Templat Makalah KBI XI 2018

KORPUS BERANOTASI: KE ARAH PENGEMBANGAN KORPUS BAHASA-BAHASA DI INDONESIA

Annotated Corpus: Toward Development of Language Corpora in Indonesia

Totok Suhardijanto[a], Arawinda Dinakaramani[b]
[a]Fakultas Ilmu Pengetahuan Budaya, Universitas Indonesia
[b] Fakultas Ilmu Komputer, Universitas Indonesia
Pos-el: totok.suhardijanto@ui.ac.id; arawinda.dinakaramani@ui.ac.id

Abstrak

Meskipun dikenal sebagai negara dengan keragaman bahasa dan budaya terbesar kedua di dunia setelah Papua Nugini, ironisnya Indonesia juga dikenal sebagai negara yang minim sumber daya bahasa elektronis. Ethnologue (Simons and Fennig 2018) menyebutkan bahwa terdapat 719 bahasa daerah di Indonesia yang tentu saja akan memakan waktu dan biaya untuk membangun sumber daya bahasa (SDB) untuk kesemuanya. Makalah ini menyajikan upaya pembangunan SDB untuk bahasa-bahasa di Indonesia dengan mengutamakan pengembangan korpus beranotasi pada bahasa-bahasa utama. Pada fase pertama, dalam SBD yang dibangun, dikembangkan korpus bahasa Jawa dan bahasa Indonesia/Melayu. Bahasa Jawa dipilih sebagai proyek perintis karena bahasa ini merupakan bahasa utama di Indonesia dengan penutur 84,3 juta orang. Sementara itu, bahasa Indonesia/Melayu dipilih karena posisinya sebagai bahasa nasional dan juga lingua franca di Indonesia. Data yang digunakan untuk mengembangkan korpus ini—pada tahap awal—difokuskan pada teks tertulis dengan mempertimbangkan keragaman variasi bahasa tulis yang ada di Indonesia. Dalam makalah ini disampaikan apa saja yang selama ini menjadi tantangan, apa yang telah dicapai, dan apa yang menjadi tujuan pengembangan pada masa yang akan datang.

Kata-kata kunci: sumber daya bahasa, korpus beranotasi, bahasa Jawa, bahasa Indonesia/ Melayu

Abstract

Although considered as the second most linguistically-diverse country, Indonesia is ironically also known as a country with many under-resourced languages. This paper presents our attempt to develop language resources for languages of Indonesia. Since there are 719 indigenous languages in Indonesia, it would be very time-consuming and costly to develop LR for all Indonesian indigenous languages. On that account, the initial phase will focus only on major languages. From these major language, Javanese and Indonesian are chosen for our pilot project. Javanese is the most important language with the largest number of speakers in Indonesia. With the total number of speakers reaching 84.3 million people, Javanese is regarded as the twelfth most spoken language in the world. Indonesian or Malay is chosen due to its national language status and widely has been known as a lingua franca in the insular region of Southeast Asian. This paper discusses the drawbacks

and opportunities in our attempt to build a Javanese annotated-corpus that is publicly accessible. At the first phase, we have developed the database system and architecture for Javanese corpus building. In this paper, we also discuss the criteria for corpus building and the design for web-based corpus management and query application.

Keywords: *language resources, annotated corpus, Javanese, Indonesian/Malay*

PENDAHULUAN

Indonesia tercatat sebagai negara kedua di dunia dalam hal kekayaan bahasanya, namun dalam dalam ketersediaan sumber daya bahasa (*language resources*) Indonesia masih sangat tertinggal (Riza 2008). Kompleksitas situasi kebahasaan di Indonesia bisa jadi merupakan salah satu faktor yang menjadi tantangan dalam pembangunan sumber daya bahasa (SDB). Selama ini, situasi kebahasaan di Indonesia dirangkum ke dalam tiga kategori: bahasa nasional, bahasa daerah, dan bahasa asing. Menurut Riza (2008), pembangunan sumber daya bahasa untuk bahasa-bahasa di Indonesia pada umumnya terfokus pada pembangunan sumber daya bahasa untuk bahasa nasional (bahasa Indonesia). Salah satu faktor utama yang menyebabkan hal ini adalah masalah keterbatasan dana dalam pembangunan sumber daya bahasa di Indonesia yang bersumber dari masih kurangnya perhatian pemerintah Indonesia menyangkut kondisi dan kedudukan bahasa-bahasa daerah di Indonesia (Lauder 2016). Prioritas pembangunan pemerintah Indonesia masih tertumpu pada pembangunan di bidang ekonomi dan politik.

Perencanaan dan pengembangan bahasa di Indonesia masih difokuskan pada bahasa nasional, yakni bahasa Indonesia. Bahasa daerah sudah dibuatkan undang-undang tersendiri yang secara eksplisit menggambarkan situasi keterancaman bahasa daerah, namun pedoman kebijakan yang dapat mengatasi masalah ini pada tingkat operasional belum tersedia (Lauder 2016). Hal ini pulalah yang menyebabkan mengapa penelitian dan pembangunan sumber daya bahasa-bahasa daerah di Indonesia tidak mengalami perkembangan signifikan sejak situasi yang disampaikan oleh Riza (2008) sepuluh tahun silam.

Dari 719 bahasa-bahasa di Indonesia, 706 bahasa merupakan bahasa yang masih digunakan penuturnya, sedangkan 13 bahasa merupakan bahasa yang tergolong ke dalam kategori terancam punah (Lauder 2016). Bahasa-bahasa daerah di Indonesia beragam dalam tipe bahasa dan hal jumlah penuturnya. Dari segi tipe bahasanya, bahasa-bahasa Indonesia dapat dikelompokkan ke dalam dua kategori besar: bahasa-bahasa Austronesia dan Non-Austronesia. Bahasa-bahasa Austronesia tersebar di wilayah barat dan timur Indonesia, sedangkan bahasa-bahasa non-Austronesia hanya tersebar di wilayah Papua, Maluku, dan Nusa Tenggara yang semuanya terletak di wilayah timur Indonesia. Dari jumlah penutur, 386 bahasa diucapkan oleh 5.000 penutur atau kurang; 233 bahasa memiliki 1.000 penutur atau kurang; 169 bahasa memiliki 500 penutur atau kurang; dan 52 bahasa memiliki 100 penutur atau kurang (Gordon 2005). Sementara itu, menurut Simons & Fennig (2018), hanya ada 20 bahasa yang dituturkan oleh lebih dari 1 juta orang, termasuk di dalamnya bahasa Jawa dengan jumlah penutur 84,3 juta.

SUMBER DAYA BAHASA

Jika melihat perkembangannya hingga saat ini, pembangunan sumber daya bahasa di Indonesia pada umumnya masih bersifat sporadis. Masing-masing mengembangkan korpusnya sendiri untuk keperluan individual. Meskipun banyak pusat penelitian melakukan pembangunan sumber daya bahasa yang berfokus pada bahasa Indonesia, pada umumnya sumber daya tersebut dibangun untuk kebutuhan terbatas, yakni penelitian atau proyek individual (Suhardijanto 2016). Kerja sama dan kolaborasi antarlembaga di Indonesia dalam pembangunan sumber daya bahasa boleh dikatakan belum terbangun dengan baik. Meskipun demikian, pada tahun-tahun terakhir ini, ada upaya ke arah kolaborasi yang dijembatani oleh organisasi profesi peneliti di bidang linguistik komputasional dan pemrosesan bahasa alami.

Ada baiknya dijelaskan terlebih dahulu apa yang disebut dengan sumber daya bahasa dan apa perbedaannya dengan korpus? Menurut ELRA (European Language Resources Association), istilah sumber daya bahasa (SDM) atau *language resources* merujuk pada set data dan deskripsi bahasa dalam bentuk terbaca mesin, digunakan khususnya untuk membangun, meningkatkan, atau mengevaluasi bahasa manusia dan algoritma atau sistem wicara, dan secara umum sebagai sumber daya utama bagi pelokalan peranti lunak dan industri pelayanan bahasa, bagi kajian kebahasaan, transaksi internasional penerbitan elektronik, area subyek spesialis dan pengguna produk akhir. Contoh dari SDB adalah korpus lisan dan tulisan, pangkalan data leksikon, tata bahasa, pangkalan data peristilahan, alat peranti dasar untuk pemerolehan, penyiapan koleksi, pengelolaan, pengaturan dan penggunaan data kebahasaan.

Dengan demikian, dapat dikatakan bahwa dalam tahun-tahun terakhir ini, perkembangan terjadi dalam SDB di Indonesia dengan munculnya Kateglo (http://kateglo.com/), tesaurus tematis (http://tesaurus.kemdikbud.go.id/tematis/), KBBI daring (https://kbbi.kemdikbud.go.id/), dan beberapa kamus berbasis web lainnya. Meskipun demikian, dalam hal korpus elektronis, dapat dikatakan belum ada SDB yang dapat diakses oleh masyarakat. Kalaupun ada, yang menyediakan dan mengembangkannya adalah pihak asing. Sebut saja *Malay Concordance* (http://mcp.anu.edu.au/), Korpus Leipzig (http://corpora.uni-leipzig.de/en?corpusId=ind_mixed_2013), Korpus Bahasa Indonesia SEAlang (http://sealang.net/indonesia/corpus.htm), Korpus Bahasa Jawa SEAlang (http://sealang.net/java/corpus.htm) dan SketchEngine (https://www.sketchengine.eu/).

Dalam makalah ini, dibahas pembangunan sumber daya bahasa-bahasa di Indonesia yang sedang kami kerjakan dalam beberapa tahap. Pada tahap pertama, kami memilih Bahasa Jawa dan Bahasa Indonesia. yang dipilih karena dua alasan. Pertama, dari segi jumlah penutur, bahasa Jawa merupakan bahasa daerah yang utama di Indonesia. Kemudian, bahasa Indonesia dipilih karena statusnya sebagai bahasa nasional. Kedua, dari segi dokumentasi, bahasa Jawa dan bahasa Indonesia—dalam hal ini sebagai lanjutan dari bahasa Melayu—merupakan bahasa yang mempunyai tradisi tulis yang panjang di Indonesia. Dengan demikian, hal itu akan memudahkan pada tahap pengumpulan data.

Selanjutnya, pembangunan sumber daya bahasa ini mempunyai tiga tujuan. Pertama, pembangunan ini bertujuan untuk menyediakan sumber daya bahasa nusantara yang berupa korpus beranotasi yang dapat dimanfaatkan oleh para peneliti dan pengembang aplikasi. Pada tahap awal, anotasi yang akan dikembangkan adalah sistem anotasi kelas kata (*part of speech* (POS) *tagging system*) untuk bahasa Jawa dan bahasa

Indonesia. Tujuan kedua adalah untuk menemukan model pengembangan yang dapat dimanfaatkan untuk membangun sumber daya bahasa bagi bahasa-bahasa daerah lain di Indonesia yang jumlahnya sangat banyak. Terakhir, pembangunan sumber daya bahasa nusantara ini merupakan salah satu upaya pelestarian bahasa daerah di Indonesia yang saat ini terus mengalami tekanan baik dari bahasa nasional—bahasa Indonesia, maupun dari bahasa asing—bahasa Inggris.

PERIHAL KORPUS

Pada saat ini, korpus—atau linguistik korpus sering diasosiasikan dengan peranti komputer dan teknik analisis yang menggunakan komputer untuk mengkaji data bahasa. Padahal, istilah korpus bukan hal baru di dalam sejarah kajian linguistik, paling tidak sudah ada sebelum tahun 1950-an (McEnery dkk. 2006: 4). Namun, dalam perkembangannya, korpus baru menjadi isu kembali setelah pelibatan komputer modern dalam pengelolaannya sejak tahun 1980-an.

Pengertian korpus sendiri, menurut Sinclair (2005: 19), adalah "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." Jadi, ada penekanan pada "bentuk elektronis" atau "terbaca machine" (*machine readable*) tentang korpus di era modern. Inilah yang membedakan dengan korpus pada masa-masa sebelumnya.

Selain terkait bentuk elektronis, pelibatan komputer modern juga telah mengubah jangkauan dan cakupan korpus. Jika sebelumnya jumlah korpus terbatas karena keterbatasan manusia untuk mengoleksi dan mengelola korpus, dengan adanya komputer, kedua hal tersebut tidak lagi menjadi isu yang membutuhkan perhatian lagi. Dengan kehadiran komputer, dimungkin juga korpus dalam bentuk multimedia, tidak hanya dalam bentuk teks.

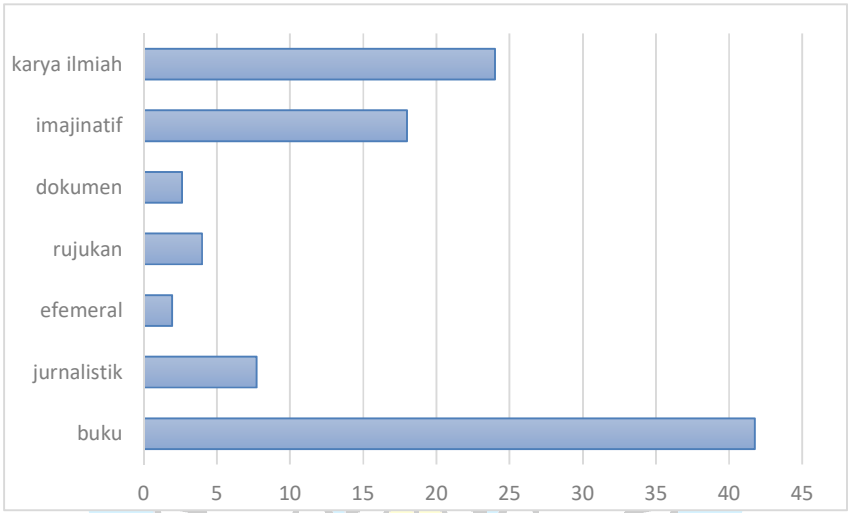
Dalam merancang korpus, ada beberapa hal yang perlu dipertimbangkan, yakni kriteria pemilihan, ukuran korpus, keautentikan data, media penyimpanan, dan manipulasi data. Pada kriteria pemilihan, ditentukan apa yang menjadi tujuan utama dalam penyusunan korpus. Sebagaimana disebutkan pada bagian sebelumnya, penyusunan korpus ini merupakan bagian dari upaya pembangunan SDB di Indonesia. Jadi, korpus yang dihasilkan adalah korpus umum yang dapat digunakan untuk segala macam keperluan. Dengan demikian, korpus ini sedapat mungkin mewakili berbagai jenis teks atau genre. Oleh karena itu, dalam perancangan korpus ini, ditentukan bahwa data kebahasaan akan dibagi menjadi tujuh kategori.

(1) Karya imaginatif
(2) Karya jurnalistik
(3) Buku
(4) Karya akademik
(5) Efemeral
(6) Dokumen
(7) Rujukan

Yang termasuk ke dalam kategori karya imaginatif adalah puisi, prosa, cerpen, naskah drama, naskah skenario, dan sebagainya. Sementara itu, semua terbitan yang digolongkan ke dalam karya jurnalistik, misalnya surat kabar, majalah, tabloid—baik daring maupun cetak—dimasukkan ke dalam kategori kedua. Kemudian, yang masuk ke

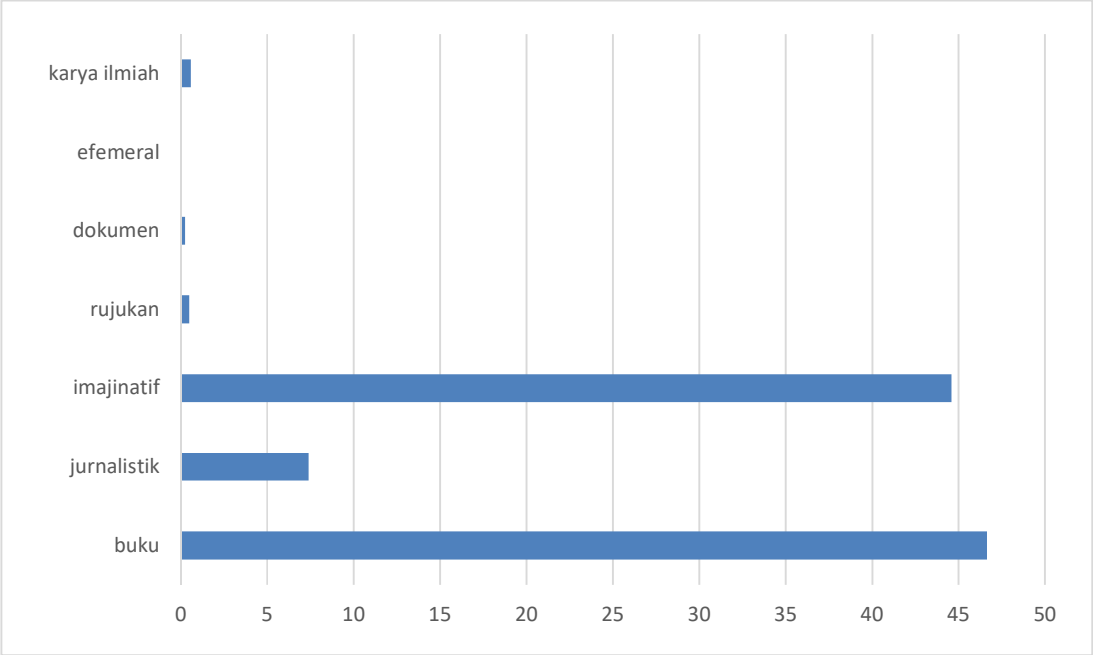
dalam kategori buku antara lain adalah buku pelajaran, buku teks, buku sekolah elektronik, buku bacaan umum, dan sebagainya. Untuk kategori karya ilmiah, termasuk di dalamnya artikel jurnal, makalah dalam prosiding, laporan penelitian, proposal penelitian, skripsi, tesis, dan disertasi. Selanjutnya, yang termasuk ke dalam kategori efemeral adalah teks cetakan atau digital dengan masa berlaku singkat, misalnya brosur, pamflet, leaflet, poster, kartu undangan, kartu ucapan, dan sebagainya. Untuk kategori dokumen, termasuk di dalamnya adalah surat-menyurat, dokumen negara, undang-undang, dan sebagainya. Untuk kategori terakhir, di dalamnya termasuk ensiklopedia, wikipedia, kamus, dan tesaurus.

Untuk korpus bahasa Indonesia, komposisi ketujuh kategori tersebut dapat dilihat pada gambar berikut ini.



Gambar 1: Genre dalam Korpus Bahasa Indonesia

Sementara itu, untuk komposisi korpus bahasa Jawa setakat ini dapat dilihat pada gambar berikut ini.



Gambar 2: Genre dalam Korpus Bahasa Jawa

Selain kategori teks, sebagaimana dalam BNC, pada korpus ini juga dipertimbangkan subyek atau bidang sebagai kategori konten korpus. Ada beberapa subyek yang menjadi kategori konten data pada korpus ini, yakni ilmu pasti alam, ilmu terapan, sosial kemasyarakatan, hubungan internasional, perdagangan dan keuangan, seni, kepercayaan dan pemikiran, urusan global, serta gaya hidup. Meskipun demikian, karena korpus ini masih dalam perkembangan, terkait kategori konten data, masih ada subyek atau topik yang belum terlengkapi. Pelengkapannya akan dilakukan seiring dengan pengembangan korpus dari segi ukuran.

Bagaimana dengan ukuran korpus? Pertanyaan ini sejatinya belum dapat dijawab saat ini karena korpus sedang terus dikembangkan baik dari segi kualitas dan kuantitas data, maupun dari segi kepraktisan penyimpanan dan pengelolaan. Paling tidak, pada saat ini, untuk korpus bahasa Jawa telah terkumpul data sebanyak 2,5 juta token. Sementara itu, untuk korpus bahasa Indonesia, terdapat lebih dari 18 juta token.

Dari segi keautentikan, materi yang digunakan sebagai data korpus merupakan bahan yang memang nyata digunakan masyarakat dalam berbagai bidang kehidupan. Jadi, tidak ada materi yang dibuat khusus untuk melengkapi korpus ini. Sementara itu, dari segi penyimpanan, seluruh data dialihbentukkan dalam format digital, khususnya berkas teks (berkas dengan ekstensi .txt) dengan kode enkoding (*encoding*) UTF-8. Pada saat ini, semua data disimpan di dalam server Universitas Indonesia. Kemudian, perihal manipulasi data, untuk mengakses dan mengelola data korpus, dibangun sistem aplikasi korpus berbasis web. Aplikasi tergolong ke dalam corpus query system (CQS) atau query management system (QMS). Contoh *corpus management system* yang terkenal antara lain adalah *Antconc* (Anthony 2006), SketchEngine (Kilgarrif *et al*. 2014), dan WordSmith (Scott 2016). Sistem Manajemen Korpus yang kami kembangkan diberi nama Korpus

Universitas Indonesia. Perihal aplikasi korpus tersebut dibahas secara detail pada bagian berikut ini.

APLIKASI MANAJEMEN KORPUS

Pengembangan aplikasi Korpus Universitas Indonesia dilakukan melalui beberapa tahap. Tahap pertama adalah pengumpulan kebutuhan (*requirement gathering*). Pada tahap pertama, semua kebutuhan, informasi, dan data awal (*initial data*) yang diperlukan untuk pengembangan aplikasi digali dan dikumpulkan. Selain yang terkait dengan pengembangan aplikasi secara langsung, pada tahap pertama ini pula ditentukan desain korpus bahasa Jawa yang akan menjadi konten dalam aplikasi Korpus Universitas Indonesia. Dari segi struktur internal korpus, secara garis besar data bahasa Jawa yang dikumpulkan terdiri atas data tertulis dan data lisan.

Aplikasi Korpus Universitas Indonesia diharapkan mempunyai sejumlah fitur. Beberapa fitur dapat diakses tanpa *login* dan beberapa lainnya hanya dapat diakses jika pengguna *login* ke aplikasi. Contoh fitur yang dapat diakses oleh pengguna tanpa *login* adalah fitur untuk menampilkan hasil pencarian kata dari korpus dalam bentuk konkordansi (*concordance*) dan fitur untuk menampilkan daftar frekuensi kata (*word list*) dari korpus. Contoh fitur yang hanya dapat diakses oleh pengguna yang *login* ke aplikasi adalah fitur untuk mengelola korpus dan fitur untuk menyumbang data ke korpus. Tahap kedua adalah analisis data awal (*initial data analysis*). Tahap ketiga adalah perancangan pangkalan data (*database*). Tahap keempat adalah perancangan aplikasi berbasis web.

DATA AWAL

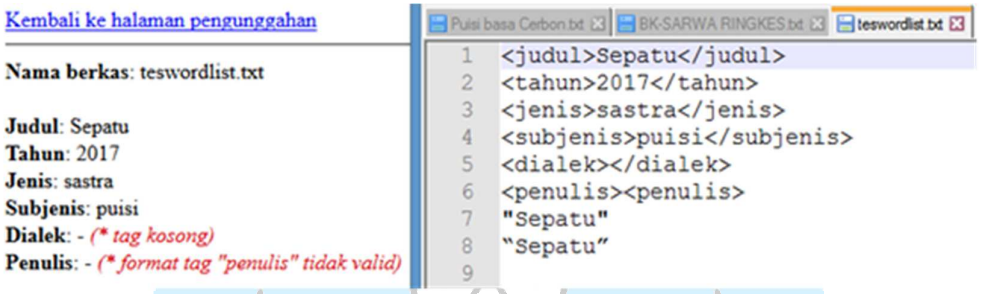
Data awal korpus yang akan diunggah ke aplikasi telah disediakan pada tahap pertama, yaitu pada tahap pengumpulan kebutuhan. Data awal korpus berupa kumpulan berkas teks (.txt). Setiap berkas teks berisi teks berbahasa Jawa yang telah dianotasi secara manual dengan memberi tambahan metadata menggunakan format bahasa pemarkah (*markup language*).

Data awal korpus yang diperoleh pada tahap pengumpulan kebutuhan kemudian dianalisis pada tahap kedua. Analisis dilakukan terutama untuk mengetahui label metadata apa saja yang ada di data awal. Hasil analisis akan digunakan sebagai dasar perancangan pangkalan data dan aplikasi web.

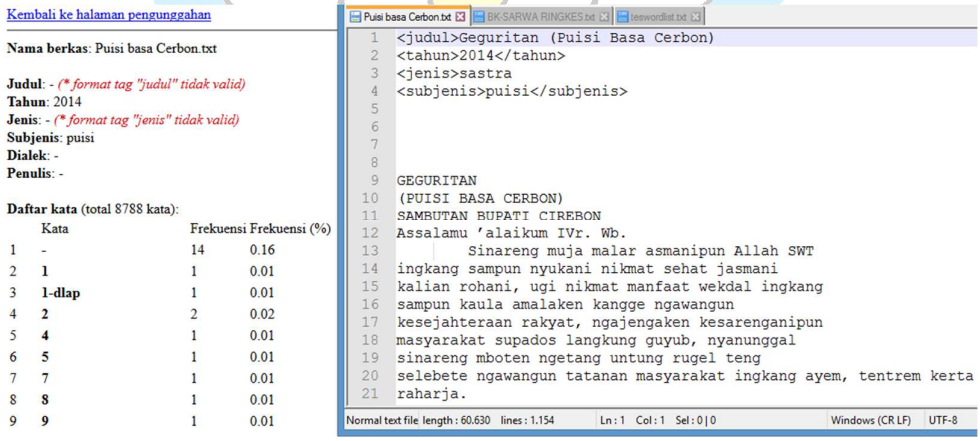
Untuk melakukan analisis data awal, kami membuat aplikasi sederhana berbasis web yang bernama Validator Berkas Korpus (*Corpus File Validator*). Fitur-fitur yang tersedia di aplikasi Validator Berkas Korpus adalah fitur unggah berkas teks, fitur validasi label metadata, fitur untuk menampilkan metadata dari berkas teks yang diunggah, fitur untuk menampilkan daftar kata dan frekuensi kemunculan tiap-tiap kata dari berkas teks yang diunggah, dan fitur untuk menampilkan keseluruhan isi teks dari berkas yang diunggah. Semua fitur yang ada di aplikasi Validator Berkas Korpus akan disertakan dalam aplikasi Korpus Universitas Indonesia.

Dalam proses analisis data awal, ada beberapa masalah yang ditemukan di data awal. Pertama, ada berkas teks yang formatnya rusak, sehingga berkas teks tersebut tidak dapat dibuka dan dibaca. Kedua, ada berkas teks yang nama berkasnya terlalu panjang, sehingga berkas teks tersebut tidak dapat dibuka di komputer. Ketiga, format enkoding berkas teks tidak seragam. Ada berkas teks yang berformat UTF-8 dan ada yang berformat

ANSI. Keempat, ada label metadata yang kosong, misalnya <penulis> </penulis>. Kelima, jumlah label metadata antara satu berkas teks dengan berkas teks lainnya tidak seragam. Ada berkas teks yang memuat enam label metadata, ada berkas teks yang memuat tiga label metadata, dan bahkan ada berkas teks yang tidak memuat label metadata sama sekali. Keenam, ada label metadata yang formatnya salah. Contoh format label metadata yang salah adalah ada label pembuka, tetapi tidak ada label penutup, misalnya di berkas teks tertulis <jenis>sastra, sedangkan seharusnya <jenis>sastra</jenis>. Contoh lain kesalahan format label metadata adalah kesalahan format tag penutup, misalnya di berkas teks tertulis <jenis>karya ilmiah<jenis>, sedangkan seharusnya <jenis>karya ilmiah</jenis>.



Gambar 3: Perbandingan Metadata yang ditampilkan pada aplikasi Validator Berkas Korpus (kiri) dan metadata dalam bentuk berkas asli (kanan)



Gambar 4: Daftar frekuensi kata yang ditampilkan pada aplikasi Validator Berkas Korpus (kiri) yang semula berasal dari sebuah berkas unggahan (kanan).

STRUKTUR PANGKALAN DATA

Entitas-entitas dalam database aplikasi Korpus Universitas Indonesia dapat dikelompokkan menjadi dua kelompok utama, yaitu kelompok entitas pengguna, yang berkaitan dengan pengguna, dan kelompok entitas korpus, yang berkaitan dengan korpus. Kelompok entitas pengguna meliputi peran (*role*), pengguna (*user*), kegiatan pengguna (*user_activity*), dan peran pengguna (*user_role*). Kelompok entitas korpus meliputi korpus, log korpus, dialek, variasi dialek (*dialect_alias*), entri, entri_metadata, entri_teks, bahasa, variasi bahasa (*language_alias*), word, word_count, subtype_meta, type_activity, dan type_meta. Untuk beberapa fitur tertentu, entitas dari kelompok entitas pengguna dapat mempunyai hubungan dengan entitas dari kelompok entitas korpus, misalnya entitas peran pengguna dengan entitas corpus untuk fitur penambahan anggota tim dan entitas pengguna dengan entitas log korpus untuk fitur pengiriman entri.

Database dibangun menggunakan MySQL karena pengembangan aplikasi Korpus Universitas Indonesia berbasis web dilakukan menggunakan XAMPP di komputer dan menggunakan panel kontrol penempatan web (web hosting control panel) cPanel di *shared hosting*.

DESAIN DAN ARSITEKTUR APLIKASI BERBASIS WEB

Aplikasi Korpus Universitas Indonesia berbasis web dirancang dengan pola arsitektur model-view-controller (MVC) dan dibuat menggunakan PHP. Default language yang digunakan dalam page interface aplikasi adalah bahasa Indonesia karena target pengunjung dan pengguna aplikasi adalah orang Indonesia dan orang yang berbahasa Indonesia. Meskipun demikian, pilihan untuk mengganti bahasa page interface menjadi bahasa Inggris juga disediakan untuk pengguna internasional.

Saat ini, aplikasi Korpus Universitas Indonesia hanya terdiri atas satu korpus, yaitu korpus Jawa. Namun, aplikasi Korpus Universitas Indonesia dirancang akan terdiri atas beberapa korpus di masa depan. Oleh karena itu, fitur utama yang ditampilkan di laman beranda aplikasi adalah fitur select corpus yang memungkinkan pengguna dapat memilih korpus yang ingin dilihat.

Gambar 5: Menu pencarian laman web Korpora Bahasa Jawa

Pengguna aplikasi dibagi menjadi tujuh tipe, yaitu admin, editor kepala (*chief editor*), editor, kontributor data (*data contributor*), anggota tertunda (*pending member*), pengguna tertunda (*pending user*), dan tak terkategori (*uncategorized*). Pengguna kategori (*uncategorized user*) adalah pengguna yang tidak login dan pengguna yang login dengan akun selain akun admin dan akun yang tidak termasuk ke dalam anggota tim sebuah korpus. Pengguna tertunda (*pending user*) adalah pengguna yang sudah mendaftarkan akun baru, tetapi masih menunggu persetujuan admin. Anggota tertunda (*pending member*) adalah pengguna yang sudah mengirim permohon berkontribusi ke sebuah proyek korpus, tetapi masih menunggu persetujuan editor kepala atau editor dari korpus tersebut. Editor kepala, editor, dan kontributor data adalah anggota tim sebuah korpus.

Ada dua modul utama aplikasi, yaitu modul pengguna dan modul korpus. Modul pengguna menangani fitur-fitur yang berkaitan dengan pengguna, misalnya fitur pendaftaran (*sign up*), ganti kata pas (*reset password*), login, edit profil (*edit profile*), dan edit akun (*edit account*). Modul korpus menangani fitur-fitur yang berkaitan dengan korpus, misalnya fitur pemilihan korpus , lihat daftar kata, dan menampilkan hasil pencarian kata dari korpus dalam bentuk konkordansi. Fitur-fitur yang berkaitan dengan korpus ada yang dapat diakses oleh semua tipe pengguna, misalnya lihat daftar kata, tetapi ada juga yang hanya dapat diakses oleh pengguna tipe tertentu, misalnya fitur tambah editor yang hanya dapat diakses oleh editor kepala.

ANOTASI

Menurut Kübler & Zinmeister (2015: 21), meskipun korpora merupakan peranti yang bermanfaat bagi penelitian kebahasaan, ada kasus-kasus yang membutuhkan lebih dari sekadar akses terhadap koleksi teks. Misalnya, dalam kasus, ketika kita mencari contoh-contoh pemakaian klausa relatif dengan *yang* yang satu mempunyai induk, sementara yang lain tidak.

(1) [Gadis]$_{induk}$ [yang mudah bosan]$_{pewatas}$ itu sudah berganti pacar lebih dari tiga kali dalam setahun ini.
(2) [Yang bosan] di ruangan penumpang silakan mencari angin di geladak kapal agar dapat menemukan udara segar dan pemandangan laut yang indah.

Tentu saja menemukan konstruksi seperti ini sulitnya bukan main jika teks korpus yang kita miliki masih berupa teks mentah. Salah satu informasi yang bisa akses secara otomatis antara lain adalah frekuensi kata. Jadi, idealnya memang ada informasi atau metadata yang kita sematkan pada teks mentah tersebut. Salah satu teknik yang lazim digunakan adalah pelabelan atau anotasi.

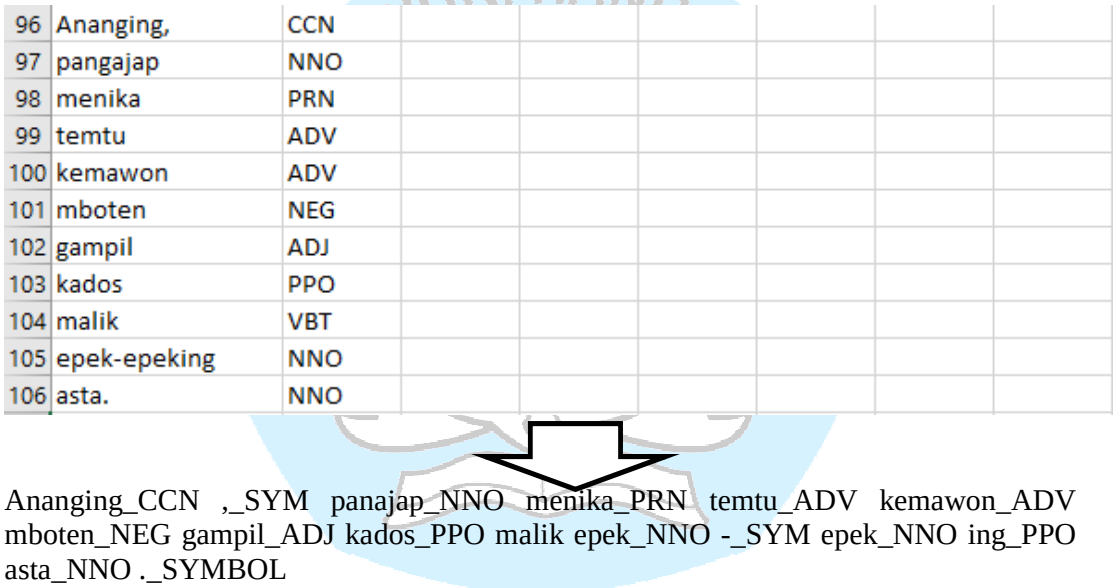
Ada cukup banyak teknik anotasi yang biasa diterapkan pada korpus, meliputi anotasi kata, anotasi sintaktis, anotasi semantis, dan anotasi wacana (Kübler & Zinmeister 2015: 45-156). Tentu saja, pekerjaan anotasi merupakan pekerjaan yang cukup menantang dari segi kerumitan sistem serta biaya dan waktu yang dibutuhkan. Oleh karena itu, untuk tahap pertama, pada Korpus Universitas Indonesia diterapkan anotasi kata terlebih dahulu. Anotasi kata yang diterapkan adalah anotasi kelas kata (*part of speech tagging*). Set pelabelan yang digunakan adalah set label milik INACL (*Indonesia Association for Computational Linguistics*) yang dikembangkan oleh Totok Suhardijanto, Ayu Purwantiari, dan Gunarso untuk kebutuhan pengembangan sumber daya bahasa Indonesia. Dengan demikian, set pelabelan inilah yang akan digunakan untuk menganotasi korpus teks bahasa Indonesia. Untuk data teks berbahasa Jawa, dikembangkan lagi set label khusus berdasarkan set label INACL untuk anotasi kelas kata bahasa Jawa pada penelitian ini (Suhardijanto 2017). Sistem pelabelan kelas kata bahasa Jawa ini memiliki 25 label seperti yang dapat dilihat pada Tabel 1 di bawah ini. Sementara itu, untuk set pelabelan kelas kata bahasa Sunda sedang dalam tahap perencanaan.

Tabel 1 Set Pelabelan Kelas Kata Bahasa Jawa

No.	Simbol	Keterangan
(1)	ADJ	Adjektiva
(2)	ADK	Penanda Kala
(3)	ADV	Adverbia
(4)	ART	Artikula
(5)	CCN	Konjungsi Koordinatif
(6)	CSN	Konjungsi Subordinatif
(7)	CUR	Penanda Mata Uang
(8)	INT	Interjeksi
(9)	NBR	Bilangan Angka (Numerik)
(10)	NEG	Penafian
(11)	NNO	Nomina Umum
(12)	NNP	Nomina Nama Diri
(13)	NUM	Numeralia
(14)	PAR	Partikel
(15)	PPO	Preposisi
(16)	PRI	Pronomina Interogatif
(17)	PRN	Pronomina
(18)	PRR	Pronomina Relatif
(19)	SYM	Tanda Baca
(20)	UNS	Satuan Pengukuran
(21)	VBI	Verba Intransitif

(22)	VBK	Verba Keadaan (*State Verbs*)
(23)	VBP	Verba Penghubung (*Linking Verbs*)
(24)	VBT	Verba Transitif
(25)	ZUK	Kata Tak Dikenal

Pada bagian ini, dilaporkan proses pelabelan anotasi kelas kata bahasa Jawa karena pelabelan inilah yang telah dikerjakan ketika tulisan ini disusun. Proses pelabelan pertama kali dilakukan dengan melakukan anotasi manual terhadap set data yang akan menjadi set rujukan (*golden set*). Pelabelan manual dilakukan secara massal oleh 20 mahasiswa Program Studi Sastra Jawa FIB Universitas Indonesia. kemudian, set data kemudian diperiksa lagi secara otomatis baik oleh ahli bahasa Jawa maupun oleh mesin. Data seperti angka, lambang mata uang, satuan pengukuran, dan tanda baca, padanya secara otomatis akan disematkan label NBR, CUR, UNS, dan SYM secara berturut-turut. Pada bentuk turunan yang merupakan kombinasi dua kata, misalnya *epeking* dan *omahe*, dilakukan pemisahan secara manual dan kemudian diberi label secara terpisah: *epek*_NNO *ing*_PPO dan *omah*_NNO *e*_PRN.



96	Ananging,	CCN
97	pangajap	NNO
98	menika	PRN
99	temtu	ADV
100	kemawon	ADV
101	mboten	NEG
102	gampil	ADJ
103	kados	PPO
104	malik	VBT
105	epek-epeking	NNO
106	asta.	NNO

Ananging_CCN ,_SYM panajap_NNO menika_PRN temtu_ADV kemawon_ADV mboten_NEG gampil_ADJ kados_PPO malik epek_NNO -_SYM epek_NNO ing_PPO asta_NNO ._SYMBOL

Gambar 6: Contoh penerapan pelabelan manual (atas) yang kemudian dielaborasi semiotomatis (bawah) pada teks bahasa Jawa

Set data acuan tersebut kemudian digunakan sebagai data pelatihan bagi sistem pembelajaran mesin (*machine learning*) untuk membangun, memvalidasi, dan meningkatkan kualitas sistem anotasi kelas kata bahasa Jawa.

KONDISI TERMUTAKHIR

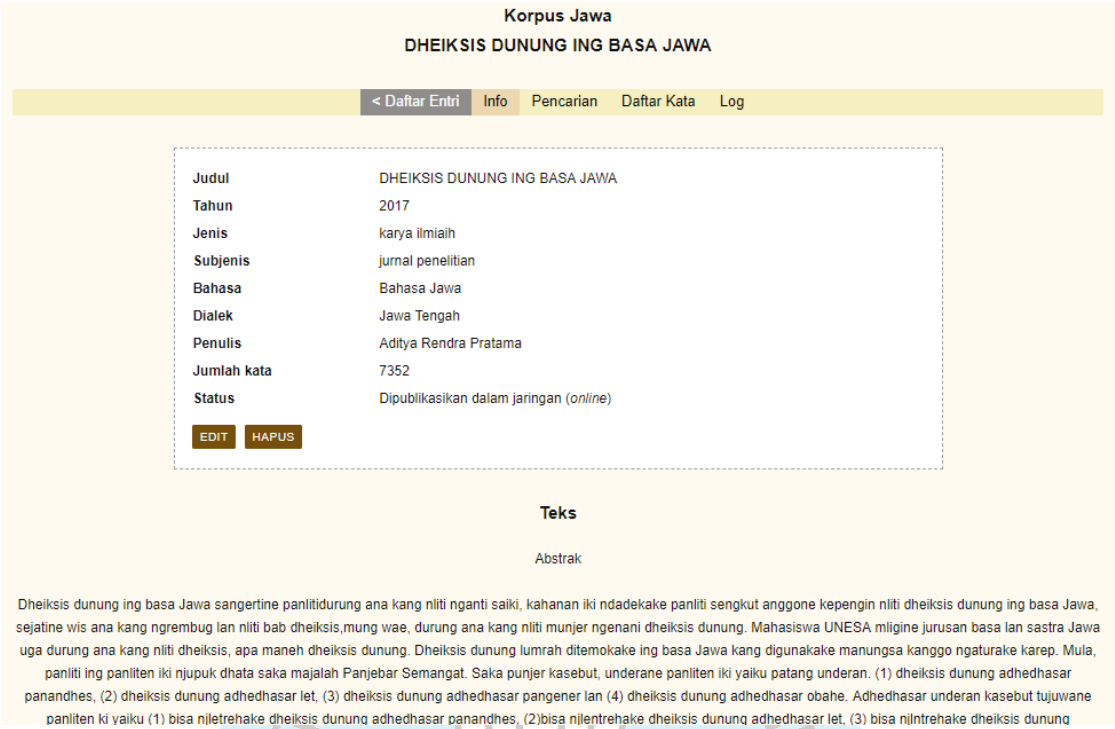
Ketika naskah ini ditulis, beberapa perkembangan terjadi pada proyek Korpus Universitas Indonesia. Untuk konten korpus, selain korpus bahasa Jawa dan Indonesia, pada saat ini sedang dikembangkan pula korpus bahasa Sunda. Setelah digitalisasi teks Sunda, dilakukan pembersihan terhadap teks untuk menghilangkan bagian-bagian yang tidak

dibutuhkan dan memperbaiki masalah tipografi yang kadang-kadang salah. Beberapa naskah telah siap untuk diunggah ke dalam pangkalan data Korpus Bahasa Sunda. Sementara itu, saat ini sedang disiapkan pula set label untuk bahasa Sunda berdasarkan set label INACL.

Dari segi aplikasi, terdapat tambahan satu fitur, yaitu fitur edit entri. Fitur edit entri tersedia untuk (1) entri yang sudah disetujui dan dipublikasikan maupun (2) entri kiriman dari kontributor yang masih harus ditinjau oleh editor. Fitur edit entri untuk entri yang sudah disetujui dan dipublikasikan tersedia di laman entri dan hanya dapat diakses oleh admin, kepala editor, dan editor. Fitur edit entri untuk entri kiriman dari kontributor tersedia di laman kelola korpus dan dapat diakses oleh (1) kontributor yang mengirim entri

tersebut, (2) admin, (3) kepala editor, dan (4) editor. Jika kontributor pengirim entri mengedit entri yang dia kirim, status entri tersebut masih berupa entri kiriman dan harus ditinjau oleh editor. Jika kepala editor, editor, dan admin yang mengedit entri kiriman dari kontributor maka setelah selesai diedit, entri tersebut akan berstatus disetujui dan dipublikasikan. Satu entri hanya dapat diedit oleh satu user dalam satu waktu.

Gambar 7: Tampilan halaman untuk pengelolaan korpus



Gambar 8: Tampilan halaman untuk pengeditan naskah yang telah diunggah

PENUTUP

Pada tahap ini, sumber daya bahasa yang dikembangkan telah cukup berkembang meskipun masih jauh dari kesempurnaan. Khusus korpus bahasa Jawa, telah dilakukan proses pelabelan kelas kata sebagai upaya meningkatkan akses terhadap informasi linguistis pada teks korpus. Titik perhatian pada tahap pertama ini adalah penentuan desain korpus dan pengembangan aplikasi korpus berbasis web. Perihal desain korpus, ditetapkan bahwa data korpus terdiri atas teks tertulis dan teks lisan dengan porsi terbesar terletak pada teks tertulis. Selanjutnya, aplikasi manajemen korpus berbasis web dikembangkan dalam empat tahap, yakni pengumpulan kebutuhan (*requirement gathering*), analisis data awal, penyiapan struktur pangkalan data serta perancangan desain dan arsitektur aplikasi berbasis web.

Sumber daya bahasa Jawa masih terus dikembangkan untuk tahap-tahap berikutnya. Setelah pengumpulan data korpus dan pembangunan aplikasi manajemen korpus berbasis web usai, dan pelabelan kelas kata tuntas, tahap selanjutnya yang akan dilakukan adalah membangun sistem anotasi berikutnya yakni pengenalan nama dan entitas, pembedaan makna ambigu, serta pemilahan struktur kalimat yang sangat bermanfaat untuk mengembangkan aplikasi manajemen korpus.

DAFTAR PUSTAKA

Semua pustaka yang dituliskan dalam daftar pustaka dikutip di dalam badan naskah. Daftar pustaka dan pengutipan menggunakan gaya APA (*American Psychological Association*).

Brooks, A. (2004). *Posfeminisme & Cultural Studies: Sebuah pengantar paling komprehensif* (S. Kunto Adi Wibowo, penerjemah dan Idi Subandy Ibrahim, editor). Yogyakarta: Jalasutra. (Karya asli diterbitkan pada 1997).

Darmawan, A. (2006). Seratus buku sastra terpilih karya perempuan. Dalam A. Kurnia (ed.), *Ensklopedia sastra dunia* (hlm. 224—227).

Ibrahim, A. Gufron. (2008). "Bahasa Terancam Punah: Sebab-sebab Gejala dan Strategi Pemecahannya". Dalam *Kongres Internasional lX Bahasa lndonesia*. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.

Kridalaksana, Harimurti. (2008). *Kamus Linguistik*. Jakarta: Gramedia.

Krisna, F.N. (2014). Studi kasus layanan pendidikan nonformal suku Baduy. *Jurnal Pendidikan dan Kebudayaan, 20*(1): 1—13.

Lumintaintang, Y.B. (2014). Industri film nasional sebagai media pelestarian bahasa ibu dalam upaya memperkuat identitas bangsa: Fenomena penggunaan alih kode. Kumpulan Makalah. *Menyelamatkan Bahasa Ibu, Seminar Internasional Hari Bahasa Ibu 2014*, 117—125.

Prihartono, Wawan. (2012). "Ciri Akustik Tuturan Modus Deklaratif Bahasa Jawa Penutur di Medan (Perbandingan dengan Ciri Akustik Tuturan Modus Deklaratif Bahasa Jawa Penutur di Solo)". Medan: Tesis USU.

Ratna, N.K. (2011). *Antropologi sastra: Peranan unsur-unsur kebudayaan dalam proses kreatif.* Yogyakarta: Pustaka Pelajar.

Sayuti, S. A. (2008). Bahasa, identitas, dan kearifan lokal dalam perspektif pendidikan. Dalam Mulyana (ed.), *Bahasa dan sastra daerah dalam kerangka budaya* (hlm. 23—44). Yogyakarta: Tiara Wacana.

Sudaryanto. (2015). *Metode dan Aneka Teknik Analisis Bahasa*. Yogyakarta: Sanata Dharma University Press.

Wiradnyana, Ketut. (2011). *Pra Sejarah Sumatra Bagian Utara: Kontribusinya pada Kebudayaan Kini.* Jakarta: Yayasan Obor Indonesia.