Developing (and utilizing) an Indonesian Treebank

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi,
Bayu Distiawan, and Ruli Manurung

Faculty of Computer Science, Universitas Indonesia

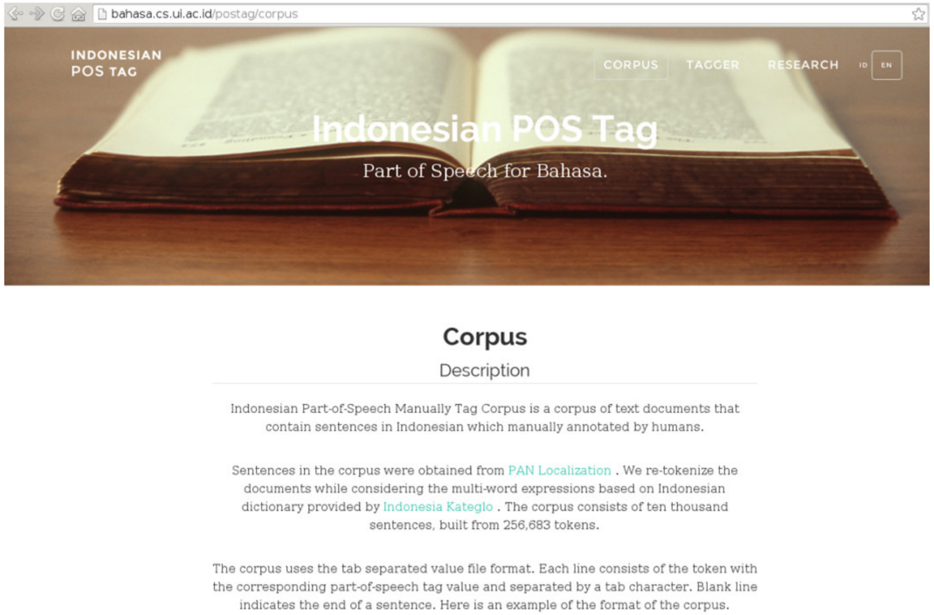
The Second Wordnet Bahasa Workshop

Nanyang Technological University, 15-16 January 2016

Outline

- **Background**
- Annotation process
- Outputs
- Making use of the treebank

At the previous workshop…



Indonesian POS Tag

Part of Speech for Bahasa.

Corpus

Description

Indonesian Part-of-Speech Manually Tag Corpus is a corpus of text documents that contain sentences in Indonesian which manually annotated by humans.

Sentences in the corpus were obtained from PAN Localization . We re-tokenize the documents while considering the multi-word expressions based on Indonesian dictionary provided by Indonesia Kateglo . The corpus consists of ten thousand sentences, built from 256,683 tokens.

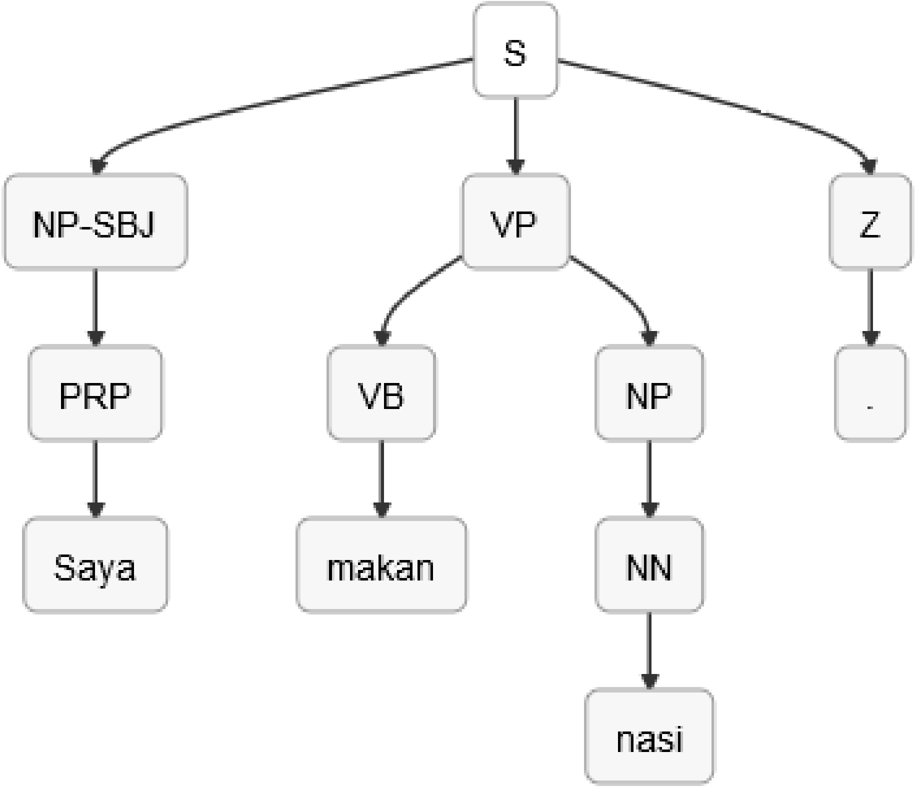
The corpus uses the tab separated value file format. Each line consists of the token with the corresponding part-of-speech tag value and separated by a tab character. Blank line indicates the end of a sentence. Here is an example of the format of the corpus.

- 10k Indonesian sentences from the PAN Localization parallel corpus (http://www.panl10n.net/indonesia)
- 23 POS tagset
- +/- 250k tokens (incl. MWE from http://kateglo.com)
- Rule-based tagger (utilizes MorphInd: http://septinalarasati.com/work/morphind)
- Released under Creative Commons BY-NC-SA 4.0

http://bahasa.cs.ui.ac.id/postag
https://github.com/famrashel/idn-tagged-corpus
https://github.com/andryluthfi/indonesian-postag

Next goal: building a treebank

- A **treebank** is a corpus of sentences complete with annotated syntactic structure.
- Useful as training data for statistical parsers.
- Example:



Bracketing Guidelines

- Our goal: treebank of the first 1000 sentences of the POS tagged corpus.
- Use POS tags as a starting point.
- Adopt Penn Treebank bracketing guidelines (Bies et al., 1995) where possible.
- Consult authoritative Indonesian grammar references (Alwi et al., 2003; Sneddon et al., 2010).

Outline

- Background
- **Annotation process**
- Outputs
- Making use of the treebank

Data preparation

- Convert from POS tagged corpus format to initial bracketing (forest of singleton POS tag trees).
- Example:

Pembahasan tadi masih dalam tahap awal.
discussion previous still in stage early

```
Pembahasan    NN
tadi          PR
masih         MD
dalam         IN
tahap         NN
awal          NN
.             Z
```

into bracketed file format:

```
(NN (Pembahasan))(PR (tadi))(MD (masih))(IN
(dalam))(NN  (tahap))(NN (awal))(Z (.))
```

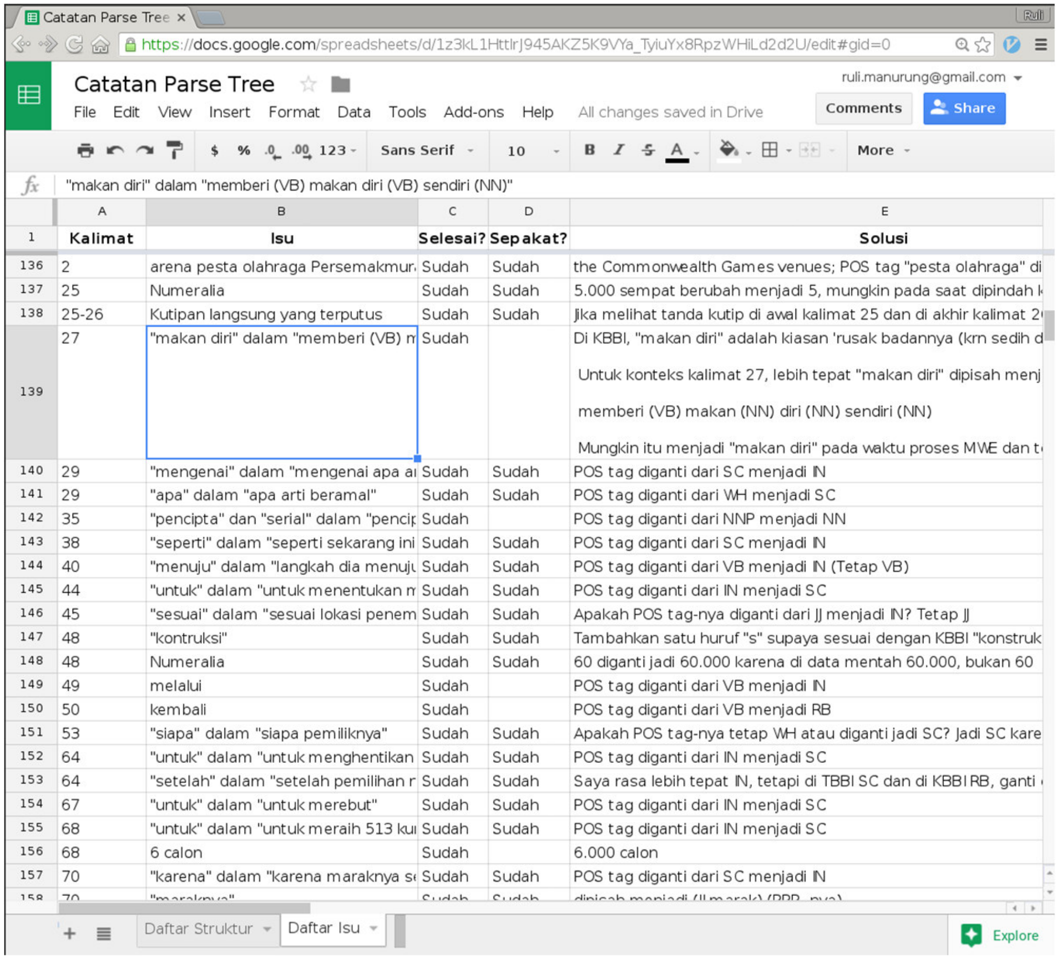
Annotation Process

- 3 annotators parsed the first 100 sentences of our corpus.
- In conjunction with development of bracketing guidelines.
- Sample:

```
(S  (PP-TMP Selama
       (NP bertahun-tahun))
    (NP-SBJ monyet)
    (VP mengganggu
       (NP warga Delhi))
    .)
```

- Keep track of all arising issues, resolve among annotators.
  - Consistent phrase structure bracketing
  - Sentence alignment (split & merge)
  - Incorrect POS tags
  - MWE

Notes of issues


Catatan Parse Tree
https://docs.google.com/spreadsheets/d/1z3kL1HttlrJ945AKZ5K9VYa_TyiuYx8RpzWHiLd2d2U/edit#gid=0
Catatan Parse Tree
ruli.manurung@gmail.com
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive
Comments
Share
Sans Serif
10
More
Daftar Struktur
Daftar Isu
Explore

fx "makan diri" dalam "memberi (VB) makan diri (VB) sendiri (NN)"

	A	B	C	D	E
1	Kalimat	Isu	Selesai?	Sepakat?	Solusi
136	2	arena pesta olahraga Persemakmur	Sudah	Sudah	the Commonwealth Games venues; POS tag "pesta olahraga" di
137	25	Numeralia	Sudah	Sudah	5.000 sempat berubah menjadi 5, mungkin pada saat dipindah k
138	25-26	Kutipan langsung yang terputus	Sudah	Sudah	Jika melihat tanda kutip di awal kalimat 25 dan di akhir kalimat 2
139	27	"makan diri" dalam "memberi (VB) m	Sudah		Di KBBI, "makan diri" adalah kiasan 'rusak badannya (krn sedih d Untuk konteks kalimat 27, lebih tepat "makan diri" dipisah menj memberi (VB) makan (NN) diri (NN) sendiri (NN) Mungkin itu menjadi "makan diri" pada waktu proses MWE dan t
140	29	"mengenai" dalam "mengenai apa a	Sudah	Sudah	POS tag diganti dari SC menjadi IN
141	29	"apa" dalam "apa arti beramal"	Sudah	Sudah	POS tag diganti dari WH menjadi SC
142	35	"pencipta" dan "serial" dalam "pencip	Sudah		POS tag diganti dari NNP menjadi NN
143	38	"seperti" dalam "seperti sekarang ini	Sudah	Sudah	POS tag diganti dari SC menjadi IN
144	40	"menuju" dalam "langkah dia menuju	Sudah	Sudah	POS tag diganti dari VB menjadi IN (Tetap VB)
145	44	"untuk" dalam "untuk menentukan m	Sudah	Sudah	POS tag diganti dari IN menjadi SC
146	45	"sesuai" dalam "sesuai lokasi penem	Sudah	Sudah	Apakah POS tag-nya diganti dari JJ menjadi IN? Tetap JJ
147	48	"kontruksi"	Sudah	Sudah	Tambahkan satu huruf "s" supaya sesuai dengan KBBI "konstruk
148	48	Numeralia	Sudah	Sudah	60 diganti jadi 60.000 karena di data mentah 60.000, bukan 60
149	49	melalui	Sudah		POS tag diganti dari VB menjadi IN
150	50	kembali	Sudah		POS tag diganti dari VB menjadi RB
151	53	"siapa" dalam "siapa pemiliknya"	Sudah	Sudah	Apakah POS tag-nya tetap WH atau diganti jadi SC? Jadi SC kare
152	64	"untuk" dalam "untuk menghentikan	Sudah	Sudah	POS tag diganti dari IN menjadi SC
153	64	"setelah" dalam "setelah pemilihan r	Sudah	Sudah	Saya rasa lebih tepat IN, tetapi di TBBI SC dan di KBBI RB, ganti
154	67	"untuk" dalam "untuk merebut"	Sudah	Sudah	POS tag diganti dari IN menjadi SC
155	68	"untuk" dalam "untuk meraih 513 ku	Sudah	Sudah	POS tag diganti dari IN menjadi SC
156	68	6 calon	Sudah		6.000 calon
157	70	"karena" dalam "karena maraknya s	Sudah	Sudah	POS tag diganti dari SC menjadi IN
158	70	"maraknya"	Sudah	Sudah	dipisah menjadi (JJ marak) (PRP -nya)

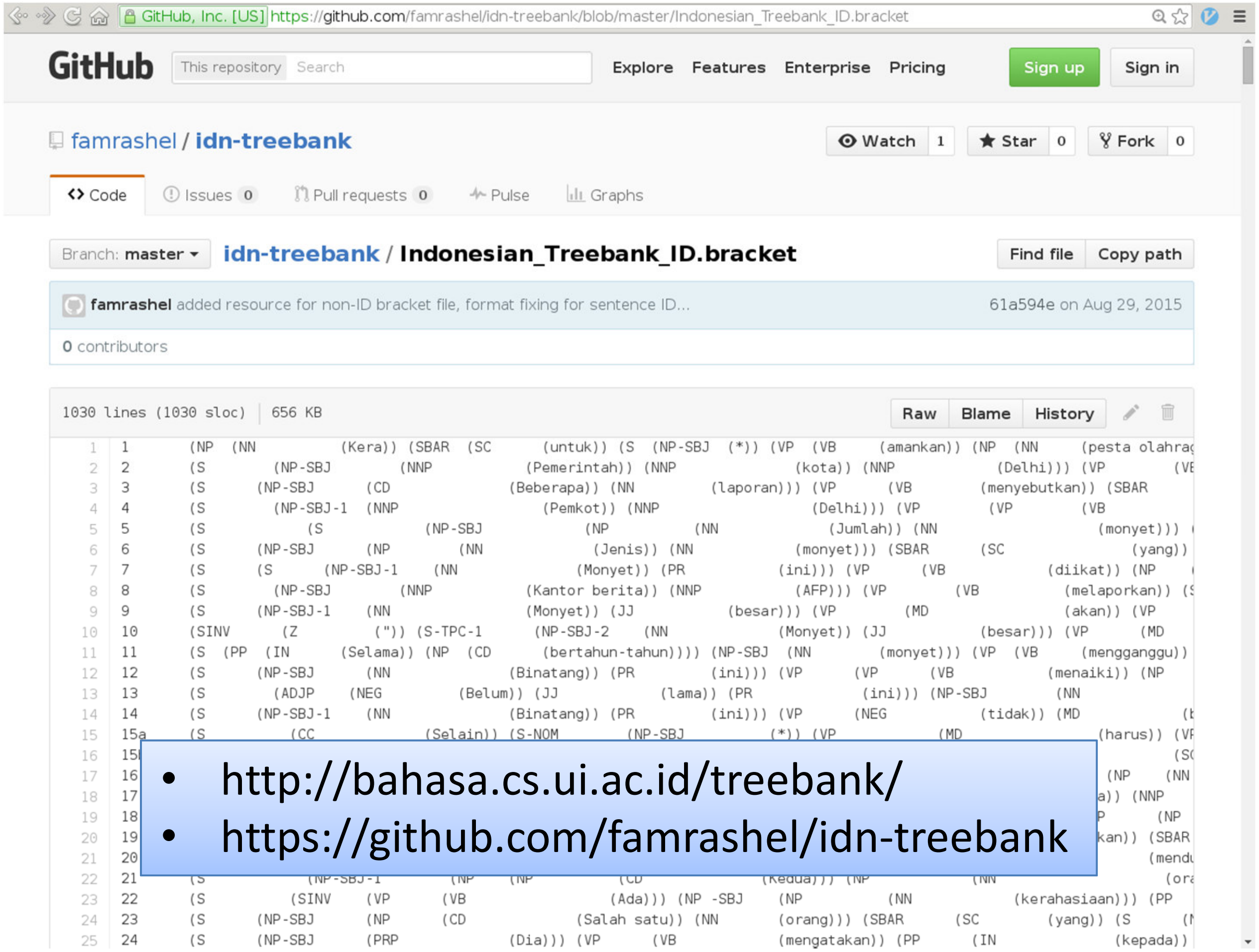
Annotation Process: Multi-phase

- Re-annotate the first 100 sentences.
- Annotate the next 100 sentences.
- Annotate remainder of the 1000 sentences.

Outline

- Background
- Annotation process
- **Outputs**
- Making use of the treebank

The Treebank



GitHub, Inc. [US] https://github.com/famrashel/idn-treebank/blob/master/Indonesian_Treebank_ID.bracket

famrashel / idn-treebank

idn-treebank / Indonesian_Treebank_ID.bracket

famrashel added resource for non-ID bracket file, format fixing for sentence ID... 61a594e on Aug 29, 2015

0 contributors

1030 lines (1030 sloc) | 656 KB

```
1       (NP  (NN          (Kera)) (SBAR  (SC      (untuk)) (S  (NP-SBJ  (*)) (VP  (VB     (amankan)) (NP  (NN     (pesta olahrag
2       (S        (NP-SBJ        (NNP             (Pemerintah)) (NNP             (kota)) (NNP            (Delhi))) (VP          (VE
3       (S      (NP-SBJ      (CD                (Beberapa)) (NN          (laporan))) (VP       (VB          (menyebutkan)) (SBAR
4       (S        (NP-SBJ-1  (NNP                   (Pemkot)) (NNP                 (Delhi))) (VP         (VP        (VB
5       (S             (S                (NP-SBJ             (NP              (NN                (Jumlah)) (NN                 (monyet)))
6       (S      (NP-SBJ      (NP         (NN                 (Jenis)) (NN               (monyet))) (SBAR      (SC                 (yang))
7       (S      (S      (NP-SBJ-1    (NN                (Monyet)) (PR             (ini))) (VP      (VB                (diikat)) (NP
8       (S        (NP-SBJ        (NNP             (Kantor berita)) (NNP             (AFP))) (VP        (VB             (melaporkan)) (S
9       (S      (NP-SBJ-1    (NN                  (Monyet)) (JJ          (besar))) (VP         (MD                  (akan)) (VP
10      (SINV       (Z          (")) (S-TPC-1       (NP-SBJ-2    (NN             (Monyet)) (JJ           (besar))) (VP       (MD
11      (S  (PP  (IN      (Selama)) (NP  (CD      (bertahun-tahun)))) (NP-SBJ  (NN        (monyet))) (VP  (VB     (mengganggu))
12      (S      (NP-SBJ      (NN                (Binatang)) (PR          (ini))) (VP      (VP      (VB             (menaiki)) (NP
13      (S        (ADJP    (NEG          (Belum)) (JJ            (lama)) (PR            (ini))) (NP-SBJ        (NN
14      (S      (NP-SBJ-1    (NN                (Binatang)) (PR          (ini))) (VP      (NEG            (tidak)) (MD              (b
15a     (S           (CC               (Selain)) (S-NOM        (NP-SBJ          (*)) (VP             (MD                (harus)) (VF
15b                                                                                                                          (SC
16                                                                                                                    (NP    (NN
17                                                                                                                   a)) (NNP
18                                                                                                                   P     (NP
19                                                                                                                   kan)) (SBAR
20                                                                                                                         (mendu
21      (S            (NP-SBJ-1    (NP     (NP           (CD              (Kedua))) (NP          (NN                  (ora
22      (S           (SINV    (VP      (VB                  (Ada))) (NP -SBJ    (NP          (NN           (kerahasiaan))) (PP
23      (S      (NP-SBJ      (NP      (CD              (Salah satu)) (NN        (orang))) (SBAR      (SC        (yang)) (S      (N
24      (S      (NP-SBJ      (PRP             (Dia))) (VP      (VB            (mengatakan)) (PP       (IN                 (kepada))
```

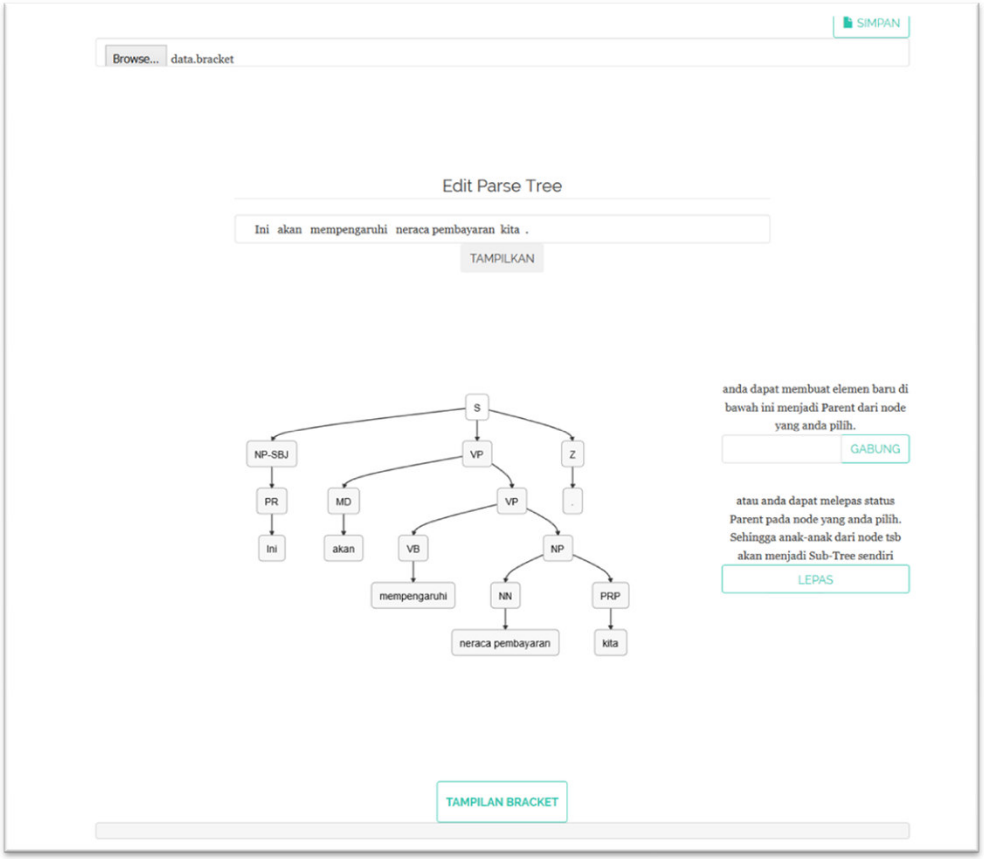
- http://bahasa.cs.ui.ac.id/treebank/
- https://github.com/famrashel/idn-treebank

- 1000 sentences
- 2 variants: with/without sentence IDs – for mapping to POS tagged corpus
- Creative Commons BY-NC-SA 4.0

Indonesian Treebank Bracketing Guidelines

- Guidelines to annotate Indonesian sentence structure in developing Indonesian Treebank.
- Rules for bracketing clauses and sentences:
  - Simple active/passive declarative, imperative, interrogative, inversion, subordinative, coordination, direct/indirect quote, etc.
- Rules for bracketing phrasal structures:
  - Phrasal structures: Adjectival phrases (ADJP), Adverbial phrases (ADVP), Conjunctor phrases (CONJP), Noun phrases (NP), Numeral phrases (QP), Prepositional phrase (PP), Verb phrase (VP), Unlike coordinated phrase (UCP)
- Syntactic category labels and function tags from the Penn Treebank bracketing guidelines.
- POS tags from our Indonesian POS tagset.

Web-Based Annotation Tool



anda dapat membuat elemen baru di bawah ini menjadi Parent dari node yang anda pilih.

atau anda dapat melepas status Parent pada node yang anda pilih. Sehingga anak-anak dari node tsb akan menjadi Sub-Tree sendiri

JavaScript only, runs locally, single user
`https://github.com/andryluthfi/annotation-tools-lightweight`

Client-server using database, multiple concurrent user, agreement checking
`https://github.com/andryluthfi/annotation-tools`

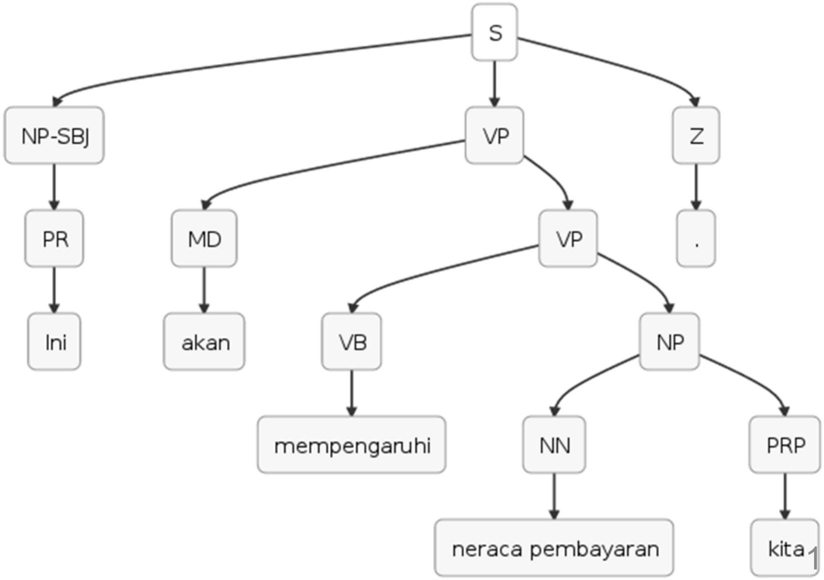
Web-Based Annotation Tool

- Direct input by user, or load from **`.bracket`** file
- Resulting annotation saved to **`.bracket`** file.
- Example:

Ini akan mempengaruhi neraca pembayaran kita.

this will impact balance payment us

```
(S (NP-SBJ (PR (Ini)))
(VP (MD (akan)) (VP (VB
(mempengaruhi)) (NP (NN
(neraca pembayaran))(PRP
(kita))))) (Z (.)))
```

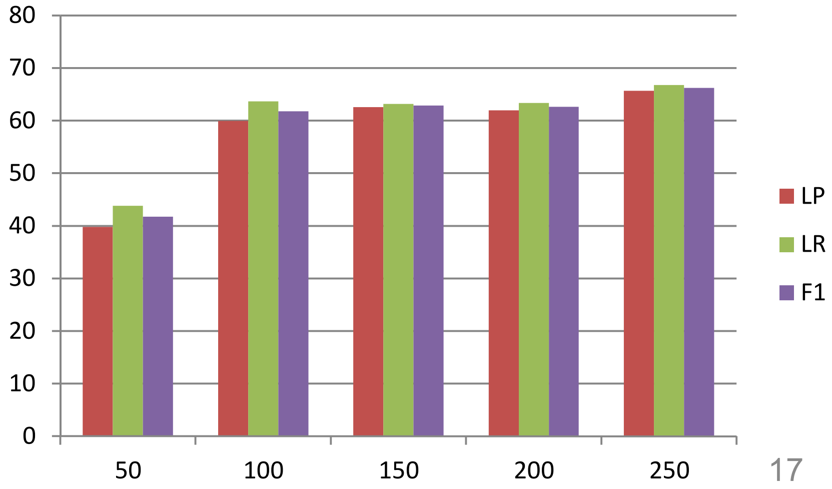


Outline

- Background
- Annotation process
- Outputs: treebank, guidelines, tools
- **Making use of the treebank**

Teaching tool

- 300 sentence treebank used for undergraduate NLP class assignment
- Each student asked to annotate 10+5 sentences ☺
- Experiment on training Stanford Parser with varying parameters



Text Mining Systemic Risk Prioritization (TM-SRP)

- Detect economic risks stated in financial news articles.
- Domain experts from macroprudential policy dept. of Indonesian central bank constructed model of 31 economic risks and related keywords.
- Baseline approach: matching of keyword occurrence in a single sentence.

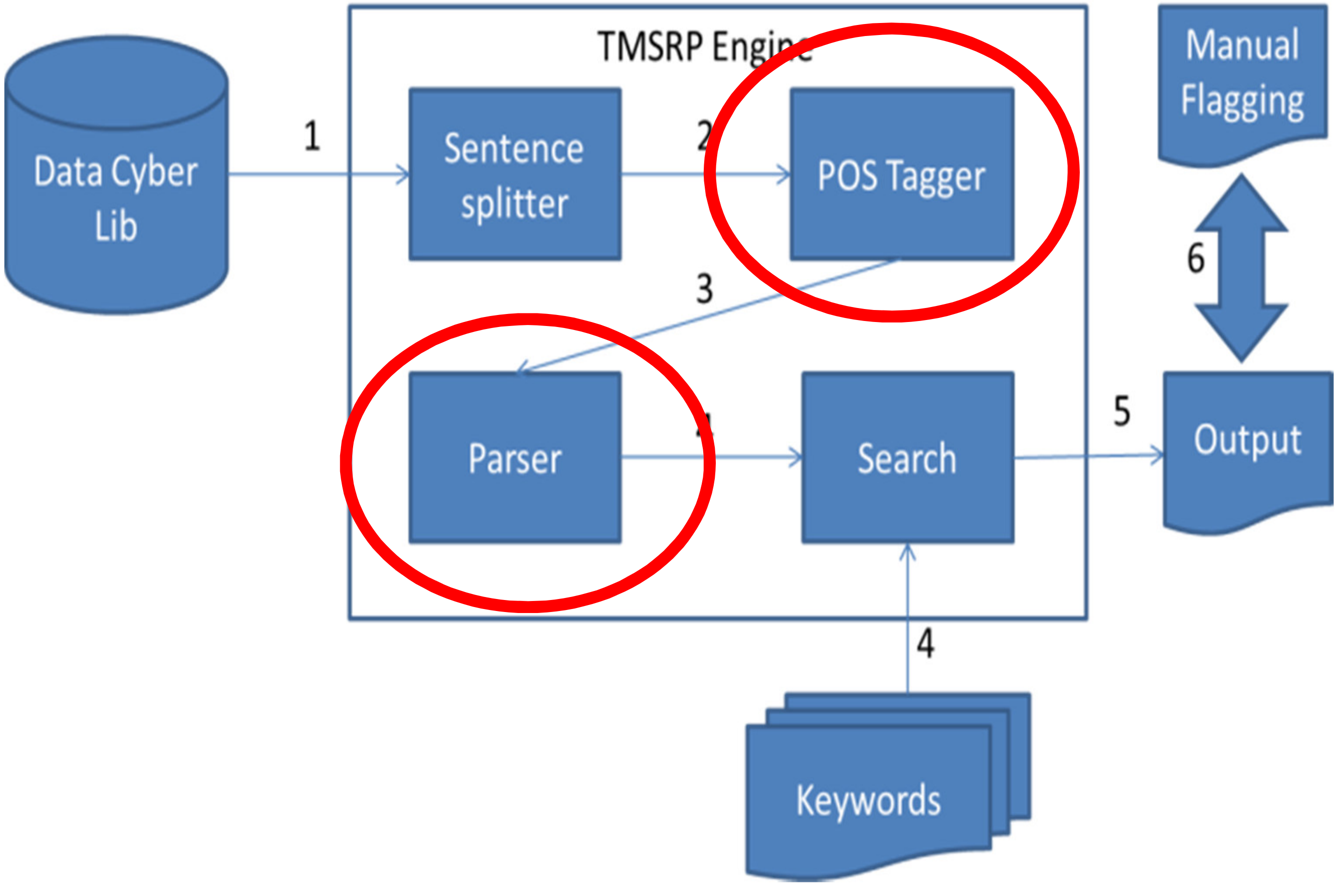
Problem with Keyword Matching

- Example risk: Global Interest Rate
  - Keyword 1: suku bunga (interest rate)
  - Keyword 2: naik (increasing)
- *Setelah naik menjadi presiden, Jokowi*
  after ascend become president, Jokowi

  memerintahkan untuk menurunkan suku_bunga BI
  instruct to lower interest rate BI

Idea: Utilize syntactic structure from probabilistic parser.
Only match keywords in corresponding syntactic relations.

Proposed Approach



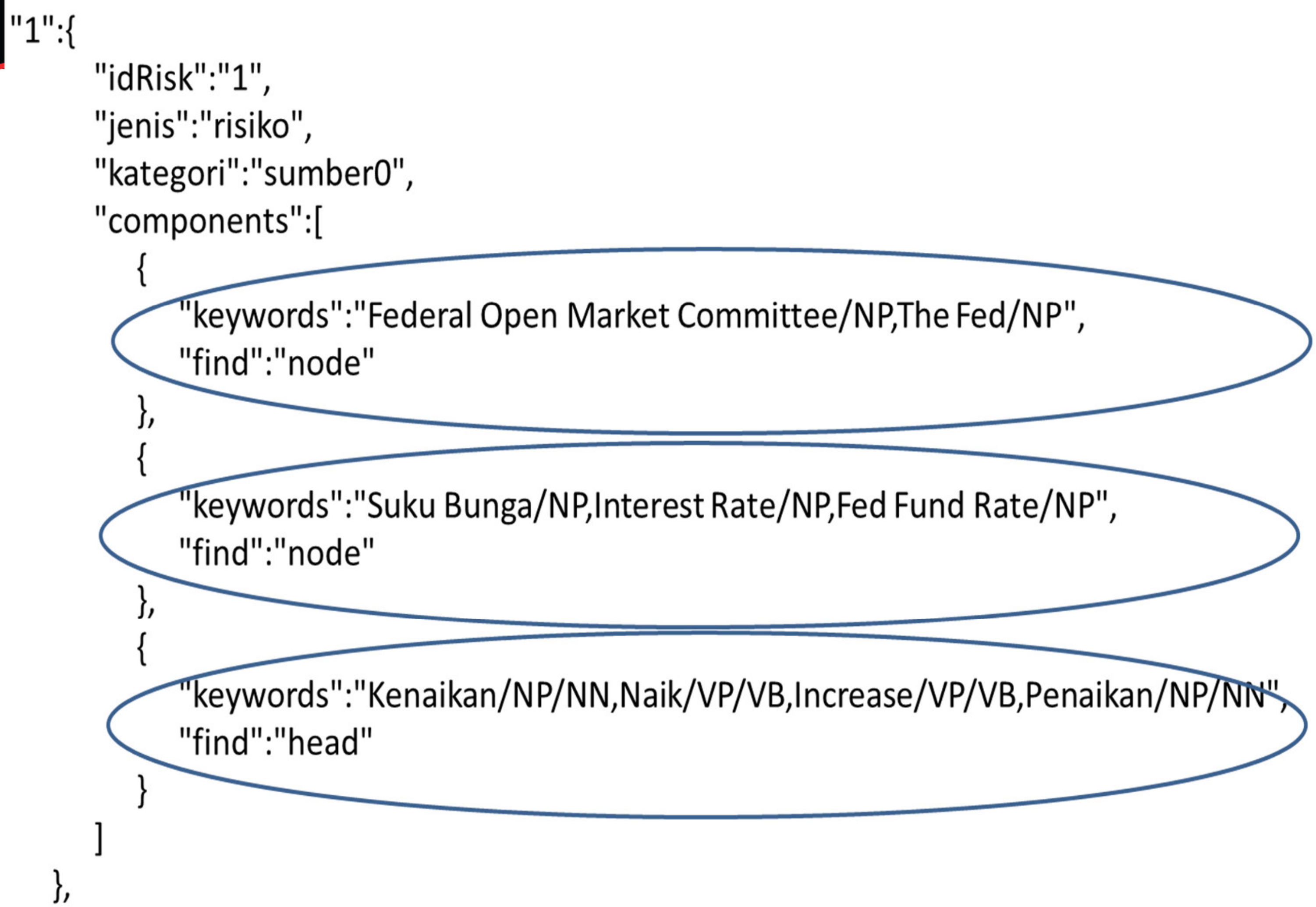
POS Tagger Domain Adaptation

- Lots of domain-specific terms not found in the training data.
  - “nilai tukar” (exchange rate)
  - “daya beli” (purchasing power)
  - etc.

Pattern matching

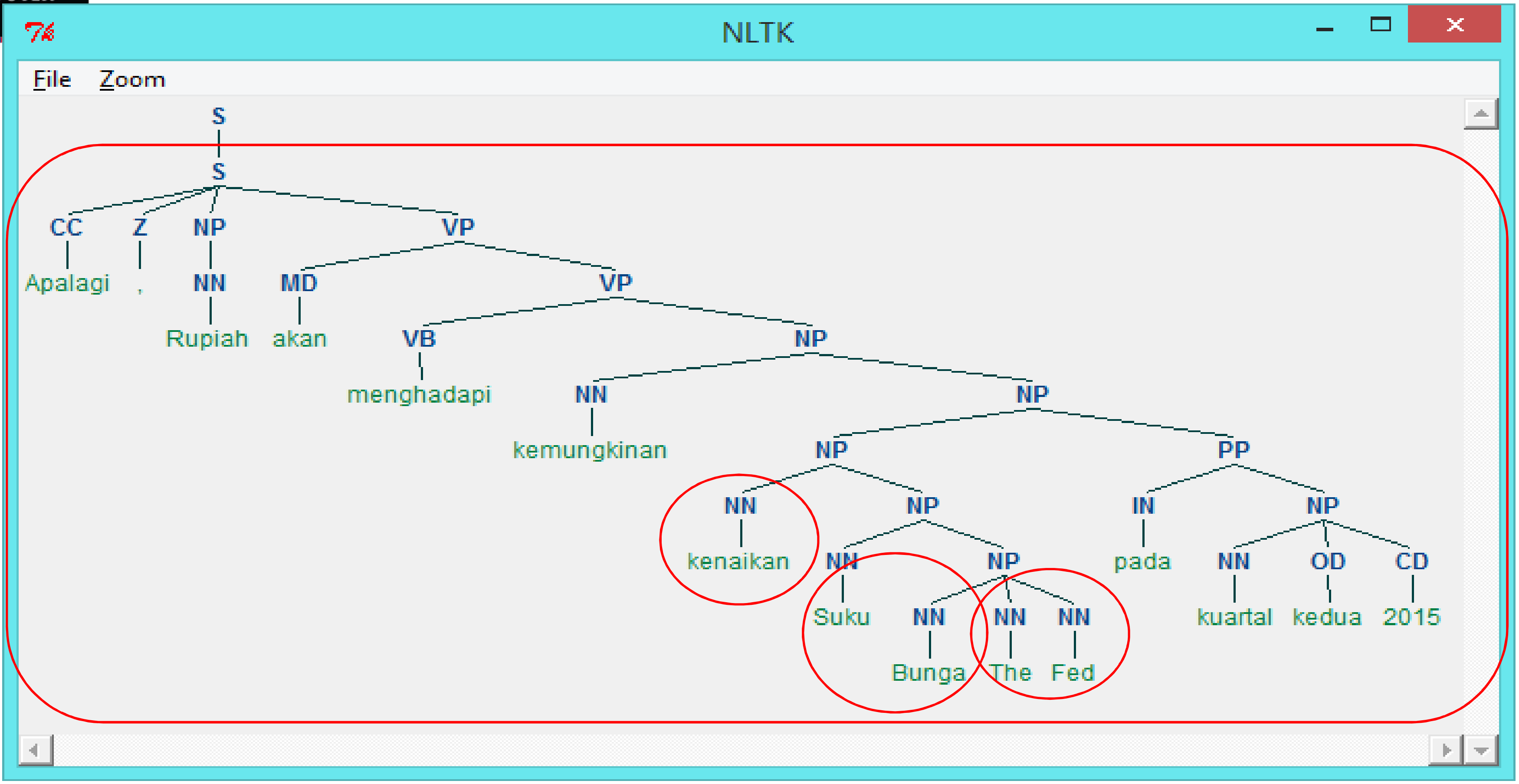
- Focus on each subtree that has root label “S”. If a sentence has several clauses, the search will focus on each clause.
- Differentiate 2 types of keywords:
  - Simple Node: Keyword can appear anywhere in a phrase. Mostly for “noun” keywords
  - Head Node: Keyword must appear at the beginning of a phrase. Mostly for “verb” keywords.
- Find a negation label on each sub-tree “S”.

Search Engine



```
"1":{
      "idRisk":"1",
      "jenis":"risiko",
      "kategori":"sumber0",
      "components":[
          {
            "keywords":"Federal Open Market Committee/NP,The Fed/NP",
            "find":"node"
          },
          {
            "keywords":"Suku Bunga/NP,Interest Rate/NP,Fed Fund Rate/NP",
            "find":"node"
          },
          {
            "keywords":"Kenaikan/NP/NN,Naik/VP/VB,Increase/VP/VB,Penaikan/NP/NN",
            "find":"head"
          }
      ]
},
```

Search Engine



keyword1: **The Fed** ; keyword2: **Suku Bunga**; keyword3: **Kenaikan**

Evaluation

- Evaluation judgments provided by domain experts → manually labelled sentences for risk
- Precision: 77.15%
- Recall: 91.76%

References

- A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. 1995. "Bracketing Guidelines for Treebank II Style Penn Treebank Project". https://catalog.ldc.upenn.edu/docs/LDC99T42/prsguid1.pdf. Last Access: September 2013.
- A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung. 2014. "Designing an Indonesian Part of Speech Tagset and Manually Tagged Indonesian Corpus". In Proceedings of the 2014 International Conference on Asian Language Processing.
- H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. Moeliono. 2003. Tata Bahasa Baku Bahasa Indonesia. Third Edition. Balai Pustaka, Jakarta.
- J. Sneddon, A. Adelaar, D. Djenar, and M. Ewing. 2010. Indonesian Reference Grammar. Second Edition. Allen & Unwin, Crows Nest.
- M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, Vol. 19, No. 2, pp. 313-330.