# ICON: Building a Large-Scale Benchmark Constituency Treebank for the Indonesian Language

**Ee Suan Lim[1*], Wei Qi Leong[1*], Ngan Thanh Nguyen[1], Dea Adhista[2], Wei Ming Kng[1], William Chandra Tjhi[1], Ayu Purwarianti[2,3]**

[1]AI Singapore, Singapore [2]Prosa.ai, Indonesia [3]Institut Teknologi Bandung, Indonesia

eesuanlim@gmail.com {weiqi, ngan, wtjhi}@aisingapore.org

dea.adhista@prosa.ai weiming.kng@gmail.com

ayu@itb.ac.id

## Abstract

Constituency parsing is an important task of informing how words are combined to form sentences. While constituency parsing in English has seen significant progress in the last few years, tools for constituency parsing in Indonesian remain few and far between. In this work, we publish ICON (Indonesian CONstituency treebank), the hitherto largest publicly-available manually-annotated benchmark constituency treebank for the Indonesian language with a size of 10,000 sentences and approximately 124,000 constituents and 182,000 tokens, which can support the training of state-of-the-art transformer-based models. We establish strong baselines on the ICON dataset using the Berkeley Neural Parser with transformer-based pre-trained embeddings, with the best performance of 88.85% F1 score coming from our own version of SpanBERT (IndoSpanBERT). We further analyze the predictions made by our best-performing model to reveal certain idiosyncrasies in the Indonesian language that pose challenges for constituency parsing.

## 1 Introduction

Constituency parsing is an important task of informing how words are combined to form sentences. It uses Context-Free Grammars (CFG) to assign a structure, usually in the form of a hierarchical syntactic parse tree, to a sentence. Parse trees can be used directly in applications such as grammar checking (Ng et al., 2013; Li et al., 2022) while linguistic features engineered through parsing can be used to boost the performance of downstream models for higher-level tasks such as semantic role labeling (Fei et al., 2021; Li et al., 2021), machine translation (Yang et al., 2020), natural language inference (Chen et al., 2017), opinion mining (Xia et al., 2021), text summarization (Xu and Durrett, 2019) and relation extraction (Jiang and Diesner, 2019).

There is another important family of grammar formalism called dependency grammar. While dependency parsing has become increasingly prevalent, this does not obviate the need for constituency parsing since the two can be used for different purposes. For span-labeling tasks such as coreference resolution, it has been argued that the explicit encoding of the boundaries of non-terminal phrases in constituency trees makes them more beneficial to the task than dependency trees (Jiang and Cohn, 2022).

The Indonesian language is the national and primary language of Indonesia, the world's fourth largest country by population at the time of writing with almost 275 million people (Aji et al., 2022). There has been mounting interest in the development of Indonesian natural language processing (NLP) although tools for constituency parsing remain few and far between. The progress in constituency parsing for the Indonesian language has been hampered by the absence of a large-scale benchmark dataset that can support the training of the current state-of-the-art (SOTA) transformer-based models, which have been pushing the envelope of English constituency parsing. In light of this, we introduce ICON (Indonesian CONstituency treebank), a 10,000-tree benchmark constituency parsing dataset for the Indonesian language. It is the hitherto largest publicly-available dataset for Indonesian constituency parsing. We also establish strong baselines on this treebank using the Berkeley Neural Parser (Kitaev and Klein, 2018) and a suite of pre-trained embeddings.

[*] Equal contribution



*Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 37 - 53

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 looks at the ICON treebank in more detail. Section 4 explains the experiments we ran on the treebank and Section 5 puts forward findings from our analyses and sheds light on the challenges in Indonesian constituency parsing. Lastly, in Section 6, we present our conclusions and lay out suggestions for future works.

## 2 Related work

### 2.1 Constituency parsing treebanks

The Penn Treebank (PTB) corpus (Marcus et al., 1993) is one of the most widely-used datasets in constituency parsing for English. It consists of over 40,000 sentences from Wall Street Journal articles and uses five clause-level, 21 phrase-level and 36 part-of-speech (POS) tags. Following the successes of the PTB in enabling the training of much more accurate English parsers than previously known ones, similar projects were initiated for other languages as well. Notably, a multilingual constituency treebank was prepared for the SPMRL 2013 Shared Task for syntactic parsing (Seddah et al., 2013), with treebanks in nine typologically-diverse languages, namely Swedish, German, French, Polish, Korean, Arabic, Hebrew, Hungarian and Basque.

While treebanks in some other languages are relatively large and cover a wide range of genres, publicly-available constituency treebanks for the Indonesian language are relatively small and domain specific (see Table 1). They are therefore not ideal for the training of end-to-end deep neural networks which most of the current SOTA models are based on.

| | Sentences | Tokens | Sources | Availability |
|---|---|---|---|---|
| INACL Treebank | 15,813 | Not available | English-translated sentences | Not available |
| IDN Treebank | 1,030 | 30,953 | Translated news from the PTB | https://github.com/famrashel/idn-treebank |
| Kethu Treebank | Same as IDN Treebank | Same as IDN Treebank | Same as IDN Treebank | https://github.com/ialfina/kethu/tree/master/kethu-2.0 |
| Cendana Treebank | 552 | 5,850 | Online chat data at Traveloka | https://github.com/davidmoeljadi/INDRA/tree/master/tsdb/gold/Cendana |
| JATI Treebank | 543 | 7,129 | Dictionary relevant to food and beverages | Not available |

Table 1: A comparison of size and sources of existing Indonesian constituency treebanks.

### 2.2 Constituency parsing models

Constituency parsing takes on two main approaches: chart-based and transition-based. There has only been a handful of papers on constituency parsing in Indonesian, and many of them took the transition-based approach. The first Indonesian constituency parser is a shift-reduce parser that utilizes an automatically-generated CFG from the treebank corpus, and it achieved an F1 score of 74.91% on the IDN treebank (Filino and Purwarianti, 2016). In a subsequent paper (Herlim and Purwarianti, 2018), another shift-reduce parser that uses beam search and structured learning was applied on the newer and larger INACL treebank but gave a lower F1 score of 50.3%. To enable a fair comparison with the first parser by Filino and Purwarianti (2016), this second shift-reduce parser was trained on the IDN treebank to give an F1 score of 74.0%. A more recent work (Arwidarasti et al., 2020) introduced an improved treebank called Kethu. The Kethu treebank resolved the compound-word problem in the IDN treebank and further adjusted the treebank to the PTB format. The Stanford CoreNLP transition-based parser (Manning et al., 2014), which employs beam search and global perceptron training, was trained on the Kethu treebank to give an F1 score of 69.97%.

We only know of one existing Indonesian constituency parser that uses the neural approach (Filino and Purwarianti, 2016). The first of two possible reasons for such a small number is that Indonesian transformer-based embeddings were previously not available. However, this has changed with the recent release of IndoBERTs (Koto et al., 2020; Wilie et al., 2020) and multilingual pre-trained language models like XLM-RoBERTa (Conneau et al., 2020) and mT5 (Xue et al., 2021). The latter have been shown to generalize well across natural language processing tasks (Devlin et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020). A second possible reason is that neural end-to-end models require a large amount of training data which existing Indonesian constituency treebanks were not able to supply. To overcome this, we built a new 10,000-tree constituency dataset which allowed us to achieve SOTA performance using neural architectures.

## 3 ICON Dataset

### 3.1 Data sources and annotation

ICON[1] is hitherto the largest publicly-available manually-annotated corpus for the task of constituency parsing in Indonesian. It contains 3,000 sentences from Indonesian Wikipedia and 7,000 sentences from news articles of various genres obtained from Tempo, an Indonesian news agency, spanning the period from 1971 to 2016. An example of a tree in the ICON dataset can be found in Figure 1.

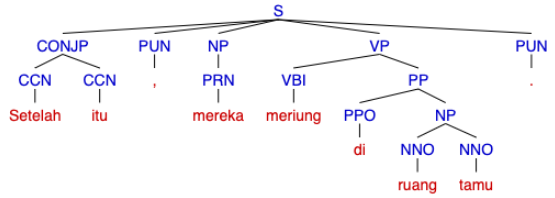


Figure 1: An example of a tree in the ICON dataset. The English equivalent of the parsed tree without POS tags would be: (S (CONJP After that) , (NP they) (VP gathered (PP in (NP the living room))) . )

The data was annotated by seven native Indonesian speakers, consisting of five annotators and two quality controllers. The annotators involved are undergraduates majoring in linguistics who have taken courses in syntax and semantics, while the quality controllers involved are linguistics graduates who have had more than two years of experience working in the field of NLP data annotation.

The annotation guidelines were formulated by the quality controllers using the PTB POS tagging (Santorini, 1990) and bracketing guidelines (Bies et al., 1995) for English as a reference with additional adaptations to account for the characteristics of the Indonesian language data. Thorough knowledge transfer sessions were then conducted by the quality controllers. Thereafter, annotators had to complete an assessment to evaluate their understanding of the guidelines. This feedback session allowed annotators to have a common understanding and deconflict any inter-annotator disagreements.

| Clause-level tag | Definition | Count |
|---|---|---|
| S | Main clause and complete clause with final intonation | 11,904 |
| SINV | Inverted clause | 1,288 |
| CP | All types of complementizer phrases and clauses | 4,057 |
| RPN | Relative clause | 3,977 |
| SBARQ | Complete interrogative clause | 64 |
| SQ | Yes-or-no question | 3 |

Table 2: Definition and count of clause-level tags.

| Phrase-level tag | Definition | Count |
|---|---|---|
| ADJP | Adjectival phrase | 3,035 |
| WHADJP | Adjectival phrase consisting of wh-premodifier and head is an adjective | 6 |
| ADVP | Adverbial phrase | 928 |
| WHADVP | Wh-adverbial phrase | 140 |
| CONJP | Conjunction spanning more than a single word | 243 |
| FRAG | Fragmented sentence | 77 |
| INTJ | Interjection | 103 |
| NP | Noun phrase | 55,736 |
| WHNP | Wh-noun phrase | 104 |
| PP | Prepositional phrase | 14,698 |
| WHPP | Wh-prepositional phrase | 8 |
| PNT | Parenthetical | 93 |
| QP | Quantifier phrase | 727 |
| UCP | Unlike coordinated phrase | 224 |
| VP | Verb phrase | 26,713 |

Table 3: Definition and count of phrase-level tags.

| POS tag | Definition | Count |
|---|---|---|
| NNO | Noun | 44,006 |
| NNP | Proper noun | 28,540 |
| PPO | Preposition | 14,233 |
| CSN | Subordinating conjunction | 3,123 |
| PRR | Relative pronoun | 3,979 |
| PRI | Interrogative pronoun | 143 |
| PRK | Clitic pronoun | 1,697 |
| PRN | Pronoun | 2,452 |
| VBI | Intransitive verb | 8,858 |
| VBT | Transitive verb | 6,292 |
| VBP | Passive verb | 4,954 |
| VBL | Linking verb (copula) | 966 |
| TAME | Tense, Aspect, Modality, Evidentiality marker | 2,859 |
| CCN | Coordinating conjunction | 5,082 |
| INT | Interjection | 103 |
| ADJ | Adjective | 6,588 |
| ADV | Adverb | 6,882 |
| NEG | Negation | 1,548 |
| NUM | Numeric value | 5,103 |
| KUA | Quantifier | 1,690 |
| ART | Article | 4,563 |
| PAR | Particle | 353 |
| SYM | Symbol | 374 |
| PUN | Punctuation | 27,727 |

Table 4: Definition and count of POS tags.

[1] https://github.com/aisingapore/seacorenlp-data/tree/main/id/constituency

| | Train | | Development | | Test | | Total |
|---|---|---|---|---|---|---|---|
| | Count | % | Count | % | Count | % | Count |
| Sentences | 8,000 | 80.00% | 1,000 | 10.00% | 1,000 | 10.00% | 10,000 |
| Tokens | 145,794 | 80.06% | 18,291 | 10.04% | 18,030 | 9.90% | 182,115 |
| Clause-level tags | 17,084 | 80.23% | 2,149 | 10.09% | 2,060 | 9.67% | 21,293 |
| Phrase-level tags | 82,357 | 80.09% | 10,349 | 10.06% | 10,129 | 9.85% | 102,835 |
| Word-level (POS) tags | 145,794 | 80.06% | 18,291 | 10.04% | 18,030 | 9.90% | 182,115 |
| | Avg (tokens) | | Avg (tokens) | | Avg (tokens) | | Avg (tokens) |
| Sentence length | 15.61 | | 15.71 | | 15.40 | | 15.43 |
| Tree depth | 8.47 | | 8.44 | | 8.38 | | 8.46 |

Table 5: Statistics of the ICON dataset.

### 3.2 Deviations from PTB guidelines

Although the annotation guidelines for ICON were based mainly on the PTB guidelines, there were several changes that were made in order to adapt them to the Indonesian language. These include changes to the POS and constituent tagsets as well as the handling of null elements and functional tags.

### 3.3 Dataset statistics and characteristics

The ICON dataset consists of six clause-level, 15 phrase-level and 24 POS tags (see Tables 2, 3 and 4) and is split into train, development and test sets using a 8:1:1 ratio (see Table 5 for the statistics for each split). The train, development and test sets were well stratified across the number of tokens, sentence length, tree depth, POS tag count and constituent label count. Distribution of the labels, tree depth and sentence length can be found in Appendix A.

### 3.4 Comparison with the Kethu treebank

The most recent Indonesian constituency parser (Arwidarasti et al., 2020) uses a treebank called Kethu. It is derived from the IDN treebank and is a publicly-available treebank which is not domain specific. There are differences between ICON and Kethu. First, their constituent and POS tagsets differ. The ICON treebank splits the SBAR label into CP and RPN while Kethu uses SBAR as per the PTB guidelines. ICON also uses CONJP whereas Kethu does not. Second, the Kethu treebank uses null elements and functional tags but the ICON treebank does not. Third, between the two, ICON, which is 9.7 times larger than Kethu, can better support the training of SOTA transformer-based models, which requires large amounts of data.

## 4 Training models with ICON

To establish a baseline on the ICON treebank, which could be used as a benchmark for future works on Indonesian constituency parsing, we trained the Berkeley Neural Parser (Kitaev and Klein, 2018) on the treebank with a suite of Indonesian and multilingual pre-trained language embeddings.

### 4.1 Model architecture

We chose the Berkeley Neural Parser (Kitaev and Klein, 2018) because it performed well for English constituency parsing on the PTB and achieved an F1 score of 95.1%. Also, the model architecture includes a POS tagger and does not require additional data like dependency treebanks to train model parameters.

Employing the chart-based method to constituency parsing, the encoder in the Berkeley Neural Parser (Kitaev and Klein, 2018) first takes in words in a sentence, embeds them by passing them through a pre-trained language model like BERT (Devlin et al., 2019) and transforms these representations using self-attention. The span vector is then constructed by subtracting the representation associated with the start of the span from the representation associated with the end of the span. The decoder part of the neural model consists of a span classifier that is used to give a score to the label in each span. To get the score for an entire parse tree, the scores of the constituent spans are summed up. Finally, a modified version of the Cocke–Younger–Kasami (CKY) algorithm (Kasami, 1965; Younger, 1967) searches over all possible trees to identify the highest-scoring tree for a given sentence.

### 4.2 Pre-trained language embeddings

**Indonesian embeddings.** In order to adapt the Berkeley Neural Parser to the Indonesian

language, we replaced the English embeddings with IndoBERT embeddings, which are Indonesian transformer-based embeddings found in the IndoLEM paper (Koto et al., 2020) and the IndoNLU paper (Wilie et al., 2020) (see Appendix B for more details).

Since the Berkeley Neural Parser looks at spans of text and it has been shown that SpanBERT (Joshi et al., 2020) produces superior results for span-based NLP tasks, we developed and added our very own version of Indonesian SpanBERT, called IndoSpanBERT, to the list of pre-trained embeddings to be used in our experiments. As the name suggests, SpanBERT focuses on spans–the Masked Language Modeling (MLM) objective of BERT is modified to mask random spans instead of random tokens. The model is then trained using span-boundary representations to predict the contents of the masked spans. We used the IndoLEM dataset (Koto et al., 2020) for pretraining and it was tokenized by IndoLEM's IndoBERT's WordPiece tokenizer. 16 A100 40GB GPUs were used for training with a maximum of 512 tokens. The base model was trained with a batch size of 8,192 and took 600,000 training steps (75 hours) to converge whereas the large model was trained with a batch size of 4,096 and took 280,000 steps (72 hours) to converge.

**Multilingual embeddings.** Multilingual masked language models have improved the state of many cross-lingual understanding tasks as well as natural language understanding tasks for each language (Devlin et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020). This is done by pre-training large Transformer models (Vaswani et al., 2017) on a single, multilingual corpus. Sub-word tokenizers like SentencePiece (Kudo and Richardson, 2018) enabled this process by facilitating the sharing of vocabulary learnt across various languages. The larger corpora used for training such models as compared to those used to train monolingual models have also contributed to the success of multilingual embeddings (Conneau et al., 2020). To see their effects on constituency parsing, we included XLM-RoBERTa (Conneau et al., 2020), BERT-Base Multilingual Uncased (Devlin et al., 2019), mT5 (Xue et al., 2021) and XGLM-1.7B (Lin et al., 2021) embeddings in our experiments (see Appendix C).

**English Embeddings.** We included English BERT embeddings (Devlin et al., 2019) in our experiments. The F1 score that can be achieved using English embeddings could be used as a baseline to compare against the F1 scores of models using Indonesian and multilingual embeddings.

### 4.3 Experiment results

We established strong baselines on the ICON treebank using the Berkeley Neural Parser (Kitaev and Klein, 2018) and a suite of pre-trained embeddings (see Table 6).

| Embedding | Language | Precision | Recall | F1 |
|---|---|---|---|---|
| **Base embeddings** | | | | |
| BERT | English | 83.67 | 83.79 | 83.73 |
| IndoLEM | Indonesian | 88.32 | **89.30** | 88.81 |
| IndoNLU | Indonesian | 86.97 | 87.90 | 87.43 |
| IndoSpanBERT | Indonesian | **88.52** | 89.19 | **88.85** |
| BERT-Base, Multilingual | Multilingual | 86.80 | 87.23 | 87.01 |
| mT5 | Multilingual | 86.81 | 88.64 | 87.71 |
| XGLM-1.7B | Multilingual | 84.81 | 85.04 | 84.92 |
| XLM-RoBERTa | Multilingual | 87.30 | 88.60 | 87.94 |
| **Large embeddings** | | | | |
| BERT | English | 83.81 | 84.22 | 84.01 |
| IndoNLU | Indonesian | 88.11 | **88.97** | **88.54** |
| IndoSpanBERT | Indonesian | 88.03 | **88.97** | 88.49 |
| mT5 | Multilingual | 88.18 | 88.77 | 88.47 |
| XLM-RoBERTa | Multilingual | **88.29** | 88.68 | 88.48 |

Table 6: Summary of experiment results.

IndoSpanBERT and IndoLEM gave comparable F1 scores on the test set of 88.85% and 88.81% respectively.

We used grid search to derive the optimum set of hyperparameters for the Berkeley Neural Parser using IndoSpanBERT and they are as follows: batch_size 32, learning_rate 0.00005, subbatch_max_tokens 1500, num_layers 8 and num_heads 8.

**Comparing against other Indonesian parsers.** For reference, prior works reported the following F1 scores when testing their parsers on their respective test sets: 74.91% (Filino and Purwarianti, 2016), 74.0% (Herlim and Purwarianti, 2018) and 69.97% (Arwidarasti et al., 2020). Since the test sets are different across the various parsers, it might not be very meaningful to compare F1 scores. We intend to perform a fairer comparison by comparing the performance of the parsers when used in a downstream task like machine translation (Meng et al., 2013; Ma et al., 2018; Yang et al., 2020), natural language

inference (Chen et al., 2017) or question answering (Zhu et al., 2022).

**Comparing across various embeddings.** Comparing the F1 scores across the various pre-trained language embeddings for the experiments we have conducted, we made the following observations, some of which merit further research and are beyond the scope of this paper.

Firstly, having IndoSpanBERT scoring the highest F1 score is in line with the English SpanBERT experiment findings (Joshi et al., 2020). This suggests that IndoSpanBERT could be used to improve the results of other Indonesian span-based tasks such as question answering, relation extraction and coreference resolution.

Secondly, the base and large versions of English BERT did not perform too badly despite being applied to Indonesian, which is from a different language family. The best Indonesian model (using IndoSpanBERT-base) achieved an F1 score of 88.85% whereas the English model (using English BERT-base) achieved an F1 score of 83.73%. This is certainly an interesting finding which could be explored further in future works.

Thirdly, when comparing across the base pre-trained embeddings, the monolingual Indonesian ones performed better than the multilingual ones. The larger Indonesian corpus used in multilingual pre-training as well as the transfer learning from other languages did not seem to benefit Indonesian constituency parsing. For example, multilingual mT5, which has the largest known number of Indonesian tokens (69 billion tokens) amongst all the pre-trained embeddings used in this paper, gave an F1 score of 87.71% whereas the model that used IndoLEM embeddings, which were pre-trained with just 220 million words, gave an F1 score of 88.81%.

## 5 Analysis

A breakdown of the performance of the best model (IndoSpanBERT-base) by constituent labeling and POS tagging can be found in Appendices D and E.

An in-depth error analysis of the predictions made by our trained parser revealed certain idiosyncrasies in the Indonesian language that pose challenges for constituency parsing. Word order is relatively flexible in Indonesian (Stack, 2005; Irmawati et al., 2017) despite the lack of morphological case markings. Furthermore, the fact that predicates in Indonesian are not only verbal, like in English, but can also be nominal, adjectival and prepositional (Sneddon et al., 2010), means that the CFG production rules are going to be much more diverse and difficult to predict for parsers. In addition, the presence of mechanisms such as topicalization as well as object voice (Arka and Manning, 1998; Sneddon et al., 2010; Djenar, 2018; Jeoung, 2020) allows verb-initial and verb-final word orders, even if the neutral word order of Indonesian is SVO (Donohue, 2007; Chung, 2008; Sneddon et al., 2010; Dryer, 2013). Other than these issues, we explore three additional problems in detail in the following sections–the ambiguity in POS in Indonesian, structural ambiguity in NPs with demonstratives as well as difficulties in parsing coordinated structures.

### 5.1 Ambiguity in POS

Categorial ambiguity is rife in Indonesian (Teeuw, 1962; Tjia, 2015), especially between adjectives and adverbs, verbs and adjectives, and prepositions and conjunctions. Depending on context, words such as *mau* and *suka* could be interpreted as auxiliaries or verbs or even both (Jeoung, 2020). We find that the parser, despite its excellent performance on POS tagging (with a F1 score of 95% and above for most categories), still falters on ADJ (86.36%), ADV (92.37%) and VBI (90.88%). This is further reflected in the low bracketing F1 scores for the ADJP (68.18%) and ADVP (71.06%) constituents. This is likely due to the fact that the parser cannot rely on morphology to distinguish reliably between categories. Certain adjectives can be used as adverbs without morphological changes (Sasangka et al., 2000; Sneddon et al., 2010), unlike in English where the suffix *-ly* can be used to distinguish ADV from ADJ. Furthermore, a single affix in Indonesian can be associated with different word classes (Sneddon et al., 2010; Mahdi, 2012; Denistia and Baayen, 2022) (see Examples 1, 2 and 3 (Sasangka et al., 2000; Sneddon et al., 2010) for the functions of *ke-/-an* circumfixation).

(1) Verb + ke-/-an → Verb/Noun
   a. Joni **kejatuhan** mangga.
      Joni **was fallen on** by a mango.
      (Passive voice/Perfective aspect)
   b. **Kejatuhan** Majapahit terjadi di awal abad ke 16.

The **fall** of Majapahit occurred in the early 16th century. (Noun formation)

(2) Adjective + ke-/-an → Adjective/Noun
   a. **Ketinggian** air mencapai satu meter.
      The water **level (height)** is up to one meter. (Abstract noun formation)
   b. Nadanya **ketinggian**. Aku tidak bisa menyanyikannya.
      The note is **too high**. I cannot sing it. (Excessive degree)

(3) Noun + ke-/-an → Noun/Adjective
   a. Jika memakai kebaya, Darni tampak sangat **keibuan**.
      When she wears a kebaya, Darni looks very **motherly**. (Adjective formation)
   b. Raja Mulawarman memerintah **Kerajaan** Hindu tertua di Indonesia.
      King Mulawarman ruled the oldest Hindu **kingdom** in Indonesia. (Noun formation)

Furthermore, there is also ambiguity between the categories of adjectives and verbs in Indonesian (Teeuw, 1962; Sasangka, 2000; Mahdi, 2012; Tjia, 2015). While literature on the subject has not gone as far as to argue for the absence of adjectives in Indonesian, as has been done for the Korean language (Kim, 2002), it has explored the notion that adjectives might be better viewed as stative verbs (Sneddon et al., 2010), a perspective that has been adopted by many a linguist for languages of Mainland Southeast Asia (MSEA), such as for the Kra-Dai languages (Pittayaporn, 2021) and Vietic languages (Alves, 2021). This ambiguity is in part due to the fact that both verbs and adjectives can be predicative in Indonesian, as well as the fact that certain affixes are common to both categories. For example, the prefixes *ter-* in terhormat, *me-* in menarik and *ber-* in berbahaya are commonly used to form both adjectives and verbs (Sasangka, 2000; Musgrave, 2013). In any case, for the initial version of the ICON dataset, we adopted the approach of distinguishing between the two categories by gradability (Keraf, 1984; Kridalaksana, 1986; Effendi, 1995). If a word is gradable, it is considered to be an adjective and not a verb.

### 5.2 Structural ambiguity in NPs with demonstratives

In Indonesian, demonstratives in a NP are preceded by all other constituents nested within the NP (Sneddon et al., 2010). This can cause structural ambiguity when there is more than one noun preceding the demonstrative or when a relative clause ending with a noun precedes the demonstrative (Sneddon et al., 2010). The demonstrative could be a modifier of the noun immediately preceding it or of the head of the entire NP (see Figure 2).

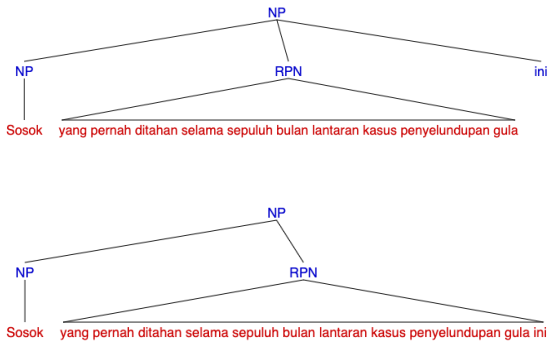


Figure 2: A case of demonstrative attachment ambiguity in which *ini* (this) can modify the head of the entire noun phrase (*Sosok*) or the NP immediately preceding it (*kasus penyelundupan gula*). POS tags and the internal structure of the relative clause have been hidden due to space constraints.

This ambiguity can usually be resolved with more discourse context (Hirst, 1984), but this is unfortunately not available in the ICON dataset (or in the Kethu dataset for that matter) since the text data comprises individual sentences that do not belong together in the same discourse. This makes it difficult even for a human annotator to decide on the most germane interpretation. A possible improvement to the dataset could therefore be to explore using entire documents for the text data, like in OntoNotes (Weischedel et al., 2013), instead of using unrelated sentences.

### 5.3 Challenges in parsing coordinated structures

Coordination has been mentioned in the literature as a major challenge in constituency parsing (Hogan, 2007; Maier et al., 2012), especially when unlike syntactic categories are involved (Prolo, 2006). We find that this is true for our model's performance on the ICON dataset as well, with a bracketing F1 score of 41.02% for UCP when evaluated on the validation dataset.

It is perhaps more complicated in Indonesian to determine the level of coordination between constituents, or indeed to determine whether there is even coordination in the first place, due to the tendency for coordinating conjunctions and even coordinating punctuations to be missing in coordinated structures. The fact that there are so many cross-categorial ambiguities (as explained in the preceding sections) and that predicates in Indonesian can be nominal, verbal, adjectival or even prepositional probably do not make this task any easier. In fact, we found that many of the UCP constituents were incorrectly annotated by the annotators due to the difficulty involved. These errors will be fixed in subsequent revisions of the treebank.

An interesting finding was that in cases where the model picked up on the coordination of unlike syntactic categories but failed to parse it as a UCP constituent, the label VP was predicted instead. While an investigation of the possible reasons behind this error, such as through an analysis of attention weights, is beyond the scope of this paper, we could venture a plausible preliminary hypothesis. As Prolo (2006) asserted, UCP coordination is not random, and coordination can only occur when two constituents fulfill the same grammatical function. It is therefore perhaps the case that when coordinating two unlike constituents which are predicative in nature (see Example 4), the model implicitly associates the coordinated structure with predication which is in turn associated with VPs given the central role of verbs in predication. This is in fact in line with suggestions in the literature to mix syntactic categories and grammatical function when dealing with UCPs (Prolo, 2006).

(4) (S (NP Tedi) (UCP (PP juga (PP di sana)) tapi (VP lolos)))
Tedi was there too but got away.

## 6 Conclusion

In conclusion, we have published ICON, the largest publicly-available manually-annotated benchmark constituency treebank for the Indonesian language with a size of 10,000 sentences and approximately 124,000 constituents and 182,000 tokens. As part of the process of building the treebank, we also re-evaluated and revamped the constituent tagset and POS tagset in use in existing treebanks to ensure that the labels are relevant and suitable for the grammatical features of the Indonesian language. In addition, we have established strong baselines on the ICON dataset using the Berkeley Neural Parser with transformer-based pre-trained embeddings, with our own IndoSpanBERT and the existing IndoLEM giving F1 scores of 88.85% and 88.81% respectively.

Moving forward, there are still certain parts of the treebank that can be improved or are worth a second look. Some possible aspects to be worked on are as follows:

1. The ambiguity between ADJ and VBI should probably be scrutinized more to arrive at a linguistically accurate rule for differentiating between the two classes.

2. SBARQ and SQ constituents are relatively lacking in the dataset (67 out of 21293 clause-level tags). In order to improve and allow for better evaluation of parsers' ability to parse questions, having more questions in the dataset might be beneficial.

Beyond improvements to the dataset, there are other research questions that could be explored as well:

1. How much do downstream tasks benefit from constituency parse trees in Indonesian? In what ways can we incorporate these syntactic features into models?

2. How much further could we push the performance of constituency parsers for the Indonesian language with other model architectures, such as using the label attention layer and head-driven phrase structure grammar (Mrini et al., 2020)?

We hope that this work will be an important catalyst for the development of better Indonesian constituency parsers and that it will enable research in linguistic phenomena and syntax-enhanced models for NLP in Indonesian.

## Acknowledgements

This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme. First and foremost, the authors would like to thank the annotation and quality control team at Prosa.ai, including Dea Adhista, Hanung Wahyuning Linuwih, Raydítya Brillian Prima, and Menik Lestari, for their professionalism and dedication in ensuring that the data is of good quality. Second, the authors would like to thank Lih Yan Wong for her contributions to the initial exploration stage of the experiments, Alvin Tan Pengshi for helping to preprocess the raw treebank data into the bracketed notation, and Haniah Wafa for helping to explore the possibility of comparing different parsers using a downstream machine translation task. Last but not least, the authors would also like to thank Datasaur.ai for providing the data annotation platform for the project.

## References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500.

Mark J. Alves. 2021. Typological profile of Vietic. In *The Languages and Linguistics of Mainland Southeast Asia: A comprehensive guide*. De Gruyter Mouton, Berlin, Boston, 469–498. https://doi.org/10.1515/9783110558142-022.

I Wayan Arka and Christopher D. Manning. 1998. Voice and grammatical relations in Indonesian: A new perspective. In *Proceedings of the LFG98 Conference*. CSLI Publications.

Jessica Arwidarasti, Ika Alfina, and Adila Krisnadhi. 2020. Adjusting Indonesian Multiword Expression Annotation to the Penn Treebank Format. In *2020 International Conference of Asian Language Processing (IALP)*. https://doi.org/10.1109/IALP51396.2020.9310479.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. *Technical Report*. University of Pennsylvania, Philadelphia, Pennsylvania.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1657–1668. https://doi.org/10.18653/v1/P17-1152.

Sandra Chung. 2008. Indonesian clause structure from an Austronesian perspective. *Lingua*, 118, 10 (2008), 1554–1582. https://doi.org/10.1016/j.lingua.2007.08.002.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.

Karlina Denistia and R. Harald Baayen. 2022. The morphology of Indonesian: Data and quantitative modeling. In *The Routledge Handbook of Asian Linguistics, 1st edition*. Routledge, London, United Kingdom, 605–634. https://doi.org/10.4324/9781003090205.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Volume 1: Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dwi Noverini Djenar. 2018. Constituent order and information structure in Indonesian discourse. In *Perspectives on information structure in Austronesian languages*. Language Science Press, Berlin, Germany, 177–205. https://doi.org/10.5281/zenodo.1402545.

Mark Donohue. 2007. Word order in Austronesian from north to south and west to east. *Linguistic Typology*, 11 (2007), 349–391. https://doi.org/10.1515/lingty.2007.026.

Matthew S. Dryer. 2013. Order of Subject, Object and Verb. In *Dryer, Matthew S. & Haspelmath, Martin (Eds.), The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81.

S. Effendi. 1995. Kata Sifat dan Kata Keterangan dalam Bahasa Indonesia. *Bahasa dan Sastra*, 12, 2 (1995), 1–53.

Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021. Better Combine Them Together! Integrating Syntactic Constituency and Dependency Representations for Semantic Role Labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 549–559. https://doi.org/10.18653/v1/2021.findings-acl.49.

Mario Filino and Ayu Purwarianti. 2016. Indonesian shift-reduce constituent parser. In *2016 International Conference on Data and Software Engineering (ICoDSE)*. 1-6. https://doi.org/10.1109/ICODSE.2016.7936118.

Robert Herlim and Ayu Purwarianti. 2018. Indonesian Shift-Reduce Constituency Parser Using Feature Templates & Beam Search Strategy. In *5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*. 54–59. https://doi.org/10.1109/ICAICTA.2018.8541292.

Graeme Hirst. 1984. A Semantic Process for Syntactic Disambiguation. In *Proceedings of the Fourth AAAI Conference on Artificial Intelligence (AAAI'84)*. AAAI, 148–152.

Deirdre Hogan. 2007. Coordinate Noun Phrase Disambiguation in a Generative Parsing Model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 680–687. https://aclanthology.org/P07-1086.

Budi Irmawati, Hiroyuki Shindo, and Yuji Matsumoto. 2017. A Dependency Annotation Scheme to Extract Syntactic Features in Indonesian Sentences. *International Journal of Technology*, 8, 5 (2017), 957–967. https://doi.org/10.14716/ijtech.v8i5.878.

Helen Jeoung. 2020. Categorial ambiguity in mau, suka, and other Indonesian predicates. *Language*, 96, 3 (2020), 157–172. https://doi.org/10.1353/lan.2020.0053.

Fan Jiang and Trevor Cohn. 2022. Incorporating Constituent Syntax for Coreference Resolution. *arXiv*. https://doi.org/10.48550/arXiv.2202.10710.

Ming Jiang and Jana Diesner. 2019. A Constituency Parsing Tree based Method for Relation Extraction from Abstracts of Scholarly Publications. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. Association for Computational Linguistics, 186–191. https://doi.org/10.18653/v1/D19-5323.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8 (2020), 64–77. https://doi.org/10.1162/tacl_a_00300.

Tadao Kasami. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. *Technical Report AFCLR-65-758*. Air Force Cambridge Research Laboratory, Bedford, MA.

Gorys Keraf. 1984. *Tatabahasa Indonesia*. Nusa Indah.

Min-joo Kim. 2002. Does Korean have adjectives? *MIT Working Papers in Linguistics*, 43, (2002), 71–89.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2676–2686. https://doi.org/10.48550/arXiv.1805.01052.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66.

Harimurti Kridalaksana. 1986. *Kelas Kata dalam Bahasa Indonesia*. Gramedia.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 66–71. https://doi.org/10.18653/v1/D18-2012.

Zuchao Li, Kevin Parnow, and Hai Zhao. 2022. Incorporating rich syntax information in Grammatical Error Correction. *Information Processing and Management*, 59, 3 (2022). https://doi.org/10.1016/j.ipm.2022.102891

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021. Syntax Role for Neural Semantic Role Labeling. *Computational Linguistics*, 47, 3 (2021), 529–574. https://doi.org/10.1162/coli_a_00408.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du,

Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot Learning with Multilingual Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2112.10668.

Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Tiejun Zhao, and Eiichiro Sumita. 2018. Forest-Based Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1253–1263. https://doi.org/10.18653/v1/P18-1116.

Waruno Mahdi. 2012. Distinguishing Cognate Homonyms in Indonesian. *Oceanic Linguistics*, 51, 2 (2012), 402–449.

Wolfgang Maier, Sandra Kübler, Erhard Hinrichs, and Julia Krivanek. 2012. Annotating Coordination in the Penn Treebank. In *Proceedings of the Sixth Linguistic Annotation Workshop*. Association for Computational Linguistics, 166–174. https://aclanthology.org/W12-3624.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 55–60. https://doi.org/10.3115/v1/P14-5010.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19, 2 (1993), 313–330. https://aclanthology.org/J93-2004.

Fandong Meng, Jun Xie, Linfeng Song, Yajuan Lü, and Qun Liu. 2013. Translation with Source Constituency and Dependency Trees. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1066–1076. https://aclanthology.org/D13-1108.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking Self-Attention: Towards Interpretability in Neural Parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 731–742. https://doi.org/10.18653/v1/2020.findings-emnlp.65.

Simon Musgrave. 2013. Functional categories in the syntax and semantics of Malay. In *Tense, aspect, mood, and evidentiality in languages of Indonesia*. PKBB Universitas Katolik Indonesia Atma Jaya, Jakarta, 135–152.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, 1–12. https://aclanthology.org/W13-3601.

Pittayawat Pittayaporn. 2021. Typological profile of Kra-Dai languages. In *The Languages and Linguistics of Mainland Southeast Asia: A comprehensive guide*. De Gruyter Mouton, Berlin, Boston, 433–468. https://doi.org/10.1515/9783110558142-021.

Carlos A. Prolo. 2006. Handling Unlike Coordinated Phrases in TAG by Mixing Syntactic Category and Grammatical Function. In *Proceedings of the 8th International Workshop on Tree Adjoining Grammar and Related Formalisms*. Association for Computational Linguistics, 137–140. https://aclanthology.org/W06-1520.

Beatrice Santorini. 1990. Part-of-speech Tagging Guidelines for the Penn Treebank Project. *Technical Report*. University of Pennsylvania, Philadelphia, Pennsylvania.

Sry Satriya Tjatur Wisnu Sasangka, Titik Indiyatini, and Nantje Harijati Widjaja. 2000. *Adjektiva dan Adverbia dalam Bahasa Indonesia*. Pusat Bahasa Departemen Pendidikan Nasional Jakarta.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2013. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*. Association for Computational Linguistics, 146–182. https://aclanthology.org/W13-4917.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar, 2nd edition*. Allen & Unwin.

Maggie Stack. 2005. Word Order and Intonation in Indonesian. In *Lexical Semantic Ontology Working*

*Papers in Linguistics 5: Proceedings of Workshop in General Linguistics*. Linguistics Student Organization, 168–182.

Alex Teeuw. 1962. Some problems in the study of word-classes in Bahasa Indonesia. *Lingua*, 11 (1962), 409–421. https://doi.org/10.1016/0024-3841(62)90050-5.

Johnny Tjia. 2015. Grammatical relations and grammatical categories in Malay; The Indonesian prefix meN- revisited. *Wacana*, 16, 1 (2015), 105–132.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. https://doi.org/10.48550/arXiv.1706.03762.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0. *Linguistic Data Consortium*. Retrieved from https://catalog.ldc.upenn.edu/LDC2013T19.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 843–857. https://aclanthology.org/2020.aacl-main.85.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 833–844. https://doi.org/10.18653/v1/D19-1077.

Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A Unified Span-Based Approach for Opinion Mining with Syntactic Constituents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1795–1804. https://doi.org/10.18653/v1/2021.naacl-main.144.

Jiacheng Xu and Greg Durrett. 2019. Neural Extractive Text Summarization with Syntactic Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3292–3303. https://doi.org/10.18653/v1/D19-1324.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.

Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou. 2020. Improving Neural Machine Translation with Soft Template Prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5979–5989. https://doi.org/10.18653/v1/2020.acl-main.531.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time $n^3$. Information and Control, 10, 2 (1967), 189–208. https://doi.org/10.1016/S0019-9958(67)80007-X.

Fangyi Zhu, Lok You Tan, See-Kiong Ng, and Stéphane Bressan. 2022. Syntax-Informed Question Answering with Heterogeneous Graph Transformer. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 17–31. https://doi.org/10.1007/978-3-031-12423-5_2.

# Appendices

## A Distribution of labels, tree depth and sentence length across splits

| | | Train | | Development | | Test | | Total |
|---|---|---|---|---|---|---|---|---|
| | | **Count** | **%** | **Count** | **%** | **Count** | **%** | **Count** |
| S | Main clause and complete clause with final intonation | 9,557 | 80.28% | 1,183 | 9.94% | 1,164 | 9.78% | 11,904 |
| SINV | Inverted clause | 1,042 | 80.90% | 126 | 9.78% | 120 | 9.32% | 1,288 |
| CP | All types of complementizer phrases and clauses | 3,238 | 79.81% | 427 | 10.53% | 392 | 9.66% | 4,057 |
| RPN | Relative clause | 3,193 | 80.29% | 407 | 10.23% | 377 | 9.48% | 3,977 |
| SBARQ | Complete interrogative clause | 51 | 79.69% | 6 | 9.38% | 7 | 10.94% | 64 |
| SQ | Yes-or-no question | 3 | 100.00% | 0 | 0.00% | 0 | 0.00% | 3 |

Table 7: Statistics of clause-level tags.

| | | Train | | Development | | Test | | Total |
|---|---|---|---|---|---|---|---|---|
| | | **Count** | **%** | **Count** | **%** | **Count** | **%** | **Count** |
| ADJP | Adjectival phrase | 2,429 | 80.03% | 284 | 9.36% | 322 | 10.61% | 3,035 |
| WHADJP | Adjectival phrase consisting of wh-premodifier and head is an adjective | 3 | 50.00% | 2 | 33.33% | 1 | 16.67% | 6 |
| ADVP | Adverbial phrase | 751 | 80.93% | 81 | 8.73% | 96 | 10.34% | 928 |
| WHADVP | Wh-adverbial phrase | 116 | 82.86% | 10 | 7.14% | 14 | 10.00% | 140 |
| CONJP | Conjunction spanning more than a single word | 192 | 79.01% | 19 | 7.82% | 32 | 13.17% | 243 |
| FRAG | Fragmented sentence | 63 | 81.82% | 6 | 7.79% | 8 | 10.39% | 77 |
| INTJ | Interjection | 85 | 82.52% | 10 | 9.71% | 8 | 7.77% | 103 |
| NP | Noun phrase | 4,4678 | 80.16% | 5,652 | 10.14% | 5,406 | 9.70% | 55,736 |
| WHNP | Wh-noun phrase | 80 | 76.92% | 9 | 8.65% | 15 | 14.42% | 104 |
| PP | Prepositional phrase | 1,1746 | 79.92% | 1,518 | 10.33% | 1,434 | 9.76% | 14,698 |
| WHPP | Wh-prepositional phrase | 2 | 25.00% | 1 | 12.50% | 5 | 62.50% | 8 |
| PNT | Parenthetical | 70 | 75.27% | 11 | 11.83% | 12 | 12.90% | 93 |
| QP | Quantifier phrase | 584 | 80.33% | 69 | 9.49% | 74 | 10.18% | 727 |
| UCP | Unlike coordinated phrase | 179 | 79.91% | 23 | 10.27% | 22 | 9.82% | 224 |
| VP | Verb phrase | 21,379 | 80.03% | 2,654 | 9.94% | 2,680 | 10.03% | 26,713 |

Table 8: Statistics of phrase-level tags.

| | | Train | | Development | | Test | | Total |
|---|---|---|---|---|---|---|---|---|
| | | **Count** | **%** | **Count** | **%** | **Count** | **%** | **Count** |
| NNO | Noun | 35,182 | 79.95% | 4,494 | 10.21% | 4,330 | 9.84% | 44,006 |
| NNP | Proper noun | 22,940 | 80.38% | 2,860 | 10.02% | 2,740 | 9.60% | 28,540 |
| PPO | Preposition | 11,369 | 79.88% | 1,469 | 10.32% | 1,395 | 9.80% | 14,233 |
| CSN | Subordinating conjunction | 2,500 | 80.05% | 324 | 10.37% | 299 | 9.57% | 3,123 |
| PRR | Relative pronoun | 3,187 | 80.10% | 416 | 10.45% | 376 | 9.45% | 3,979 |
| PRI | Interrogative pronoun | 108 | 75.52% | 14 | 9.79% | 21 | 14.69% | 143 |
| PRK | Clitic pronoun | 1,378 | 81.20% | 151 | 8.90% | 168 | 9.90% | 1,697 |
| PRN | Pronoun | 1,987 | 81.04% | 254 | 10.36% | 211 | 8.61% | 2,452 |
| VBI | Intransitive verb | 7,088 | 80.02% | 888 | 10.02% | 882 | 9.96% | 8,858 |
| VBT | Transitive verb | 5,033 | 79.99% | 624 | 9.92% | 635 | 10.09% | 6,292 |
| VBP | Passive verb | 3,969 | 80.12% | 510 | 10.29% | 475 | 9.59% | 4,954 |
| VBL | Linking verb (copula) | 777 | 80.43% | 109 | 11.28% | 80 | 8.28% | 966 |
| TAME | Tense, Aspect, Modality, Evidentiality marker | 2,267 | 79.29% | 284 | 9.93% | 308 | 10.77% | 2,859 |
| CCN | Coordinating conjunction | 4,038 | 79.46% | 509 | 10.02% | 535 | 10.53% | 5,082 |
| INT | Interjection | 86 | 83.50% | 9 | 8.74% | 8 | 7.77% | 103 |
| ADJ | Adjective | 5,296 | 80.39% | 640 | 9.71% | 652 | 9.90% | 6,588 |
| ADV | Adverb | 5,520 | 80.21% | 652 | 9.47% | 710 | 10.32% | 6,882 |
| NEG | Negation | 1,242 | 80.23% | 153 | 9.88% | 153 | 9.88% | 1,548 |
| NUM | Numeric value | 4,079 | 79.93% | 480 | 9.41% | 544 | 10.66% | 5,103 |
| KUA | Quantifier | 1,347 | 79.70% | 186 | 11.01% | 157 | 9.29% | 1,690 |
| ART | Article | 3,624 | 79.42% | 449 | 9.84% | 490 | 10.74% | 4,563 |
| PAR | Particle | 292 | 82.72% | 33 | 9.35% | 28 | 7.93% | 353 |
| SYM | Symbol | 291 | 77.81% | 36 | 9.63% | 47 | 12.57% | 374 |
| PUN | Punctuation | 22,194 | 80.04% | 2,747 | 9.91% | 2,786 | 10.05% | 27,727 |

Table 9: Statistics of POS tags.

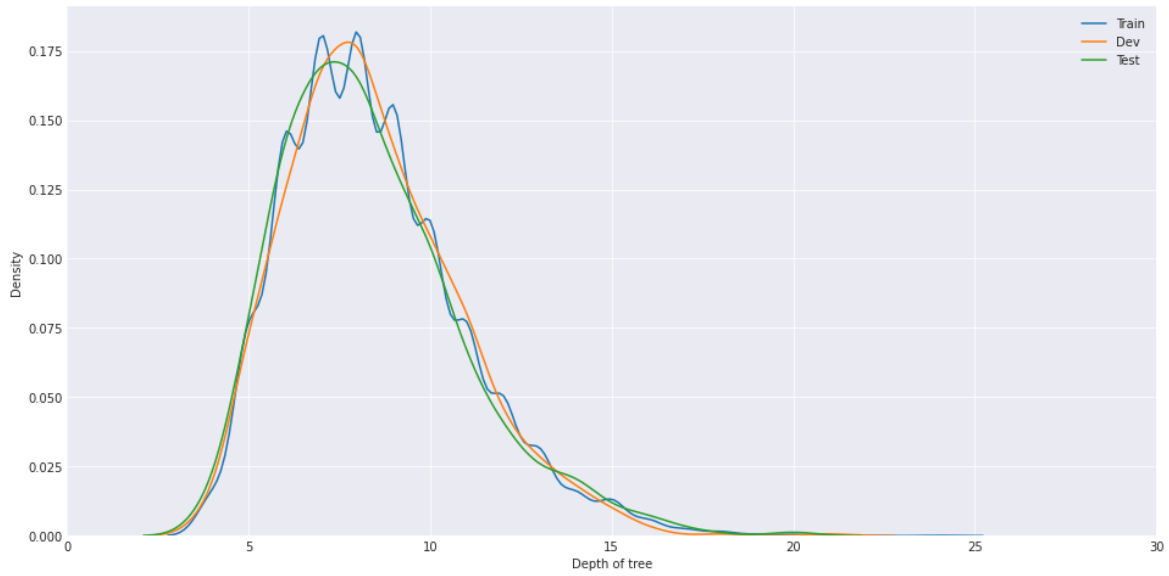


Figure 3: Distribution of tree depth in train, development and test sets.

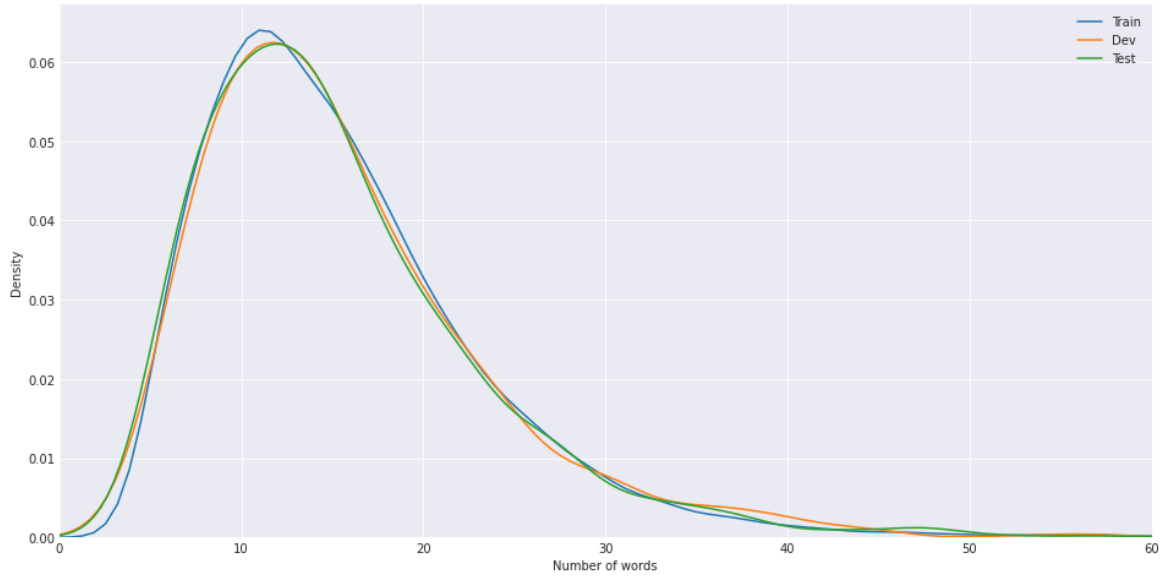


Figure 4: Distribution of sentence length in train, development and test sets.

## B A comparison of Indonesian embeddings

| | **IndoNLU IndoBERT (Wilie et al., 2020)** | **IndoLEM IndoBERT (Koto et al., 2020)** | **IndoSpanBERT (ours)** |
|---|---|---|---|
| Data sources | News, web corpus, Wikipedia, Twitter, etc. | News, web corpus, Wikipedia | Same as IndoLEM IndoBERT |
| Data size | 3.6B words (23GB) | 220M words (3.9GB) | |
| Tokenization algorithm | SentencePiece | WordPiece | |
| Vocabulary size | 30,522 | 31,923 | |
| Number of parameters | Base: 125M<br>Large: 335M | Base: 110M | Base: 108M<br>Large: 334M |

Table 10: Table comparing Indonesian pre-trained language embeddings used in our experiments. M stands for million, B stands for billion and GB stands for gigabytes.

## C A comparison of multilingual embeddings

| | **XLM-RoBERTa (Conneau et al., 2020)** | **BERT-Base, Multilingual Uncased (Devlin et al., 2019)** | **mT5 (Xue et al., 2021)** | **XGLM (Lin et al., 2021)** |
|---|---|---|---|---|
| Data sources | CC-100, a filtered version of CommonCrawl, covering 100 languages | Wikipedia, covering 102 languages | mC4, a version of CommonCrawl, covering 101 languages | A subset of CC100-XL, covering 68 CommonCrawl snapshots and 134 languages |
| Overall data size | Number of tokens not available (2.5TB) | Data size not available | 6.3T tokens (size not available) | 1.9T tokens (8.4TB) |
| Indonesian data size | 22.7B tokens (148.3GB) | Data size not available | 69B tokens (size not available) | 15B tokens (67.51GB) |
| Tokenization algorithm | SentencePiece | WordPiece | SentencePiece | SentencePiece |
| Vocabulary size | 250,000 | 110,000 | 250,000 | 250,000 |
| Number of parameters | Base: 270M<br>Large: 550M | Base: 120M | Base: 580M<br>Large: 1.2B | XGLM-1.7B: 1.7B |

Table 11: Table comparing multilingual pre-trained language embeddings used in our experiments. M stands for millions, B stands for billions, T stands for trillions, GB stands for gigabytes and TB stands for terabytes.

## D Model performance by constituent labeling

| Constituent | Count | Recall | Precision | F1 score |
|---|---|---|---|---|
| ADJP | 284 | 68.66 | 67.71 | 68.18 |
| ADVP | 81 | 66.67 | 76.06 | 71.06 |
| CONJP | 19 | 84.21 | 88.89 | 86.49 |
| CP | 427 | 89.23 | 85.62 | 87.39 |
| FRAG | 6 | 66.67 | 66.67 | 66.67 |
| INTJ | 10 | 80.00 | 72.73 | 76.19 |
| NP | 5,652 | 90.06 | 89.46 | 89.76 |
| PP | 1,518 | 92.09 | 90.66 | 91.37 |
| PRN | 11 | 72.73 | 100.00 | 84.21 |
| QP | 69 | 76.81 | 67.95 | 72.11 |
| RPN | 407 | 93.61 | 89.44 | 91.48 |
| S | 1,183 | 93.15 | 94.19 | 93.67 |
| SBARQ | 6 | 66.67 | 100.00 | 80.00 |
| SINV | 126 | 87.30 | 84.62 | 85.94 |
| UCP | 23 | 34.78 | 50.00 | 41.02 |
| VP | 2,654 | 90.99 | 89.44 | 90.21 |
| WHNP | 9 | 44.44 | 80.00 | 57.14 |

Table 12: Model performance by constituent labeling.

## E Model performance by POS tagging

| POS tag | Count | Recall | Precision | F1 score |
|---|---|---|---|---|
| ADJ | 640 | 86.56 | 86.16 | 86.36 |
| ADV | 652 | 92.79 | 91.95 | 92.37 |
| ART | 449 | 91.76 | 94.06 | 92.90 |
| CCN | 509 | 97.25 | 97.25 | 97.25 |
| CSN | 324 | 95.99 | 91.20 | 93.53 |
| INT | 9 | 88.89 | 72.73 | 80.00 |
| KUA | 186 | 94.62 | 93.62 | 94.12 |
| NEG | 153 | 99.35 | 98.06 | 98.70 |
| NNO | 4,494 | 95.68 | 96.22 | 95.95 |
| NNP | 2,860 | 96.43 | 95.43 | 95.93 |
| NUM | 480 | 96.67 | 97.27 | 96.97 |
| PAR | 33 | 96.97 | 91.43 | 94.12 |
| PPO | 1,469 | 98.16 | 98.97 | 98.56 |
| PRI | 14 | 92.86 | 100.00 | 96.30 |
| PRK | 151 | 90.07 | 83.95 | 86.90 |
| PRN | 254 | 96.06 | 95.69 | 95.87 |
| PRR | 416 | 98.56 | 99.03 | 98.79 |
| PUN | 2,747 | 99.93 | 99.89 | 99.91 |
| SYM | 36 | 88.89 | 88.89 | 88.89 |
| TAME | 284 | 98.94 | 98.94 | 98.94 |
| VBI | 888 | 89.19 | 92.63 | 90.88 |
| VBL | 109 | 100.00 | 100.00 | 100.00 |
| VBP | 510 | 98.24 | 97.08 | 97.66 |
| VBT | 624 | 94.87 | 94.27 | 94.57 |

Table 13: Model performance by POS tagging.