PENGEMBANGAN TATA BAHASA BAKU BAHASA INDONESIA (TBBI) DARING TERPADU

Development of An Integrated Online Standard Grammar of Indonesian

David Moeljadi
Universitas Teknologi Nanyang, Singapura
Pos-el: davidmoeljadi@gmail.com

Abstrak

Badan Pengembangan dan Pembinaan Bahasa (Badan Bahasa) di bawah naungan Kementerian Pendidikan dan Kebudayaan Republik Indonesia, sebagai instansi pemerintah yang ditugaskan untuk menangani masalah kebahasaan dan kesastraan di Indonesia, menerbitkan berbagai produk kebahasaan. Dua produk yang sering dimanfaatkan para pemelajar bahasa Indonesia adalah *Kamus Besar Bahasa Indonesia* (KBBI) dan *Tata Bahasa Baku Bahasa Indonesia* (TBBI). KBBI terbaru edisi kelima (Amalia 2016) diluncurkan pada tahun 2016 dalam tiga versi: cetak, daring, dan luring (Moeljadi *et al*. 2017). Sejak diluncurkan pada 28 Oktober 2016, KBBI Daring mendapat sambutan hangat masyarakat, baik dari dalam maupun luar negeri. KBBI Daring memudahkan pemelajar bahasa Indonesia dan masyarakat umum menggunakan kamus pada era digital ini. Hal yang serupa dapat dilakukan untuk TBBI. Makalah ini membahas tahap awal pengembangan pangkalan data dan laman TBBI Daring Terpadu dengan menggunakan tata bahasa komputasional bahasa Indonesia INDRA (*Indonesian Resource Grammar*) (Moeljadi *et al*. 2015) yang dikembangkan dengan metode rekayasa tata bahasa dengan mengacu pada buku-buku referensi tata bahasa baku bahasa Indonesia, terutama TBBI (Alwi *et al*. 2014) dan Indonesian *Reference Grammar* (Sneddon *et al*. 2010). TBBI Daring Terpadu akan memuat aturan-aturan tata bahasa bahasa Indonesia baku, dipadukan dengan leksikon dan contoh-contoh dari korpus bahasa Indonesia baku yang telah dianotasi secara sintaksis dan semantis. Penulis berharap TBBI Daring Terpadu dapat menjadi acuan utama tata bahasa baku bahasa Indonesia yang dapat diakses dengan mudah oleh para penggunanya, misalnya pemelajar Bahasa Indonesia bagi Penutur Asing (BIPA), dan dapat memperkaya KBBI Daring dalam penggolongan kelas kata yang lebih spesifik, serta mendorong kemajuan bidang linguistik komputasional dan pemrosesan bahasa alami bahasa Indonesia, misalnya dalam penerjemahan mesin dan pengembangan sistem pemeriksaan gramatika dan leksikon bahasa Indonesia baku.

Kata-kata kunci: TBBI, tata bahasa, bahasa Indonesia, pangkalan data, daring

Abstract

The Language Development and Cultivation Agency or Badan Bahasa *under the Ministry of Education and Culture of the Republic of Indonesia, as a government agency assigned to deal with matters related to Indonesian language and literature, publishes language-related products. Two products which are often used by Indonesian language learners are* Kamus Besar Bahasa Indonesia (KBBI) *dictionary* and Tata Bahasa Baku Bahasa Indonesia (TBBI) *reference grammar. The latest, fifth edition of KBBI (Amalia 2016) was launched in 2016 in three versions: online, printed, and mobile applications (Moeljadi et al. 2017). Since its launch on October 28, 2016, the*

online KBBI has helped Indonesian learners and others in this digital era. A similar thing can be done for TBBI. This paper discusses the initial stage of the development of an online integrated TBBI page and database. It employs an Indonesian computational grammar called the Indonesian Resource Grammar (INDRA) (Moeljadi et al. 2015) which has been developed using grammar engineering method, referring to the existing reference grammars of Indonesian, especially TBBI (Alwi et al. 2014) and the Indonesian Reference Grammar (Sneddon et al. 2010). The online integrated TBBI will contain linguistic documentation of phenomena, integrated with lexicon and examples from a standard, syntactically-and-semantically-annotated Indonesian corpus. I hope the online integrated TBBI can become the main reference for standard Indonesian grammar which can be easily accessed by its users, e.g. Indonesian for Foreign Speakers (BIPA) learners, can enrich the online KBBI in a refined part-of-speech subcategorization, and can promote the advance of Indonesian computational linguistics and natural language processing fields, e.g. in the development of a tool for grammar and lexicon check for standard Indonesian and machine translation.

Keywords: *TBBI, grammar, Indonesian, database, online*

PENDAHULUAN

Pendaringan produk kebahasaan dan kesastraan Badan Pengembangan dan Pembinaan Bahasa (Badan Bahasa) telah dimulai dari *Kamus Besar Bahasa Indonesia* (KBBI) Daring. KBBI Daring[1] adalah produk Badan Bahasa yang paling banyak digunakan saat ini. Alexa[2] mencatat bahwa KBBI Daring menduduki peringkat ke-45 situs *web* yang paling banyak diakses di Indonesia dan peringkat pertama untuk situs pemerintah yang diakhiri dengan nama domain ‘go.id’. Hal ini menunjukkan antusiasme masyarakat dalam menggunakan produk kebahasaan, terutama kamus. Selain KBBI Daring, Badan Bahasa juga telah meluncurkan *Tesaurus Tematis Bahasa Indonesia*,[3] Glosarium,[4] dan Ensiklopedia Sastra Indonesia.[5] Produk Badan Bahasa lainnya yang belum didaringkan, tetapi sering digunakan para pemelajar bahasa Indonesia dan berpotensi untuk didaringkan adalah *Tata Bahasa Baku Bahasa Indonesia* (TBBI). Dengan adanya TBBI Daring, tata bahasa baku bahasa Indonesia dapat didokumentasikan secara digital. Selain itu, karena tata bahasa tidak dapat dipisahkan dari leksikon dan penggunaannya di masyarakat (korpus), TBBI Daring memadukan atau mengintegrasikan tata bahasa, leksikon, dan korpus.

[1] https://kbbi.kemdikbud.go.id/
[2] https://www.alexa.com/topsites/countries/ID, diakses pada 29 Agustus 2018
[3] http://tesaurus.kemdikbud.go.id/tematis/
[4] http://118.98.223.79/glosarium/
[5] http://ensiklopedia.kemdikbud.go.id/sastra/

Pendaringan atau dokumentasi tata bahasa secara daring memerlukan tata bahasa komputasional (*computational grammar*), yaitu kumpulan aturan-aturan tata bahasa dan leksikon yang telah dirumuskan secara detail dan eksplisit dan "diterjemahkan" ke dalam bahasa yang dipahami oleh komputer sehingga dapat diproses secara otomatis, dengan menggunakan metode rekayasa tata bahasa (*grammar engineering*). Tata bahasa komputasional bahasa Indonesia berlisensi sumber terbuka yang berpotensi digunakan dalam pengembangan TBBI Daring Terpadu adalah *Indonesian Resource Grammar* (INDRA) (Moeljadi *et al*. 2015). INDRA dikembangkan di dalam kerangka teori sintaks Tata Bahasa Struktur Frasa Berbasis Induk atau *Head-driven Phrase Structure Grammar* (HPSG) (Pollard dan Sag 1994; Sag *et al*. 2003) dan model semantik bernama Semantik Rekursi Minimal atau Minimal Recursion Semantics (MRS) (Copestake et al. 2005), dengan menggunakan alat-alat komputasional atau perkakas (*tools*) yang dikembangkan oleh kelompok peneliti *DEep Linguistic Processing with HPSG-INitiative* (DELPH-IN). INDRA telah digunakan dalam aplikasi *treebank* berlisensi sumber terbuka, bernama JATI (Moeljadi 2017) dan berpotensi digunakan dalam aplikasi lainnya seperti penerjemahan bahasa Indonesia-Inggris dengan mesin, pemelajaran bahasa Indonesia dengan bantuan komputer, pengecekan tata bahasa secara otomatis, dan tentu saja pendokumentasian bahasa Indonesia secara daring.

Makalah ini membahas aspek-aspek pengembangan pangkalan data dan laman TBBI Daring tahap awal. Untuk pengembangan tahap lanjut, TBBI Daring dapat dipadukan dengan leksikon (KBBI Daring) dan korpus (Korpus Indonesia) (Kwary 2018) menggunakan INDRA.

Rekayasa Tata Bahasa

Praktik umum yang biasanya dilakukan di bidang dokumentasi bahasa atau tata bahasa deskriptif meliputi bidang-bidang berikut ini: fonologi, morfologi, sintaks, semantik, dan pragmatik. Bidang rekayasa tata bahasa (*grammar engineering*) mirip dengan dokumentasi tata bahasa karena bidang ini mencoba mendeskripsikan bahasa sebagaimana digunakan oleh penutur jati, tetapi berfokus pada sintaks dan semantik. Selain itu, rekayasa tata bahasa memanfaatkan komputer dalam pengecekan konsistensi analisis dan pemodelan tata bahasa dan pengujiannya terhadap berbagai contoh-contoh yang ada di korpus secara luas (Bender dan Fokkens 2010). Sag et al. (2003) menulis bahwa sintaks berperan penting dalam pemrosesan bahasa manusia karena sintaks

mengenakan batasan-batasan bagaimana kalimat-kalimat dapat atau tidak dapat dibentuk dan menentukan satu set aturan-aturan yang memprediksi keberterimaan kalimat-kalimat dalam suatu bahasa. Beberapa tujuan rekayasa tata bahasa adalah sebagai berikut.

1. untuk menentukan apakah sebarang kalimat gramatikal/berterima atau tidak dan untuk memberikan berbagai kemungkinan interpretasi sintaks dan representasi semantik
2. untuk meninjau bagaimana tata bahasa suatu bahasa berbeda dengan tata bahasa bahasa lainnya
3. untuk mengetahui kemampuan berbahasa manusia secara umum

Flickinger et al. (2010) menyebutkan komponen-komponen penting dalam rekayasa tata bahasa, sebagai berikut.

1. Teori linguistik. Teori linguistik yang solid yang memiliki fondasi matematis yang kukuh dan model yang mudah diimplementasikan secara komputasional, serta bersifat universal (berlaku untuk bahasa-bahasa yang berbeda). Teori ini akan dijabarkan dalam bab Landasan Teori.
2. Platform rekayasa tata bahasa, yang digunakan untuk implementasi deskripsi bahasa secara formal. Platform tersebut harus memiliki editor tata bahasa, prosesor yang memiliki sistem pengurai kalimat dan pembentuk kalimat, antarmuka pengguna, dan perkakas aplikasi *treebank*.
3. Sumber-sumber linguistik, seperti korpus, *treebank*, dan buku-buku referensi tata bahasa, termasuk tata bahasa komputasional.
4. Metode penelitian, yang akan dijelaskan dalam bab Metode Penelitian.

LANDASAN TEORI

Bab ini berisi pengantar tentang kerangka teori HPSG (Pollard dan Sag 1994; Sag et al. 2003) dan MRS (Copestake et al. 2005) yang digunakan dalam pengembangan INDRA.

Tata Bahasa Struktur Frasa Berbasis Induk

Kebanyakan model formal sintaks bahasa alami adalah Tata Bahasa Bebas Konteks atau Context-Free Grammars (CFG), juga disebut Tata Bahasa Struktur Frasa atau Phrase-Structure Grammars. Kerangka tata bahasa ini berdasarkan struktur

konstituen yang dirumuskan oleh Chomsky (1956). CFG terdiri dari sebuah set aturan-aturan tata bahasa dan leksikon simbol-simbol (kelas kata) dan kata-kata, seperti ditunjukkan pada (1). Set aturan-aturan tata bahasa ini mengelompokkan dan mengurutkan simbol-simbol. Leksikon menggabungkan simbol-simbol dengan kata-kata.

(1) a. Contoh set aturan-aturan tata bahasa:

S → NP VP

NP → N

VP → V

b. Contoh leksikon simbol (kelas kata) dan kata:

N: anjing

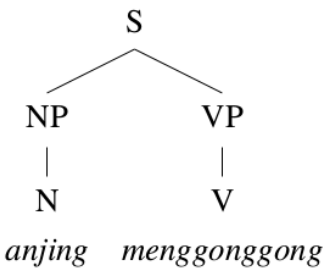
V: menggonggong

Dari set aturan-aturan dan leksikon yang ditunjukkan pada (1) di atas, sebuah kalimat dapat dibentuk, seperti yang ditunjukkan pada (2).

(2) Anjing menggonggong.

Kalimat bentukan tersebut juga dapat disajikan dalam pohon, seperti yang digambarkan pada (3).

(3)



Tata Bahasa Struktur Frasa Berbasis Induk atau Head-driven Phrase Structure Grammar (HPSG) berorientasi pada teks yang ada di permukaan dan tidak mengasumsikan adanya struktur abstrak. Karena itu, HPSG menyajikan struktur yang cukup sederhana yang berhubungan langsung dengan untaian kata-kata yang membentuk kalimat. HPSG bersifat mono-stratal, yaitu ortografi, sintaks, semantik,

pragmatik semuanya dalam sebuah struktur tunggal atau sebuah tanda. Tanda, yang merupakan pasangan bentuk dan makna, adalah satuan dasar atau primer dalam HPSG yang dimodelkan melalui Struktur Fitur Bertipe atau Typed Feature Structures (TFS). Tanda dalam HPSG meliputi kata, frasa, kalimat, dan ujaran. HPSG bersifat leksikalis, yaitu sebagian besar properti sintaksis dan semantis didefinisikan di dalam leksikon. Karena itu, informasi yang terdapat dalam struktur fitur sebuah tanda meliputi baik sintaks maupun semantik. Informasi tentang HPSG selebihnya dapat dilihat di Pollard dan Sag (1994) dan Sag et al. (2003).

Semantik Rekursi Minimal

Semantik Rekursi Minimal atau Minimal Recursion Semantics (MRS) adalah model representasi semantik yang datar dan nonrekursif, sesuai untuk struktur bertipe yang digunakan HPSG dan untuk pemecahan struktur sintaks dan pembentukan kalimat. MRS bukan teori semantik, melainkan sistem representasi semantik. Representasi MRS didesain untuk mengatasi masalah-masalah dalam pendekatan transfer semantik untuk penerjemahan mesin, khususnya untuk membuat model ambiguitas yang sering ada pada kalimat dengan kuantifikasi, misalnya 'setiap anjing mengejar kucing putih',[6] dengan menggunakan prinsip hubungan lingkup semantis yang kurang spesifik. MRS dapat dikonversi ke dalam sistem yang lebih dikenal, seperti kalkulus predikat (Copestake 2002).

Tujuan utama representasi MRS adalah menemukan leksem-leksem yang tepat dan hubungan-hubungan di antara leksem-leksem tersebut yang dilisensikan oleh sintaks. Inti representasi MRS adalah kumpulan predikat dasar atau *elementary predications* (EP). EP menunjukkan hubungan-hubungan dengan argumen-argumen terkait. Misalnya, makna kata 'anjing' dapat direpresentasikan dalam logika predikat, sebagaimana ditunjukkan pada (4a). Demikian pula 'menggonggong' dan 'anjing menggonggong' dapat direpresentasikan sebagai berikut.

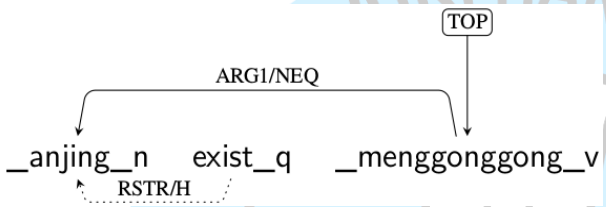
(4) a. anjing(x)
b. menggonggong(x)
c. menggonggong(x), anjing(x)

[6] Kalimat ambigu ini dapat berarti: (1) setiap anjing mengejar satu kucing putih yang berbeda, atau (2) setiap anjing mengejar satu kucing putih yang sama.

Struktur (4c) adalah senarai (*list*) EP yang digabungkan. Urutan anggota-anggota yang ada di senarai tersebut bersifat arbitrer. Dalam HPSG, EP direpresentasikan sebagai struktur fitur bertipe atau Typed Feature Structure (TFS). Dalam TFS, pengodean semantik dilakukan bersamaan dengan sintaks.

Representasi MRS untuk kalimat ‘anjing menggonggong’ dapat diilustrasikan dengan grafik dependensi (*dependency graph*), disebut “Semantik Rekursi Minimal Dependensi” atau Dependency Minimal Recursion Semantics (DMRS), seperti yang digambarkan pada (5). Struktur DMRS bersifat minimal, predikat-predikat yang ada ditunjukkan dengan tautan-tautan sederhana dan tanpa variabel.

(5)



Hal paling penting yang perlu diperhatikan di sini adalah predikat utamanya (TOP) terletak pada kata kerja ‘menggonggong’ yang memiliki argumen pertama (ARG1) dan satu-satunya, yaitu ‘anjing’. Informasi tentang MRS selebihnya dapat dibaca di Copestake et al. (2005).

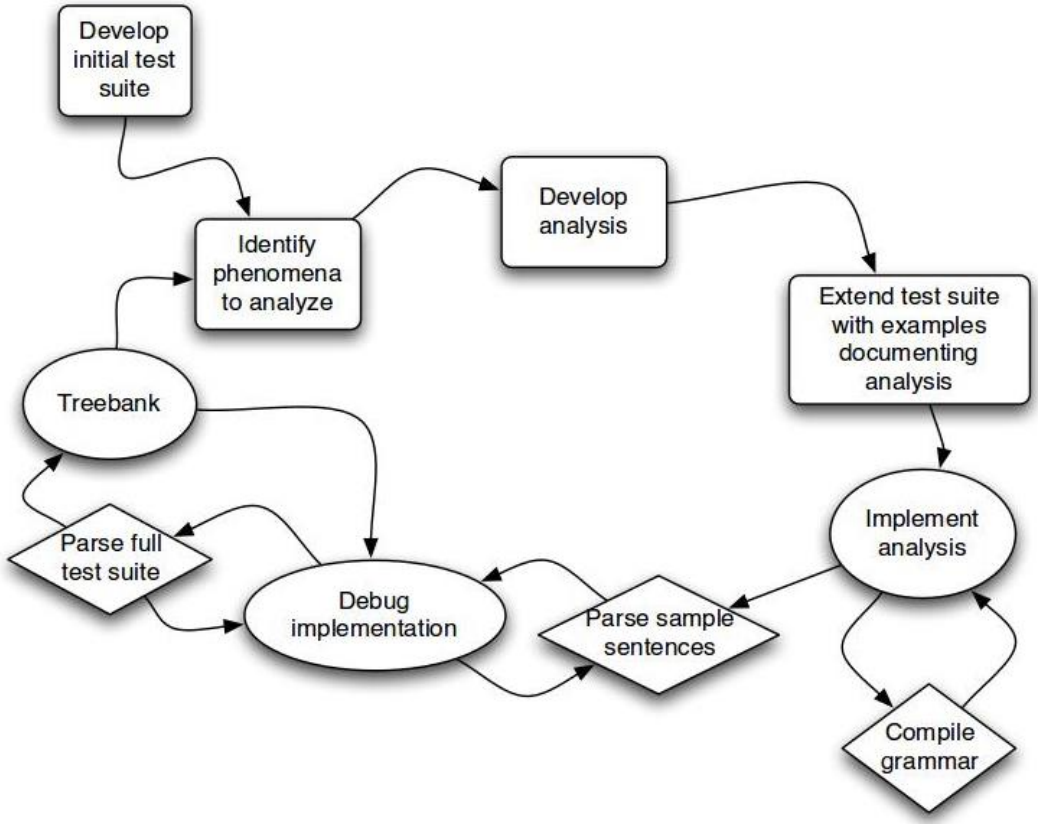
METODE PENELITIAN

Bab ini membahas metode penelitian dalam rekayasa tata bahasa. Pada umumnya, metode penelitian yang digunakan adalah perpaduan analisis linguistik dan implementasi komputasional. Dengan mengembangkan tata bahasa komputasional, setiap detail fenomena kebahasaan yang mungkin tidak terpikirkan saat kita mendokumentasikan atau menganalisis bahasa harus dipertimbangkan.

Metode penelitian rekayasa tata bahasa didorong oleh data di korpus (*corpus-driven*). Pertama-tama, contoh teks kalimat-kalimat gramatikal (dan juga tak gramatikal) dipilih dan diformat dalam bentuk teks yang telah ditokenisasi dan dianotasi dengan glos interlinear ke dalam satu atau beberapa berkas contoh yang disebut *test-suite*. Test-suite dapat dibagi menjadi dua tipe: *test-suite* berdasarkan fenomena tata bahasa yang berisi fenomena-fenomena tertentu dan *test-suite* alami yang diambil dari korpus atau

teks yang dikutip langsung dari sumbernya. Setelah itu, fenomena kebahasaan tertentu yang akan dianalisis diidentifikasi, misalnya konstruksi kopula, kalimat pasif, atau kata penggolong. Setelah menganalisis suatu fenomena kebahasaan berdasarkan buku-buku referensi tata bahasa dan kepustakaan linguistik lainnya, analisis dibuat menurut model HPSG dan diimplementasikan secara manual dengan menambahkan atau mengedit berkas kode komputasional.

Gambar 1
Proses pengembangan tata bahasa komputasional



Sumber: Bender et al. (2011, hlm. 10)

Implementasi fenomena yang mirip di tata bahasa komputasional bahasa-bahasa lainnya seperti English Resource Grammar (ERG) (Flickinger 2000) untuk bahasa Inggris, Jacy (Siegel et al. 2016) untuk bahasa Jepang, dan Zhong (Fan et al. 2015) untuk bahasa Mandarin dapat menjadi rujukan. Setelah itu, tata bahasa komputasional tersebut dikompilasi dan dites dengan menguraikan kalimat-kalimat contoh atau *test-suite* (baik yang telah ada sebelumnya maupun yang baru dibuat). Biasanya tata bahasa komputasional tersebut tidak dapat menguraikan beberapa kalimat contoh, tidak dapat

menghasilkan representasi semantik (MRS) yang tepat, atau tidak mendapatkan cakupan sempurna dari *test-suite*. Karena itu, pengembang tata bahasa komputasional harus menyelidiki kalimat-kalimat bermasalah yang tidak dapat diuraikan dengan baik dengan MRS yang tepat atau kalimat-kalimat yang memiliki jumlah hasil penguraian yang tinggi (jauh lebih tinggi daripada semua kemungkinan ambiguitas yang diprediksi). Jika masalahnya telah ditemukan, pengembang akan mengawakutu (*debug*) tata bahasa komputasional tersebut hingga semua fenomena yang ada, baik yang baru maupun yang sebelumnya, dapat tercakup dengan benar. Kadang-kadang proses pengawakutuan ini memerlukan waktu lama karena analisis baru harus dipikirkan, dibuat modelnya dalam HPSG, dan diimplementasikan setelahnya. Kemudian, kalimat-kalimat contoh atau *test-suite* diuraikan kembali dan dibuat *treebank*-nya dengan menggunakan tata bahasa komputasional yang baru, lalu profil yang baru akan dibuat. Profil yang baru dapat dibandingkan dengan yang sebelumnya dari segi cakupan (*coverage*) dan efisiensi. Proses ini berjalan berulang-ulang, seperti yang ditunjukkan pada Gambar 1. Jika masalah yang ada telah dapat diatasi atau proses pengawakutuan telah selesai dengan hasil baik, tata bahasa komputasional yang baru tersebut akan diunggah di GitHub.[7]

Tata Bahasa Komputasional Bahasa Indonesia (INDRA)

Indonesian Resource Grammar (INDRA) (Moeljadi et al. 2015) adalah tata bahasa komputasional bahasa Indonesia baku[8] pertama yang dikembangkan di dalam kerangka teori HPSG dan MRS dengan pendekatan analisis korpus yang memiliki cakupan fenomena kebahasaan yang luas yang terdapat dalam korpus, dengan menggunakan perkakas yang dibuat oleh DELPH-IN. INDRA berlisensi sumber terbuka dan dapat diunduh di GitHub.[9] INDRA dikembangkan dengan menggunakan metode penelitian yang disebutkan di atas. Sejauh ini, INDRA telah digunakan dalam pengembangkan aplikasi *treebank* berlisensi sumber terbuka, bernama JATI (Moeljadi 2017) yang berisi 2.003 frasa nominal. INDRA berpotensi digunakan dalam pengembangan *treebank* Korpus Indonesia (Kwary 2018). Penelitian sebelumnya mengenai tata bahasa komputasional bahasa Indonesia sebagian besar dikerjakan dalam kerangka teori Tata

[7] GitHub adalah layanan penginangan web bersama untuk proyek pengembangan perangkat lunak yang menggunakan sistem pengontrol versi Git dan layanan hosting internet.
[8] Hingga makalah ini ditulis (5 September 2018), INDRA hanya berisi leksikon dan aturan-aturan tata bahasa untuk bahasa Indonesia baku. Di kemudian hari, INDRA akan dikembangkan menjadi tata bahasa komputasional bahasa Indonesia, baik yang baku maupun yang takbaku.
[9] https://github.com/davidmoeljadi/INDRA

Bahasa Leksikal-Fungsional atau Lexical-Functional Grammar (LFG) (Kaplan dan Bresnan 1982, Dalrymple 2001). Tata bahasa komputasional bahasa Indonesia bernama IndoGram (Arka 2012) dikembangkan dalam kerangka Parallel Grammar (ParGram) berdasarkan LFG, menggunakan pengurai (*parser*) Xerox Linguistic Environment (XLE).

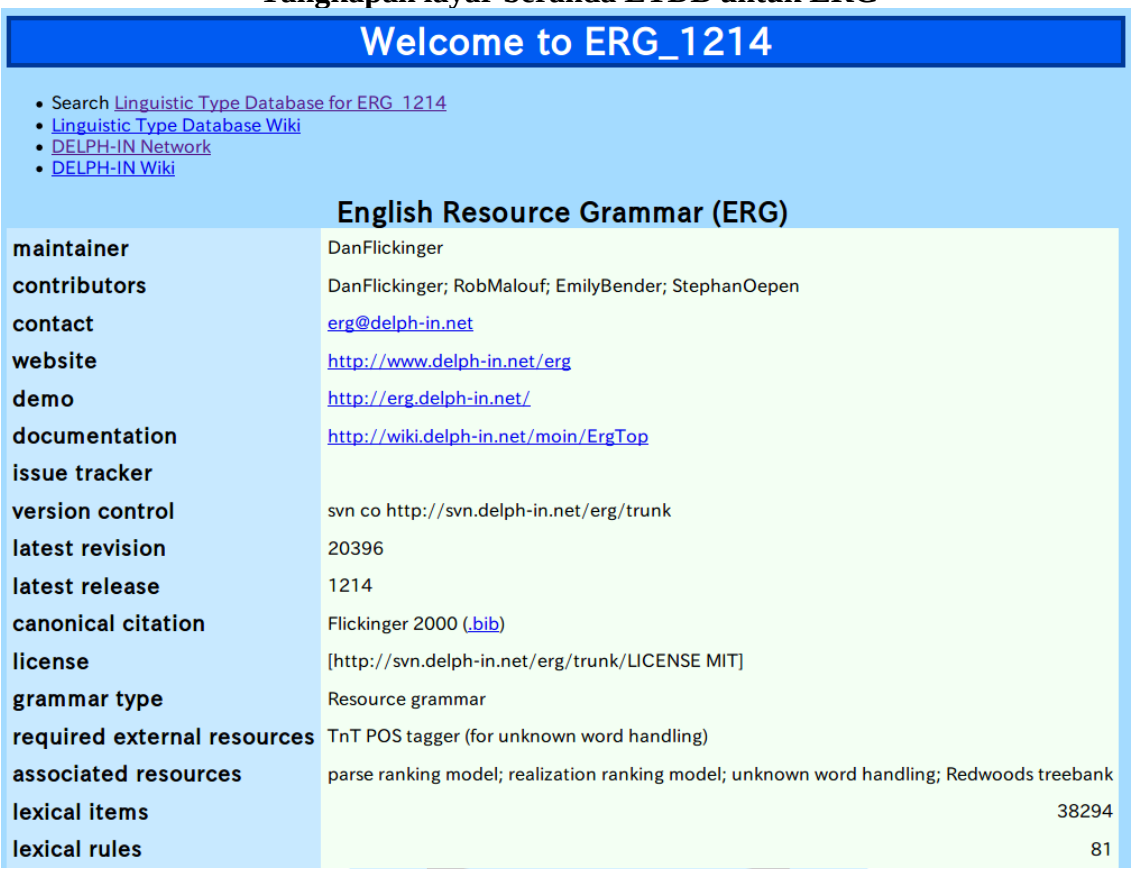
Sejauh ini, INDRA dapat menguraikan dan menghasilkan frasa nominal kompleks dengan klitik, pronomina penunjuk, numeralia, kata penggolong, dan klausa yang dimulai dengan 'yang'; frasa verbal dengan kata kerja bantu dan penanda diatesis aktif dan pasif yang berupa imbuhan 'meN-' dan 'di-'; konstruksi kopula; kata majemuk; klausa koordinatif dengan kata dan frasa berkelas kata sama; dan klausa subordinatif. Karena INDRA masih dalam tahap pengembangan, saat ini INDRA masih belum dapat menguraikan frasa adjektival ekuatif, komparatif, dan superlatif; klausa koordinatif dengan kata dan frasa berkelas kata berbeda; kalimat seru; dan kalimat tanya dengan kata tanya. Walaupun INDRA masih memiliki keterbatasan-keterbatasan, dibandingkan dengan IndoGram, INDRA memiliki ketepatan yang lebih baik dalam analisis beberapa fenomena kebahasaan dan memiliki *treebank* berlisensi sumber terbuka dengan ukuran lima belas kali lebih besar. Selain itu, INDRA berpotensi digunakan dalam berbagai aplikasi, misalnya penerjemahan multibahasa dengan mesin dan pemelajaran bahasa dengan bantuan komputer. Karena INDRA dikembangkan dalam komunitas DELPH-IN bersama dengan tata bahasa komputasional bahasa-bahasa lainnya seperti ERG yang menggunakan representasi semantik yang sama, yaitu MRS, sistem penerjemahan mesin berbasis transfer semantik dapat dibuat dengan mudah.

Pangkalan Data Tipe Linguistik

Setelah tata bahasa komputasional yang dikembangkan telah mencakup banyak fenomena kebahasaan dan memiliki *treebank*, Pangkalan Data Tipe Linguistik atau The Linguistic Type Database (LTDB) (Hashimoto et al. 2007) dapat mulai dikembangkan. LTDB adalah pangkalan data terstruktur yang berisi dokumentasi tipe-tipe leksikal (penggolongan kelas kata yang terperinci) dan aturan-aturan tata bahasa yang ada di tata bahasa komputasional beserta dengan contoh-contoh dan informasi frekuensi penggunaan tipe-tipe leksikal dan aturan-aturan tersebut, yang diperoleh dari *treebank*. LTDB terdiri dari sistem manajemen pangkalan data dan antarmuka pengguna. LTDB dibuat dan dikembangkan secara semi-otomatis. Kode program LTDB berlisensi sumber

terbuka dan dapat diunduh di GitHub.[10] Dokumentasi LTDB ada di laman DELPH-IN.[11] Saat ini, LTDB telah diaplikasikan ke tata bahasa komputasional bahasa Inggris[12] dan bahasa Jepang.[13] Gambar 2 adalah tangkapan layar laman beranda LTDB untuk tata bahasa komputasional bahasa Inggris ERG.

Gambar 2
Tangkapan layar beranda LTDB untuk ERG

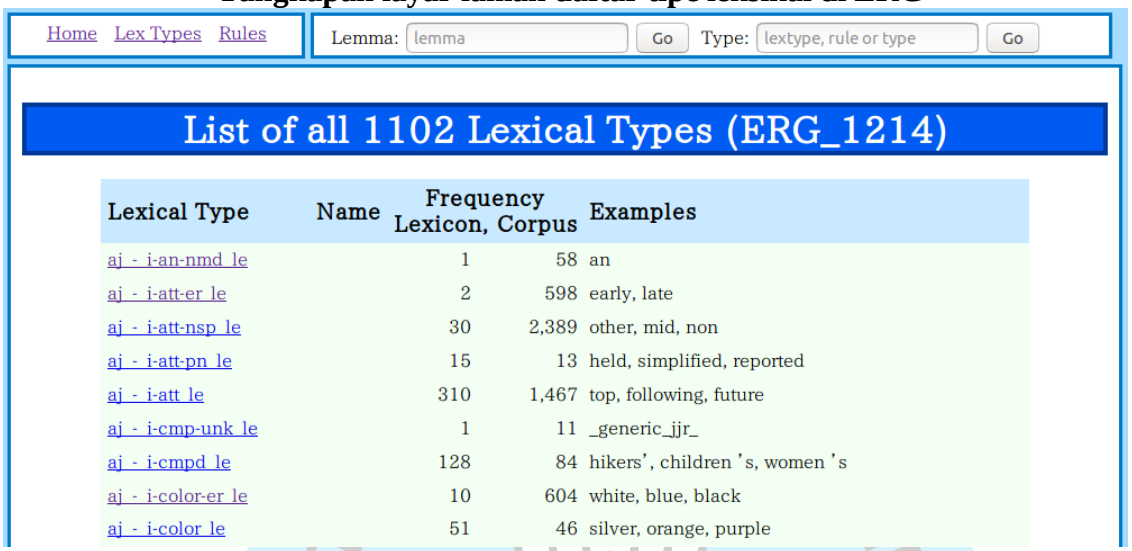


Beranda ini berisi metadata ERG, misalnya nama pengembang, kontributor, alamat laman, tanggal revisi terakhir, lisensi, serta jumlah aturan-aturan leksikal dan gramatikal. Jika ingin mencari informasi terperinci yang ada di pangkalan data, pengguna dapat memilih tautan pencarian yang ada di bagian kiri atas laman beranda. Setelah itu, pengguna akan dibawa ke laman yang pojok kiri atasnya berisi pilihan ‘Beranda’, ‘Tipe Leksikal’, dan ‘Aturan’. Jika ‘Tipe Leksikal’ dipilih, laman berisi daftar semua tipe leksikal yang ada akan dimunculkan (lihat Gambar 3). Tipe leksikal

[10] https://github.com/fcbond/ltdb
[11] http://moin.delph-in.net/LkbLtdb
[12] http://compling.hss.ntu.edu.sg/ltdb/ERG_1214//index.html
[13] http://compling.hss.ntu.edu.sg/ltdb/Jacy_1301/

adalah penggolongan kelas kata yang terperinci, dapat juga disebut sebagai subkategorisasi kelas kata.

Gambar 3
Tangkapan layar laman daftar tipe leksikal di ERG

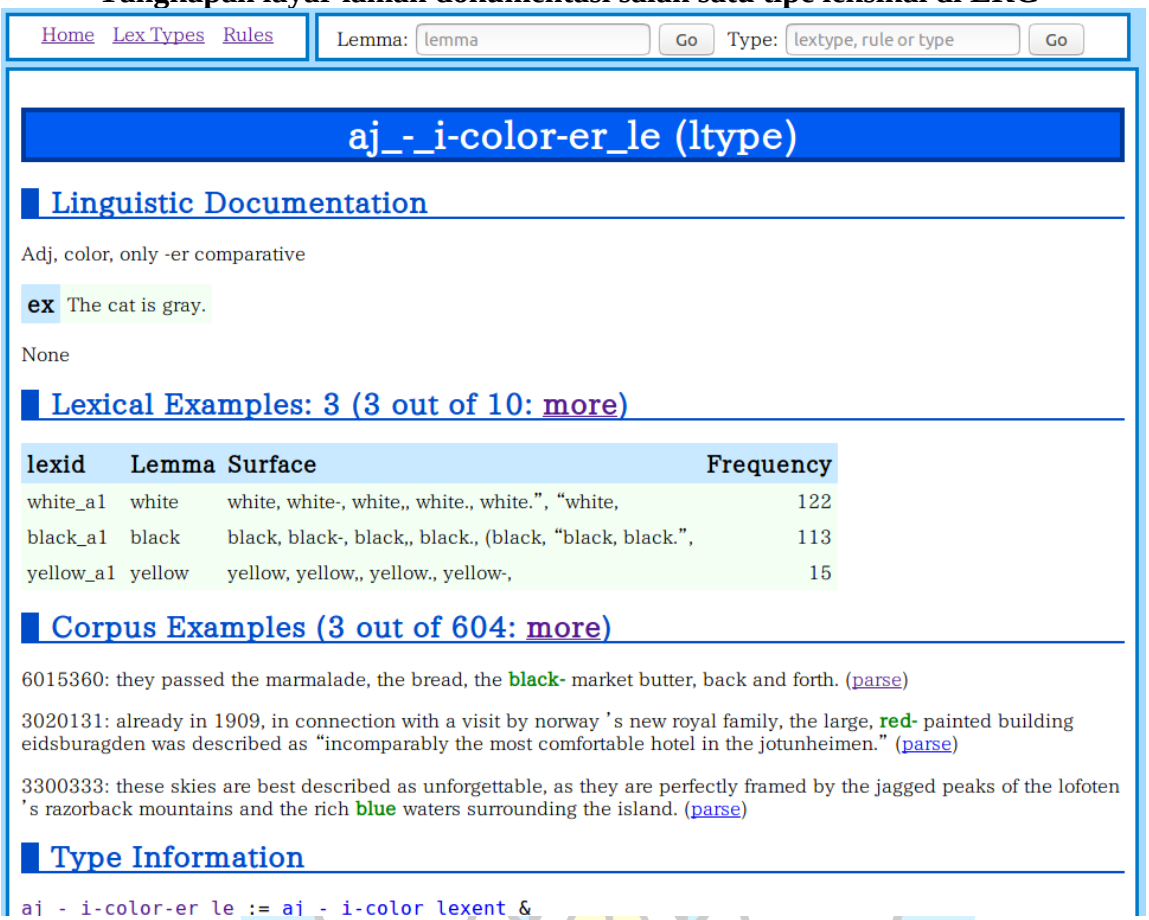


Lexical Type	Name	Frequency Lexicon, Corpus	Examples
aj - i-an-nmd_le		1 58	an
aj - i-att-er_le		2 598	early, late
aj - i-att-nsp_le		30 2,389	other, mid, non
aj - i-att-pn_le		15 13	held, simplified, reported
aj - i-att_le		310 1,467	top, following, future
aj - i-cmp-unk_le		1 11	_generic_jjr_
aj - i-cmpd_le		128 84	hikers', children 's, women 's
aj - i-color-er_le		10 604	white, blue, black
aj - i-color_le		51 46	silver, orange, purple

Gambar 3 menunjukkan bahwa dalam tata bahasa komputasional bahasa Inggris ERG, adjektiva warna seperti putih (*white*), biru (*blue*), dan hitam (*black*) digolongkan dalam satu kategori bernama 'aj_-_i-color-er_le' dengan jumlah total 10 kata dan frekuensi 604 (muncul 604 kali di korpus), sementara perak (*silver*), jingga (*orange*), dan ungu (*purple*) digolongkan dalam kategori yang berbeda bernama 'aj_-_i-color_le' dengan jumlah total 51 kata dan frekuensi 46 (muncul 46 kali di korpus). Penggolongan ini berdasarkan aturan bahwa dalam bahasa Inggris, adjektiva seperti putih (*white*), biru (*blue*), dan hitam (*black*) mendapat akhiran *-er* dalam bentuk komparatif.

Hal ini dapat diketahui dari laman dokumentasi (lihat Gambar 4) yang berisi dokumentasi linguistik, contoh leksikal beserta lema (bentuk dasar) dan bentuk permukaannya (bentuk yang muncul di korpus) serta frekuensi (jumlah kemunculan di korpus) masing-masing lema, contoh-contoh yang diambil dari korpus, dan informasi tipe (implementasi yang terdapat di tata bahasa komputasional). Di bagian contoh korpus, terdapat tautan 'uraikan'. Jika pengguna memilih tautan tersebut, maka laman berisi pohon sintaks dan representasi semantik kalimat yang dipilih akan dimunculkan. Informasi pohon sintaks dan representasi semantik tersebut diambil dari *treebank* yang ada.

Gambar 4
Tangkapan layar laman dokumentasi salah satu tipe leksikal di ERG



Tahap awal pengembangan TBBI Daring Terpadu adalah dengan menggunakan LTDB ini yang berisi dokumentasi tipe-tipe leksikal dan aturan-aturan tata bahasa yang ada di INDRA. Bab berikut ini membahas fenomena kebahasaan konstruksi kopula dalam bahasa Indonesia, analisis dan model HPSG yang telah dibuat beserta implementasinya di INDRA, serta tampilan laman LTDB (TBBI Daring).

PEMBAHASAN

Bab ini membahas salah satu fenomena kebahasaan yang telah dianalisis dan diimplementasikan di INDRA, yaitu konstruksi kopula dasar (Moeljadi et al. 2016). Analisis kopula dalam bahasa Indonesia telah ditulis di berbagai buku tata bahasa, misalnya Alwi et al. (2014) dan Sneddon et al. (2010). Alwi et al. (2014) menulis bahwa 'kalimat berpredikat nominal' atau 'kalimat persamaan' atau 'kalimat ekuatif' dapat menggunakan 'adalah' untuk memisahkan subjek dari predikat, dan 'adalah' dapat disulih dengan 'ialah' atau 'merupakan'. Sneddon et al. (2010) menambahkan bahwa

‘ialah’ dapat digunakan jika subjeknya orang ketiga karena ‘ialah’ berasal dari ‘ia’. Contoh (6) menunjukkan bahwa ‘saya’ tidak dapat menjadi subjek konstruksi kopula dengan ‘ialah’.

(6) Saya (adalah/*ialah/merupakan) guru.

Moeljadi et al. (2016) menyatakan bahwa ‘merupakan’ adalah verba yang sedang dalam proses menjadi kopula. Verba ‘merupakan’ tidak dapat digunakan jika predikat nominal bersifat spesifik, misalnya nama diri, demonstrativa, atau pronomina (lihat Contoh (7)). Dengan kata lain, ‘merupakan’ digunakan jika predikat nominal adalah kata benda umum (*common noun*).

(7) Orang itu (adalah/ialah/*merupakan) Budi.

Verba ‘merupakan’ dapat didahului oleh pewatas depan, sementara ‘adalah’ dan ‘ialah’ tidak dapat (lihat Contoh (8)).

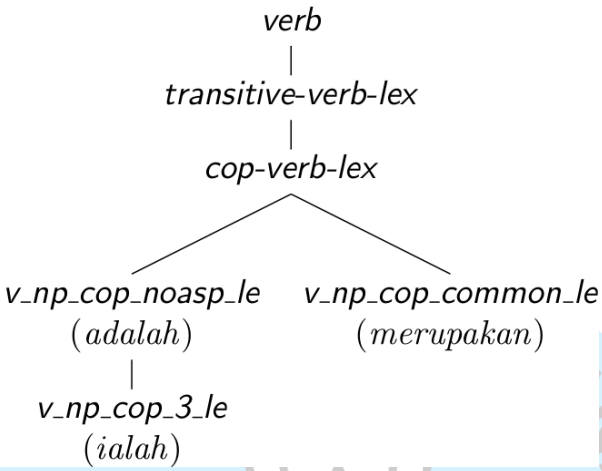
(8) Ini sudah/akan *adalah/*ialah/merupakan hal yang luar biasa.

Analisis di atas didapat berdasarkan data korpus bahasa Indonesia yang ada di Nanyang Technological University Multilingual Corpus (NTU-MC) (Tan dan Bond 2012) yang berisi 2.975 kalimat.

Verba kopulatif ‘adalah’, ‘ialah’, dan ‘merupakan’ memiliki dua argumen, mirip verba transitif secara sintaksis. Verba-verba kopulatif tersebut dapat digolongkan secara terperinci, seperti yang ditunjukkan melalui hierarki tipe leksikal pada Gambar 5. Semua verba kopulatif berasal dari tipe atau kelas kata verba (*verb*), khususnya verba transitif (*transitive-verb-lex*). Karena verba kopulatif berbeda dengan verba transitif lainnya, misalnya tidak memiliki bentuk pasif, verba kopulatif memiliki tipe sendiri (*cop-verb-lex*). Karena ‘adalah’ dan ‘ialah’ tidak dapat didahului pewatas depan, sementara ‘merupakan’ dapat, tipe verba kopulatif perlu dibagi lagi menjadi dua tipe: (1) yang dapat didahului pewatas depan (*v_np_cop_common_le*), yaitu ‘merupakan’, dengan batasan predikat nominal harus bertipe kata benda umum (*common noun*); dan (2) yang tidak dapat didahului pewatas depan (*v_np_cop_noasp_le*), yaitu ‘adalah’ dan

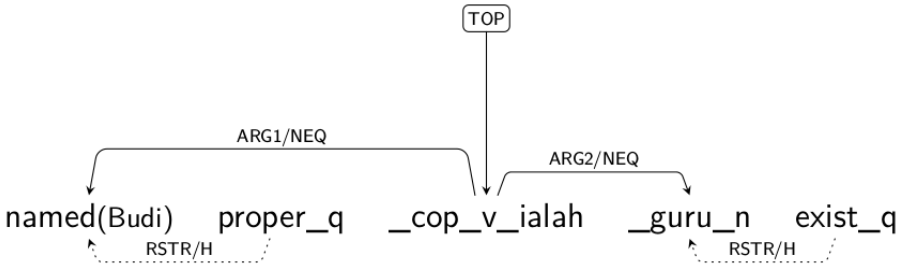
‘ialah’, dengan batasan tanpa pewatas depan. Karena ‘ialah’ hanya dapat digunakan jika subjeknya orang ketiga, ‘ialah’ memiliki tipe sendiri, yaitu *v_np_cop_3_le*, dengan batasan subjek harus orang ketiga.

Gambar 5
Pembagian kelas kata verba kopulatif

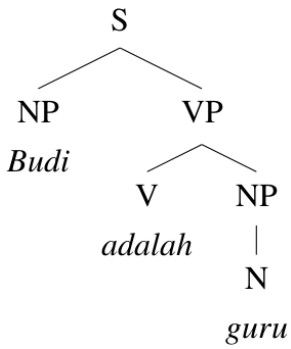


Representasi MRS kalimat kopula mirip dengan MRS kalimat transitif, seperti ditunjukkan pada (9) untuk kalimat ‘Budi ialah guru’ dengan argumen 1 (ARG1) merujuk pada Budi dan argumen 2 (ARG2) merujuk pada guru. Pohon sintaks kalimat ‘Budi adalah guru’ ditunjukkan pada (10).

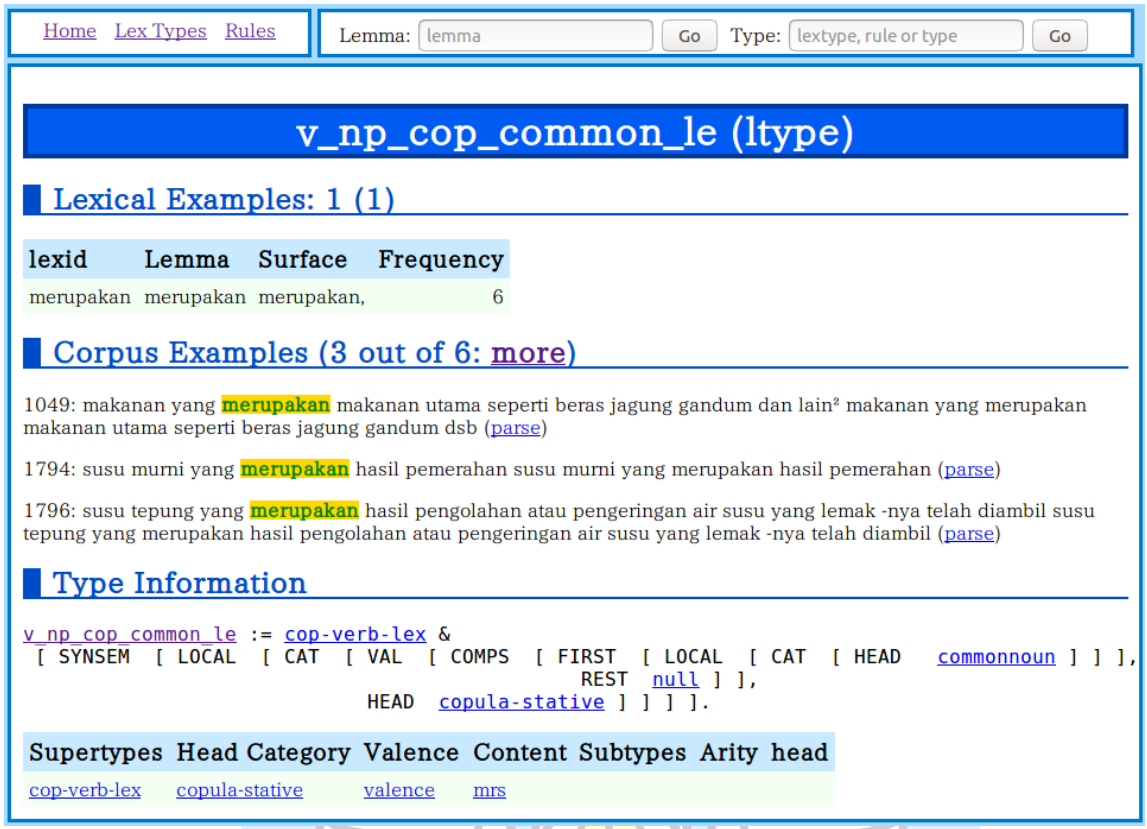
(9)



(10)



Gambar 6
Tangkapan layar laman salah satu tipe verba kopulatif di INDRA



Gambar 6 menunjukkan laman LTDB untuk INDRA (yang berpotensi dikembangkan menjadi TBBI Daring Terpadu). Laman ini berisi dokumentasi tipe leksikal untuk verba kopulatif ‘merupakan’. Tampilan yang ada saat ini menggunakan menu dalam bahasa Inggris, tetapi akan diubah ke dalam bahasa Indonesia. Akan ditambahkan pula dokumentasi linguistik di bagian awal laman ini. Desain laman juga dapat diperbaiki dan dipercantik. Saat ini laman ini berisi: (1) contoh leksikal: bentuk dasar atau lema, bentuk permukaan yang ada di korpus, dan frekuensi atau jumlah kemunculan di korpus; (2) contoh-contoh klausa atau frasa yang mengandung kata ‘merupakan’ yang ditemukan di korpus beserta pilihan untuk menguraikan contoh klausa atau frasa tersebut (jika dipilih, akan muncul pohon sintaks dan representasi semantiknya); (3) informasi tipe leksikal, yang berisi implementasi kode di INDRA dengan batasan-batasannya, misalnya predikat nominal harus berupa kata benda umum (*common noun*) dan diturunkan dari tipe *cop-verb-lex*. Informasi frekuensi dan contoh klausa atau frasa, serta pohon sintaks dan representasi semantik diambil dari *treebank* JATI.

Informasi tipe leksikal atau kelas kata yang terperinci seperti ini (verba kopula, verba transitif dll.) dapat memperkaya informasi kelas kata yang ada di KBBI. Saat ini, informasi kelas kata di KBBI hanya berupa nomina, verba, adjektiva, adverbia, pronomina, numeralia, dan partikel. Bukan tidak mungkin bila di kemudian hari informasi kelas kata yang terperinci, misalnya nomina bernyawa, nomina tak bernyawa, nomina terbilang, nomina tak terbilang, nomina kolektif, verba intransitif, verba transitif, dan verba kopulatif dapat diwujudkan melalui TBBI Daring ini. Selain untuk kepentingan dokumentasi bahasa dan pengayaan kamus (KBBI), TBBI Daring berpotensi dimanfaatkan untuk pengolahan data teks bahasa Indonesia, pengecekan tata bahasa dan kebakuan kalimat, pemelajaran bahasa Indonesia dengan bantuan komputer, dan penerjemahan dengan mesin karena kemampuannya dalam menguraikan dan menghasilkan kalimat-kalimat baku bahasa Indonesia dan hubungannya dengan tata bahasa komputasional bahasa-bahasa lainnya yang menggunakan representasi semantik yang sama.

PENUTUP

Makalah ini telah membahas aspek-aspek pengembangan pangkalan data dan laman TBBI Daring Terpadu tahap awal, yang dimulai dari pembahasan tentang rekayasa tata bahasa, landasan teori HPSG dan MRS, metode penelitian, tata bahasa komputasional bahasa Indonesia INDRA, pangkalan data tipe linguistik LTDB, hingga pembahasan salah satu fenomena kebahasaan konstruksi kopula, dari analisis sampai tampilan laman LTDB (TBBI Daring). TBBI Daring Terpadu berpotensi menjayakan bahasa Indonesia, yaitu dalam penggunaan teknologi informasi dan komunikasi untuk pengembangan bahasa Indonesia, melalui dokumentasi tata bahasa baku bahasa Indonesia dan aplikasi-aplikasi pengolahan teks bahasa Indonesia. TBBI Daring Terpadu perlu dikembangkan lebih lanjut dari segi linguistik, komputasional, serta desain dan tampilan laman. Dari segi linguistik, analisis dan implementasi fenomena kebahasaan, seperti konstruksi adjektival, perlu ditambah. Dari segi komputasional, leksikon yang ada dapat ditambah dan diintegrasikan dengan pangkalan data KBBI Daring. Selain itu, *treebank* yang ada dapat dikembangkan lebih lanjut dengan menganotasi kalimat-kalimat yang terdapat dalam Korpus Indonesia. Karena begitu luasnya fenomena kebahasaan yang ada, pengembangan TBBI Daring Terpadu ini selayaknya dilakukan secara bertahap.

DAFTAR PUSTAKA

Alwi, Hasan, Soenjono Dardjowidjojo, Hans Lapoliwa dan Anton M. Moeliono. (2014). *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka.

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

Arka, I Wayan. (2012). Developing a deep grammar of Indonesian within the ParGram framework: theoretical and implementational challenges. Dalam *26th Pacific Asia Conference on Language, Information and Computation* (hlm. 19—38).

Bender, Emily M., Dan Flickinger dan Stephan Oepen. (2011). Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis. Dalam *Language from a Cognitive Perspective: Grammar, Usage and Processing* (hlm. 5—29). Stanford: CSLI Publications.

Bender, Emily M. dan Antske Sibelle Fokkens. (2010). The LinGO Grammar Matrix: Rapid Grammar Development for Hypothesis Testing. URL http://www.delph-in.net/matrix/hpsg2010/hpsg-tutorial.pdf .

Chomsky, Noam. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3): 113—124.

Copestake, Ann. (2002). *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

Copestake, Ann, Dan Flickinger, Carl Pollard dan Ivan A. Sag. (2005). Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4): 281—332.

Dalrymple, Mary. (2001). Lexical-Functional Grammar. *Syntax and Semantics* 34. Academic Press.

Fan, Zhenzhen, Sanghoun Song dan Francis Bond. (2015). An HPSG-based Shared-Grammar for the Chinese Languages: Zhong [|]. Dalam *Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP* (hlm. 17—24).

Flickinger, Dan. (2000). On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering*, 6(1): 15—28.

Flickinger, Dan, Yi Zhang dan Valia Kordoni. (2010). Grammar Engineering for Deep Linguistic Processing: Project Seminar 2010. URL http://www.coli.uni-saarland.de/~yzhang/ge-ss10/lecture-01.pdf .

Hashimoto, Chikara, Francis Bond dan Dan Flickinger (2007) The lextype DB: A web-based framework for collaborative multilingual grammar and treebank development. Dalam *The First International Workshop on Intercultural Collaboration (IWIC-2007)* (hlm. 44—58). Kyoto.

Kaplan, Ronald dan Joan Bresnan. (1982). Lexical Functional Grammar: A formal system for grammatical representation. Dalam *The Mental Representation of Grammatical Relations* (hlm. 173—281). Cambridge: the MIT Press.

Kwary, Deny A. (2018). Towards the First Online Indonesian National Corpus. Makalah akan diterbitkan dalam *Proceedings of Fourth Asia Pacific Corpus Linguistics Conference (APCLC 2018).*

Moeljadi, David. (2017). Building JATI: A Treebank for Indonesian. Dalam *Proceedings of The 4th Atma Jaya Conference on Corpus Studies (ConCorps 4)* (hlm. 1—9). Jakarta.

Moeljadi, David, Francis Bond dan Luís Morgado da Costa. (2016). Basic copula clauses in Indonesian. Dalam *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar* (hlm. 442—456). Warsaw.

Moeljadi, David, Francis Bond dan Sanghoun Song. (2015). Building an HPSG-based Indonesian Resource Grammar (INDRA). Dalam *Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP* (hlm. 9—16).

Moeljadi, David, Ian Kamajaya dan Dora Amalia. (2017). Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications. Dalam Hai Xu (ed.), *Proceedings of the 11th International Conference of the Asian Association for Lexicography* (hlm. 64—80). Guangzhou: The Asian Association for Lexicography Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

Pollard, Carl dan Ivan A Sag. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Sag, Ivan A., Thomas Wasow dan Emily M. Bender. (2003). *Syntactic Theory: A Formal Introduction*. Stanford: CSLI Publications.

Siegel, Melanie, Emily M. Bender dan Francis Bond. (2016). *Jacy: An Implemented Grammar of Japanese*. Stanford: CSLI Publications.

Sneddon, James Neil, Alexander Adelaar, Dwi Noverini Djenar dan Michael C. Ewing. (2010). *Indonesian Reference Grammar*. New South Wales: Allen dan Unwin.

Tan, Liling dan Francis Bond. (2012). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4): 161—174.