Building JATI: A Treebank for Indonesian

David Moeljadi
Nanyang Technological University, Singapore
D001@e.ntu.edu.sg

ABSTRACT
This paper introduces and describes the ongoing construction of a new lexical resource for Indonesian: the JATI treebank. It is being built from a subset of parsed dictionary definition sentences. The main data for this study comes from the fifth edition of Kamus Besar Bahasa Indonesia (KBBI) (Amalia, 2016), the official and the most comprehensive dictionary for the Indonesian language. The dictionary definition sentences are parsed using the Indonesian Resource Grammar (INDRA) (Moeljadi, Bond, and Song, 2015), a computational grammar for Indonesian in the Head-Driven Phrase Structure Grammar (HPSG) framework (Sag, Wasow, and Bender, 2003). JATI will be employed to build an ontology, in which knowledge is extracted from the semantic representation in Minimal Recursion Semantics (MRS) (Copestake et al., 2005).

Keywords: Indonesian; treebank; computational grammar; dictionary corpus; ontology

Introduction

A treebank is a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech level (Nivre, 2008). It is used for empirical linguistic research, as well as Natural Language Processing (NLP). It enables more precise queries and reduces the noise in the answer set. The treebank data can be used in qualitative research, such as finding an example of a certain linguistic construction or a counter-example to a claim about syntactic structure, as well as in quantitative research, as a source of information about frequencies and cooccurences. In addition, treebanks are indispensable in order to achieve robust broad-coverage parsing.

However, to date, there are few open-source treebanks for Indonesian, annotated with both syntactic and semantic information. One such is the Indonesian Treebank in ParGramBank (Sulger et al., 2013), a parallel treebank which is based on Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Dalrymple, 2001). Similar to the Indonesian Treebank in ParGramBank, the new treebank is being built based on a computational grammar for Indonesian. The motivation is to develop a broad-coverage grammar together with the treebank. Treebanking allows us to immediately identify problems in the grammar and improving the grammar directly improves the quality of the treebank (Oepen, Dan Flickinger, and Bond, 2004). The broad-coverage grammar being used to annotate syntactic and semantic structure in the new treebank is the Indonesian Resource Grammar (INDRA) (Moeljadi, Bond, and Song, 2015), a computational grammar for Indonesian in the Head-Driven Phrase Structure Grammar (HPSG) framework. It is being developed using open-source tools in the Deep Linguistic Processing with HPSG Initiative (DELPH-IN), a research collaboration between linguists and computer scientists which builds and develops open-source grammar and tools for grammar development and NLP applications using HPSG and Minimal Recursion Semantics (MRS). The DELPH-IN research community has built one treebank for English, called The LinGO Redwoods Treebank (Oepen, Dan Flickinger, Toutanova, et al., 2002), and one treebank for Japanese, called Hinoki (Bond et al., 2004).[1] The new Indonesian treebank introduced and described in this paper is named JATI, the Indonesian word for "teak", which is the national tree of Indonesia.[2]

[1] The DELPH-IN tradition is to name the treebank after a species of tree (Francis Bond p.c.).
[2] https://en.wikipedia.org/wiki/List_of_national_trees

A subset of the fifth edition of Kamus Besar Bahasa Indonesia (KBBI) dictionary database (Moeljadi, Kamajaya, and Amalia, 2017) is employed as the corpus for JATI. The fifth edition of KBBI (Amalia, 2016) is the latest edition of the official and the most comprehensive dictionary of the Indonesian language. It is a machine-tractable dictionary consisting of more than 100,000 entries with more than 120,000 definitions (as of 15 June 2017). The KBBI entries and definitions related to food and beverages were extracted and made as a corpus for JATI.

In this paper I present the details of the treebank construction. Section 2 mentions some related work in treebanks, especially for Indonesian. Section 3 contains a brief explanation about the source of the corpus, i.e. the KBBI dictionary data. Section 4 describes INDRA, the computational grammar used to annotate JATI. Section 5 explains the method for building JATI from the parsed KBBI dictionary definitions. Finally, Section 6 summarizes and outlines the future research.

Related Work

Most previous work on Indonesian treebanks focuses on syntactic annotation, rather than semantic annotation, for example: the Indonesian Dependency Treebank developed by Charles University in Prague (Green, Larasati, and Žabokrtský, 2012), a treebank with manually annotated dependency structures for Indonesian; the Indonesian treebank developed by the Faculty of Computer Science of Universitas Indonesia (Dinakaramani et al., 2016) which uses a part-of-speech (POS) tagged corpus as a starting point and adopts Penn Treebank bracketing guidelines; and the Indonesian treebank in the Asian Language Treebank (ALT) which was built by the Agency for the Assessment and Application of Technology (BPPT) (Riza et al., 2016), comprises about 20,000 sentences originally sampled from the English Wikinews in 2014, and uses tools such as POS tagger, syntax tree generator, shallow parser, and word alignment.

The Indonesian Treebank in the ParGram Parallel Treebank (ParGramBank) is based on a not-open-source LFG computational grammar called "IndoGram" and covers a diverse set of phenomena, implemented using XLE that includes a parser, a generator, and a transfer system. It is publicly available via the INESS treebanking environment[3] (Sulger et al., 2013) and contains 79 sentences and 433 words. The building of JATI is similar to this in terms of syntactic and semantic annotation using a computational grammar for Indonesian. However, the annotation process and the tools used are similar to the ones employed in the LinGO Redwoods Treebank of English (Oepen, Dan Flickinger, Toutanova, et al., 2002) and Hinoki (Bond et al., 2004), in which utterances are parsed and the annotator selects the best parse from the full analyses derived by the grammar. Moreover, JATI uses an open-source HPSG computational grammar where the syntax and semantics are represented in the same structure, same as in LinGO Redwoods Treebank and Hinoki.

The corpus: Kamus Besar Bahasa Indonesia (KBBI)

The main corpus data for JATI comes from the fifth edition of KBBI, the official Indonesian dictionary published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency) or Badan Bahasa under Ministry of Education and Culture, Republic of Indonesia (Amalia, 2016). The KBBI database is a machine-tractable dictionary (Moeljadi, Kamajaya, and Amalia, 2017) and contains 108,240 entries, 126,643 definitions, and 29,260 examples (as of 15 June 2017).[4]

KBBI is the most comprehensive Indonesian dictionary and contains a wealth of linguistic information, biodiversity (flora and fauna), and cultural diversity of Indonesia, which consists of various domains or fields such as food, clothing, and weapons. For the purpose of building the initial stage of JATI, only a subset of the KBBI database containing entries related to a certain domain, i.e. food and beverages, is extracted by looking up definitions started with genus terms such as makanan "food", masakan "dish", kue "snack", and minuman "drink". By restricting to a certain domain, an ontology can be relatively easy to obtain. The definition sentences of the extracted entries are semi-automatically edited

[3] http://clarino.uib.no/iness

[4] https://kbbi.kemdikbud.go.id/Beranda/Statistik

and rewritten to make them more consistent. An example is given in Table 1. Definition sentences in KBBI paper dictionary use abbreviations such as yg, dr, and dng. These abbreviations can be automatically changed to their full forms: yang "rel", dari "from", and dengan "with".[5] The second example in Table 1 shows the omission of the relative clause marker yang. To make consistent, the relative clause marker should be manually added. The definition of the entry bakso "meatball" in the third example uses terbuat instead of dibuat and there is no conjunction which connects the ingredients. To make consistent, terbuat should be changed to dibuat and a conjunction atau "or" should be added. As of 15 June 2017, the total number of sentences is 543, containing 7,129 words (1,086 unique words). The average length is 13.1 words per sentence. Around 1,500 more sentences will be extracted in order to build JATI.

Table 1: Some KBBI definition sentences, before and after editing

Before	After
minuman keras *yg* dibuat *dr* nira *yg* telah disuling	minuman keras *yang* dibuat *dari* nira *yang* telah disuling
kue kering, dibuat *dr* sagu dan dibungkus *dng* daun nipah	kue kering *yang* dibuat *dari* sagu dan dibungkus *dengan* daun nipah
makanan *terbuat dr* daging, udang, ikan *yg* dicincang	makanan *yang dibuat dari* daging, udang, *atau* ikan *yang* dicincang

Comparing with other commonly used text for corpora such as newspaper text, dictionary sentences are shorter, contain more fragments, especially noun phrases as single utterances, and fewer quoted sentences and proper names. However, they are valid examples of naturally occurring texts and a native speaker can read and understand them without special training (Fujita et al., 2007).

The parser: Indonesian Resource Grammar (INDRA)

INDRA is an open-source computational grammar which can parse and generate Indonesian text (Moeljadi, Bond, and Song, 2015). It is being developed using open-source tools in the Deep Linguistic Processing with HPSG Initiative (DELPH-IN).[6] Its development is documented in the DELPH-IN page.[7] The present state of INDRA is available to be examined and can be downloaded from GitHub.[8]

The analysis of the Indonesian grammar in INDRA uses the theoretical framework of Head-Driven Phrase Structure Grammar (HPSG) (Sag, Wasow, and Bender, 2003), a unification- and constraint-based context free grammar (or phrase structure grammar) which consists of a set of rules and a lexicon of symbols (parts-of-speech) and words, with the idea that the lexical head of each constituent or the grammatically most important word in a phrase is passed up the parse tree (Jurafsky and Martin, 2009). HPSG is surface oriented, positing no additional abstract structures, and mono-stratal, in which orthography, syntax, semantics are all handled in a single structure or the sign, the primary or elementary unit modeled through *typed feature structures*. Feature structures are sets of feature or attribute and value pairs which represent objects. Lexical entries, lexical rules, and phrase structure rules are all feature structures. As of 15 June 2017, INDRA contains 1,885 types, 15,099 lexical items, and 38 rules. INDRA is being extended in order to parse the dictionary definition sentences by adding the defining vocabulary, some new rules, and lexical types. The lexical acquisition process will be done semi-automatically: the defining vocabulary not yet included in INDRA will be automatically extracted and manually checked and inputted.

[5] Another approach would be to add the abbreviations to INDRA's lexicon with the same predicate as the full form (Francis Bond p.c.).

[6] http://www.delph-in.net/wiki/index.php/Home

[7] http://moin.delph-in.net/IndraTop

[8] https://github.com/davidmoeljadi/INDRA

INDRA uses Minimal Recursion Semantics (MRS) (Copestake et al., 2005) as its semantic framework, which is adaptable for HPSG typed feature structure and suitable for parsing and generation. The semantic structures in MRS are underspecified for scope and thus suitable for representing ambiguous scoping. The primary interest is in finding the correct lexemes and the relationships between them that are licensed by the syntax. This semantic framework is important for semantic annotation in JATI. By using INDRA, syntactic and semantic structure can be simultaneously annotated without overburdening the annotator.

INDRA uses open-source tools for grammar development, provided by the DELPH-IN community. Especially for treebanking, INDRA uses ITSDB or [incr tsdb()] (pronounced *tee ess dee bee plus plus*) (Oepen and Daniel Flickinger, 1998) and Full Forest Treebanker (FFTB) (Packard, 2014). ITSDB is a tool for testing, profiling the performance of the grammar (analyzing the coverage and performance) and treebanking. ITSDB stores all its information in plain text files. Thus, it is possible to make notational changes in the grammar and then apply the changes directly to the database with any tools that manipulate text. FFTB allows the selection of an arbitrary tree from the "full forest" without enumerating/unpacking all analyses in the parsing stage. It is partly integrated with the ITSDB.

Treebank development

Treebanking is a part of the grammar development process (Bender, Dan Flickinger, and Oepen, 2011). In the first stage, a collection of sentences called test-suite is developed. JATI uses a natural test-suite from KBBI, as mentioned in Section 3. Linguistic phenomena in the test-suite are then identified and analyzed based on reference grammars and other linguistics literature. The analyses are modeled in HPSG and implemented in INDRA. Afterwards, the grammar is compiled and tested by parsing the sentences in test-suite. The grammar is debugged if some gaps or problems are found according to the parse results until both the new and the previous phenomena are covered correctly. Afterwards, the sentences in test-suite are parsed again and treebanked. If a sentence is ungrammatical, no parse tree will be found. However, if a sentence is grammatical and no correct tree is found, all the possible trees should be rejected and the grammar has to be modified or debugged. Sentences for which no analysis had been implemented in the grammar or which fail to parse are left unannotated. The test-suite can be extended if needed. This process goes repetitively, as shown in Figure 1.

The construction of the treebank is a two stage process. First, the corpus is parsed using INDRA and then the annotator selects the correct analysis (or rejects all analyses) using FFTB. Selection is done through a choice of discriminants. The system selects features that distinguish between different parsers and the annotator selects or rejects the features until only one parse is left. The choices made by the annotators are saved and thus, it is possible to update the treebank when the grammar changes (Oepen, Dan Flickinger, and Bond, 2004). Figure 2 shows the FFTB page for JATI using a KBBI test-suite. Using FFTB, we can note some interesting findings or linguistic analyses item by item.

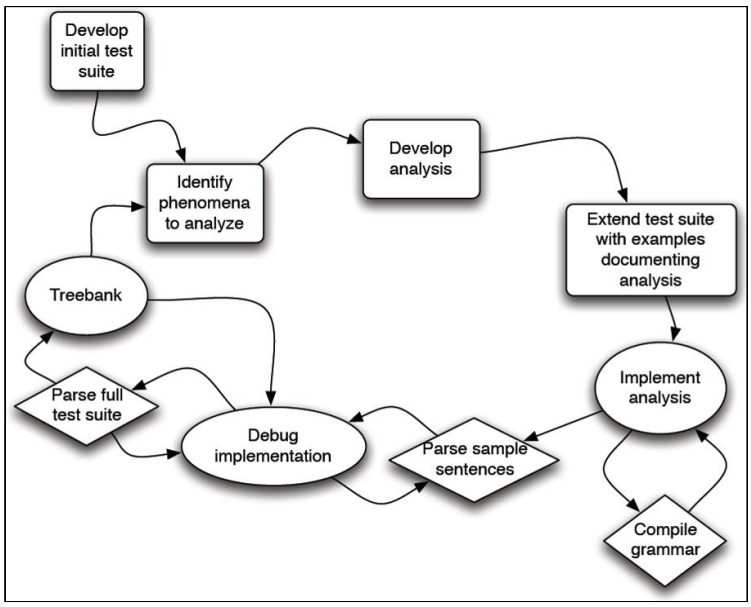

Fig. 1: The process of grammar development (Bender, Dan Flickinger, and Oepen, 2011)

Fig. 2: Screenshot of FFTB main page for JATI

Figure 3 shows the annotation page for the third sentence in the KBBI test-suite *makanan yang dibuat dari daging, udang, atau ikan yang dicincang* "food made from minced meat, shrimp, or fish". This sentence is syntactically ambiguous: the first interpretation is that the relative clause *yang dicincang* "minced" modifies *ikan* "fish", the second interpretation is that it modifies *daging* "meat", *udang* "shrimp", and *ikan* "fish". This sentence is the definition sentence for *bakso* "meatball" and the second interpretation is the correct one, illustrated as a phrase structure tree in the annotation result (see Figure 4). After the correct parse tree is selected, the annotator selects "accept" to save the result and continues annotating the next sentence. The annotator may select "list" to go back to the FFTB main page or "exit"

to exit the FFTB. If "Show MRS" is selected, the MRS semantic representation will be shown (see Figure 5).

Fig. 3: Screenshot of the annotation page for sentence number 3 in the KBBI test-suite

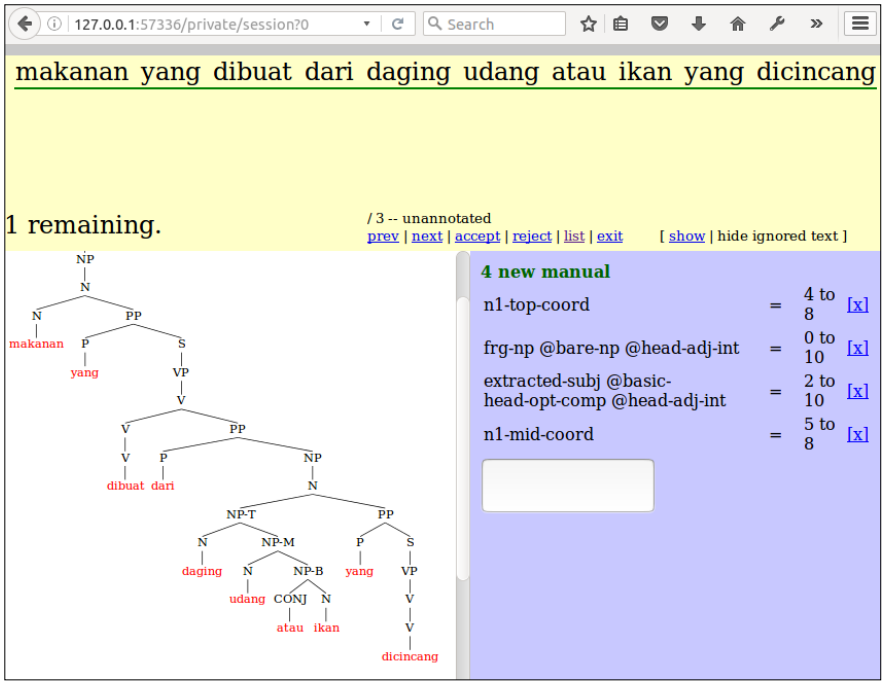

Fig. 4: Screenshot of the annotation result with parse tree for sentence number 3 in the KBBI test-suite

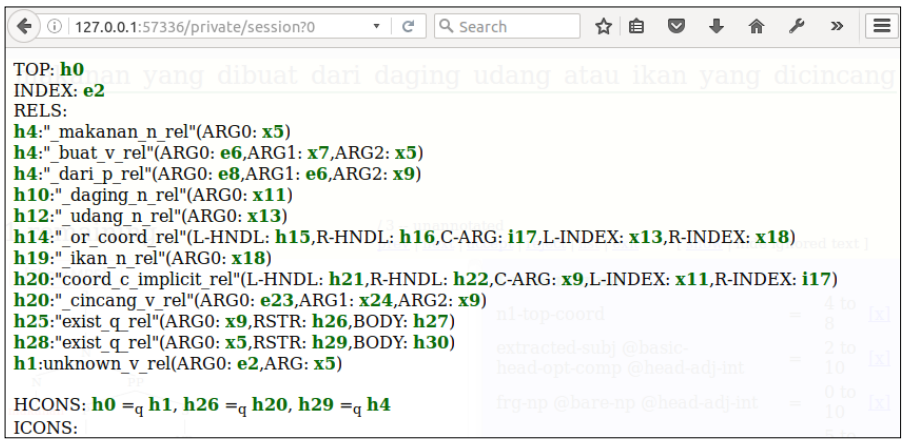

Fig. 5: Screenshot of the MRS semantic representation of the annotation result

At this initial stage, testing INDRA on the set of 543 defining sentences from KBBI gave a coverage of 17.1% (93 out of 543 sentences could be parsed). As stated in Bond et al. (2004), the first big increase in coverage in Hinoki (from the initial coverage of around 10% to 55%) came from automatically expanding the lexicon and with the addition of some new rules, the coverage increased to over 70%. Thus, lexical acquisition and rule addition are important to increase the coverage. The next work is to expand the lexicon and add more rules to INDRA.

To simplify grammar development, a snapshot of the grammar used to treebank in each development cycle will be taken. From this, information about lexical items and their types from both the grammar and treebank can be extracted and converted into an electronically accessible structured database (the lexical-type database) (Hashimoto et al., 2005). This allows grammar developers and treebankers to see comprehensive up-to-date information about lexical types, including documentation, syntactic properties, usage examples from the treebank, and links to other dictionaries.

Summary and Future Work

This paper introduces the building of JATI treebank in its initial stage. The main source of data is from a subset of KBBI definition sentences. The sentences are parsed using a computational grammar called INDRA. The use of the grammar as a base enforces consistency, i.e. all sentences annotated have well-formed parses. Besides INDRA, this treebank construction is supported by the availability of a wide range of open-source tools from the DELPH-IN. The immediate work is to expand the lexicon and refine the analyses by improving INDRA, adding more rules for the uncovered linguistic phenomena in the definition sentences, in order to increase the grammatical coverage.

Well-documented annotation guidelines should be written in order to ensure consistency, i.e. the same or similar linguistic phenomena are annotated in the same or similar ways throughout the corpus, as noted in Nivre (2008). The guidelines can also serve as a source of information for future users of JATI. In the future, JATI will be used for training a statistical model in order to parse other definition sentences and extracting an ontology.

Acknowledgements

Thanks to Francis Bond for his inspiration and advice to build JATI. Thanks to Dora Amalia who gave permission to use a part of the fifth edition of KBBI data.

References

Amalia, Dora, ed. 2016. *Kamus Besar Bahasa Indonesia*. 5th ed. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.

Bender, Emily M., Dan Flickinger, and Stephan Oepen 2011. "Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis". In: *Language from a Cognitive Perspective: Grammar, Usage and Processing*. Stanford: CSLI Publications, pp. 5–29.

Bond, Francis et al. 2004. "The Hinoki Treebank: A Treebank for Text Understanding". In: *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Springer Verlag Lecture Notes in Computer Science, pp. 158–167.

Copestake, Ann et al. 2005. "Minimal Recursion Semantics: An Introduction". In: *Research on Language and Computation* 3.4, pp. 281–332.

Dalrymple, Mary. 2001. "Lexical-Functional Grammar". In: *Syntax and Semantics* 34. Academic Press.

Dinakaramani, Arawinda et al. 2016. "Developing (and Utilizing) an Indonesian Treebank". In: *The Second Wordnet Bahasa Workshop*. Nanyang Technological University, Singapore.

Fujita, Sanae et al. 2007. "Exploiting Semantic Information for HPSG Parse Selection". In: *Proceedings of the Workshop on Deep Linguistic Processing*. Association for Computational Linguistics, pp. 25–32.

Green, Nathan, Septina Dian Larasati, and Zdenek Žabokrtský 2012. "Indonesian Dependency Treebank: Annotation and Parsing". In: *26th Pacific Asia Conference on Language, Information and Computation*, pp. 137–145.

Hashimoto, Chikara et al. 2005. "Integration of a lexical type database with a linguistically interpreted corpus". In: *6th International Workshop on Linguistically Integrated Corpora (LINC-2005)*. Cheju, Korea, pp. 31–40.

Jurafsky, Daniel and James H. Martin 2009. *Speech and Language Processing*. 2nd ed. New Jersey: Pearson Education, Inc.

Kaplan, Ronald and Joan Bresnan. 1982. "Lexical Functional Grammar: A formal system for grammatical representation". In: *The Mental Representation of Grammatical Relations*. Cambridge: the MIT Press, pp. 173–281.

Moeljadi, David, Francis Bond, and Sanghoun Song. 2015. "Building an HPSG-based Indonesian Resource Grammar (INDRA)". In: *Proceedings of the GEAF Workshop, ACL 2015*, pp. 9–16. url: http://aclweb.org/anthology/W/W15/W15-3302.pdf.

Moeljadi, David, Ian Kamajaya, and Dora Amalia 2017. "Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications". In: *Proceedings of the 11th International Conference of the Asian Association for Lexicography*. Ed. by Hai Xu. the Asian Association for Lexicography. Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, pp. 64–80.

Nivre, Joakim. 2008. "Treebanks". In: Corpus Linguistics: *An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. Berlin: Walter de Gruyter. Chap. 13, pp. 225–241.

Oepen, Stephan, Dan Flickinger, and Francis Bond. 2004. "Towards Holistic Grammar Engineering and Testing–Grafting Treebank Maintenance into the grammar revision cycle". *In: Beyond Shallow Analyses–Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*. Hainan Island.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova, et al. 2002. "LinGO Redwoods: A Rich and Dynamic Treebank for HPSG". In: *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria.

Oepen, Stephan and Daniel Flickinger. 1998. "Towards systematic grammar profiling: Test suite technology ten years after". In: *Journal of Computer Speech and Language*. Vol. 12. 4, pp. 411–436.

Packard, Woodley. 2014. *FFTB: the full forest treebanker*. url: http://moin.delphin.net/FftbTop (visited on 04/24/2015).

Riza, Hammam et al. 2016. "Introduction of the Asian Language Treebank". In: *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*. IEEE, pp. 1–6.

Sag, Ivan A., Thomas Wasow, and Emily M. Bender 2003. *Syntactic Theory: A Formal Introduction.* 2nd ed. Stanford: CSLI Publications.

Sulger, Sebastian et al. 2013. "ParGramBank: The ParGram Parallel Treebank." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Sofia, Bulgaria, pp. 550–560.