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

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

- **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

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

---

**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.

\(^1\) Nomoto et al. (2018), https://github.com/matbahasa/MALINDO_Morph


\(^3\) Rashel et al. (2014), https://github.com/andryluthfi/indonesian-postag

\(^4\) Moeljadi, Bond & Song (2015), http://moin.delph-in.net/IndraTop

\(^5\) http://wortschatz.uni-leipzig.de/en/

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). 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 & Zeeman 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. 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.
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 tangan_. 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). This line shows that the surface form _menanda-tangani_ is a spelling variant of the lemma _menandatangani_. However, identifying _menanda tangan_ as a variant of _menandatangani_ requires a tokenizer.

\[(2) \text{ ec-42406 tanda tangan } \text{ menanda-tangani } \text{ meN- } -i \text{ 0 0 Leipzig } \text{ tandatangani } \text{ menandatangani} \]

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’.

\[(3) \text{ a. cc-14809 } \text{ cam } \text{ mengecam } \text{ meN- } 0 0 0 \text{ Kamus cam mengecam } \text{ mengecam} \]
\[\text{ b. cc-36912 } \text{ kecam } \text{ mengecam } \text{ meN- } 0 0 0 \text{ Kamus kecam mengecam} \]

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 Con (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


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


