Linguistic studies using large annotated corpora: Introduction

Hiroki NOMOTO and David MOELJADI

Tokyo University of Foreign Studies and Palacký University Olomouc
nomoto@tufs.ac.jp, david.moeljadi@upol.cz

1. Background[1]

Corpora have been used widely in modern linguistic research. Two notable features of corpus development in recent years are a significant increase in size and various kinds of annotations. Billion-size corpora are not uncommon nowadays. Efforts have been made to enrich raw texts with linguistic information, such as morphology, parts of speech (POS), constituent structure, semantic dependency, information and discourse structural status and so on. However, these developments, which took place primarily in the field of natural language processing, have not been maximally utilized in the linguistic research of languages in Nusantara.

This *NUSA* special issue was planned to encourage researchers to explore the available resources and share ways of using them to investigate old and new empirical and theoretical topics. We solicited submissions openly by means of an official call for papers.

In the call for papers, we provided the list of available resources (1) and requested that all manuscripts explicitly state what resource(s) they used and how they utilized the annotations. Besides the suggested annotated corpora, the authors were also allowed to build their own corpus by annotating a raw corpus using a morphological dictionary (e.g. MALINDO Morph[2]), a POS tagger (e.g. MorphInd,[3] Rule-Based POS Tagger Bahasa Indonesia[4], an HPSG grammar (e.g. INDRA[5]) and so on.

(1) Examples of large annotated corpora

a. **MALINDO Conc** (Nomoto, Akasegawa & Shiohara 2018a)
(`https://malindo.aa-ken.jp/conc/`)
Reclassified version of the Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012; Nomoto, Akasegawa & Shiohara 2018b)[6]
morphological annotation; Malay, Indonesian; 3 million words

b. **Korpus Indonesia (KOIN)** (Kwary 2018)
(`https://korpusindonesia.kemdikbud.go.id/`)

[1] **Acknowledgements** This work was supported in part by JSPS KAKENHI Grant Number JP18K00568. We would like to express our sincere gratitude to all those involved in the publication process, especially the eight peer reviewers who devoted their precious time and expertise in their respective fields to improve the quality of this volume. All remaining errors are ours.

[2] Nomoto et al. (2018), `https://github.com/matbahasa/MALINDO_Morph`

[3] Larasati, Kuboň & Zeman (2011), `http://septinalarasati.com/morphind/`

[4] Rashel et al. (2014), `https://github.com/andryluthfi/indonesian-postag`

[5] Moeljadi, Bond & Song (2015), `http://moin.delph-in.net/IndraTop`

[6] `http://wortschatz.uni-leipzig.de/en/`

NOMOTO, Hiroki and David MOELJADI, 2019. 'Linguistic studies using large annotated corpora: Introduction'. In Hiroki Nomoto and David Moeljadi, eds. *Linguistic studies using large annotated corpora*. NUSA 67: 1–6. [Permanent URL: http://repository.tufs.ac.jp/handle/10108/94450] [doi: 10.15026/94450]

POS annotation; Indonesian; 3.7 million words

c. **One Million POS Tagged Corpus of Bahasa Indonesia**
(`http://www.panl10n.net/indonesia/#Linguistic_Resources_`)
POS annotation; Indonesian; 1 million words

d. **Data from the Jakarta Field Station, Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, 1999–2015**
(`http://jakarta.shh.mpg.de/data.php`)
word gloss annotation; Jakarta Indonesian and other languages in Indonesia; 4.5 million words

To the best of our knowledge, these are the major large annotated corpora of the languages used in Nusantara that were openly available at the time of writing this introduction. Other large-size corpora also exist, but they are either raw texts or accessible to only those who belong to a certain institution or who (are affiliated with an institution that) can afford to pay a subscription fee (e.g. Korpus Dewan by Dewan Bahasa dan Pustaka, Malaysia,[7] Sketch Engine[8]). Use of open resources is recommended to ensure the replicability of the findings and equality amongst researchers from different financial backgrounds.

2. Articles in this volume

Siaw-Fong Chung and Meng-Hsien Shih use a Malay corpus containing 35,767 newspaper articles collected from *Utusan Malaysia*, a national Malaysian newspaper, between December 2010 and June 2011. They annotate the corpus using the following (combination of) tools: (i) the morphology analyzer provided by Tan et al. (2017) and the Malay NLP tool provided by University of Malaya and (ii) MorphInd (Larasati, Kuboň & Zeman 2011), a POS tagger for Indonesian. Although the first approach has the advantage that the tools are based on Malaysian Malay and thus have fewer unknown words, the second approach is chosen due to the high cost of computing and the need to check all POS tags manually in the first approach. After annotating the corpus, they examine the frequency and percentage of all lemma and POS tags, suffixes, prefixes and morphological combinations in the corpus. They also list the top ten roots without affixes, foreign words and unknown words. The latter two are those commonly found in Malaysia but are not part of the vocabulary in the MorphInd dictionary. In addition, they use a smaller portion of the annotated corpus to generate a wordlist for tokens prefixed by *ber-* using the AntConc Concordancer (Anthony 2005). They do semantic categorization and annotation by adding semantic information to the tags provided by the MorphInd POS tagger.

Gede Primahadi Wijaya Rajeg, Karlina Denistia and Simon Musgrave use the Indonesian component of the original Leipzig Corpora Collection (LCC; Goldhahn, Eckart & Quasthoff 2012) and investigate similarities among three kinds of denominal verbs, i.e. those with (i) the prefix *meN-*, (ii) the prefix *meN-* plus the suffix *-kan* and (iii) the prefix *meN-* and the suffix *-i*.[9] LCC itself is not annotated in any way. However, its huge

[7] `http://sbmb.dbp.gov.my/korpusdbp/SelectUserCat.aspx`

[8] `https://www.sketchengine.eu`

[9] Choi (2019) also investigates their usage patterns using LCC data. She annotates sentences containing *-kan* and *-i* verbs with the semantic roles of the verbs' arguments and examined the annotated data.

Table 1. Corpora and techniques used in this volume

Authors	Corpus	Size (tokens)	Techniques etc.
Chung & Shih	own data	13,979,859	MorphInd, AntConc, Python
Rajeg, Denistia & Musgrave	LCC	180,769,204	vector space model, hierarchical agglomerative clustering, MorphInd, R
Shiohara, Sakon & Nomoto	LCC	16,602,778	MALINDO Conc, Python

size allows us to apply the state-of-the-art computational technique of capturing the property of a word numerically, more specifically in terms of vectors ('vector space model'). Such vectors are normally regarded as representing lexical meanings. Thus, obtaining a vector for each word in a corpus amounts to annotating it in terms of lexical meanings. With words expressed as vectors, it is possible to calculate the relative distances between multiple words and identify clusters based on them. The authors conduct a hierarchical agglomerative clustering analysis. They show that '*meN-*/*meN-*+*-kan*/*meN-*+*-i*' triplets are not uniform. Some (e.g. *menyusu* '(of a baby/young animal) to suckle', *menyusukan* 'to let sb. suckle, to breast-feed sb.', *menyusui* 'to breast-feed sb.') belong to a cluster and are similar to each other, whereas others (e.g. *mengata* 'to say', *mengatakan* 'to say', *mengatai* 'to rebuke') belong to separate clusters, pointing to their dissimilarity.

Asako Shiohara, Yuta Sakon and Hiroki Nomoto also use the Indonesian component of LCC, but use the reclassified version (Nomoto, Akasegawa & Shiohara 2018b), which has recently been made publicly available from the LCC website. They use the XML files with morphological annotations used in the MALINDO Conc concordancer (Nomoto, Akasegawa & Shiohara 2018a) to obtain the three inflectional forms of a transitive verb, i.e. *meN-*, *di-* and bare forms, and their frequencies with different agent expressions. With third person agents, all three verb forms are grammatical, unless the prefix *meN-* is banned for syntactic reasons (e.g. Saddy 1991). They thus investigate what determines the choice among them. They choose the six most frequent stems, i.e. *miliki* 'to possess', *lakukan* 'to do', *buat* 'to make', *lihat* 'to look', *gunakan* 'to use' and *katakan* 'to say', and search the corpus for sentences containing those stems using MALINDO Conc. They then examine the usage patterns of the three clause types, mainly focusing on the two non-active clauses, and report frequencies in different conditions for individual stems.

3. State of affairs, future prospects

Although the three studies deal with different topics, there is one element they have in common. They discuss the behaviours of individual lexical items. This is often neglected in traditional grammatical descriptions, sometimes resulting in overgeneralizations. While careful attention to individual lexical items is a strength of the corpus-based approach, it cannot be denied that what has been attained is still not as general as attained by studies based on speaker intuitions and/or small corpora. That is to say, its predictive power beyond individual lexical items is rather limited. We know how individual lexical items actually behave. However, it is not necessarily obvious from our direct findings that the same specific items do not behave otherwise and how items that were not examined behave. To achieve the latter level of understanding, it is important to aim not just at revealing the detailed behaviours of individual lexical items but also at discovering general

patterns underlying behind them, ultimately in a semi- or fully-automated fashion based on large corpora combined with some sophisticated computational techniques.

Unfortunately, however, it is also true that we are not equipped with the necessary resources to do so yet. Chung & Shih attempted to annotate their Malay texts with POS's. Currently, no POS annotated corpora of Malay are available. A couple of POS taggers have been developed, but their accuracies are not sufficiently high for linguistic research. At the end, they chose MorphInd (Larasati, Kuboň & Zeman 2011), which was designed for Indonesian and hence not a perfect choice for them.

MorphInd is not free from problems, even for analysing Indonesian. Rajeg, Denistia & Musgrave obtained unwanted results because MorphInd cannot handle spelling variation. Like words are pronounced differently depending on the speed, formality, etc. in speech, they are spelt differently in writing. Thus, *kantung* 'pocket' is also spelt as *kantong* in the corpus. Repetition of a letter is a common strategy to express emotion (e.g. *aduhhh* instead of *aduh* 'ouch'). In the standard orthography, some compounds and reduplicated words contain a hyphen while other do not. This inconsistency often confuses speakers. Thus, *menandatangani* 'to sign on', which is derived from *tanda tangan* (notice the white space), is misspelt as *menanda-tangani* or *menanda tangani*. The latter two instances should be treated together with *menandatangani*, which MorphInd is not able to.

In fact, the problems they faced can be solved by proper pre-processing, in particular normalization and tokenization. However, no reliable tools exist for these processes as far as we know. MALINDO Morph (Nomoto et al. 2018) can be used to some extent because it contains words with non-standard spellings. For example, it has an entry for *menanda-tangani*, as in (2).[10] This line shows that the surface form *menanda-tangani* is a spelling variant of the lemma *menandatangani*. However, identifying *menanda tangani* as a variant of *menandatangani* requires a tokenizer.

(2) ec-42406 tanda tangan **menanda-tangani** meN- -i 0 0 Leipzig tandatangani *menandatangani*

The most challenging issue in developing a real tokenizer, that is, one capable of identifying token boundaries that are not white spaces accurately, is ambiguity. As Rajeg, Denistia & Musgrave note, while most instances of *menanda* is a part of the lemma *menandatangani* in Indonesian (but not in Malay), some are indeed instances of the lemma *menanda*. Ambiguity abounds in the MALINDO Morph morphological dictionary too. First, some surface forms have more than one morphological analysis, and hence appear more than once in MALINDO Morph. For example, *mengecam* has the two analyses in (3). They are derived from different roots. The one derived from *cam* (3a) means 'to recognize', whereas the one derived from *kecam* (3b) means 'to condemn'.[11]

(3) a. cc-14809 *cam* **mengecam** meN- 0 0 0 Kamus cam mengecam

b. cc-36912 *kecam* **mengecam** meN- 0 0 0 Kamus kecam mengecam

[10] Each entry in MALINDO Morph consists of the following ten items: ID, root, surface form, prefix/proclitic, suffix/enclitic, circumfix, reduplication type, source, stem and lemma.

[11] Tomita (2020) examines morphological ambiguities present in MALINDO Morph like this in detail.

Second, ambiguity can also arise in a single line. Consider the surface form *Me-nariknya* in (4). This form is ambiguous in three ways, as reflected in the stem information, where the '@' and '+' signs indicate disjunction and token boundary, respectively. It can be (i) the exclamatory and nominalized form of the adjective *menarik* 'interesting' (e.g. *Menariknya cerita itu!* 'How interesting the story is!'; *Cerita ini menariknya di mana?* 'Speaking of this story, where is the interesting part?'), (ii) the adjective *menarik* followed by the enclitic form of the pronoun *dia* 's/he' (e.g. *cerita menariknya* 'his/her interesting story') or (iii) the morphological active form of the verb *tarik* followed by the enclitic form of the pronoun *dia*, meaning 'to pull him/her' (cf. Nomoto 2020)

(4) ec-42593 tarik Me-nariknya meN- -nya 0 0 Leipzig
menariknya@menarik+dia@tarik+dia menariknya@menarik+dia

Currently, disambiguation must be done manually, as was the case with the morphological annotation of the data in MALINDO Conc (Nomoto, Akasegawa & Shiohara 2018a; Tomita 2020). However, the process needs to be automated in the future to handle larger data with less time and cost. The automation will involve machine learning, which requires annotated corpora based on which learning can take place. Therefore, automation brings about a good circulation. An annotated corpus is created by annotating a corpus, and the resulting corpus after necessary manual corrections can be used to improve the annotation process, which will in turn be used for annotating another corpus.

To summarize, annotated corpora are important in two interconnected ways. They are essential for corpus-based linguistic research to go beyond observations about individual lexical items. They are also vital to automate the production of annotated corpora. Therefore, linguists and natural language processing researchers need to work closely, especially given that the number of researchers is much smaller in Malay/Indonesian compared to languages such as English, Mandarin and Japanese.

Finally, we hope more people will use annotated corpora for their research and, if possible, develop open annotated corpora and annotation tools. The articles in this volume give ideas about what kind of tools are available, what their strong and weak points are and how they can be used to investigate a specific research question. Furthermore, we look forward to seeing studies on languages other than Malay/Indonesian, which are unfortunately not included in this volume. The techniques used for Malay/Indonesian can be modelled for other languages. The Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012) offers large raw corpora for free download in the following languages: Balinese, Banjar, Javanese, Madurese, Minangkabau and Sundanese.[12]

References

Anthony, Laurence. 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of the International Professional Communication Conference*, 729–737.

Choi, Hannah Yun Jung. 2019. *A corpus based analysis of* -kan *and* -i *in Indonesian*: Nanyang Technological University MA thesis. `https://hdl.handle.net/10356/136955`.

[12] `https://wortschatz.uni-leipzig.de/en/download`

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Kwary, Deny A. 2018. Towards the first online Indonesian National Corpus. In *The 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018)*.

Larasati, Septina Dian, Vladislav Kuboň & Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In Cerstin Mahlow & Michael Piotrowski (eds.), *Systems and frameworks for computational morphology*, 119–129. Verlag: Springer. doi:10.1007/978-3-642-23138-4_8.

Moeljadi, David, Francis Bond & Sanghoun Song. 2015. Building an HPSG-based Indonesian resource grammar (INDRA). In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 workshop*, 9–16. Beijing, China: Association for Computational Linguistics. doi:10.18653/v1/W15-3302.

Nomoto, Hiroki. 2020. Towards genuine stemming and lemmatization in Malay/Indonesian. In *Proceedings of the Twenty-Sixth Annual Meeting of the Association for Natural Language Processing*, 1033–1036. http://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/F4-3.pdf.

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018a. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL).

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018b. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA* 65. 47–66. doi:10.15026/92899.

Nomoto, Hiroki, Hannah Choi, David Moeljadi & Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In Kiyoaki Shirai (ed.), *Proceedings of the LREC 2018 workshop "The 13th Workshop on Asian Language Resources"*, 36–43. http://lrec-conf.org/workshops/lrec2018/W29/pdf/8_W29.pdf.

Rashel, Fam, Andry Luthfi, Arawinda Dinakaramani & Ruli Manurung. 2014. Building an Indonesian rule-based part-of-speech tagger. In *International Conference on Asian Language Processing (IALP 2014)*, IEEE. doi:10.1109/IALP.2014.6973521.

Saddy, Douglas. 1991. WH scope mechanism in Bahasa Indonesia. In Lisa L. S. Cheng & Hamida Demirdash (eds.), *MIT working papers in linguistics 15: More Papers on Wh-movement*, 183–218.

Tan, Tien-Ping, Bali Ranaivo-Malançon, Laurent Besacier, Yin-Lai Yeong, Keng Hoon Gan & Enya Kong Tang. 2017. Evaluating LSTM networks, HMM and WFST in Malay part-of-speech tagging. *Technology Transforming Lives I* 9(2-9). 79–83.

Tomita, Nanami. 2020. Mareego no keitaitekiaimaigo nitaisuru kyoukigo o mochiita hanbetsu [Distinguishing morphologically ambiguous words in Malay using their collocates]. Tokyo University of Foreign Studies BA thesis.