KONGRES BAHASA INDONESIA XII

# PROSIDING

## Literasi dalam Kebinekaan untuk Kemajuan Bangsa

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi
2023

# PROSIDING
## KONGRES BAHASA INDONESIA XII
### Literasi dalam Kebinekaan untuk Kemajuan Bangsa

**Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi**

XII
2023

Editor
Adi Budiwiyanto dkk.

**PROSIDING KONGRES BAHASA INDONESIA XII**
Diterbitkan pada tahun 2023
Penerbit:
Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi

Dikeluarkan oleh:
Badan Pengembangan dan Pembinaan Bahasa
Jalan Daksinapati Barat IV
Rawamangun
Jakarta Timur



**ISBN:**

# PENYUSUNAN KOPER:
# KORPUS PEMELAJAR BAHASA INDONESIA BERANOTASI EROR

*Building Koper: an Error-Annotated Indonesian Learner Corpus*

**David Moeljadi**
Kanda University of International Studies
(Universitas Bahasa Asing Kanda, Jepang)
moeljadi-d@kanda.kuis.ac.jp

**Abstrak**

Makalah ini membahas proses penyusunan korpus tahap awal pemelajar bahasa Indonesia bagi penutur asing (BIPA) beranotasi eror yang diberi nama Koper, singkatan dari Korpus Pemelajar Bahasa Indonesia Beranotasi Eror. Data korpus sementara ini diperoleh dari karangan mahasiswa-mahasiswa yang mengambil mata kuliah Bahasa Indonesia di beberapa universitas di Jepang, yaitu Kanda University of International Studies (Universitas Bahasa Asing Kanda) atau KUIS, Ritsumeikan Asia Pacific University (Universitas Asia Pasifik Ritsumeikan) atau APU, dan Tokyo University of Foreign Studies (Universitas Kajian Asing Tokyo) atau TUFS. Koper tidak hanya berisi kumpulan karangan-karangan mahasiswa, tetapi juga berisi anotasi eror. Daftar kesalahan disusun, dikelompokkan, dan disempurnakan oleh para dosen BIPA di Jepang dan digunakan untuk menganotasi korpus. Data korpus terus bertambah dan akan diperkaya dengan tulisan mahasiswa dari universitas lainnya, baik di Jepang maupun di luar Jepang. Koper akan berguna bagi berbagai bidang, misalnya pengajaran BIPA, linguistik korpus, dan bidang teknologi informasi dan komunikasi dalam hal pengembangan aplikasi penyuntingan ejaan bahasa Indonesia serta pembuatan sistem pemelajaran bahasa berbantuan komputer (*computer-assisted language learning* atau CALL).

**Kata kunci:** korpus pemelajar, anotasi eror, karangan tertulis, mahasiswa di Jepang, pemelajar BIPA

***Abstract***

*This paper discusses the process of building Koper, an acronym of Korpus Pemelajar Bahasa Indonesia Beranotasi Eror (an error-annotated learner corpus of Indonesian), in its early stage. At the present stage, the corpus data was obtained from the essays of students in Japan who took Indonesian language courses at several universities in Japan, namely Kanda University of International Studies (KUIS), Ritsumeikan Asia Pacific University (APU), and Tokyo University of Foreign Studies (TUFS). Koper contains not only a collection of student essays, but also error annotations. The errors were analyzed, compiled, and classified by BIPA lecturers in Japan and used to annotate the corpus. The data is increasing and will be enriched by student essays from other universities in Japan and other countries. Koper will be useful for various fields, such as Indonesian language teaching for*

*non-native speakers, corpus linguistics, and information and communication technology in terms of developing a grammatical error correction system and creating a computer-assisted language learning system (CALL).*

***Keywords:*** *learner corpus, error annotation, written essays, undergraduates in Japan, BIPA students*

## PENDAHULUAN

Korpus merupakan kumpulan teks atau data bahasa yang dikumpulkan dan disusun secara sistematis untuk keperluan analisis lingustik atau pemelajaran bahasa. Kajian korpus pemelajar berfokus pada pengumpulan dan analisis data pemelajar bahasa asing (L2), khususnya pada eror atau kesalahan yang dibuat penutur jati karena pengaruh bahasa ibunya (L1). Pengaruh bahasa ibu ini berperan penting dalam pemelajaran bahasa asing (Gass, 1988). Penggunaan korpus pemelajar dalam penelitian linguistik dan pemelajaran bahasa asing makin meningkat. Korpus pemelajar berperan membantu pemelajar mencapai kompetensi yang lebih baik dan memfasilitasi pengembangan bahan ajar yang lebih efektif. Namun, meskipun berpotensi memberikan manfaat yang signifikan, penelitian tentang korpus pemelajar bahasa Indonesia masih terbatas. Kebanyakan penelitian bahasa Indonesia dengan korpus menggunakan data yang dikumpulkan dari teks hasil tulisan penutur jati walaupun ada beberapa hasil penelitian mengenai analisis kesalahan berbahasa pemelajar bahasa Indonesia, antara lain Nugroho dkk. (2018), Yahya dkk. (2018), dan Hanifah dkk. (2020).

Korpus pemelajar berguna jika semua eror dalam data yang dikumpulkan sudah teridentifikasi dan teranotasi (Granger, 2003). Oleh karena itu, pertama-tama kajian mengenai jenis eror atau kesalahan berbahasa sangat diperlukan. Namun, hingga awal Mei 2023, belum ditemukan korpus pemelajar bahasa Indonesia dengan label eror yang dapat diunduh dan diakses secara gratis.[1] Dengan demikian, penyusunan korpus pemelajar bahasa Indonesia dengan label eror perlu dilakukan untuk mengisi kesenjangan ini dan memberikan kontribusi berharga bagi perkembangan pendidikan bahasa. Makalah ini berfokus pada proses penyusunan tahap awal Koper atau Korpus Pemelajar Bahasa Indonesia bagi Penutur Asing (BIPA) Beranotasi Eror.[2] Sementara ini, data Koper bersumber dari karangan mahasiswa yang mengambil mata kuliah Bahasa Indonesia di beberapa universitas di Jepang. Koper diharapkan dapat menjadi alat bantu pengembangan materi ajar dan perancangan strategi pengajaran yang lebih adaptif dan relevan dengan kebutuhan pemelajar BIPA.

## LANDASAN TEORETIS

### Korpus Pemelajar

Korpus pemelajar adalah koleksi tulisan pemelajar suatu bahasa (L2) yang dapat dibaca komputer yang mewakili bahasa tersebut dan mengandung data pemelajar bahasa

[1]Korpus BIPA (Suhardijanto & Putra, 2019) tidak dapat diakses dan diunduh secara gratis dan tidak beranotasi eror.

[2]Penelitian ini didanai oleh JSPS KAKENHI Nomor 23K12235.

tersebut. Korpus ini berguna untuk mengetahui seberapa jauh bahasa pertama (L1) memengaruhi pemelajaran bahasa kedua (L2). Korpus pemelajar memuat kategori-kategori yang dapat digunakan untuk penelitian pemelajaran bahasa (Prihantoro, 2022). Misalnya, korpus pemelajar bahasa Inggris lisan akademik MICASE (Michigan Corpus of Academic Spoken English) (Simpson dkk., 1999) mengandung informasi jenis kelamin pemelajar, umur, bahasa ibu, serta kompetensi pemelajar, baik yang merupakan penutur jati maupun yang bukan, termasuk yang kemampuannya mendekati penutur jati. Korpus pemelajar bahasa Inggris untuk pemelajar di Asia ICNALE (The International Corpus Network of Asian Learners of English) (Ishikawa, 2013) memuat informasi identitas pemelajar, seperti negara asal dan kompetensi bahasa Inggris. Dengan adanya informasi tersebut, pengguna korpus dapat memilih dan mencari data korpus pemelajar dari negara atau kompetensi tertentu saja, juga dapat membandingkan data korpus pemelajar dari negara berbeda dengan kompetensi sama.

Korpus pemelajar yang ada saat ini sebagian besar adalah korpus pemelajar bahasa-bahasa di Eropa, terutama bahasa Inggris, dan bahasa-bahasa di Asia yang penelitiannya sudah banyak dilakukan, seperti bahasa Jepang dan bahasa Mandarin (Morgado Da Costa, 2021). Penyusunan korpus pemelajar bahasa Indonesia (yang merupakan anggota rumpun bahasa Austronesia) sangat penting dari sisi linguistik dan pengajaran bahasa karena bahasa Indonesia memiliki fenomena tata bahasa yang unik dan jenis-jenis eror atau kesalahan yang berbeda dengan bahasa lainnya. Selain itu, ada kebutuhan untuk menyusun korpus pemelajar bahasa Indonesia karena jumlah pemelajar bahasa Indonesia di seluruh dunia makin meningkat. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia melalui Badan Pengembangan dan Pembinaan Bahasa telah membuka banyak kelas BIPA di berbagai universitas mancanegara. Di Jepang bahasa Indonesia diajarkan di berbagai universitas negeri dan swasta dari tingkat S-1 hingga S-3, SMA, dan lembaga-lembaga kursus bahasa asing.

**Anotasi Eror**

Korpus yang berisi data teks mentah yang komponen lingualnya sudah diimbuhi dengan label analisis disebut korpus beranotasi (Prihantoro, 2022). Anotasi adalah pemberian label pada satuan bahasa di dalam teks dalam korpus yang berisi informasi analisis linguistik, misalnya pemberian label kelas kata untuk tiap kata. Dalam penyusunan korpus pemelajar, biasanya jenis anotasinya adalah anotasi eror (van Rooy, 2017). Proses anotasi pada umumnya memerlukan satu set label yang terperinci dan terklasifikasi. Proses anotasi memberi informasi tambahan dan tidak mengubah data teks mentah korpus. Biasanya, format anotasi ini menggunakan XML (Extensible Markup Language) yang didesain secara khusus sehingga teks dapat diakses dalam berbagai platform dan mudah dibaca. Proses anotasi dapat dilakukan secara otomatis, misalnya dalam pemberian label kelas kata. Akan tetapi, proses anotasi eror biasanya dilakukan secara manual. Jenis-jenis eror, seperti kesalahan pada ejaan, bentuk kata, dan urutan kata harus ditemukan, dikumpulkan, dan diklasifikasikan terlebih dahulu sebelum masuk pada tahap anotasi atau pemberian label eror. Selain pemberian label eror, proses anotasi eror juga dapat melibatkan pengoreksian eror tersebut yang juga dilakukan secara manual (van Rooy, 2017).

Ludeling dan Hirschmann (2017) merekomendasikan sistem anotasi eror dengan banyak lapisan (*multi-layer corpus standoff architecture*) pada korpus pemelajar. Sistem anotasi eror dengan banyak lapisan ini memungkinkan banyak anotator menganotasi data korpus yang sama dengan menggunakan perangkat yang berbeda dan hasil anotasi para anotator tersebut dapat digabung menjadi satu dan dibandingkan. Ada kemungkinan satu satuan bahasa (dapat berupa satu kata, gabungan kata, atau bahkan satu kalimat) mengandung lebih dari satu jenis eror, misalnya kesalahan ortografi atau ejaan dan kesalahan morfologi atau pemberian imbuhan. Selain itu, ada juga kemungkinan satu jenis eror dapat dikoreksi dengan lebih dari satu cara. Sistem anotasi eror yang baik harus dapat mengakomodasi pemberian lebih dari satu label eror untuk satu satuan bahasa yang sama dalam satu lapisan anotasi. Ludeling dan Hirschmann (2017) mencatat dua jenis pemberian label eror. Yang pertama disebut pemberian label eror berdasarkan hasil koreksi (*edit-distance-based error tagging*) dan yang kedua disebut pemberian label eror berdasarkan informasi linguistik (*linguistically-based error tagging*). Pemberian label eror didasari hasil koreksi dengan menggunakan label-label eror, seperti *ubah*, *hapus*, dan *sisipkan*. Sebaliknya, pemberian label eror yang didasari informasi linguistik menggunakan label-label eror, seperti *urutan kata*, *tanda baca*, dan *penggunaan kata yang lewah*. Karena anotasi eror biasanya dilakukan secara manual, hasil anotasi tersebut perlu dievaluasi. Evaluasi dapat dilakukan dengan dua cara. Yang pertama adalah dengan membandingkan hasil anotasi manual dan kunci jawaban (korpus yang hasil anotasinya sempurna). Yang kedua adalah dengan melibatkan beberapa anotator yang menganotasi teks yang sama dengan petunjuk anotasi dan satu set label eror yang sama. Hasil anotasi para anotator tersebut kemudian dibandingkan dan dievaluasi (disebut kesepakatan antaranotator atau *inter-annotator agreement*). Evaluasi pada umumnya dilakukan secara berulang-ulang hingga hasilnya konsisten. Winder dkk. (2017) membahas analisis dan klasifikasi eror, proses anotasi, dan evaluasi pada korpus pemelajar bahasa Inggris NTUCLE (Nanyang Technological University Corpus of Learner English).

## METODE PENELITIAN

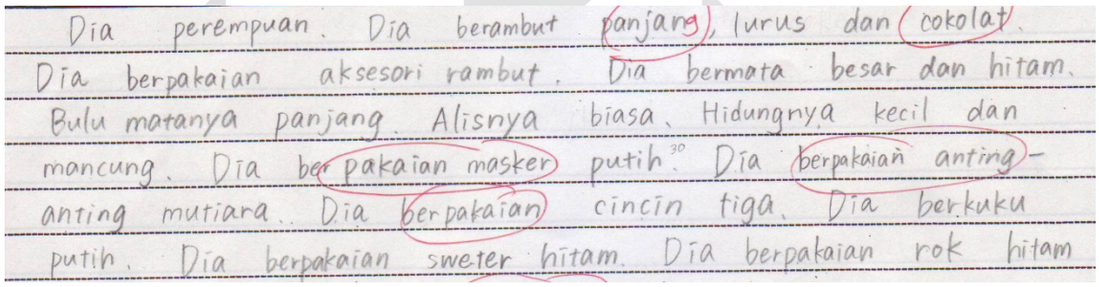
Bagian ini membahas metode penyusunan Koper yang meliputi (1) data dan sumber data; (2) teknik dan prosedur pengumpulan data; (3) pengumpulan, analisis, dan klasifikasi eror; (4) tahap pemrosesan awal data, serta (5) metode analisis dan anotasi data.

### Data dan Sumber Data

Data Koper bersumber dari karangan mahasiswa yang mengambil mata kuliah Bahasa Indonesia di tiga universitas di Jepang, yaitu Kanda University of International Studies (Universitas Bahasa Asing Kanda) atau KUIS, Ritsumeikan Asia Pacific University (Universitas Asia Pasifik Ritsumeikan) atau APU, dan Tokyo University of Foreign Studies (Universitas Kajian Asing Tokyo) atau TUFS. KUIS adalah universitas swasta bahasa asing di Perfektur Chiba, Jepang yang memiliki program studi Bahasa Indonesia. Mahasiswa Prodi Bahasa Indonesia KUIS setiap angkatan berjumlah 17 hingga 30 orang. Semuanya adalah penutur jati bahasa Je-

pang. Mereka mengambil mata kuliah Bahasa Indonesia wajib dan pilihan selama 4 tahun menempuh pendidikan S-1. APU adalah universitas swasta di Pulau Kyushu di sebelah selatan Jepang yang menawarkan mata kuliah bahasa-bahasa Asia Pasifik sebagai mata kuliah pilihan. Salah satunya adalah mata kuliah Bahasa Melayu/Indonesia yang memiliki empat tingkat, dari tingkat dasar (tingkat 1) hingga tingkat mahir (tingkat 4). Mahasiswa yang mengambil mata kuliah Bahasa Melayu/Indonesia di APU sebagian berbahasa ibu bahasa Jepang dan sebagian lainnya berbahasa ibu non-Jepang, misalnya bahasa Thai, bahasa Vietnam, dan bahasa Myanmar. Mahasiswa tingkat dasar (tingkat 1) berjumlah lebih dari 50 orang dan dibagi ke dalam 7 kelas, mahasiswa tingkat menengah dasar (tingkat 2) berjumlah 15 orang dan dibagi ke dalam 2 kelas, mahasiswa tingkat menengah lanjut (tingkat 3) berjumlah 5 orang, dan mahasiswa tingkat mahir (tingkat 4) berjumlah 2 orang. TUFS adalah universitas negeri kajian asing di Tokyo, Jepang yang menawarkan program studi bahasa Indonesia selama 4 tahun untuk pendidikan S-1. Mahasiswa Prodi Bahasa Indonesia TUFS setiap angkatan berjumlah 15 hingga 30 orang. semuanya adalah penutur jati bahasa Jepang.

Mahasiswa di setiap universitas tersebut menulis karangan sebagai bagian dari tugas mata kuliah Bahasa Indonesia. Karangan tersebut berupa karangan narasi, deskripsi, argumentasi, dan surat. Tema dan panjang karangan disesuaikan dengan silabus mata kuliah dan kebijakan tiap-tiap dosen. Untuk mahasiswa tingkat dasar dan menengah dasar, dosen biasanya memberi contoh dan mengajarkan pola kalimat dalam mengarang. Karangan mahasiswa ada yang berupa tulisan tangan (lihat Gambar 1), ada yang ditik menggunakan aplikasi Padlet[3] (lihat Gambar 2), dan ada yang menggunakan format teks digital lainnya (.doc atau .txt). Padlet adalah aplikasi daring gratis berupa papan tulis digital yang dapat digunakan oleh pemelajar dan pengajar untuk menulis teks di halaman yang sama sehingga dapat dimanfaatkan sebagai media kolaborasi antarpemelajar dan antara pemelajar dan pengajar. Semua karangan mahasiswa dalam Padlet dapat diekspor ke dalam format .xls yang terdiri atas kolom nama pengarang, judul, dan teks karangan.


Dia perempuan. Dia berambut panjang, lurus dan cokolat.
Dia berpakaian aksesori rambut. Dia bermata besar dan hitam.
Bulu matanya panjang. Alisnya biasa. Hidungnya kecil dan
mancung. Dia berpakaian masker putih. Dia berpakaian anting-
anting mutiara. Dia berpakaian cincin tiga. Dia berkuku
putih. Dia berpakaian sweter hitam. Dia berpakaian rok hitam

**Gambar 1**
**Karangan Deskripsi yang Ditulis Tangan dan Sudah Dikoreksi**

[3]https://padlet.com

**ANONYMOUS** JAN 16, 2023 01:59AM UTC

**Shinji Ikari**

Dia bernama Shinji Ikari, dia adalah siswa SMP.
Dia berbadan rendah, 157 cm.
Dia berambut hitam.
Dia bermata besar.
Dia beramah.

**Gambar 2**
**Karangan Deskripsi yang Ditik Menggunakan Aplikasi Padlet**

## Teknik dan Prosedur Pengumpulan Data

Penulis bekerja sama dengan setiap dosen Bahasa Indonesia di KUIS, APU, dan TUFS. Sebelum karangan mahasiswa dikumpulkan sebagai data untuk penyusunan Koper, mahasiswa diberikan penjelasan mengenai tujuan penelitian dan dibagikan formulir persetujuan oleh dosen masing-masing. Penulis merujuk pada formulir persetujuan proyek penyusunan korpus pemelajar bahasa Inggris di Universitas Kansai[4] dan menyesuaikannya. Formulir persetujuan yang sudah disesuaikan berisi nama mahasiswa, umur, jenis kelamin, jurusan, bahasa ibu, lama belajar bahasa Indonesia, lama belajar bahasa asing lainnya, ujian kemampuan berbahasa Indonesia yang pernah diambil, lama tinggal di Indonesia, serta penilaian tentang tingkat kesulitan mengarang dalam bahasa Indonesia dan alasannya. Formulir persetujuan dalam bahasa Jepang dibagikan kepada mahasiswa Jepang dan formulir persetujuan dalam bahasa Inggris dibagikan kepada mahasiswa non-Jepang (lihat Gambar 3). Mahasiswa boleh tidak mengumpulkan formulir persetujuan tersebut. Jika demikian, karangan mahasiswa tersebut tidak digunakan sebagai data Koper. Mahasiswa juga boleh tidak menjawab semua pertanyaan yang ada di formulir persetujuan. Informasi dalam formulir persetujuan ini digunakan sebagai informasi metalinguistik yang berguna untuk membuat penelusuran informasi korpus menjadi lebih spesifik.

**Consent Form**

I agree that the writing assignments I have written and will write this semester, together with other students' writing assignments, will be compiled into a database for the purpose of Indonesian language education and research. The writing assignments will be used to build an error-annotated learner corpus for Indonesian (JSPS research number: 23K12235, principal investigator: David Moeljadi, research institution: Kanda University of International Studies, e-mail: moeljadi-d@kanda.kuis.ac.jp). I also agree to provide the following personal information for the purpose of organizing the data, provided that my personal information will not be disclosed to any third party.

Date: _____

Signature: _____

---

1. Name: _____
2. Age: _____
3. Gender: [ Male / Female ]
4. Grade: _____ year

**Gambar 3**
**Salah Satu Bagian Formulir Persetujuan dalam Bahasa Inggris**

[4]http://someya-net.com/cgi-bin/agreement-1.pdf

Untuk tahap awal penyusunan Koper, karangan mahasiswa tingkat dasar dan menengah dasar (mahasiswa tingkat 1 dan 2) diprioritaskan, dianalisis kesalahannya, dan dianotasi atau diberi label eror. Karangan mahasiswa tingkat 3 dan 4 disimpan untuk dianalisis dan dianotasi kemudian. Karangan mahasiswa yang berupa tulisan tangan ditik sehingga semua data karangan mahasiswa ada dalam format teks digital.

**Pengumpulan, Analisis, dan Klasifikasi Eror**

Jenis-jenis eror atau kesalahan berbahasa L2 dapat bersifat universal (misalnya saltik) dan dapat bersifat khusus (misalnya afiksasi verba). Hal itu bergantung pada bahasa L2 masing-masing dan pengaruh bahasa ibu penutur jati (L1). Analisis eror ini diperlukan untuk pembuatan dan pengelompokan label eror yang digunakan dalam anotasi Koper. Jenis-jenis eror dikumpulkan dari penelitian-penelitian sebelumnya tentang analisis eror pemelajar BIPA, antara lain Nugroho dkk. (2018), Yahya dkk. (2018), dan Hanifah dkk. (2020). Setelah itu, label eror diberikan untuk setiap jenis eror berdasarkan informasi linguistik. Label eror tersebut dibagi ke dalam beberapa kelompok, misalnya Kesalahan Verba, Kesalahan Urutan Kata, dan Kesalahan Leksikal, dengan mengacu pada kumpulan label eror NTUCLE (Winder dkk., 2017) yang disesuaikan dengan bahasa Indonesia. Label-label eror yang ada mencakup kesalahan tata bahasa dan juga kesalahan yang berhubungan dengan gaya menulis, misalnya penggunaan kata-kata informal/cakapan dalam karangan formal, penulisan kalimat yang terlalu panjang dan/atau berbelit-belit, serta penulisan anak kalimat tanpa induk kalimat. Kumpulan label eror yang sudah dikelompokkan ini disempurnakan seiring dengan proses anotasi Koper. Label eror yang perlu ditambah akan dibubuhkan dan label eror yang tidak digunakan sama sekali akan dihapus. Anotator menghubungi penulis jika ada label eror yang perlu ditambahkan. Setelah itu, penulis menentukan apakah label tersebut perlu ditambahkan dan memberi tahu anotator lainnya jika ada label eror baru yang ditambahkan. Kumpulan label eror yang ada hingga awal Mei 2023 dapat dilihat di bagian Analisis dan Diskusi.

**Tahap Pemrosesan Awal Data**

Karangan mentah dalam bentuk berkas digital dengan format .doc, .txt, atau .xls (dari aplikasi Padlet) diseragamkan dan diubah ke dalam format .txt. Setelah itu, informasi pribadi mahasiswa (nama, nomor telepon, pos-el, alamat tempat tinggal, dll.) yang ada di dalam karangan dianonimkan atau disamarkan atau diganti dengan label, seperti <NAMA>, <NOMOR>, dan <POS-EL> untuk melindungi privasi mahasiswa tersebut seperti yang tertulis dalam formulir persetujuan.

Semua karangan yang terkumpul dikelompokkan dan dibagikan kepada lima orang anotator untuk dianotasi. Lima orang anotator tersebut adalah dosen-dosen bahasa Indonesia yang bekerja sama dengan penulis dalam tahap pengumpulan data. Mereka penutur jati bahasa Indonesia dan memiliki pengalaman mengajar bahasa Indonesia selama bertahun-tahun. Karangan yang terkumpul dikelompokkan dan dipilah terlebih dahulu sehingga setiap anotator mendapatkan jumlah karangan yang sama dan merata. Artinya, dosen KUIS tidak hanya menganotasi sebagian

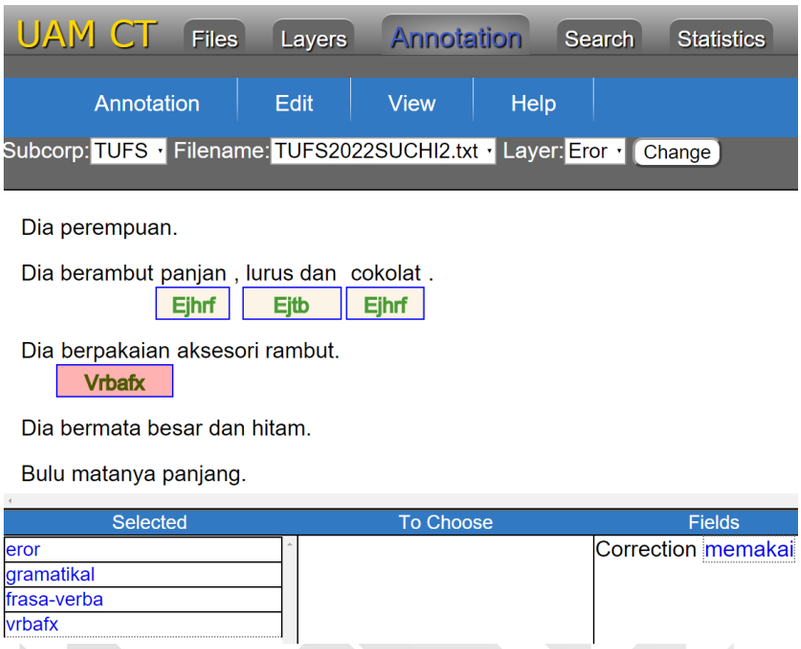
karangan mahasiswa KUIS, tetapi juga sebagian karangan mahasiswa APU dan TUFS, demikian juga sebaliknya.

**Metode Analisis dan Anotasi Data**

Dalam proses anotasi atau pemberian label eror, setiap anotator menganotasi teks masing-masing dan tidak mengetahui teks mana yang dianotasi oleh anotator lain. Dengan kata lain, tidak ada kerja sama antaranotator dalam proses anotasi eror. Proses anotasi ini menggunakan perangkat lunak anotasi korpus UAM CorpusTool[5] (O'Donnell, 2008). UAM CorpusTool dipilih karena sudah banyak digunakan dalam proses anotasi korpus pemelajar lainnya, mudah untuk diinstal, antarmuka penggunanya mudah digunakan, dan menghasilkan keluaran data dalam format XML. Selain itu, UAM CorpusTool memiliki berbagai fungsi, misalnya anotasi manual dengan banyak lapisan, input metadata, pembuatan skema set label, pencarian dalam korpus, dan statistika data korpus.

Anotasi dapat dilakukan setelah data teks dimasukkan ke dalam UAM CorpusTool. UAM CorpusTool secara otomatis memecah paragraf ke dalam kalimat-kalimat (satu kalimat satu baris). Hal ini memudahkan anotator dalam proses anotasi karena anotator dapat memusatkan perhatian pada setiap kalimat. Meskipun demikian, anotator dapat melihat seluruh teks karangan saat menganotasi sehingga memudahkan anotator memahami konteks karangan. Gambar 4 menampilkan layar anotasi UAM CorpusTool dengan data salah satu karangan mahasiswa. Pada Gambar 4 dapat dilihat satu baris berisi satu kalimat. Anotator diberi petunjuk agar memberi label eror sespesifik mungkin. Dalam proses anotasi eror, anotator dapat memilih kata tunggal, gabungan kata, atau keseluruhan kalimat. Selain itu, anotator dapat menganotasi eror yang tumpang tindih dalam satu kalimat yang sama. Anotator juga dapat memberi lebih dari satu label eror pada satu kata yang memiliki lebih dari satu eror. Anotator dapat memberi label eror untuk kata atau tanda baca yang kurang lengkap dengan memilih kata-kata yang ada di sekeliling kata atau tanda baca yang seharusnya ada tersebut. Pada Gambar 4 dapat dilihat kata *panjan* dan *cokolat* mendapat label eror "Ejhrf" (label "Ejhrf" adalah label eror ejaan huruf dalam penulisan kata; kata *panjan* seharusnya ditulis *panjang* dan kata *cokolat* seharusnya ditulis *cokelat*; penjelasan terperinci mengenai label eror ada pada bagian "Analisis dan Diskusi"), kata *berpakaian* mendapat label eror "Vrbafx" (label "Vrbafx" adalah label eror afiks verba; kata dasar *pakai* seharusnya mendapat afiks *meng*-, bukan *ber-…-an*), dan *lurus dan* mendapat label eror "Ejtb" (label "Ejtb" adalah label eror ejaan tanda baca; seharusnya ada tanda koma setelah kata *lurus*). Salah satu keterbatasan proses anotasi eror ini adalah anotasi eror morfologi (afiks) tidak dapat dilakukan. Jika ada eror dalam pengimbuhan afiks, kata berafiks tersebut dipilih dan diberi label eror seperti yang ditunjukkan pada Gambar 4.

[5]http://www.corpustool.com/




**Gambar 4**
**Tampilan Layar Anotasi Koper dengan UAM CorpusTool**

Untuk setiap label eror, anotator dapat menulis catatan, komentar, perbaikan, atau koreksi eror tersebut. Pada Gambar 4 bagian kanan bawah tertulis kata *memakai* sebagai koreksi kata *berpakaian*. Meskipun demikian, anotator tidak perlu menghabiskan waktu terlalu banyak untuk memperbaiki atau mengoreksi eror. Sakaguchi dkk. (2017) menyatakan bahwa sangat susah atau tidak mungkin seseorang dapat menemukan semua cara atau pilihan koreksi setiap eror yang ada dengan tetap mempertahankan makna yang dimaksud oleh penulis karangan tersebut, dalam hal ini mahasiswa BIPA. Hal tersebut akan memerlukan lebih banyak waktu dan tenaga.

Setelah proses anotasi untuk satu fail atau berkas karangan selesai dilakukan, anotator menyimpan hasil anotasi tersebut dan dapat mengekspornya ke dalam format XML. Penulis mengumpulkan semua hasil anotasi para anotator dan dapat membandingkan hasil anotasi anotator yang satu dengan hasil anotasi anotator lainnya. Dengan demikian, hasil anotasi para anotator yang dibandingkan dapat dievaluasi dan diketahui nilai kesepakatan antaranotator.

## ANALISIS DAN DISKUSI

Hingga awal Mei 2023 karangan mahasiswa yang sudah terkumpul berjumlah 60 karangan (30 karangan dari setiap universitas). Beberapa karangan sudah dianotasi dan sebagian besar lainnya masih dalam proses anotasi. Jumlah data teks yang ada sekarang masih sangat sedikit karena proyek ini baru berjalan satu bulan sejak April 2023. Meskipun demikian, jumlah karangan terus bertambah setiap bulan. Ada kemungkinan universitas lain di Jepang, seperti Universitas Keio, Universitas Chuo, dan Universitas Sophia juga turut berpartisipasi dalam pengumpulan data

teks karangan untuk penyusunan Koper ini. Penulis sudah menghubungi dosen Bahasa Indonesia di universitas tersebut. Penulis menargetkan setidaknya ada seribu teks karangan dengan anotasi eror hingga akhir Maret 2025. Selama proses anotasi eror berlangsung, label-label eror diperiksa dan disesuaikan dengan jenis eror yang ada. Kumpulan label eror yang sudah disusun berjumlah 42 label eror yang terdiri atas empat kategori berikut.

1. Kategori leksikal berisi 12 label eror yang berkenaan dengan pemilihan kata, frasa, atau ungkapan (lihat Tabel 1).
2. Kategori ejaan berisi 4 label eror yang berhubungan dengan tanda baca, penggunaan huruf kapital dan huruf kecil, spasi, dan penulisan huruf dalam kata (lihat Tabel 2).
3. Kategori gramatikal berisi 25 label eror yang berkaitan dengan aturan tata bahasa, misalnya urutan kata, kekurangan dan kelebihan kata, serta penggunaan bentuk aktif dan pasif. Kategori gramatikal ini dibagi ke dalam subkategori frasa nominal, frasa adjektival, frasa preposisional, frasa verbal, konjungsi, dan kalimat (lihat Tabel 3).
4. Kategori lainnya berisi 1 label eror yang mencakup jenis-jenis eror yang tidak dapat dikelompokkan menurut kategori di atas.

**Tabel 1**
**Kumpulan Label Eror Kategori Leksikal**

| Subkategori | Label | Keterangan |
|---|---|---|
| Kata | NomS | Kesalahan pemilihan nomina (kata dasarnya salah)<br>***Kasus*** *smartphone-nya bagus.* |
| | PronS | Kesalahan pemilihan pronomina<br>*Ketika* ***kita*** *sampai di sana, tokonya sudah tutup.* |
| | GolS | Kesalahan pemilihan kata penggolong<br>*Di sana ada tiga* ***buah*** *kucing.* |
| | AdjS | Kesalahan pemilihan adjektiva<br>*Adik saya tidak tinggi, dia* ***rendah****.* |
| | PrepS | Kesalahan pemilihan preposisi<br>*Saya membayar minuman* ***atas*** *kasir.* |
| | VrbS | Kesalahan pemilihan verba (kata dasarnya salah)<br>*Kakak* ***membuat*** *pendaftaran.* |
| | KonjS | Kesalahan pemilihan konjungsi<br>***Dan*** *makanannya juga enak.* |
| | KtAs | Penggunaan kata atau istilah asing yang ada padanannya dalam bahasa Indonesia<br>*Saya lupa membawa* ***passport****.* |
| | KtInF | Penggunaan kata/ungkapan informal/takbaku dalam teks formal atau sebaliknya<br>*Kamar saya kecil,* ***tapi*** *bersih.* |
| | KtVar | Pilihan kata kurang bervariasi<br>*Itu terbuat dari kaca.* ***Itu*** *diterangi oleh lampu.* ***Itu*** *berkilau dan cantik.* |
| Frasa/Ungkapan | UkpA | Frasa/ungkapan yang maknanya aneh/tidak biasa, tetapi masih bisa dipahami<br>*Saya makan udang* ***dengan mata-mata masih****.* |
| | UkpS | Frasa/ungkapan yang maknanya tidak jelas<br>*Saya bisa memasak dan tahu banyak alkohol* ***sangat minuman****.* |

**Tabel 2**
**Kumpulan Label Eror Kategori Ejaan**

| Label | Keterangan |
|---|---|
| EjK | Kesalahan penggunaan huruf kapital atau huruf nonkapital<br>*Saya ingin belajar **Bahasa indonesia** lebih banyak.* |
| EjTB | Penggunaan tanda baca (titik, koma, dll.) yang salah atau kurang<br>*Makanan ini murah, **manis dan** enak.* |
| EjSp | Kekurangan atau kelebihan spasi<br>*Dia tidak **membeli buku** apapun.* |
| EjHrf | Kekurangan, kelebihan, atau kesalahan penulisan huruf<br>*Saya **purgi** ke **Prefekter** Chiba.* |

**Tabel 3**
**Kumpulan Label Eror Kategori Gramatikal**

| Subkategori | | Label | Keterangan |
|---|---|---|---|
| Frasa Nomina | Urutan kata | UrtFN | Kesalahan urutan kata dalam frasa nominal<br>*Di **ini saya kamar** ada **menarik foto banyak**.* |
| | Nomina | NomAfx | Kesalahan afiksasi nomina (kata dasarnya benar)<br>*Mereka fokus pada **belajaran** bahasa Indonesia.* |
| | | Nom0 | Kekurangan nomina<br>*Saya suka **pelajaran Indonesia**.* |
| | | NomX | Kelebihan nomina<br>*Saya mau bermain di sana sampai **jam** sore.* |
| | | NomJ | Penjamakan nomina yang ganda<br>*Dia membeli **beberapa buku-buku**.* |
| | | NomR | Reduplikasi nomina yang tidak tepat<br>*Di sana ada banyak **rumah sakit-sakit**.* |
| | Pronomina | Pron0 | Kekurangan pronomina<br>*Dia telah bersama kami sejak lahir. Dia adalah anggota **keluarga**.* |
| | | PronX | Kelebihan pronomina<br>*Dia mengambil gitar dan **dia** memberikannya kepada saya.* |
| | Klitik *=nya* | NyaItu | Kesalahan penggunaan *itu* dan enklitik *–nya*<br>*Kalau mau makan, nasi **itu** di meja.* |
| | | Nya0 | Kekurangan enklitik *-nya*<br>*Saya ingin makan, tetapi **nasi** sudah habis.* |
| | | NyaX | Kelebihan enklitik *-nya*<br>***Alatnya** musiknya bagus sekali.* |
| | Kata *yang* | Yg0 | Kekurangan kata *yang*<br>*Ada empat patung **emas berbaris**.* |
| Frasa Adjektiva | Urutan kata | UrtFAj | Kesalahan urutan kata dalam frasa adjektival<br>*Uangnya **banyak sangat**.* |
| Frasa Preposisi | Preposisi | Prep0 | Kekurangan preposisi<br>*Dia **berkuliah Universitas** Kanda.* |
| | | PrepX | Kelebihan preposisi<br>*Saya sangat ingin **untuk** makan soto.* |

| | | | |
|---|---|---|---|
| Frasa Verba | Urutan kata | UrtFV | Kesalahan urutan kata dalam frasa verbal<br>*Dia **makan juga** tempe.* |
| | Verba | VrbPas | Kesalahan pemilihan bentuk verba aktif dan pasif<br>*Ada makanan yang **sering makan orang Jepang**.* |
| | | VrbAfx | Kesalahan afiksasi verba (kata dasarnya benar)<br>*Saya akan **deskripsi** rumah saya.* |
| | | Vrb0 | Kekurangan verba<br>*Cita-cita saya **adalah orang** kaya.* |
| | | VrbX | Kelebihan verba<br>*Saya pergi **berkunjung** mengunjungi rumah paman.* |
| Konjungsi | | Konj0 | Kekurangan konjungsi<br>*Di dalam kamar saya ada meja, kursi, **lemari, tempat** tidur.* |
| | | KonjX | Kelebihan konjungsi<br>*Karena masakan itu dimasak ibu, **sehingga** rasanya enak.* |
| Kalimat | Struktur kalimat | KalSub | Kalimat tidak bersubjek<br>***Mudah** untuk dimasak ketika Anda lapar.* |
| | | KalObj | Kalimat tidak berobjek<br>*Saya bisa **menghabiskan** di kamar saya.* |
| | | KalBel | Kalimat yang berbelit-belit<br>*Alasan cerita mengapa topik saya "Belajar Bahasa Inggris di Jepang" karena orang Jepang menghabiskan waktu dalam wajib belajar dan sambil menyinggung, mereka ragu mengapa tidak pandai.* |
| | | KalPen | Dua atau lebih kalimat yang seharusnya menjadi satu<br>*Mereka pergi ke kampus. Meskipun hujan deras.* |

Gambar 5 menunjukkan keluaran data teks beranotasi eror dalam format XML. Dapat dilihat kata-kata yang memiliki eror diapit oleh label segmen `<segment>...</segment>` yang di dalamnya berisi nomor identifikasi (`id`), label eror dan kategorinya (`features`), dan koreksi (`correction`). Data Koper akan diunggah ke GitHub dengan lisensi CC BY 4.0 dan dapat diunduh secara gratis. Selain itu, Koper akan dimasukkan ke MALINDO Conc (Nomoto dkk., 2018) dan CQPWeb (Hardie, 2012).

```
Dia berambut
<segment id='7'
features='eror;ejaan;ejhrf' Correction='pan-
jang'>panjan</segment>,
<segment id='15'
features='eror;ejaan;ejtb'
Correction='lurus, dan'>lurus dan</segment>
<segment id='10'
features='eror;ejaan;ejhrf' Correction='coke-
lat'>cokolat</segment>.
Dia
<segment id='11'
features='eror;gramatikal;frasa-verba;vrbafx'
Correction='memakai'>berpakaian</segment>
aksesori rambut.
```

**Gambar 5**
**Data Teks Beranotasi Eror dalam Format XML**

## PENUTUP

Koper (Korpus Pemelajar BIPA Beranotasi Eror) berisi data teks karangan tertulis (karangan narasi, deskripsi, argumentasi, dan surat) pemelajar BIPA dengan berbagai tingkat kemampuan bahasa Indonesia (L2) yang sebagian besar merupakan penutur jati bahasa Jepang (L1). Hingga awal Mei 2023 Koper berisi data karangan mahasiswa yang mengambil mata kuliah Bahasa Indonesia di tiga universitas di Jepang, yaitu KUIS, APU, dan TUFS. Data karangan mahasiswa di universitas lainnya di Jepang akan segera ditambahkan dan data karangan mahasiswa di luar Jepang dapat ditambahkan pada kemudian hari. Data karangan diambil sejak tahun 2022 dan terus bertambah sehingga diharapkan Koper dapat menjadi korpus diakronis yang dapat memantau perkembangan pemelajar BIPA. Kesalahan penggunaan kata (leksikal), tata bahasa (gramatikal), dan ejaan yang ada dalam teks karangan dianalisis, dikumpulkan, dikelompokkan, dan diberi label untuk setiap jenis kesalahan atau eror. Kumpulan label eror yang sudah disusun berdasarkan informasi linguistik ini berjumlah 42 label eror yang terdiri atas 12 label eror yang berkenaan dengan pemilihan kata, frasa, atau ungkapan dan tidak memengaruhi tata bahasa; 4 label eror yang berhubungan dengan tanda baca, penggunaan huruf kapital dan huruf nonkapital, spasi, dan penulisan huruf dalam kata; 25 label eror yang berkaitan dengan aturan tata bahasa, misalnya urutan kata, kekurangan dan kelebihan kata, serta penggunaan bentuk aktif dan pasif; dan 1 label eror untuk jenis-jenis eror lainnya. Label-label eror ini digunakan dalam proses anotasi yang dilakukan oleh para pengajar BIPA di Jepang dengan menggunakan UAM CorpusTool yang menghasilkan keluaran data korpus berlabel eror dalam format XML.

Penyusunan Koper ini baru berjalan satu bulan sejak April 2023. Hingga awal Mei 2023 karangan mahasiswa yang sudah terkumpul berjumlah 60 karangan (30 karangan dari setiap universitas). Beberapa karangan sudah dianotasi dan sebagian besar lainnya masih dalam proses anotasi. Penulis menargetkan setidaknya ada seribu teks karangan dengan anotasi eror hingga akhir Maret 2025. Penelitian ini diharapkan dapat menghasilkan dua keluaran, yaitu Koper dan daftar label eror yang dapat digunakan untuk menganotasi korpus pemelajar BIPA. Koper akan diunggah ke GitHub dengan lisensi CC BY 4.0 dan menjadi korpus pemelajar BIPA beranotasi eror terbuka pertama yang datanya dapat diakses dan diunduh secara gratis. Selain itu, Koper akan dimasukkan ke dalam MALINDO Conc dan CQPWeb. Penulis sudah meminta izin kepada para pengembangnya.

Koper diharapkan berguna dalam bidang pengajaran BIPA, misalnya dalam pembuatan bahan ajar mengarang dan membaca yang disesuaikan dengan tingkat kemampuan pemelajar. Selain itu, Koper juga dapat digunakan dalam pengembangan aplikasi (Meurers, 2017), misalnya aplikasi penyuntingan ejaan bahasa Indonesia, aplikasi penilaian karangan otomatis, dan aplikasi pemelajaran bahasa berbantuan komputer (*computer-assisted language learning* atau CALL) untuk bahasa Indonesia.

## DAFTAR PUSTAKA


Gass, Susan M. (1988). *Second language acquisition and linguistic theory: The role of language transfer* (hlm. 384–403). Springer Netherlands.

Granger, Sylviane. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538–546.

Hanifah, Rifa, Anang Santoso, & Gatut Susanto. (2020). Kesalahan klausa dalam karangan mahasiswa BIPA tingkat pemula. *Jurnal Pendidikan*, *5*(4), 447–453.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. Dalam *International Journal of Corpus Linguistics*, *17*(3), 380–409.

Ishikawa, Shin'ichiro. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Dalam S. Ishikawa, *Learner Corpus Studies in Asia and The World 1* (hlm. 91–118). Kobe University Press.

Ludeling, Anke & Hagen Hirschmann. (2017). Error annotation systems. Dalam Sylviane Granger, Gaetanelle Gilquin, & Fanny Meunier (ed) *The Cambridge handbook of learner corpus research* (hlm. 135–157). Cambridge University Press.

Meurers, Detmar. (2017). Learner corpora and natural language processing. Dalam Sylviane Granger, Gaetanelle Gilquin, & Fanny Meunier (ed) *The Cambridge handbook of learner corpus research* (hlm. 537–566). Cambridge University Press.

Morgado da Costa, Luis. (2021). *Using rich models of language in grammatical error detection*. Disertasi Nanyang Technological University.

Nomoto, Hiroki, Hannah Choi, David Moeljadi, & Francis Bond. (2018). MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. Dalam Kiyoaki Shirai (ed.) *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"* (hlm. 36–43).

Nugroho, Rahadiyan Duwi, Cicilia Tantri Suryawati, & Hendri Zuliastutik. (2018). Analisis kesalahan dalam penulisan karya ilmiah mahasiswa Jepang dalam pembelajaran BIPA. *Jurnal Pendidikan Bahasa dan Sastra 18*(2), 193–210.

O'Donnell, M. (2008). "The UAM CorpusTool: Software for corpus annotation and exploration". Dalam Bretones Callejas, Carmen M. dkk. (ed) *Applied linguistics now: Understanding language and mind/la lingüística aplicada hoy: Comprendiendo el lenguaje y la mente* (hlm. 1433–1447). Universidad de Almería.

Prihantoro. (2022). *Buku referensi pengantar linguistik korpus: Lensa digital data bahasa*. Undip Press.

Sakaguchi, Keisuke, Courtney Napoles, & Joel Tetreault. (2017). GEC into the future: Where are we going and how do we get there? Dalam *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (hlm. 180–187). Association for Computational Linguistics.

Simpson, R. C., S. L. Briggs, J. Ovens, & J. M. Swales. (1999). *The Michigan corpus of academic spoken English*. The Regents of the University of Michigan.

Suhardijanto, Totok & Deodatus Perdana Putra. (2019) Acquiring extended units of meaning: The role of learner corpus in teaching Indonesian as a foreign language. Dalam *Proceedings of the 2nd Konferensi BIPA Tahunan by Postgraduate Program of Javanese Literature and Language Education in Collaboration with Association of Indonesian Language and Literature Lecturers*, KEBIPAAN, 9 November 2019, Surakarta.

Van Rooy, Bertus. (2017). Annotating learner corpora. Dalam Sylviane Granger, Gaetanelle Gilquin, & Fanny Meunier (ed) *The Cambridge handbook of learner corpus research* (hlm. 79–105). Cambridge University Press.

Winder, Roger V.P., Joe MacKinnon, Shu Yun Li, Benedict Lin, Carmel Heah, Luis Morgado da Costa, Takayuki Kuribayashi, & Francis Bond. (2017). NTU-CLE: Developing a corpus of learner English to provide writing support for engineering students. Dalam *Proceedings of the 4th Workshop on NLP Techniques for Educational Applications (NLPTEA 2017)*. Taipei, Taiwan.

Yahya, Mokh., Andayani, & Kundharu Saddhono. (2018). Tendensi kesalahan sintaksis bahasa tulis pembelajar bahasa Indonesia bagi penutur asing (BIPA). *SUKMA: Jurnal Pendidikan 2*(1).