MERDEKA BELAJAR

KONGRES BAHASA INDONESIA XII

Makalah
Pembicara Seleksi

Penyusunan Korpus Paralel Bahasa Indonesia–Bahasa Melayu Ambon, Melayu Kupang, Beaye, dan Uab Meto

Joanito Agili Lopo, David Moeljadi, Samuel Cahyawijaya, Alham Fikri Aji, Carly J. Sommerlot, dan June Jacob
Universitas Kristen Satya Wacana

Subtema
Revitalisasi Bahasa dan Sastra Daerah

Jakarta, 25—28 Oktober 2023

"Literasi dalam Kebinekaan untuk Kemajuan Bangsa"

Adibasa Adiwangsa

Badan Pengembangan dan Pembinaan Bahasa
Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi

kbi.kemdikbud.go.id

PENYUSUNAN KORPUS PARALEL BAHASA INDONESIA–BAHASA MELAYU AMBON, MELAYU KUPANG, BEAYE, DAN UAB METO

Building a Parallel Corpus of Indonesian–Ambon Malay, Kupang Malay, Beaye, and Uab Meto

Joanito Agili Lopo[a], David Moeljadi[b], Samuel Cahyawijaya[c], Alham Fikri Aji[d], Carly J. Sommerlot[e], dan June Jacob[f]
[a]Universitas Kristen Satya Wacana
[b]Kanda University of International Studies
[c]Hong Kong University of Science and Technology
[d]MBZUAI
[e]National University of Singapore
[f]Universitas Kristen Artha Wacana
amalopo99@gmail.com

Abstrak

Teknologi informasi memiliki peran yang penting dalam upaya revitalisasi bahasa-bahasa daerah di Indonesia yang penuturnya sangat sedikit dan terancam punah. Salah satu peran teknologi informasi tersebut adalah sebagai sarana pembelajaran bahasa dan mempermudah akses bahasa-bahasa tersebut melalui teknologi digital, seperti pemrosesan bahasa alami. Namun, pembangunan teknologi digital untuk bahasa-bahasa daerah tersebut seringkali terhambat karena permasalahan ketersediaan korpus di bahasa-bahasa daerah tersebut. Untuk menanggulangi permasalahan tersebut, makalah ini membahas penyusunan korpus paralel dengan bahasa Indonesia sebagai bahasa sumber yang diterjemahkan ke dalam empat bahasa daerah di Indonesia sebagai bahasa sasaran, yaitu bahasa Melayu Ambon di Maluku, bahasa Melayu Kupang di NTT, bahasa Beaye di Kalimantan, dan bahasa Uab Meto di NTT. Bahasa Melayu Ambon dan bahasa Melayu Kupang digunakan dalam komunikasi regional (*wider communication*) dan bahasa Uab Meto memiliki status berkembang (*developing*), sedangkan bahasa Beaye belum terdokumentasi. Belum ada informasi jumlah penutur dan status bahasa tersebut. Penerjemahan kalimat dari bahasa Indonesia ke bahasa sasaran dilakukan oleh penutur asli bahasa sasaran. Untuk menjamin kualitas penerjemahan, dilakukan juga proses kontrol kualitas yang dilakukan secara independen. Hasil penerjemahan menunjukkan bahwa bahasa-bahasa daerah cenderung menghasilkan kalimat yang lebih panjang karena terdapat beberapa kata atau frasa yang tidak ada padanannya, terutama untuk kosakata yang tidak sering digunakan dalam bahasa daerah tersebut. Korpus paralel ini dapat diakses melalui GitHub dan diunduh secara gratis untuk memudahkan proses menerjemahan, membantu penelitian linguistik dan dokumentasi bahasa-bahasa daerah di Indonesia, serta meningkatkan aksesibilitas informasi bagi masyarakat yang menggunakan bahasa daerah sebagai bahasa sehari-hari.

Kata kunci: korpus digital, korpus paralel, bahasa daerah di Indonesia, revitalisasi bahasa

Abstract

Information technology has an important role in the revitalization of regional languages in Indonesia, which have very few speakers and are endangered. One of the roles of information technology is as a means of language learning and facilitating access to these languages through digital technology such as natural language processing. However, the development of digital technology for these regional languages is often hampered due to the problem of corpus availability in these regional languages. To overcome this problem, this paper discusses the construction of a parallel corpus with Indonesian as the source language translated into four local languages in Indonesia as target languages, namely Ambon Malay in Maluku, Kupang Malay in NTT, Beaye in Kalimantan, and Uab Meto in NTT. Ambon Malay and Kupang Malay are used in wider communication and Uab Meto has a developing status while Beaye is underdocumented, with no information on the number of speakers and language status. The translation of sentences from Indonesian to the target language was carried out by native speakers of the target language. To ensure the quality of the translation, an independent quality control process was also conducted. The translation results show that local languages tend to produce longer sentences because there are some words or phrases that have no equivalent, especially for vocabulary that is not often used in the local language. This parallel corpus can be accessed through GitHub and downloaded for free to facilitate the translation process, assist linguistic research and documentation of local languages in Indonesia, and improve information accessibility for people who use local languages as their daily language.

Keywords*: digital corpus, parallel corpus, local languages in Indonesia, language revitalization*

PENDAHULUAN

Indonesia merupakan negara keempat dengan jumlah penduduk terbanyak di dunia yang mencapai 270 juta jiwa dan tersebar di 17.508 pulau. Dengan jumlah penduduk yang begitu besar, Indonesia memiliki ekologi bahasa asli terbesar kedua setelah Papua Nugini yang setara dengan 10% dari total bahasa di dunia (Aji *et al*., 2022). Hal itu mencerminkan kekayaan budaya yang unik, keragaman linguistik yang sangat kaya dan kompleks. Menurut pemetaan bahasa yang dilakukan oleh Kementerian Pendidikan dan Kebudayaan Indonesia, telah diidentifikasi sebanyak 718 bahasa daerah yang ada di Indonesia sejak 1991 hingga 2019 (Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan, 2019). Namun, jumlah tersebut tidak termasuk dialek dan subdialek yang ada. Hal ini membuat keragaman bahasa asli di Indonesia sebagai salah satu elemen tak terpisahkan dari nilai-nilai budaya yang dapat disampaikan di antara masyarakat (Alamsyah, 2018).

Meskipun demikian, 361 di antara bahasa-bahasa tersebut terancam punah; 80 bahasa dalam kondisi hampir punah; dan 10 bahasa belum terdaftar (Eberhard *et al*., 2019). Selain itu, kebanyakan bahasa tersebut tidak tercatat secara baik dalam karya tulis, kurang diajarkan secara resmi, memiliki kesenjangan digital dan isolasi akibat hambatan budaya, serta tidak memiliki pedoman yang diterapkan di

kalangan penuturnya (Aji *et al*., 2022; Novitasari *et al*., 2020). Hal-hal tersebut menghambat perkembangan riset global untuk mengeksplorasi bahasa-bahasa ini. Tambahan pula, pembangunan sumber daya bahasa di Indonesia umumnya lebih berfokus pada pembangunan sumber daya bahasa untuk bahasa nasional, yaitu bahasa Indonesia (Suhardijanto dan Dinakaramani, 2018).

Oleh karena itu, untuk meningkatkan inklusivitas dan aksesibilitas data bahasa daerah di Indonesia, kami mengembangkan korpus paralel dengan bahasa Indonesia sebagai bahasa sumber dan empat bahasa daerah sebagai bahasa sasaran. Keempat bahasa daerah tersebut (bahasa Melayu Ambon, bahasa Melayu Kupang, bahasa Beaye, dan bahasa Uab Meto) belum memiliki korpus digital terbuka, bahkan salah satunya, yaitu bahasa Beaye, belum terdokumentasi. Tujuan penyusunan korpus paralel ini adalah membantu penelitian linguistik dan dokumentasi bahasa-bahasa daerah di Indonesia dan juga dapat digunakan untuk pemrosesan bahasa alami, seperti program penerjemahan otomatis. Selain itu, korpus paralel ini dapat meningkatkan aksesibilitas informasi bagi masyarakat yang menggunakan bahasa daerah sebagai bahasa sehari-hari, serta sebagai salah satu upaya pelestarian bahasa daerah di Indonesia.

LANDASAN TEORETIS

Korpus Paralel

Korpus paralel adalah korpus multibahasa yang memiliki segmen paralel bahasa sumber dan bahasa sasaran. Korpus paralel biasanya dipakai untuk penelitian terjemahan dan sangat bermanfaat bagi peneliti riset terjemahan karena dapat mengungkap strategi penerjemahan yang dilakukan oleh penerjemah. Selain itu, data ini dapat digunakan untuk pangkalan data program penerjemahan otomatis, seperti Google Translate (Prihantoro, 2022). Pengembangan korpus paralel sudah dilakukan untuk berbagai pasangan bahasa, termasuk untuk bahasa Indonesia (Wahyu Guntara *et al*., 2020). Dalam lingkup bahasa-bahasa daerah di Indonesia, terdapat beberapa korpus paralel yang sudah dibangun, seperti bahasa Minang (Koto dan Koto, 2020) dan bahasa Bugis (Wahyuni *et al*., 2019). Selain itu, terdapat beberapa penelitian yang berfokus pada pembangunan korpus paralel multibahasa yang juga mencakup beberapa bahasa daerah di Indonesia (Cahyawijaya *et al*., 2021; Winata *et al*., 2022).

Bahasa Indonesia

Bahasa Indonesia [ind] adalah bahasa negara Republik Indonesia, seperti yang tertulis pada Pasal 36 UUD 1945. Bahasa Indonesia termasuk dalam cabang bahasa Melayu-Polinesia Barat (MPB) dalam rumpun bahasa Austronesia yang dituturkan oleh lebih dari 198 juta orang di Indonesia (Sensus Penduduk 2010). Bahasa Indonesia memiliki 19 konsonan, 6 vokal, dan 3 diftong. Bahasa Indonesia tidak memiliki tona. Proses pembentukan kata dalam bahasa Indonesia meliputi afiksasi (penambahan prefiks, sufiks, konfiks), reduplikasi, dan pemajemukan kata. Bahasa Indonesia memiliki urutan kata subjek-predikat-objek dalam kalimat. Dalam penulisan resmi, bahasa Indonesia menggunakan aksara Latin dan sebagian besar bahasa daerah di Indonesia saat ini juga mengadopsi aksara Latin. Bahasa Indonesia berkembang dari bahasa Melayu Klasik Kesultanan Riau-Johor (Sneddon, 2003).

B**ahasa Melayu Kupang**

Bahasa Melayu Kupang [mkn] adalah sebuah bahasa Melayu Kreol yang digunakan oleh sekitar 200.000 penutur di bagian barat Pulau Timor, NTT dan sekitarnya. Bahasa ini memegang peranan yang sangat penting dalam komunikasi antarmasyarakat penuturnya. Sebagai salah satu bahasa Austronesia yang ada di sekitar Kota Kupang, bahasa ini diklasifikasikan sebagai bahasa Melayu-Polinesia Tengah (Central Malayo-Polynesian atau CMP) (Grimes, 1997). Bahasa Melayu Kupang berkembang dari varietas Melayu perdagangan yang digunakan di pusat perdagangan dan sebagai bahasa komunikasi interetnik. Perlu diketahui bahwa orang-orang yang tinggal di sekitar Kupang bukanlah penutur asli berbagai bahasa Melayu yang ada. Mereka justru mempelajarinya secara tidak sempurna dan digunakan pada lingkup tuturan yang terbatas. Bahasa Melayu Kupang memiliki kemiripan struktur dan kosakata dengan bahasa Melayu Ambon, tetapi terdapat beberapa perbedaan kosakata dan logat.

Bahasa Melayu Ambon

Bahasa Melayu Ambon [abs] adalah bahasa Melayu Kreol yang dituturkan oleh sekitar 1,6 juta orang di Indonesia dengan jumlah penutur jati (L1) sekitar 200.000 (Collins, 1987) dan jumlah penutur bahasa Melayu Ambon sebagai bahasa kedua (L2) sekitar 1,4 juta orang (Connor, 2013). Bahasa Melayu Ambon dituturkan di Provinsi Maluku. Bahasa ini berkembang sejak abad ke-16, pertama kali digunakan sebagai bahasa perdagangan dalam perdagangan rempah-rempah, dan sekarang menjadi *lingua franca* untuk komunikasi antaretnis. Bahasa ini memiliki kemiripan leksikal sebesar 81% dengan bahasa Indonesia (Eberhard *et al*., 2019) dan ditulis dalam aksara Latin.

Bahasa Beaye

Bahasa Beaye [day][1] adalah bahasa yang kurang terdokumentasi yang dituturkan di sekitar ujung utara Kalimantan Barat. Bahasa ini termasuk dalam cabang bahasa Benyadu-Bekati' dalam rumpun bahasa Dayak Darat (Land Dayak) (Sommerlot, 2020). Bahasa Beaye memiliki kemiripan struktur, fonem, dan kosakata dengan Mali dan Ba'aje, dua bahasa lainnya dalam cabang bahasa Benyadu-Bekati'. Bahasa Beaye memiliki urutan kata subjek-predikat-objek, mirip dengan bahasa Indonesia, tetapi afiksasi kata kerjanya lebih sedikit. Tidak ada aksara standar untuk bahasa ini.

Bahasa Uab Meto

Bahasa Uab Meto [aoz] dituturkan di Pulau Timor bagian barat, Nusa Tenggara Timur oleh sekitar 700.000 orang (Sensus Penduduk 2009). Bahasa ini termasuk dalam cabang bahasa Melayu-Polinesia Tengah-Timur (MPTT) dalam rumpun bahasa Austronesia. Bahasa ini memiliki banyak dialek, seperti Dialek Amfoan-Fatule'u-Amabi (Amabi, Amfoan, Amfuang, Fatule'u), Amanuban-Ama-

[1]Makalah ini menggunakan kode bahasa [day] untuk bahasa Beaye. Sebenarnya kode bahasa ini adalah kode bahasa Dayak Darat (Land Dayak), tetapi karena bahasa Beaye belum terdokumentasi dan belum memiliki kode bahasa sendiri, makalah ini meminjam kode bahasa Dayak Darat.

natun (Amanatun, Amanuban, Amanubang), Mollo-Miomafo (Miomafo, Mollo), Biboki-Insana (Biboki, Insanao), Kusa-Manlea (Kusa, Manea, Manlea). Bahasa ini menggunakan aksara Latin dan memiliki terjemahan Alkitab.

METODE PENELITIAN

Metode yang digunakan pada penelitian ini terdiri atas beberapa tahapan. Setiap tahapan merujuk pada Winata *el al*. (2022) yang memulai penyusunan korpus paralel untuk bahasa-bahasa daerah di Indonesia dengan penutur dan sumber daya yang sedikit (*low-resource*).[2] Namun, terdapat beberapa modifikasi untuk mengikuti keterbatasan dan tantangan yang ada pada penelitian ini.

Perekrutan Sukarelawan (*Volunteer Recruitment*)

Penelitian ini melibatkan sukarelawan yang bersedia membantu dalam penyusunan korpus paralel untuk empat bahasa daerah. Kami merekrut 5–10 sukarelawan untuk setiap bahasa, baik para tetua adat, keluarga, teman, masyarakat lokal, maupun tokoh agama yang memiliki pengetahuan tentang budaya dan bahasa setempat. Para sukarelawan harus fasih dalam bahasa daerahnya dan memiliki kemampuan berkomunikasi dengan baik agar memudahkan pengumpulan data. Mereka harus memiliki pengetahuan yang cukup tentang budaya setempat, termasuk penggunaan bahasa sehari-hari dan kosakata yang digunakan. Pendekatan ini dipilih karena kesulitan dalam mencari sumber daya yang memadai dan penganotasi ahli dalam bahasa-bahasa daerah tersebut.

Untuk responden penelitian ini, kami mencari orang asli yang telah menetap di daerah empat bahasa daerah selama paling sedikit 20 tahun. Setelah itu, untuk mendapatkan responden berikutnya, kami meminta para responden awal untuk merekomendasikan orang lain yang sesuai (Waters, 2015). Para sukarelawan tersebut dibagi menjadi dua kelompok: kelompok 1 menerjemahkan bahasa (penganotasi internal) dan kelompok 2 melakukan pengecekan kualitas (penganotasi eksternal). Setiap penganotasi diberikan panduan dan instruksi untuk memastikan pemahaman dan ketersediaan perangkat yang diperlukan.

Penyaringan Data *(Data Filtering)*

Sumber data yang digunakan pada penelitian ini diambil dari Tatoeba Project (https://tatoeba.org/id), yaitu sebuah proyek komunitas daring yang menyediakan kumpulan kalimat beserta terjemahannya dalam berbagai bahasa dari seluruh dunia. Kalimat-kalimat yang ada di Tatoeba Project berasal dari kontribusi sukarelawan dan cenderung berfokus pada penggunaan bahasa sehari-hari dalam konteks dan topik yang beragam. Kalimat-kalimatnya sederhana, tidak terlalu kompleks, dan mudah dipahami. Ini menjadi pendekatan awal yang baik, terutama terhadap bahasa-bahasa yang memiliki penutur dan sumber daya yang sedikit karena apa yang ada di Tatoeba Project dapat memberikan manfaat pada proses terjemahan dan analisis bahasa. Kami mengambil korpus paralel bahasa Inggris–Indonesia untuk diterjemahkan ke dalam empat bahasa daerah. Data tersebut terdiri atas 8.816

[2]Sumber daya bahasa (*language resource*) meliputi sumber data digital dan alat atau perangkat yang dapat digunakan untuk memproses data digital tersebut.

kalimat, tetapi yang digunakan pada penelitian ini berjumlah 2.000 kalimat bahasa Indonesia. Kami melakukan penyaringan data terlebih dahulu untuk menghapus duplikasi kalimat dan pengecekan saltik (*typo*).

Penerjemahan Manual *(Human Translation)*

Keterbatasan infrastruktur dan aksesibilitas masih menjadi masalah di beberapa daerah di Indonesia, terutama bagi bahasa Beaye dan bahasa Uab Meto yang masih berada di daerah terpencil. Oleh karena itu, kami mengadopsi berbagai media penerjemahan yang disesuaikan dengan kebutuhan dan tingkat ketersediaan teknologi setiap penganotasi di setiap daerah. Ini termasuk penggunaan *spreadsheet* daring untuk penganotasi dengan akses internet dan teknologi yang memadai dan dokumen penerjemahan melalui pos-el atau aplikasi WhatsApp dalam format Excel atau Word, bahkan melalui cetakan fisik untuk penganotasi dengan keterbatasan akses teknologi.

Untuk memastikan bahwa terjemahan yang dihasilkan tetap akurat dan konsisten dalam menyampaikan makna kalimat, kami memberikan instruksi kepada para penganotasi untuk tetap mempertahankan makna asli teks tanpa secara langsung menerjemahkan entitas, seperti orang, organisasi, lokasi, dan waktu ke dalam empat bahasa tersebut. Meskipun demikian, kami meminta para penganotasi untuk tetap memperhatikan keunikan dan kekhasan dari bahasa daerah tersebut. Lebih lanjut, kami memberi mereka instruksi untuk (1) mempertahankan penggunaan dialek setiap bahasa; (2) mencari padanan kalimat yang tepat jika tidak ada terjemahan yang cocok dan jika tetap tidak ada, kalimat asalnya harus tetap digunakan; (3) menjaga entitas dalam teks, dan (4) mempertahankan tanda baca dan penggunaan huruf besar dan huruf kecil pada teks aslinya.

Validasi dan Evaluasi *(Human-Assisted Quality Assurance)*

Untuk memastikan hasil terjemahan yang akurat dan konsisten, kami memvalidasi dan mengevaluasi setiap hasil terjemahan penganotasi. Kami melibatkan paling tidak empat hingga lima orang penganotasi eksternal yang tidak terlibat dalam proses terjemahan sebelumnya untuk memastikan keobjektifan hasil evaluasi. Penganotasi tersebut dipilih secara independen tanpa rekomendasi dari responden awal dan menjaga kerahasian identitas penganotasi untuk memastikan hasil evaluasi yang asli dan independen. Agar para penganotasi dapat melakukan validasi dan evaluasi secara efektif, kami menyediakan panduan selama 30 menit untuk setiap penganotasi. Selain itu, kami juga memberikan contoh-contoh terjemahan yang baik dan buruk untuk membantu para penganotasi memahami kriteria evaluasi secara lebih jelas dan efektif. Contoh terjemahan baik dan buruk dapat dilihat pada Tabel 1.

Tabel 1
Contoh Terjemahan Baik dan Buruk

Bahasa	Terjemahan Buruk	Terjemahan Baik
Melayu Kupang	~~Tom ada perbaiki itu.~~	Tom ada bekin itu bae-bae.
Melayu Ambon	Tom biking bae itu~~?~~	Tom biking bae itu.
Beaye	~~tom~~ mait labe ngen	Tom mait labe ngen.
Uab Meto	Tom ~~nalekonane~~.	Na Tom naleko nane.

Secara garis besar, kami meminta agar para penganotasi melakukan pengecekan terhadap hasil terjemahan dan mengubahnya jika terdapat kesalahan terjemahan. Namun, jika tidak terdapat kesalahan, kami meminta para penganotasi untuk membiarkan terjemahan tersebut apa adanya. Selanjutnya, kami memberikan instruksi yang lebih spesifik kepada para penganotasi untuk memastikan kualitas terjemahan yang baik, yaitu (1) memperbaiki terjemahan jika tidak cocok dengan konteks penggunaan sehari-hari; (2) memastikan bahwa hasil terjemahan tetap mempertahankan makna asli teks sumber; (3) memastikan kelengkapan dan kejelasan terjemahan; dan (4) memberikan variasi terjemahan jika diperlukan.

ANALISIS DAN DISKUSI

Penyusunan korpus paralel bahasa Indonesia–bahasa Melayu Ambon, Melayu Kupang, Beaye, dan Uab Meto berlangsung selama satu tahun, dimulai pada bulan Februari 2022 hingga Maret 2023. Ada 35 sukarelawan yang terlibat dalam penyusunan dan evaluasi, termasuk tetua adat, keluarga, teman, masyarakat lokal, dan tokoh agama yang memiliki pengetahuan tentang budaya dan bahasa setempat. Mereka semua adalah penutur asli dan/atau penduduk setempat sekurang-kurangnya selama 20 tahun. Setiap bahasa paling tidak memiliki dua hingga lima penganotasi internal yang menerjemahkan bahasa Indonesia ke dalam bahasa sasaran dan empat hingga lima penganotasi eksternal yang melakukan validasi dan evaluasi hasil terjemahan. Jumlah kalimat yang berhasil diterjemahkan dan dievaluasi adalah 2.000 kalimat untuk setiap bahasa. Namun, kami mengalami kesulitan dalam merekrut sukarelawan untuk bahasa Uab Meto karena bahasa ini memiliki lima dialek yang tersebar di berbagai wilayah di Pulau Timor bagian barat. Oleh karena itu, kami memutuskan untuk mencari penganotasi hanya dari dua dialek, yaitu Amanuban-Amanatun (5 penganotasi) dan Mollo-Miomafo (4 penganotasi), untuk mewakili variasi dialek yang ada dalam bahasa Uab Meto.

Selanjutnya, penganotasi yang memiliki akses internet dan teknologi yang memadai menggunakan *spreadsheet* daring untuk penerjemahan. Penganotasi tersebut paling banyak berasal dari Kota Ambon dan Kota Kupang. Kami juga melibatkan beberapa komunitas di kedua bahasa tersebut, seperti komunitas mahasiswa Ambon di Kota Salatiga dan relawan dari yayasan non-profit di Kota Kupang dan sekitarnya. Sebaliknya, penganotasi bahasa Uab Meto dan bahasa Beaye beragam dan kebanyakan menerima dokumen penerjemahan melalui pos-el atau aplikasi WhatsApp dalam format Excel atau Word. Cetakan fisik juga digunakan untuk penganotasi yang tidak memiliki akses internet atau kurang bisa menggunakan

internet dan teknologi informasi. Dalam bagian ini, akan dijelaskan secara lebih terperinci analisis hasil terjemahan, analisis korpus paralel, dan tantangan yang dihadapi ketika melakukan penyusunan korpus paralel.

Analisis Hasil Validasi dan Evaluasi

Kami mengambil 100 sampel hasil evaluasi dari para penganotasi pada setiap bahasa, kemudian melakukan analisis terhadap sampel tersebut. Sesuai dengan instruksi dan panduan yang diberikan kepada para penganotasi, kami mengategorikan beberapa bagian untuk dievaluasi, yaitu kesalahan pengetikan, variansi, perubahan minor, dan perubahan mayor.

Kesalahan Pengetikan atau Saltik

Kesalahan pengetikan merujuk pada kesalahan dalam ejaan, penggunaan tanda baca, huruf kapital, spasi, tanda hubung, dan format numerik.

Variansi

Variansi dalam bahasa terjemahan merujuk pada penambahan variasi atau variasi alternatif dalam penerjemahan kalimat oleh para penganotasi, khususnya kalimat, frasa, atau kata dalam bahasa asal ke dalam bahasa tujuan. Hal itu tidak menunjukkan kesalahan terjemahan para penganotasi internal, tetapi bisa disebabkan oleh ragam dialek dalam bahasa asli.

Perubahan Minor

Perubahan minor yang dievaluasi pada bagian ini mencakup berbagai jenis perubahan pada kata, tanda baca, struktur kalimat, dan tata bahasa. Meskipun perubahan minor tidak signifikan dalam arti mengubah makna teks secara keseluruhan, itu membantu membuat teks terjemahan lebih jelas dan mudah dipahami, misalnya, penggantian kata yang jarang digunakan dengan kata yang lebih umum dipakai dan penggunaan kata seru atau interjeksi yang ada dalam kalimat terjemahan (mkn: *do, ko, le, dolo, e*; day: *ngen*).

Perubahan Mayor

Perubahan mayor meliputi perubahan signifikan yang mengubah makna dan konteks teks. Hal itu dapat terjadi karena kesalahan penerjemahan, interpretasi yang berbeda terhadap makna asli teks, atau kesalahan dalam memilih kata atau frasa yang tepat, misalnya penambahan atau penghilangan informasi penting dan penyisipan informasi yang tidak ada dalam teks asli.

Analisis dilakukan dengan menelusuri satu per satu sampel yang dipilih, kemudian menghitung berapa banyak perubahan yang ada per kategori. Tabel 2 memberikan informasi mengenai jumlah perubahan yang dilakukan oleh penganotasi pada 100 sampel acak. Meskipun 100 sampel acak tidak bisa merepresentasikan keseluruhan validasi dan evaluasi, ini dapat menjadi gambaran awal proses validasi. Terlihat bahwa pada kategori kesalahan pengetikan, bahasa Beaye dan Uab Meto memiliki jumlah yang signifikan. Hal itu disebabkan oleh media penerjemahan yang digunakan sangat beragam sehingga potensi kesalahan penggunaan

tanda baca, huruf kapital, dan ejaan sangat besar. Selain itu, bahasa Beaye dan Uab Meto belum memiliki standar penulisan baku sehingga setiap orang memiliki standar sendiri-sendiri dalam menulis. Hal ini berbanding terbalik dengan bahasa Melayu Ambon dan Melayu Kupang yang telah memiliki standar penulisan yang baku sehingga kesalahan-kesalahan tersebut dapat diminimalkan. Selain itu, para penganotasi bahasa Melayu Ambon dan bahasa Melayu Kupang kebanyakan menggunakan *spreadsheet* daring sehingga memudahkan mereka mengubah dan memperbaiki kesalahan-kesalahan tersebut.

Kategori perubahan mayor dan variansi menjadi sangat problematik karena permasalahan ragam dialek dan pemahaman konteks yang berbeda di antara para penganotasi tiap-tiap bahasa. Untuk mengatasi permasalahan tersebut, kami meminta para penganotasi eksternal untuk mempertimbangkan penggunaan kalimat yang dapat dimengerti oleh beragam dialek yang ada. Hal itu dapat membuat para penganotasi dapat mengurangi perbedaan dalam pemahaman konteks antardialek yang berbeda. Melalui perubahan minor, banyak dilakukan penggantian pada kata-kata yang jarang digunakan dengan kata yang lebih umum dipakai dan juga penambahan kata sisipan atau partikel untuk memperjelas kalimat.

Tabel 2
Analisis Hasil Validasi dan Evaluasi

Kategori	abs	mkn	day	aoz
Kesalahan Pengetikan	3	0	35	20
Variansi	11	6	2	4
Perubahan Mayor	11	12	10	16
Perubahan Minor	16	11	3	3

Analisis Korpus Paralel

Kalimat yang berhasil diterjemahkan berjumlah 2.000 kalimat setiap bahasa. Rata-rata panjang kalimat setiap bahasa daerah adalah 4 kata, sedangkan untuk bahasa sumbernya (bahasa Indonesia) adalah 3 kata. Hal itu menunjukkan bahwa terjemahan ke dalam empat bahasa sasaran memiliki kalimat yang sedikit lebih panjang jika dibandingkan dengan bahasa sumbernya. Misalnya, kalimat Sampai jumpa! diterjemahkan ke dalam bahasa Beaye (**day**) menjadi '*Jaja adep bedapet!*', ke dalam bahasa Uab Meto (**aoz**) menjadi '*Na tal natatef!*' dan ke dalam bahasa Melayu Kupang (**mkn**) menjadi '*Sampai katumu lai!*'. Beberapa contoh kalimat terjemahan dapat dilihat pada Tabel 3.

Tabel 3
Contoh Kalimat Paralel

ind	abs	mkn	day	aoz
Mari, kita lakukan dengan caraku.	Mari biking pake beta pung cara.	Mari ko katong bekin deng beta pung akal.	Yok adep nge-lakukan ngen cara ken.	Om hit het moe nek au lomo it.
Kamarmu yang mana?	Se kamar yang mana?	Lu pung kamar di?	Kamar ko neng pe?	Ho ke ne es le me?
Aturannya sangatlah jelas.	Dia pung aturan paleng jelas.	Dia pung atoran talalu jalas.	Aturan ngen sangatlah jelas.	Plenat na naknino.
Tom memperbaiki itu.	Tom biking bae itu.	Tom ada bekin babae tu.	Tom mai laba ngen.	Na Tom naleko nane.
Kami ambisius.	Katong talalu usaha.	Katong talalu usaha.	Kadi ambisius.	Haim ambisius.

Dalam usaha untuk melakukan terjemahan ke dalam bahasa daerah, tidak dapat terelakkan lagi kita akan berhadapan dengan kondisi *untranslatable words* atau kata-kata yang tidak dapat diterjemahkan. Contohnya pada Tabel 3, terdapat kata yang tidak dapat diterjemahkan ke dalam bahasa Beaye dan Uab Meto, yaitu kata *ambisius* dan *aturan*. Kata-kata ini sebenarnya memiliki makna yang sangat spesifik dan konteks yang sangat terbatas pada kedua bahasa tersebut. Meskipun terdapat usaha untuk mencari padanan kata, sering kali kata tersebut tidak dapat sepenuhnya diwakili oleh satu kata saja. Misalnya, pada bahasa Melayu Ambon dan Melayu Kupang yang menambahkan kata *talalu* (**ind** 'terlalu') di depan kata *usaha* untuk menunjukkan keinginan kuat untuk mencapai sesuatu (**ind** 'ambisius').

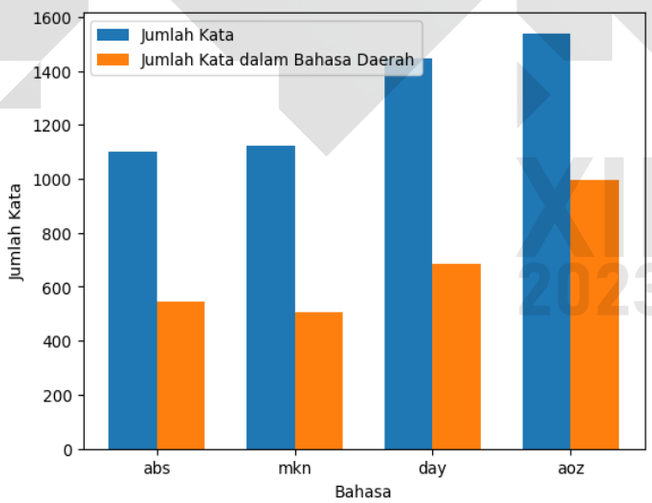
Tabel 4 menunjukkan jumlah kata unik atau kata-kata tanpa duplikasi, termasuk kata-kata pinjaman yang terdapat dalam data. Bahasa Uab Meto dan bahasa Beaye memiliki jumlah kata unik yang lebih banyak jika dibandingkan dengan bahasa Melayu Ambon dan Melayu Kupang. Bahasa Uab Meto dan bahasa Beaye cenderung memiliki kosakata sendiri yang berbeda dengan bahasa Indonesia sehingga satu kata bahasa Indonesia dapat dijelaskan dengan beberapa kata dalam bahasa tersebut, sedangkan bahasa Melayu Ambon dan Melayu Kupang dipengaruhi oleh asal-usul bahasa tersebut yang dapat disebut sebagai varian atau dialek bahasa Melayu.

Tabel 4
Statistik Korpus

Bahasa	Panjang Kalimat	Jumlah Kata Unik
Indonesia	3.417	1.504
Melayu Ambon	3.704	1.102
Melayu Kupang	3.920	1.121
Beaye	3.676	1.447
Uab Meto	4.488	1.539

Bahasa Melayu Ambon dan Melayu Kupang adalah varian dari bahasa Melayu. Bahasa Indonesia sendiri pada mulanya adalah bahasa Melayu yang mengalami perkembangan sehingga struktur dan tata bahasanya berkembang menjadi sedikit berbeda dari bahasa Melayu yang asli. Kemiripan antara bahasa Indonesia dan bahasa Melayu Kupang dan Melayu Ambon ini dapat dilihat, misalnya pada kalimat bahasa Indonesia *Tom memberitahuku tentang hal itu*. yang diterjemahkan ke dalam bahasa Melayu Kupang (**mkn**) menjadi '*Tom kasi tau beta itu soal.*' dan ke dalam bahasa Melayu Ambon (**abs**) menjadi '*Tom kastau beta soal itu*.' Kata *memberitahuku* dalam bahasa Indonesia memiliki awalan atau prefiks meN- yang tidak digunakan dalam bahasa Melayu Kupang dan Melayu Ambon, kata kerja *beri tahu* menjadi '*kasi tau*' dalam bahasa Melayu Kupang atau '*kastau*' dalam bahasa Melayu Ambon dan enklitik *-ku* dalam bahasa Indonesia menjadi kata '*beta*' yang berdiri sendiri dalam bahasa Melayu Kupang dan Melayu Ambon. Contoh lainnya adalah pada kalimat *Siapa orang-orang ini*? dalam bahasa Indonesia yang diterjemahkan menjadi '*Dong sapa?*' dalam bahasa Melayu Kupang (**mkn**) dan Melayu Ambon (**abs**). Kalimat terjemahan tersebut menggunakan kata *dong* yang memiliki arti 'mereka' atau 'mereka semua' dalam bahasa Indonesia yang mengacu pada kelompok orang yang dimaksud, bukan hanya satu orang.

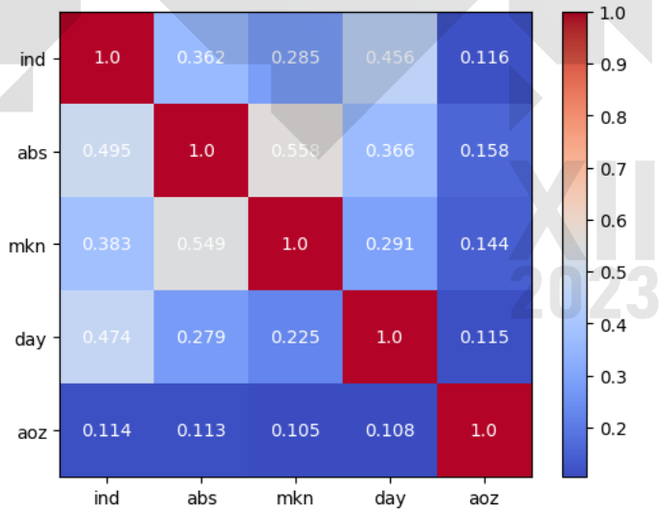
Kami juga melakukan analisis mengenai pengaruh bahasa Indonesia terhadap bahasa daerah dari segi kosakata dan tata bahasa. Banyak kata bahasa Indonesia masuk ke dalam bahasa daerah, bahkan memengaruhi tata bahasa dan struktur kalimat. Agar dapat melihat pengaruh tersebut, kami memisahkan kata-kata serapan dan kata-kata asli bahasa daerah yang tidak terdapat dalam bahasa Indonesia. Dalam Gambar 1, dapat dilihat perbedaan jumlah kata-kata unik secara keseluruhan dengan kata-kata yang tidak terdapat dalam bahasa Indonesia. Kata-kata yang tidak terdapat dalam bahasa Indonesia meliputi kata asli dalam bahasa daerah, kata serapan dari bahasa Indonesia, dan kata-kata yang tidak dapat diterjemahkan (*untranslatable words)*.



Gambar 1
Jumlah Kata dalam Bahasa Daerah

Jumlah kata asli (kata-kata yang tidak terdapat dalam bahasa Indonesia) dalam bahasa Melayu Ambon, Melayu Kupang, Beaye, dan Uab Meto sebanyak 545, 505, 689, dan 994 kata. Bahasa Uab Meto memiliki lebih banyak kata jika dibandingkan dengan tiga bahasa lainnya. Untuk kata-kata yang sulit diterjemahkan, penganotasi lebih memilih menjelaskan maknanya daripada mencari kata yang sepadan. Hal ini disebabkan oleh beberapa kata dalam bahasa Indonesia memiliki nuansa atau makna yang kompleks sehingga sulit untuk diterjemahkan langsung ke dalam bahasa daerah. Contohnya, pada kalimat *Ini bukan lelucon*. dalam bahasa Indonesia diterjemahkan ke dalam bahasa Uab Meto (**aoz**) menjadi '*I het ma kaif*.' Terjemahan kata *I* dalam bahasa Indonesia adalah 'ini'. Kata *het* adalah kata sisipan atau partikel dan frasa *ma kaif* berarti 'jangan garuk/jangan ganggu/jangan korek', yang diinterpretasikan sebagai sebuah peringatan untuk tidak memperhatikan atau mengganggu situasi yang sedang terjadi. Hasil terjemahan tersebut memberikan gambaran situasi yang secara tidak langsung berarti 'Ini bukan lelucon'. Demikian juga pada bahasa Beaye (**day**) yang menerjemahkan kalimat Sampai jumpa! menjadi '*Jaja adep bedapet!*' yang secara literal berarti 'Nanti kita bertemu!'. Kalimat tersebut diartikan sebagai ucapan perpisahan yang menyiratkan harapan untuk bertemu lagi pada masa depan.

Selanjutnya, kami juga melakukan analisis terhadap kata-kata yang tumpang tindih (*overlapping words*) antarbahasa. Gambar 2 menunjukkan ketumpangtindihan kata di setiap bahasa terhadap bahasa lainnya. Angka-angka yang terdapat dalam gambar juga sudah dinormalkan dengan jumlah kata unik setiap bahasa sehingga makin tinggi angka tersebut makin banyak ketumpangtindihan kata dalam kedua bahasa tersebut. Bahasa Melayu Kupang dan Melayu Ambon memiliki nilai yang sangat signifikan, dapat dilihat bahwa jumlah ketumpangtindihan kata dalam kedua bahasa tersebut lebih besar jika dibandingkan dengan jumlah kata unik yang dimiliki. Hal itu disebabkan oleh fakta bahwa kedua bahasa tersebut adalah varian dari bahasa Melayu sehingga banyak kata dan frasa yang mirip atau sama antara kedua bahasa tersebut.



Gambar 2
Overlapping Words Setiap Bahasa

Selain itu, bahasa Beaye juga memiliki keterkaitan yang cukup signifikan dengan bahasa Indonesia. Hal itu bisa disebabkan oleh bahasa Beaye yang belum terdokumentasi secara lengkap sehingga masih banyak aspek dari bahasa tersebut yang belum diteliti secara mendalam. Bahasa Uab Meto tidak memiliki signifikansi yang tinggi terhadap bahasa-bahasa lain, yang berarti bahwa kosakatanya tidak dipengaruhi oleh bahasa-bahasa lain tersebut. Meskipun demikian, karena letak geografisnya yang dekat dengan daerah penutur bahasa Melayu Kupang, terdapat beberapa kosakata yang sama atau mirip meskipun tidak signifikan.

Kami juga melakukan analisis frekuensi kata untuk mengidentifikasi kata-kata yang paling sering muncul dalam data. Analisis ini bertujuan untuk mengetahui topik-topik utama yang dibahas dalam data dan hubungannya dengan variabel lain. Gambar 3 memberikan gambaran mengenai kata-kata yang sering muncul dalam data di setiap bahasa. Dari hasil analisis frekuensi kata pada keempat bahasa tersebut, terlihat bahwa kata-kata yang paling sering muncul dalam teks adalah kata ganti orang pertama (*beta, au, ken*), kata ganti orang kedua (*se, lu, ko, ho*), serta kata-kata yang menunjukkan orang atau benda tertentu (*dia, Tom*). Selain itu, terdapat kata *ke*, *ka*, dan *pung* yang menyatakan kepemilikan. Dapat disimpulkan bahwa kalimat-kalimat yang ada dalam data berkaitan dengan percakapan atau kalimat sehari-hari.

Gambar 3
Word Cloud Bahasa Melayu Ambon, Melayu Kupang, Beaye, dan Uab Meto

Sebagai tambahan, terdapat beberapa kata dalam bahasa Uab Meto dan Beaye yang digunakan sebagai pelengkap atau partikel, yaitu *na*, *on*, *le*, *in*, dan *ngen*. Kata-kata tersebut juga dihitung karena kami melakukan tokenisasi atau memisahkan setiap kata dalam kalimat berdasarkan spasi, tanda baca, dan simbol-simbol tertentu dalam pemrosesan data tahap awal.

Tantangan dan Umpan Balik Penyusunan Korpus Paralel

Menyusun korpus paralel untuk bahasa-bahasa dengan sumber daya dan penutur yang sedikit merupakan proses yang kompleks dan memakan waktu. Kami mengalami kesulitan dalam mencari penganotasi sukarelawan yang sesuai dengan kriteria dan standar yang kami tetapkan serta dalam memilih terjemahan yang baik dan akurat di tengah variasi dialek yang ada. Meskipun kami meminta penganota-

si eksternal untuk mempertimbangkan penggunaan kalimat yang dapat dimengerti oleh penutur bahasa daerah dari berbagai dialek, menemukan penganotasi dengan pengetahuan komprehensif tentang berbagai ragam dialek dalam suatu bahasa tetaplah sulit. Keterbatasan akses internet dan teknologi informasi juga menjadi tantangan dalam pengumpulan dan proses pengolahan data. Hal itu menyebabkan variasi media yang digunakan oleh penganotasi dalam penerjemahan, seperti penggunaan media cetak fisik yang memperlambat proses pengolahan data dan sulit diakses oleh penganotasi. Kami kesulitan dalam melakukan pengecekan data dan menerima umpan balik dari penganotasi sehingga sering kali kami harus menyalin hasil terjemahan ke dalam format digital untuk mempermudah pengecekan dan evaluasi.

Meskipun penelitian ini menghadapi berbagai tantangan, kami menerima umpan balik positif dari para penganotasi terkait penggunaan kata-kata serapan dari bahasa asing dalam bahasa Indonesia yang tidak sesuai dengan konteks lokal. Contohnya, kata-kata Tom, Ski, DJ, Mary, MP3, dan Red Sox sering digunakan tanpa perlu diterjemahkan, tetapi penggunaannya tidak sesuai dengan konteks budaya lokal. Pada proses penerjemahan dan evaluasi, kami selalu menerima tanggapan atau umpan balik dari para penganotasi tentang seberapa banyak mereka terinspirasi dari penelitian ini untuk mempertimbangkan dan merefleksikan kembali pentingnya penggunaan bahasa daerah atau bahasa ibu mereka. Bahkan, pada beberapa kasus, proses penerjemahan dilakukan dengan bantuan keluarga besar penganotasi yang berkumpul dan menerjemahkan bersama-sama. Hal itu dapat mempererat kebersamaan dalam komunitas. Meskipun hal ini mengakibatkan kesulitan validasi dan evaluasi hasil terjemahan, kualitas hasil terjemahan yang memperhitungkan konteks budaya setempat memiliki dampak yang lebih nyata dan signifikan. Hal ini menjadi masukan berharga untuk pengembangan penelitian selanjutnya.

PENUTUP

Dalam penelitian ini, telah disajikan langkah-langkah penting dalam menyusun korpus paralel bahasa Indonesia–bahasa Melayu Ambon, Melayu Kupang, Beaye, dan Uab Meto. Penggunaan korpus paralel ini diharapkan dapat mempercepat dan memudahkan proses penelitian linguistik dan pengembangan teknologi bahasa, seperti program terjemahan otomatis. Selain itu, kami telah membuat repositori terbuka di Github[3] agar korpus paralel ini dapat diakses secara gratis dan mudah oleh masyarakat umum dan peneliti.

Kami berharap bahwa korpus paralel ini dapat memberikan kontribusi positif pada pelestarian bahasa-bahasa daerah di Indonesia dan menjadi referensi bagi penelitian selanjutnya di bidang terkait. Dalam rangka meningkatkan aksesibilitas informasi terhadap bahasa-bahasa daerah di Indonesia, langkah-langkah lebih lanjut akan kami lakukan untuk memperbanyak jumlah bahasa daerah dalam korpus paralel digital terbuka.

[3] https://github.com/joanitolopo/bhinneka-korpus

DAFTAR PUSTAKA

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., dan Ruder, S. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. Dalam *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500.

Alamsyah, A. (2018). Local language, bahasa Indonesia, or foreign language? 125, 61–66. https://doi.org/10.2991/icigr-17.2018.15.

Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan (2019). Bahasa dan peta bahasa di Indonesia edisi keenam, diperoleh melalui situs internet: https://petabahasa.kemdikbud.go.id/digital.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., dan Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699.

Eberhard, D., Simons, G., dan Fennig, C. (2019). *Ethnologue: Languages of the world, 22nd Edition*.

Grimes, C. E. (1997). *A guide to the people and languages of Nusa Tenggara*. Artha Wacana Press.

Koto, F., dan Koto, I. (2020). Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation, diperoleh melalui situs internet: http://arxiv.org/abs/2009.09309.

Novitasari, S., Tjandra, A., Sakti, S., dan Nakamura, S. (2020). Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, diperoleh melalui situs internet: https://aclanthology.org/2020.sltu-1.18.pdf, 131–138.

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

Sneddon, J. N. (2003). *The Indonesian language: Its history and role in modern society*. UNSW Press.

Sommerlot, C. J. (2020). *On the syntax of West Kalimantan: Asymmetries and a'-movement in Malayic and Land Dayak languages*, diperoleh melalui situs internet: https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/29430/SOMMERLOT-DISSERTATION-2020.pdf?sequence=1.

Suhardijanto, T., dan Dinakaramani, A. (2018). *Korpus beranotasi: Ke arah pengembangan korpus bahasa-bahasa di Indonesia*, diperoleh melalui situs internet: https://kbi.kemdikbud.go.id/kbi_back/file/dokumen_makalah/dokumen_makalah_1540364204.pdf.

Tatoeba.org (n.d.): Tatoeba: Collection of sentences and translations., diperoleh 1 Mei 2023, melalui situs internet: https://tatoeba.org/id.

Wahyu Guntara, T., Fikri Aji, A., Eko Prasojo, R., dan Kemang Raya No, J. (2020). Benchmarking multidomain English-Indonesian machine translation. *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, 35–43.

Wahyuni, M., Sujaini, H., dan Muhardi, H. (2019). Pengaruh kuantitas korpus monolingual terhadap akurasi mesin penerjemah statistik. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 7(1), 20. https://doi.org/10.26418/justin.v7i1.27241.

Waters, J. (2015). Snowball sampling: A cautionary tale involving a study of older drug users. *International Journal of Social Research Methodology*, 18(4), 367–380. https://doi.org/10.1080/13645579.2014.953316.

Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Fung, P., Baldwin, T., Lau, J. H., Sennrich, R., dan Ruder, S. (2022). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834.