnamese text. The tagset used by JVnTextpro sets the standards for Vietnamese NLP as they pioneered the VLSP project (2006-2010) to "building basic resources and tools for Vietnamese language and speech processing", a five year long project from 2006 – 2010.

The Stanford Chinese POS tagger tags the Chinese data with the `chinese.tagger` model; the Chinese Penn Treebank tagset were used by the Stanford tagger (Tseng et al, 2005).

The primary issues with multilingual corpus POS annotation is the difference in encoding of the sources and the encoding that the POS tagger accepts as input and produce as output. When feeding data into the English (HunPos) and the Korean (POSTAG/Sejong) tagger, the encoding needed to be changed to the respective encoding that the tagger accepts (ISO-8859-1 and EUC-KR respectively). This caused some problems for Korean, as the input text contained characters that cannot be represented in the EUC-KR encoding used by POSTAG/Sejong (such as the – , é and © characters). We mapped them to - , e and (C) during the POS-tagging task for the Korean texts. We hope that more systems will produce UTF-8 versions of their morphological analyzers in the future.

Table 2: A sample of the monolingual annotation from the NTU-MC

Language	Segemented, Part of Speech tagged Text
English	<s>If_IN you_PRP only_RB have_VBP time_NN for_IN one_CD club_NN in_IN Singapore_NN ,_, then_RB it_PRP simply_RB has_VBZ to_TO be_VB zouk_JJ ._.</s>
Japanese	<s>シンガポール_名詞-固有名詞-地域-国 で_助詞-格助詞-一般 一つ_名詞-一般 の_助詞-連体化 クラブ_名詞-一般 に_助詞-格助詞-一般 しか_助詞-係助詞 行く_動詞-自立 時間_名詞-副詞可能 が_助詞-格助詞-一般 なかっ_形容詞-自立 た_助動詞 と_助詞-格助詞-引用 し_動詞-自立 たら_助動詞 、_記号-読点 間違い_名詞-ナイ形容詞語幹 なく_助動詞 、_記号-読点 この_連体詞 ズーク_名詞-一般 に_助詞-格助詞-一般 行く_動詞-自立 べき_助動詞 です_助動詞 。_記号-句点 </s>
Korean	<s>싱가포르_NNP 에서_JKB 클럽_NNP 한_NNP 군데_NNB 밖에_JX 가_VV 르_ETM 시간_NNG 이_JKS 없_VA 다면_EC ,_SP Zouk_SL 를_JKO 선택_NNG 하_XSV 시_EP 어요_EF ._SF</s>
Vietnamese	<s>Nếu_C bạn_N chỉ_R có_V thời_gian_N ghé_V thăm_V một_M câu_lạc_bộ_N ở_E Singapore_Np ,_, hãy_R đến_V Zouk_Np ._.</s>
Chinese	<s>如果_CS 您_PN 在_P 新加坡_NR 只_AD 能_VV 前往_VV 一_CD 间_M 俱乐部_NN ， _PU 祖卡_NN 酒吧_NN 必然_AD 是_VC 您_PN 的_DEG 不二_JJ 选择_NN 。_PU </s>

2.6 Cross-lingual Annotation - Sentence-level Alignment

As machine-readable dictionaries are only available for certain languages in the NTU-MC, the dictionary and length based hunalign tool is suitable for aligning the NTU-MC as the algorithm "remains completely meaningful even in total absence of a dictionary" (Varga et al. 2005). The alignments generated by hunalign are bi-directionally equivalent. The sentence-level alignment task was carried out with four different conditions:

–dic – hunalign outputs without language pair dictionary,
+dic – hunalign outputs with language pair dictionary,
+human – manually aligned Gold Standard,
+pivot – alignments generated by transitive relation using 2 +human alignments

Only sentences from the textfiles that were available in all 6 languages were sentence-aligned. Two native Chinese and Japanese speakers were enlisted to correct the +dic alignments for the English-Chinese and English-Japanese data. The English-Chinese, English-Japanese and English-Korean were generated with the `CC-CEDICT` (MDBG, 2011), `JMDICT` (Breen, 2004) and enhanced `engdic` (Paik and Bond, 2003) respectively.

By extending the idea of exploiting existing resources to building and extending valency dictionaries, we used the +human alignments to produce +pivot alignments. Using English as the pivot language, we aligned Chinese-English-Japanese.

3 Corpus Evaluation

The corpus evaluation is based on the data availability, corpus outputs and its monolingual and cross-lingual annotations. The monolingual annotations were evaluated extrinsically by measuring Inter-annotator Agreement (IAA) between the POS-taggers and human annotators. The lack of in-depth knowledge about the tagsets deters the human annotators to use sophisticated tags thus intrinsic evaluation (i.e. using human Gold Standard) is not viable. The quality of the parallel text alignments was intrinsically evaluated by computing the F-score of the hunalign outputs against manually aligned data.

Corpus Availability

For a corpus to be a valuable resource, it must be both useful and accessible (Ishida et al. 2006). The owners of the source data (Singapore Tourism Board) have allowed the redistribution of this data, licensed by the Creative Commons (CC) Attribution 3.0 Unported License. Users of the corpus are able to share (i.e. copy, distribute and transmit) and remix (i.e. to adapt) the corpus under the condition of attributing the work to the NTU-MC project. The data is available from http://linguistics.hss.ntu.edu.sg/ResearchinLMS/Pages/NTUMultilingualCorpus.aspx

Corpus Outputs

The NTU-MC project compiled a foundation text of 375,000 words (15,000 sentences) for the NTU-MC in 6 languages from 6 language family trees. The breakdown of the monolingual annotation is as followed (the number. of tokens excludes punctuations and symbols):

Table 3: Monolingual Annotation Outputs

Language (language code)	Language Family	#Texts	#Sentences	#Tokens	POS Tagged
English (en)	Indo-European	398	3,255	76,339	✓
Japanese (ja)	Japonic	267	2,648	72,797	✓
Korean (ko)	Language Isolate	266	2,407	67,341	✓
Vietnamese (vi)	Austro-Asiatic	269	2,236	56,535	✓
Chinese (zh)	Sino-Tibetan	280	2,365	52,047	✓
Indonesian (id)	Austronesian	270	2,185	50,315	✗
Total:	6 Families	1750	15,096	375,374	

The main alignment task for NTU-MC focused on the English-Asian Languages alignments due to the amount of lexical resources available for English bitext. The corpus produced 2 Gold Standard (+human) alignments, 3 +dic alignments, 1 +pivot alignment and 11 -dic alignments generated with the `null.dic` option on hunalign.

Table 4: Cross-lingual Annotation Outputs

	en	id	ja	ko	vi	zh
en						
id	–dic					
ja	+human / +dic	–dic				
ko	+dic	–dic	–dic			
vi	–dic	–dic	–dic	–dic		
zh	+human / +dic	–dic	+pivot	–dic	–dic	

3.1 Monolingual Annotation Evaluation

The `fish-head-curry.txt` from the NTU-MC was selected at random for human annotators to verify the POS-taggers' accuracy. The human annotators were assigned to verify the POS tags and mis-segmented tokens. The accuracy of the human annotation might be primed by what the POS tagger had tagged. Therefore the human verifications were not treated as the "gold standard" but an inter-annotation agreement (IAA) score that was derived from the annotators' identification of the mis-segmented and mis-tagged tokens[3]. For the Japanese POS

[3] This excludes punctuation and both the number of mis-segments and mis-tagged tokens.

evaluation, there was no human annotator available. Thus a different POS tagger, ChaSen morphemic analyzer, was used to calculate IAA. Both programs uses the ipadic POS, but the noticeable difference is that ChaSen is more conservative when tagging unknown words: ChaSen applied the 未知語 *michigo* "unknown word" tag to tokens for unseen words whereas MeCab forces the closest fit POS to the unknown tokens. The 12 instances of 未知語 tags in `fish-head-curry.txt` were not included in the IAA calculation.

Table 5: Summary of Segmentation and POS Annotation Task

Lang-uage	Sentence Order	#Tokens	#Sent-ences	#Mis-segments	#Mis-tagged	IAA	Reported accuracy
en	SVO	235	7	-	18	92.23%	96.58% (Halacsy et al, 2009)
ja	SOV	293	14	3	8	96.25%	97.66 % (Kudo et al, 2004)
ko	SOV	374	14	44	27	81.02%	90.7% (Lee et al, 2002)
vi	SVO	225	7	14	10	89.33%	93.32% (Nguyen et al, 2010)
zh	SVO	249	9	19	16	85.94%	93.65% (Tseng et al, 2005)

The IAA reported in table 4 serves as a gauge, an error bar, of the reported accuracy reported by the individual taggers. The IAA is measured as such:

$$\text{non-matches} = \text{no. of mis-segment} + \text{no. of mis-tagged} \quad (1)$$

$$\text{matches} = \text{no. of tokens} - \text{non-matches} \quad (2)$$

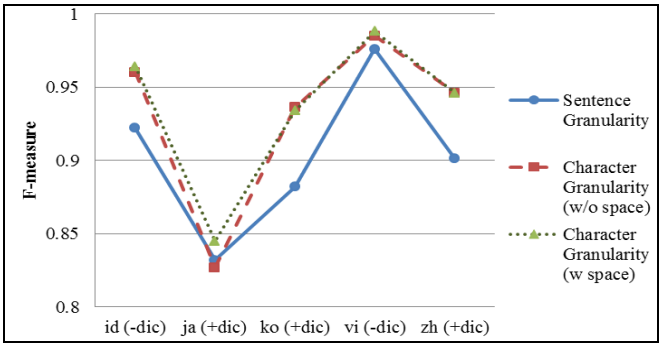
$$\text{IAA} = \text{matches} / (\text{matches} + \text{non-matches}) * 100\% \quad (3)$$

3.2 Cross-lingual Annotation Evaluation

A subset of 9 text files was selected to evaluate the quality of the hunalign outputs for language pairs with English sentences. The evaluation metrics adheres to standards set by the ARCADE II project (Chiao et al. 2006); the recall, precision and F-score is computed on the hunalign output of word segmented sentences. F-scores were computed using sentence and character granularity (with and without space).

From Figure 1, the alignment task on Japanese, Korean and Chinese is a much more difficult task than aligning Indonesian or Vietnamese data; even with the dictionaries' input, alignments for non-Latin character-based languages are poorer in alignments. Possibly, it is the difference in sentence order (refer table 4) that affected the lexicon quality of the Japanese-English and Korean-English alignments. Nevertheless, +human alignments were manually crafted for English-Japanese and English-Chinese sentences and the English-Korean alignment is reasonably good in terms of character granularity.

Figure 1: F-measure of `hunalign` on English-Asian Language alignments



The primary advantage of pivoting alignments to generate other language-pairs alignments is the simplicity to leverage on Gold Standard alignments to produce alignments where the

bilinguals of the language pairs are scarce. Similar to the idea of increasing number of language pairs quadratically by sourcing parallel sources with more languages (Eisele & Yu, 2010), +pivot alignments can produce +human like alignments quadratically with each +human alignments. Although it is possible to create more alignments through other pivoting permutations, generating pivoted alignments from crude -dic alignments will be perpetuating the original mis-alignments that hunalign had produced. Thus only the pivoted Gold Standard alignments was worth the effort as it can be able to produce word-level alignments of similar quality to the +human alignments.

4 Discussion and Future work

A comparable corpus to the NTU-MC is OPUS which taps open source parallel text (Tiedemann, 2009). The OPUS is representative of a global open source enthusiast's community, while the NTU-MC targets data from a specific cosmopolitan society. The OPUS covers a wider range of domains with large sub-corpora and it provides automated monolingual (POS tags and syntactic parses) and cross-lingual (sentence and word level alignments) annotations; whereas the NTU-MC is a corpus of a smaller size but more diverse in Asian language data with deeper annotations. Over time we intend to achieve Gold Standard annotations beneficial for NLP tasks.

The NTU-MC is an ongoing effort to add content, layers of annotation and usability as it continues to make multilingual resources machine readable for NLP tasks. Future work on the NTU-MC involves increasing the amount of data, the layers of monolingual annotations and cross-lingual annotations. The immediate expansion of the corpus would be to use the parallel texts (English, Malay, Chinese and Tamil) distributed by National Environment Agency of Singapore (NEA) and Sembawang Town Council (SBTC).[4] 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)[5] and semantic annotation with the Global WordNet Association (GWA).[6] 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).

For cross-lingual annotation, sentence-level and word-level alignments will be carried out as resources permits; then alignment pivoting will be done in a mesh manner to achieve Gold Standard sentence alignments for all language pairs and proceed with word-level alignments. These word alignments from the hitherto under-represented language pairs should provide rich data for language technologies like MT and IR.

5 Conclusion

This project has produced the initial text collection of the NTU Multilingual Corpus, small in size but rich in language diversity. The NTU-MC contains a layer of monolingual annotation

[4] The authors thank Ms Dorothy Cheung, Public Relations Manager of Sembawang Town Council (SBTC) and Mr Edrick Chua, Assistant Director of Corporate Communications from National Environment Agency (NEA) for their permission and aid in providing access to their data. Though the data from SBTC and NEA is not used for the current phase of NTU-MC compilation, we hope to use it for the future extension of the corpus.

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

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

(POS tags) on all language data except Indonesian and a layer of cross-lingual annotation (sentence-level alignments) valuable for cross-lingual NLP tasks. The texts and annotation will be released under an open license (CC by). In any cosmopolitan city like Singapore, the wealth of parallel text remains untapped for corpora building. 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 the state-of-the-art NLP techniques through more robust cross-lingual training (Matsumoto et al. 1993).

References

Bird, S., Klein, E., Loper, E., 萩原正人 (Hagiwara, M.), 中山敬広 (Nakayama, T.) and 水野貴明 (Mizuno, T.) (translation). 2010. *入門 自然言語処理 (Introduction to Natural Language Processing)*. O'Reilly, Japan (translation, with one extra chapter, of Bird *et al.* 2009).

Bird, S., Ewan, K., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Breen, J.W. 2004. JMDict: a Japanese-multilingual dictionary. In *COLING 2004 Workshop on Multilingual Linguistic Resources*, Geneva. pp, 71–78.

Chiao, Y.C., Kraif, O., Laurent, D., Nguyen, T.M.H., Semmar, N., Stuck, F., Veronis, J. and Zaghouani,W. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. *Proceedings of the LREC 2006 Conference*.

Eisele, A. and Chen, Y. 2010. MultiUN: A multilingual corpus from United Nation documents. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).*

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Fourth International Conference on Language Resources and Evaluation, LREC'04, (ELRA*), pp. 1535-1538.

Halácsy, P., Kornai, A. and Oravecz, C. 2007. HunPos - an open source trigram tagger In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume *Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics*, pp.209--212.

Hwa, R., Resnik, R., Weinberg, A., Cabezas, C., and Kolak,C. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering, 11(3)*, pp.311–325

Kudo T., Yamamoto, K., and Matsumoto, Y. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. pp.230–237.

Lewis, P.M. 2009. *Ethnologue: Languages of the World, Sixteenth edition*. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Marek, M., Pecina, P. and Spousta, M. 2007. Web page cleaning with conditional random fields. In *Proceedings of the Web as Corpus Workshop (WAC3), Cleaneval Session*.

Matsumoto, Y., Ishimoto, H. and Utsuro, T. 1993. Structural matching of parallel texts. In *31st Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp.23–30.

Matsumoto, Y., Takaoka, K. and Asahara, M. 1999. *ChaSen Morphological Analyzer version 2.4.0 User's Manual*. NAIST Technical Report, Nara Institute of Science and Technology

Technical Report 99009. Retrieved on 07 Jan 2011 from *http://sourceforge.jp/projects/chasen-legacy/docs/chasen-2.4.0-manual-en.pdf/en/1/chasen-2.4.0-manual-en.pdf.pdf*

MDGB. 2011. *CC-CEDICT [Machine-Readable Dictionary]*. Netherlands : MDGB, Retreived May 03, 2011 from http://www.mdbg.net/chindict/chindict.php?page=cedict

McCallum, B. 2011. Translation Technology at the United Nations. *MultiLingual Computing & Technology, 15(2)*, pp. 62.

Mitamura, T., Lin, F., Shima, H., Wang, M., Ko, J., Betteridge, J., Bilotti, M., Schlaikjer, A., and Nyberg, E. 2007. JAVELIN III: Cross-Lingual Question Answering from Japanese and Chinese Documents. *Proceedings of NTCIR-6 Workshop Meeting*, Tokyo, Japan.

Paik, K. and Bond, F. 2003. Enhancing an English/Korean Dictionary. In *Papillon 2003 Workshop on Multilingual Lexical Databases*, Sapporo, Japan.

Nguyen, C.T. and Phan, X.H. 2007. *JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool*. Retrieved on 30 Jan 2011 from *http://jvnsegmenter.sourceforge.net/*

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*,Genoa, Italy.

Resnik, P. 2004. Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation. In Alexander Gelbukh (ed.), *Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing*, pp. 283-299.

Roche, X. 2007. *Httrack Website Copier - Offline Browser*. [Computer Software]. Retrieved Jan 30, 2011. Available from *http://www.httrack.com/*

Sarrafzadeh, B., Yakovets, N., Cercone, N., & An, A. 2011. *Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources. (Technical Report CSE-2011-01).* Ontario: Department of Computer Science and Engineering.

Saravanan, K., Udupa, R., and Kumaran, A. 2010. Crosslingual Information Retrieval System Enhanced with Transliteration Generation and Mining, In *Forum for Information Retrieval Evaluation (FIRE-2010) Workshop*, Kolkata, India.

Snyder, B. and Barzilay, R. 2008. Cross-lingual propagation for morphological analysis. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pp. 848–854.

Ishida, T. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp.96-100, keynote.

Tiedemann, J. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of RANLP'09*, Borovets, Bulgaria. pp.237–248.

Tseng, H., Jurafsky, D. and Manning, C. 2005. Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop*, Jeju Island, Korea.

Varga, D., Nemeth, L., Halacsy, P., Kornai, A., Tron, V. and Nagy, V. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, Borovets. Bulgaria. pp.590–596.

VLSP Project (2006-2010). *VLSP Project – Vietnamese Language Processing*. Retrieved on 02 Jan 2010 from http://vlsp.vietlp.org:8080/demo/?page=about .

Xiao, Z., McEnery, A., Baker, P. and Hardie, A. 2004. Developing Asian language corpora: standards and practice. In *Proceedings of the 4th Workshop on Asian Language Resources, Sanya, Hainan Island.* pp. 1-8.

Zhang, Y., Uchimoto, K., Ma, Q. and Isahara, H. 2005. Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multilingual Corpora. *Second International Joint Conference on Natural Language Processing*, Jeju Island, Republic of Korea. pp 85-90.