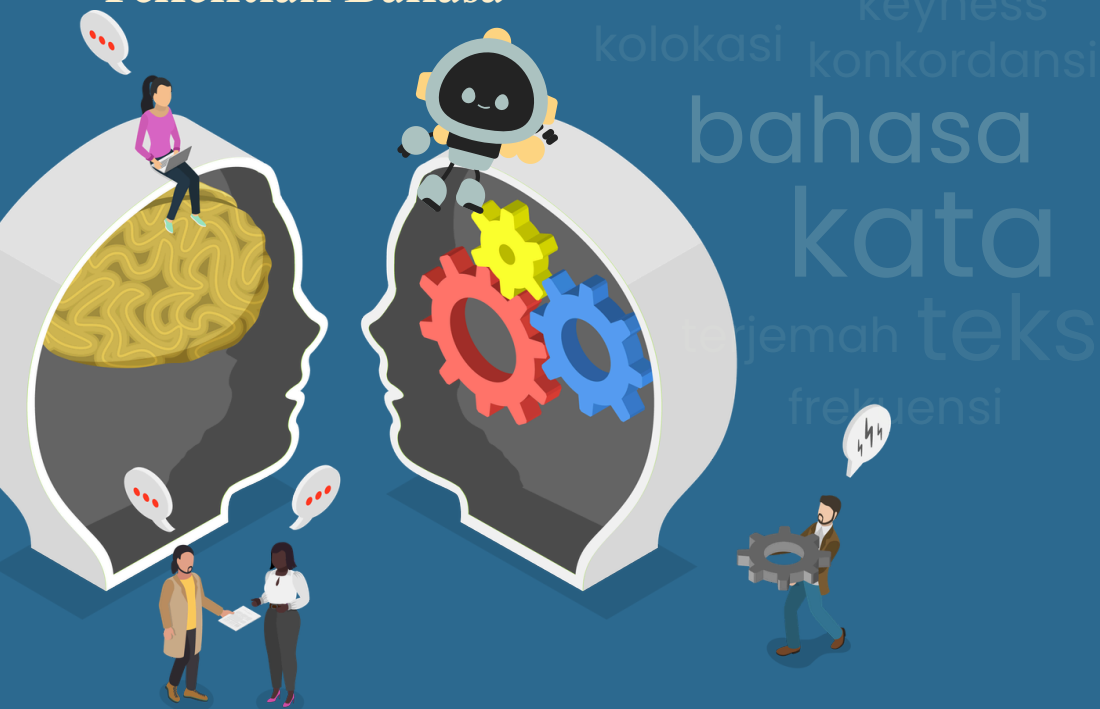


EDITOR: Dr. Baiatun Nisa, M.Pd.

 FUTURE SCIENCE

LINGUISTIK KORPUS

Alat dan Aplikasi untuk
Penelitian Bahasa



Pengajaran bahasa
berbasis korpus



Penyusunan kamus
dan glosarium



Analisis wacana
dan media



Variasi bahasa
dalam masyarakat



Kajian terjemahan dan
perbandingan bahasa

Sardian Maharani Asnur | Nurhayati S. | Bejo Sutrisno | Akhmad Hairul Umam
Dwi Indarti | Stevanus Ngenget | Cipto Wardoyo | Hanafi Pelu | Loso Judijanto
Siska Eka Syafitri | Yanuarius Yanu Dharmawan | Ria Saraswati
Abd. Hakim Muslim | Ika Apriani Fata

LINGUISTIK KORPUS:
Alat dan Aplikasi untuk Penelitian Bahasa

UU No 28 tahun 2014 tentang Hak Cipta

Fungsi dan sifat hak cipta Pasal 4

Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Pelindungan Pasal 26

Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

- i. Penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan informasi aktual;
- ii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
- iii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
- iv. Penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

Sanksi Pelanggaran Pasal 113

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

LINGUISTIK KORPUS:

Alat dan Aplikasi untuk Penelitian Bahasa

Penulis:

Sardian Maharani Asnur
Nurhayati S.
Bejo Sutrisno
Akhmad Hairul Umam
Dwi Indarti
Stevanus Ngenget
Cipto Wardoyo
Hanafi Pelu
Loso Judijanto
Siska Eka Syafitri
Yanuarius Yanu Dharmawan
Ria Saraswati
Abd. Hakim Muslim
Ika Apriani Fata

Editor:

Dr. Baiatun Nisa, M.Pd.



LINGUISTIK KORPUS:

Alat dan Aplikasi untuk Penelitian Bahasa

Penulis:

**Sardian Maharani Asnur
Nurhayati S.
Bejo Sutrisno
Akhmad Hairul Umam
Dwi Indarti
Stevanus Ngenget
Cipto Wardoyo
Hanafi Pelu
Loso Judijanto
Siska Eka Syafitri
Yanuaris Yanu Dharmawan
Ria Saraswati
Abd. Hakim Muslim
Ika Apriani Fata**

Editor: Dr. Baiatun Nisa, M.Pd.

Desain Cover: Nada Kurnia, S.I.Kom.

Tata Letak: Samuel, S.Kom.

Ukuran: A5 Unesco (15,5 x 23 cm)

Halaman: xiv, 266

e-ISBN: 978-634-7592-70-5

Terbit Pada: Juni 2026

Hak Cipta 2026, Pada Penulis

Isi diluar tanggung jawab percetakan

Copyright © 2026 by Future Science Publisher
All Right Reserved

Hak cipta dilindungi undang-undang
Dilarang keras menerjemahkan, memfotokopi, atau
memperbanyak sebagian atau seluruh isi buku ini
tanpa izin tertulis dari Penerbit.

**PENERBIT FUTURE SCIENCE
(CV. FUTURE SCIENCE)**

Anggota IKAPI (348/JTI/2022)

Perum Sarimadu II B3 No.09 Pakisaji, Kab. Malang,
Jawa Timur, Indonesia 65162
www.futuresciencepress.com

KATA PENGANTAR

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas terselesaikannya buku berjudul “*Linguistik Korpus: Alat dan Aplikasi untuk Penelitian Bahasa*”. Buku ini disusun untuk menjembatani kebutuhan akademisi, mahasiswa, peneliti, dan praktisi bahasa yang ingin memahami serta mengaplikasikan linguistik korpus dalam berbagai bidang kajian bahasa.

Perkembangan teknologi digital telah mengubah cara bahasa diproduksi, digunakan, dan disebarkan. Data bahasa kini tersedia dalam jumlah besar melalui media sosial, berita digital, komunikasi daring, dan berbagai platform berbasis teks. Dalam konteks tersebut, linguistik korpus hadir sebagai pendekatan ilmiah yang memungkinkan analisis bahasa berbasis data nyata, terukur, sistematis, dan dapat direplikasi.

Buku ini tidak hanya menghadirkan konsep-konsep dasar linguistik korpus, tetapi juga memperkenalkan teknik, perangkat lunak, analisis, serta berbagai bidang penerapannya seperti pengajaran bahasa, leksikografi, analisis wacana, sosiolinguistik, terjemahan, hingga integrasinya dengan teknologi *Artificial Intelligence* dan *Big Data*.

Kami berharap buku ini dapat:

1. Menjadi bahan ajar yang komprehensif untuk mahasiswa.
2. Menjadi panduan praktis bagi peneliti dalam merancang penelitian berbasis korpus.
3. Menjadi referensi bagi praktisi bahasa dalam pengembangan kamus, materi ajar, dan riset kebahasaan terapan.

Ucapan terima kasih kami sampaikan kepada rekan peneliti, mahasiswa, dan komunitas akademik yang turut memberikan inspirasi dan kontribusi ide dalam penyusunan karya ini. Kritik dan saran sangat kami harapkan untuk penyempurnaan karya di masa mendatang.

Semoga buku ini dapat memberikan manfaat dan menjadi landasan kuat dalam mengembangkan penelitian bahasa berbasis data.

Malang, Mei 2026

Editor

DAFTAR ISI

KATA PENGANTAR.....	v
DAFTAR ISI	vii
BAB 1 PENGANTAR LINGUISTIK KORPUS	1
Sardian Maharani Asnur	1
PENDAHULUAN	1
SEJARAH DAN PERKEMBANGAN LINGUISTIK KORPUS.....	3
KONSEP KORPUS DALAM KAJIAN BAHASA.....	6
PERAN DATA OTENTIK DALAM ANALISIS BAHASA.....	9
HUBUNGAN LINGUISTIK KORPUS DENGAN LINGUISTIK MODERN.....	12
KELEBIHAN DAN KETERBATASAN PENDEKATAN KORPUS.....	15
KESIMPULAN.....	17
BAB 2 JENIS DAN KARAKTERISTIK KORPUS.....	21
Nurhayati S.	21
PENDAHULUAN	21
KORPUS UMUM DAN KORPUS SPESIFIK.....	22
KORPUS LISAN DAN KORPUS TULISAN.....	25
KORPUS SINKRONIS DAN DIAKRONIS	28
KORPUS DIAKRONIS.....	29
KORPUS BERBAHASA TUNGGA DAN MULTIBAHASA	30
KORPUS PARALEL DAN KORPUS KOMPARABEL ...	33

	KRITERIA KUALITAS DAN KEREPRESENTATIFAN KORPUS	34
	ANOTASI DAN METADATA	35
	KESIMPULAN	36
BAB 3	TEKNIK PENGUMPULAN DAN PENYUSUNAN KORPUS	41
	Bejo Sutrisno	41
	PENDAHULUAN	41
	PENGUMPULAN DATA TEKS DARI SUMBER DIGITAL	43
	TRANSKRIPSI TUTURAN UNTUK KORPUS LISAN ..	44
	NORMALISASI DAN PEMBERSIHAN FORMAT DATA	46
	PENANDAAN METADATA DAN KLASIFIKASI	48
	PENYIMPANAN, PENGELOLAAN, DAN BACKUP DATA	50
	KESIMPULAN	52
BAB 4	PENANDAAN (ANNOTATION) DALAM LINGUISTIK KORPUS	61
	Akhmad Hairul Umam	61
	PENDAHULUAN	61
	PENGERTIAN PENANDAAN (ANNOTATION)	62
	FUNGSI DAN TUJUAN	63
	JENIS-JENIS PENANDAAN	66
	PRINSIP-PRINSIP PENANDAAN YANG BAIK	72
	KESIMPULAN	74
BAB 5	ALAT DAN PERANGKAT LUNAK LINGUISTIK KORPUS	79

Dwi Indarti.....	79
PENDAHULUAN	79
KLASIFIKASI ALAT LINGUISTIK KORPUS	80
PERANGKAT LUNAK POPULER DALAM LINGUISTIK KORPUS.....	82
INTEGRASI NLP DALAM LINGUISTIK KORPUS	83
TANTANGAN DAN KETERBATASAN ALAT LINGUISTIK KORPUS	84
PANDUAN PRAKTIS: PENGGUNAAN ANTCONC (STEP-BY-STEP).....	86
PANDUAN PRAKTIS: PENGGUNAAN #LANCSBOX (STEP-BY-STEP).....	88
KESIMPULAN.....	89
BAB 6 ANALISIS FREKUENSI DAN DISTRIBUSI KORPUS ..	95
Stevanus Ngenget	95
PENDAHULUAN	95
KONSEP DASAR FREKUENSI DALAM KAJIAN KORPUS.....	97
PERANGKAT LUNAK ANALISIS FREKUENSI	98
TEKNIK IDENTIFIKASI DAN PENGUKURAN FREKUENSI	99
PENGHITUNGAN FREKUENSI KATA	101
VISUALISASI FREKUENSI KATA MENGGUNAKAN WORD CLOUD	103
HUBUNGAN LEKSIKAL DAN POLA KEMUNCULAN KATA	104
IDENTIFIKASI KEKHASAN LEKSIKAL DALAM KORPUS.....	106

	PENYEBARAN KATA DAN FRASA DALAM BERBAGAI KONTEKS.....	107
	MEMAKNAI DATA FREKUENSI DALAM PENELITIAN BAHASA	108
	KESIMPULAN	108
BAB 7	KONKORDANSI DAN INTERPRETASI DATA.....	113
	Cipto Wardoyo	113
	PENDAHULUAN	113
	KONSEP KONKORDANSI.....	115
	MEMBACA BARIS KONKORDANSI (KWIC).....	118
	ANALISIS POLA PENGGUNAAN KATA	127
	KESIMPULAN	130
BAB 8	LINGUISTIK KORPUS DAN PENGAJARAN BAHASA	135
	Hanafi Pelu.....	135
	PENDAHULUAN	135
	FONDASI LINGUISTIK KORPUS	137
	PRINSIP-PRINSIP DESAIN KORPUS	138
	METODOLOGI ANALISIS KORPUS	139
	LINGUISTIK KORPUS DALAM PENGAJARAN BAHASA	140
	TRANSFORMASI PARADIGMA PENGAJARAN BAHASA MELALUI KORPUS.....	142
	IMPLIKASI UNTUK PENGEMBANGAN MATERI DAN KURIKULUM	142
	TANTANGAN DAN PROSPEK MASA DEPAN.....	143
	TEORI USAGE-BASED DAN LINGUISTIK KOGNITIF.....	144

TEORI KONSTRUKSI DAN PEMBELAJARAN POLA	145
TEORI INPUT DAN INTERAKSI DALAM AKUISISI BAHASA KEDUA	146
TEORI FORMULAIC LANGUAGE DAN CHUNKING	147
TEORI REGISTER DAN VARIASI BAHASA	148
TEORI PEMBELAJARAN INDUKTIF DAN DISCOVERY LEARNING	149
TEORI KOMUNIKATIF DAN AUTENTISITAS INPUT.....	150
INTEGRASI TEKNOLOGI DALAM PEMBELAJARAN BAHASA.....	152
KESIMPULAN.....	152
BAB 9 LINGUISTIK KORPUS DALAM LEKSIKOGRAFI	163
Loso Judijanto.....	163
PENDAHULUAN	163
LANDASAN OPERASIONAL LINGUISTIK KORPUS DALAM LEKSIKOGRAFI.....	164
PEMANFAATAN DATA KORPUS DALAM PENYUSUNAN ENTRI KAMUS	168
INOVASI LEKSIKOGRAFI BERBASIS LINGUISTIK KORPUS.....	174
KESIMPULAN.....	179
BAB 10 LINGUISTIK KORPUS DALAM KAJIAN WACANA .	185
Siska Eka Syafitri.....	185
PENDAHULUAN	185
IDENTIFIKASI TEMA DAN TOPIK.....	185
ANALISIS POLA BAHASA DALAM DISKURUS	187
KORPUS DALAM ANALISIS WACANA KRITIS	190

	PERBANDINGAN WACANA ANTAR KELOMPOK/MEDIA	193
	KESIMPULAN	198
BAB 11	LINGUISTIK KORPUS DALAM SOCIOLINGUISTICS	201
	Yanuarius Yanu Dharmawan	201
	PENDAHULUAN	201
	VARIASI BAHASA BERDASARKAN WILAYAH	204
	VARIASI BERDASARKAN GENDER, USIA, DAN SOSIAL	207
	KORPUS LISAN UNTUK ANALISIS VARIASI	209
	ANALISIS PERUBAHAN BAHASA DI KOMUNITAS	210
	KORPUS DAN IDENTITAS SOSIAL	211
	KESIMPULAN	213
BAB 12	LINGUISTIK KORPUS DALAM TERJEMAHAN DAN INTERPRETASI	225
	Ria Saraswati	225
	PENDAHULUAN	225
	KORPUS PARALEL DAN KORPUS KOMPARATIF	226
	IDENTIFIKASI POLA TERJEMAHAN	228
	PENERJEMAHAN BERBASIS DATA KORPUS	229
	ANALISIS KUALITAS TERJEMAHAN	230
	PENERAPAN KORPUS DALAM PELATIHAN PENERJEMAH	231
	KESIMPULAN	232
BAB 13	BIG DATA DAN KORPUS DIGITAL BERSKALA BESAR	239

Abd. Hakim Muslim	239
PENDAHULUAN	239
WEB SEBAGAI KORPUS DAN PEMROSESAN DATA SKALA BESAR	246
PEMBELAJARAN MESIN DAN PENAMBANGAN TEKS	247
PENGAMBILAN DATA DAN OTOMATISASI PENGUMPULAN DATA WEB	249
TANTANGAN DALAM ANALISIS KORPUS SKALA BESAR.....	250
INTEGRASI LINGUISTIK KORPUS DAN KECERDASAN BUATAN	251
KESIMPULAN.....	252
BAB 14 MASA DEPAN LINGUISTIK KORPUS	257
Ika Apriani Fata	257
PENDAHULUAN	257
PERKEMBANGAN TEKNOLOGI DAN DATA BAHASA	258
KORPUS MULTIMODAL DAN STANDARISASI ETIKA PEMANFAATAN DATA	259
PROSPEK RISET DAN KOLABORASI GLOBAL	261
KESIMPULAN.....	262

BAB 1

PENGANTAR LINGUISTIK KORPUS

Sardian Maharani Asnur
Universitas Islam Negeri Alauddin, Makassar
E-mail: sardian.maharani@uin-alauddin.ac.id

PENDAHULUAN

Bahasa merupakan sistem yang hidup, dinamis, dan terus berkembang seiring perubahan sosial, budaya, teknologi, serta kebutuhan komunikasi manusia. Karena sifatnya yang terus berubah, kajian bahasa tidak cukup hanya mengandalkan intuisi penutur atau contoh-contoh kalimat yang dibuat secara artifisial. Penelitian bahasa modern menuntut data nyata yang merepresentasikan bagaimana bahasa benar-benar digunakan dalam kehidupan sehari-hari. Dalam konteks inilah linguistik korpus hadir sebagai salah satu pendekatan paling berpengaruh dalam studi kebahasaan kontemporer.

Secara umum, linguistik korpus adalah cabang ilmu linguistik yang mempelajari bahasa melalui kumpulan data teks atau ujaran yang tersusun secara sistematis dan dapat dianalisis dengan bantuan komputer. Kumpulan data tersebut disebut korpus, yaitu himpunan bahasa autentik yang dikumpulkan berdasarkan tujuan tertentu sehingga dapat digunakan untuk mengamati pola penggunaan bahasa secara empiris (McEnery & Hardie, 2019; Pérez-Paredes & Curry, 2024). Melalui korpus, peneliti tidak lagi hanya bertanya bagaimana bahasa seharusnya digunakan, tetapi juga bagaimana bahasa sesungguhnya digunakan oleh penutur dalam berbagai konteks.

Perkembangan teknologi digital telah menjadikan linguistik korpus semakin relevan. Jika pada masa lalu pengumpulan data bahasa membutuhkan waktu panjang dan tenaga besar, kini

jutaan kata dapat dihimpun, disimpan, dan dianalisis dalam waktu relatif singkat. Perangkat lunak seperti AntConc, Sketch Engine, dan berbagai platform berbasis web memungkinkan peneliti menelusuri frekuensi kata, kolokasi, konkordansi, variasi gramatikal, hingga perubahan makna secara lebih objektif dan terukur. Hal ini menjadikan linguistik korpus sebagai metode yang sangat penting dalam era big data kebahasaan.

Linguistik korpus juga memiliki cakupan aplikasi yang luas. Pendekatan ini digunakan dalam penelitian leksikografi untuk penyusunan kamus modern, pengajaran bahasa, penerjemahan, analisis wacana, kajian pragmatik, forensik linguistik, hingga pengembangan teknologi bahasa seperti mesin penerjemah dan kecerdasan buatan berbasis Artificial Intelligence (Biber et al., 1998; Lin & Adolphs, 2023). Dengan kata lain, linguistik korpus tidak hanya penting secara teoritis, tetapi juga sangat berguna secara praktis.

Bagi peneliti pemula, linguistik korpus menawarkan jalan masuk yang menarik ke dunia penelitian bahasa karena berbasis data nyata dan dapat diverifikasi. Temuan penelitian korpus cenderung lebih transparan karena sumber data dapat diperiksa ulang, prosedur analisis dapat direplikasi, dan hasilnya dapat diuji secara kuantitatif maupun kualitatif. Pendekatan ini sekaligus membantu mengurangi bias subjektivitas yang sering muncul ketika analisis hanya bergantung pada intuisi peneliti (Sinclair, 1991).

Namun demikian, memahami linguistik korpus tidak cukup hanya dengan mengenal perangkat lunaknya. Peneliti juga perlu memahami prinsip dasar korpus, representativitas data, desain korpus, teknik anotasi, serta interpretasi hasil analisis. Tanpa pemahaman metodologis tersebut, data yang besar justru dapat menghasilkan kesimpulan yang keliru. Karena itu, bab ini disusun sebagai pengantar fundamental untuk mengenalkan

konsep, sejarah, fungsi, serta posisi linguistik korpus dalam perkembangan ilmu bahasa modern.

Melalui pembahasan pada bab ini, pembaca diharapkan memperoleh gambaran awal bahwa linguistik korpus bukan sekadar penggunaan komputer dalam linguistik, melainkan paradigma penelitian yang menempatkan data autentik sebagai pusat analisis bahasa. Dengan fondasi ini, pembaca akan lebih siap memasuki pembahasan lanjutan mengenai alat, metode, dan aplikasi linguistik korpus pada bab-bab berikutnya.

SEJARAH DAN PERKEMBANGAN LINGUISTIK KORPUS

Linguistik korpus sebagai bidang kajian ilmiah tidak muncul secara tiba-tiba. Perkembangannya merupakan hasil evolusi panjang dari tradisi dokumentasi bahasa, analisis teks, kemajuan teknologi komputasi, serta perubahan paradigma dalam ilmu linguistik. Secara historis, gagasan dasar linguistik korpus berangkat dari keyakinan bahwa bahasa sebaiknya dipelajari melalui bukti penggunaan nyata, bukan semata-mata melalui spekulasi teoritis atau intuisi penutur.

Pada masa awal, sebelum istilah *corpus linguistics* dikenal luas, para filolog dan ahli bahasa telah lama menggunakan kumpulan teks sebagai sumber kajian bahasa. Pada abad ke-19 dan awal abad ke-20, penelitian bahasa banyak bertumpu pada manuskrip kuno, karya sastra, dokumen hukum, dan teks keagamaan untuk menelusuri sejarah bahasa, perubahan bunyi, serta perkembangan makna kata. Dalam konteks ini, “korpus” masih dipahami sebagai kumpulan teks tertulis yang dijadikan bahan analisis manual (Biber et al., 1998; McEnery & Hardie, 2019). Tradisi tersebut tampak jelas dalam penyusunan kamus besar seperti Oxford English Dictionary, yang menggunakan ribuan kutipan autentik dari teks nyata untuk

mendokumentasikan penggunaan kata. Pendekatan berbasis bukti ini menjadi fondasi penting bagi linguistik korpus modern.

Pada pertengahan abad ke-20 (era strukturalisme), linguistik struktural memberi kontribusi besar terhadap perkembangan linguistik korpus. Tokoh seperti Zellig Harris menekankan pentingnya distribusi unsur bahasa dalam data nyata. Harris (1951) memperkenalkan metode analisis distribusional yang mempelajari pola kemunculan unsur linguistik dalam konteks tertentu. Gagasan ini sangat dekat dengan prinsip linguistik korpus masa kini, terutama dalam analisis kolokasi dan pola gramatikal. Pada periode ini, data bahasa masih dikumpulkan dan dianalisis secara manual, sehingga ukuran korpus relatif kecil. Meski demikian, pendekatan empiris terhadap data bahasa mulai memperoleh pijakan metodologis yang kuat.

Namun, pada tahun 1950-an hingga 1970-an, perkembangan linguistik korpus sempat melambat karena dominasi linguistik generatif yang dipelopori Noam Chomsky. Chomsky (2014) berargumen bahwa tujuan linguistik adalah menjelaskan kompetensi mental penutur ideal, bukan sekadar mendeskripsikan performa bahasa dalam data aktual. Dalam pandangan ini, korpus dianggap terbatas karena hanya merekam ujaran yang terbatas, mengandung kesalahan, dan tidak cukup untuk menjelaskan kreativitas bahasa manusia. Akibatnya, penggunaan korpus sempat dipandang kurang sentral dalam teori linguistik arus utama. Namun, kritik terhadap pendekatan yang terlalu bergantung pada intuisi kemudian membuka jalan bagi kebangkitan kembali metode berbasis data.

Kebangkitan linguistik korpus modern dimulai ketika komputer mulai digunakan dalam penelitian Bahasa. Salah satu tonggak penting adalah pembuatan *the Brown University Standard Corpus of Present-Day American English* pada tahun 1960-an oleh Henry Kučera dan W. Nelson Francis. Brown Corpus berisi sekitar satu juta kata bahasa Inggris Amerika dari

berbagai genre tertulis dan menjadi korpus elektronik berskala besar pertama yang banyak digunakan dalam penelitian Bahasa (Francis & Kucera, 1964). Kehadiran Brown Corpus menunjukkan bahwa komputer mampu mempercepat pengolahan data bahasa secara signifikan.

Keberhasilan Brown Corpus mendorong lahirnya berbagai korpus penting lainnya, seperti Lancaster-Oslo/Bergen Corpus (LOB Corpus) untuk bahasa Inggris Britania, British National Corpus (BNC) yang berisi sekitar 100 juta kata, dan Corpus of Contemporary American English (COCA) yang menjadi salah satu korpus paling populer saat ini. Dengan hadirnya korpus-korpus tersebut, peneliti dapat membandingkan variasi bahasa, perubahan kosakata, serta perbedaan register secara lebih akurat.

Pada akhir abad ke-20, linguistik korpus tidak lagi dipandang sekadar alat bantu teknis, tetapi berkembang menjadi pendekatan metodologis yang mapan. John Sinclair melalui karyanya *Corpus, Concordance, Collocation* (1991) menekankan bahwa makna bahasa sering muncul dari pola penggunaan berulang, bukan hanya dari aturan abstrak. Pada masa ini berkembang konsep penting seperti concordance, collocation, frequency analysis, dan semantic prosody, yang hingga kini menjadi dasar analisis korpus.

Memasuki abad ke-21, perkembangan internet membawa perubahan besar terhadap linguistik korpus. Data bahasa kini tersedia dalam jumlah sangat besar melalui berita daring, blog, media sosial, forum diskusi, subtitle film, dan komunikasi digital lainnya. Korpus tidak lagi terbatas pada teks tertulis formal, tetapi juga mencakup ujaran spontan dan interaksi digital. Perangkat lunak seperti AntConc dan Sketch Engine memudahkan peneliti menganalisis jutaan kata secara cepat dan efisien.

Saat ini, linguistik korpus memiliki peran penting dalam berbagai bidang, seperti pengajaran bahasa, penerjemahan,

leksikografi, analisis wacana, forensik linguistik, hingga pengembangan teknologi bahasa berbasis Artificial Intelligence. Dengan kata lain, linguistik korpus telah berkembang dari tradisi pengumpulan teks sederhana menjadi disiplin ilmiah modern yang sangat relevan di era digital.

Sejarah linguistik korpus menunjukkan bahwa kemajuan ilmu bahasa sangat dipengaruhi oleh perubahan teknologi dan metodologi. Dari analisis manual hingga pemrosesan data besar, linguistik korpus terus memperluas cara manusia memahami bahasa secara lebih objektif, empiris, dan aplikatif.

KONSEP KORPUS DALAM KAJIAN BAHASA

Istilah korpus berasal dari bahasa Latin *corpus* yang berarti “tubuh” atau “kumpulan yang utuh.” Dalam kajian linguistik modern, istilah ini digunakan untuk merujuk pada kumpulan data bahasa yang dihimpun secara sistematis dan digunakan sebagai dasar analisis ilmiah. Data tersebut dapat berupa teks tertulis, transkrip ujaran lisan, maupun bentuk komunikasi digital yang mewakili penggunaan bahasa dalam konteks nyata. Berbeda dengan kumpulan teks biasa, korpus disusun berdasarkan prinsip ilmiah tertentu seperti representativitas, keseimbangan data, kejelasan sumber, dan tujuan analisis yang spesifik (McEnery & Hardie, 2019).

Secara sederhana, jika seseorang menyimpan seribu artikel berita di laptop, itu belum tentu disebut korpus. Namun, apabila artikel-artikel tersebut dipilih dengan kriteria tertentu—misalnya berasal dari lima media nasional, rentang waktu tertentu, genre tertentu, dan dikonversi ke format yang dapat dianalisis—maka kumpulan tersebut dapat dikategorikan sebagai korpus. Jadi, korpus bukan sekadar banyaknya data, tetapi kualitas desain datanya.

Stepanov (2022) mendefinisikan korpus sebagai kumpulan bahasa autentik yang disimpan dalam bentuk elektronik dan

dipilih untuk merepresentasikan suatu bahasa atau variasi bahasa tertentu. Definisi ini menegaskan dua unsur penting: keotentikan data dan representasi penggunaan bahasa.

Salah satu konsep utama dalam korpus adalah representativitas, yaitu sejauh mana data yang dikumpulkan mampu mewakili karakteristik bahasa atau variasi bahasa yang diteliti. Sebuah korpus tidak harus memuat seluruh bentuk bahasa, karena hal tersebut hampir mustahil dilakukan. Sebaliknya, korpus harus dirancang agar cukup mencerminkan pola penggunaan bahasa pada komunitas, periode, atau konteks tertentu. Misalnya, penelitian tentang bahasa akademik sebaiknya menggunakan data dari jurnal, buku ilmiah, atau tugas mahasiswa. Sebaliknya, penelitian tentang bahasa percakapan sehari-hari akan lebih tepat menggunakan transkrip dialog atau percakapan spontan. Dengan demikian, kualitas korpus tidak hanya ditentukan oleh jumlah data, tetapi juga oleh kesesuaian data dengan tujuan penelitian (Sevilla-Requena, 2025)

Korpus yang baik umumnya memiliki beberapa ciri utama. Pertama, data bersifat otentik (*authentic*), yaitu berasal dari penggunaan bahasa nyata. Misalnya percakapan sehari-hari, artikel surat kabar, pidato, unggahan media sosial, atau karya akademik; Kedua, data dikumpulkan secara sistematis (*systematically*), dengan kriteria yang jelas mengenai sumber data, jumlah teks, periode waktu, penutur, genre, dan variabel lainnya ditentukan sejak awal; Ketiga, korpus biasanya tersedia dalam format digital (*machine-readable*) sehingga dapat dianalisis menggunakan perangkat lunak seperti AntConc atau Sketch Engine; Keempat, korpus sering dilengkapi *metadata*, seperti informasi tambahan tentang data, misalnya nama sumber, tahun publikasi, jenis teks, jenis kelamin penutur, usia, lokasi geografis, tingkat pendidikan, yang berguna untuk analisis lebih lanjut; Kelima, korpus dapat diverifikasi yang memungkinkan

peneliti melakukan pencarian kata, frasa, pola gramatikal, kolokasi, frekuensi, dan distribusi secara cepat; Keenam, korpus bersifat *representative* atau diharapkan mampu mewakili karakteristik bahasa yang diteliti. Jika meneliti bahasa Inggris akademik, maka isi korpus sebaiknya berasal dari artikel jurnal, esai mahasiswa, buku akademik, dan genre sejenis.

Dalam praktiknya, korpus berkembang ke dalam berbagai bentuk sesuai kebutuhan ilmiah. Beberapa jenis korpus yang umum digunakan antara lain: 1) korpus umum (*General corpus*), dirancang untuk merepresentasikan bahasa secara luas dari berbagai genre. Contohnya British National Corpus yang memuat data bahasa Inggris lisan dan tulisan. 2) Korpus spesifik (*specialized corpus*), berfokus pada bidang tertentu, seperti bahasa hukum, bahasa medis, atau bahasa akademik. Yang mana tentunya akan sangat berguna untuk penelitian khusus. 3) Korpus lisan (*spoken corpus*), yang berisi transkrip percakapan, wawancara, pidato, diskusi kelas, atau interaksi verbal lainnya. Korpus ini penting untuk analisis pragmatik, fonologi, dan wacana. 4) Korpus tertulis (*written corpus*) yang berisi artikel, buku, atau surat kabar. 5) Korpus sinkronik (*synchronic corpus*), yang menggambarkan bahasa pada periode waktu tertentu. 6) Korpus diakronik (*diachronic corpus*) digunakan untuk meneliti perubahan bahasa lintas waktu, misalnya perubahan kosakata atau struktur gramatikal dari abad ke abad. 7) Korpus pembelajar (*learner corpus*) yang berisi produksi bahasa pelajar bahasa kedua. Korpus ini banyak digunakan dalam penelitian pemerolehan bahasa dan pengajaran EFL/ESL (Dash, 2008; Meyer, 2023)

Penting dipahami bahwa korpus bukan sekadar kumpulan data besar. Korpus kecil yang dirancang dengan baik sering kali lebih bermanfaat daripada data besar yang tidak terarah. Misalnya, penelitian tentang kesalahan tata bahasa mahasiswa dapat dilakukan dengan korpus tugas menulis yang jumlahnya

terbatas tetapi relevan. Sebaliknya, jutaan data dari internet belum tentu berguna jika tidak sesuai dengan fokus penelitian.

Dalam kajian bahasa, korpus memiliki dua fungsi utama. Pertama, sebagai sumber data, yaitu menyediakan bukti nyata tentang penggunaan bahasa. Kedua, sebagai pendekatan metodologis, yaitu cara meneliti bahasa melalui observasi empiris, frekuensi, distribusi, dan konteks penggunaan. Oleh karena itu, linguistik korpus bukan hanya penggunaan teknologi, tetapi juga cara pandang ilmiah yang menempatkan data sebagai pusat analisis.

Perkembangan teknologi digital semakin memperluas konsep korpus. Saat ini, peneliti dapat membangun korpus dari berita daring, media sosial, subtitle film, forum internet, hingga percakapan kelas online. Hal ini menjadikan korpus sangat relevan dalam penelitian bahasa kontemporer, termasuk studi komunikasi digital dan pengembangan teknologi bahasa berbasis *Artificial Intelligence*.

Dengan demikian, konsep korpus dalam kajian bahasa menegaskan bahwa penelitian linguistik yang kuat membutuhkan data yang tepat, representatif, dan dapat diverifikasi. Pemahaman terhadap konsep ini menjadi dasar penting sebelum peneliti melangkah ke tahap analisis yang lebih teknis, seperti frekuensi kata, kolokasi, konkordansi, atau analisis wacana berbasis korpus.

PERAN DATA OTENTIK DALAM ANALISIS BAHASA

Dalam linguistik korpus, data otentik merupakan fondasi utama dalam proses analisis bahasa. Data otentik merujuk pada bahasa yang digunakan secara nyata oleh penutur atau penulis dalam situasi komunikasi sesungguhnya, bukan kalimat yang dibuat secara artifisial untuk kepentingan teori atau latihan tata bahasa. Bentuk data ini dapat berupa percakapan sehari-hari, artikel berita, pidato, karya sastra, tulisan akademik, pesan

digital, maupun interaksi di media sosial. Penggunaan data otentik sangat penting karena memberikan gambaran yang lebih realistis mengenai bagaimana bahasa benar-benar dipakai dalam kehidupan sosial (Biber et al., 1998).

Dalam tradisi linguistik sebelumnya, analisis bahasa sering bergantung pada intuisi penutur ahli atau contoh kalimat buatan. Pendekatan tersebut bermanfaat untuk menjelaskan aturan gramatikal, tetapi sering kali tidak cukup untuk menunjukkan variasi dan kompleksitas bahasa dalam penggunaan nyata. Melalui data otentik, peneliti dapat melihat bahwa bahasa sehari-hari tidak selalu mengikuti bentuk ideal sebagaimana tertulis dalam buku tata bahasa. Penutur sering menggunakan singkatan, pengulangan, jeda, kalimat tidak lengkap, atau ungkapan tertentu yang sangat bergantung pada konteks komunikasi.

Salah satu peran utama data otentik adalah menyediakan bukti empiris dalam penelitian bahasa. Dengan data nyata, peneliti dapat mengetahui frekuensi penggunaan kata, kecenderungan struktur kalimat, serta hubungan antarkata dalam konteks tertentu. Misalnya, analisis korpus menunjukkan bahwa ungkapan *make a decision* jauh lebih umum digunakan daripada *do a decision*. Demikian pula, frasa *strong coffee* lebih alami dibanding *powerful coffee*. Informasi seperti ini sangat penting dalam studi kolokasi, semantik, dan pengajaran bahasa.

Data otentik juga berperan penting dalam mengungkap variasi bahasa. Bahasa yang digunakan dalam artikel akademik tentu berbeda dari bahasa media sosial. Bahasa dalam pidato resmi berbeda dari bahasa percakapan santai. Dengan menganalisis data dari berbagai sumber, peneliti dapat membandingkan penggunaan bahasa berdasarkan register, usia penutur, profesi, wilayah, atau media komunikasi. Oleh karena itu, data otentik sangat bermanfaat dalam bidang sosiolinguistik dan analisis wacana.

Dalam pengajaran bahasa, penggunaan data otentik memberikan manfaat praktis yang besar. Pembelajar bahasa asing sering mengalami kesulitan karena materi ajar hanya berisi kalimat buatan yang kurang alami. Melalui data korpus, guru dapat menyajikan contoh penggunaan bahasa yang benar-benar lazim dipakai penutur asli. Misalnya, siswa dapat belajar bahwa ungkapan *How's it going?* umum digunakan dalam percakapan informal, sedangkan *How are you?* lebih netral atau formal. Pendekatan ini mendukung metode data-driven learning, yaitu pembelajaran bahasa melalui eksplorasi data nyata (Godwin-Jones, 2021)

Dalam bidang leksikografi, data otentik juga menjadi sumber utama penyusunan kamus modern. Definisi kata, contoh kalimat, makna baru, dan tingkat frekuensi penggunaan biasanya ditentukan berdasarkan korpus. Banyak kamus besar seperti *Oxford English Dictionary* dan *Longman Dictionary of Contemporary English* memanfaatkan jutaan data autentik dalam proses penyusunannya.

Meskipun memiliki banyak kelebihan, penggunaan data otentik juga menghadapi tantangan. Jumlah data yang sangat besar memerlukan keterampilan teknis dan perangkat lunak untuk mengelolanya. Selain itu, data nyata kadang mengandung kesalahan ketik, struktur tidak lengkap, atau konteks sosial yang terbatas. Dalam penggunaan data digital, peneliti juga perlu memperhatikan aspek etika dan privasi. Namun demikian, tantangan tersebut tidak mengurangi nilai penting data otentik sebagai sumber analisis bahasa yang paling dekat dengan realitas penggunaan.

Di era digital, ketersediaan data otentik semakin luas. Peneliti kini dapat mengumpulkan data dari berita online, forum diskusi, komentar pengguna, *subtitle film*, *podcast*, hingga percakapan kelas daring. Bahkan, banyak sistem bahasa berbasis Artificial Intelligence dilatih menggunakan data autentik dalam

jumlah besar agar mampu memahami pola bahasa manusia secara lebih akurat.

Dengan demikian, data otentik memiliki peran sentral dalam analisis bahasa karena memungkinkan penelitian yang lebih objektif, kontekstual, dan relevan. Melalui data nyata, peneliti dapat memahami bahasa sebagaimana digunakan dalam masyarakat, bukan sekadar sebagai sistem aturan abstrak. Oleh karena itu, dalam linguistik korpus, data otentik menjadi dasar utama untuk menghasilkan kajian bahasa yang ilmiah dan aplikatif.

HUBUNGAN LINGUISTIK KORPUS DENGAN LINGUISTIK MODERN

Linguistik modern berkembang melalui berbagai pendekatan yang berupaya menjelaskan bahasa sebagai sistem, perilaku sosial, proses kognitif, dan alat komunikasi. Dalam perkembangan tersebut, linguistik korpus memiliki posisi penting karena menyediakan data autentik yang dapat digunakan untuk menguji teori, menjelaskan pola kebahasaan, dan memperluas metode penelitian. Jika sebagian cabang linguistik berfokus pada konsep teoretis tentang bahasa, maka linguistik korpus menekankan bagaimana bahasa benar-benar digunakan dalam kehidupan nyata.

Salah satu hubungan paling jelas terlihat pada linguistik deskriptif. Linguistik deskriptif bertujuan menggambarkan bahasa sebagaimana dipakai oleh masyarakat, bukan menentukan benar atau salah secara preskriptif. Pendekatan korpus sangat mendukung tujuan ini karena menyediakan bukti empiris mengenai frekuensi kata, variasi struktur kalimat, serta perbedaan penggunaan bahasa dalam berbagai konteks. Misalnya, data korpus dapat menunjukkan bahwa bentuk kontraksi seperti *I'm*, *don't*, dan *it's* jauh lebih umum dalam percakapan dibandingkan dengan bentuk penuh seperti *I am* atau

do not (Biber et al., 1998). Temuan semacam ini membantu menghasilkan deskripsi bahasa yang lebih realistis.

Hubungan linguistik korpus dengan linguistik generatif juga menarik untuk dicermati. Linguistik generatif yang dipelopori Noam Chomsky menekankan kemampuan mental manusia dalam menghasilkan kalimat yang tidak terbatas jumlahnya. Fokus utamanya adalah kompetensi bahasa, bukan data aktual. Pada awalnya, pendekatan korpus dianggap kurang memadai karena hanya merekam performa bahasa yang terbatas. Namun, perkembangan mutakhir menunjukkan bahwa keduanya dapat saling melengkapi. Linguistik generatif menawarkan penjelasan teoretis tentang struktur bahasa, sedangkan linguistik korpus menyediakan bukti penggunaan nyata untuk menguji atau memperkaya teori tersebut (Chomsky, 2014).

Dalam bidang Sociolinguistics, linguistik korpus memiliki kontribusi besar karena memungkinkan analisis variasi bahasa berdasarkan faktor sosial. Peneliti dapat membandingkan penggunaan bahasa menurut usia, gender, wilayah, kelas sosial, atau medium komunikasi. Misalnya, korpus media sosial dapat menunjukkan perbedaan pilihan kosakata antara generasi muda dan generasi tua, atau antara bahasa formal dan informal. Baker (2010) menegaskan bahwa korpus menjadi alat penting dalam sociolinguistik modern karena mampu menggabungkan data besar dengan analisis sosial.

Hubungan berikutnya tampak pada bidang pragmatik, yaitu studi tentang makna bahasa dalam konteks. Linguistik korpus membantu pragmatik dengan menyediakan data nyata mengenai penggunaan ungkapan, strategi kesantunan, tindak tutur, dan penanda wacana. Kata seperti *well*, *actually*, *just*, dan *you know* sering berfungsi pragmatis dalam percakapan, tetapi sulit dijelaskan hanya melalui tata bahasa formal. Karena itu, muncul cabang baru yang dikenal sebagai corpus pragmatics, yaitu

kajian pragmatik berbasis data korpus (Aijmer & Rühlemann, 2015; Landert et al., 2023)

Dalam analisis wacana, khususnya *Critical Discourse Analysis*, linguistik korpus juga memiliki peran strategis. Peneliti dapat menelusuri bagaimana media, institusi, atau aktor politik merepresentasikan kelompok tertentu melalui pilihan kata dan pola bahasa. Misalnya, kata apa yang sering muncul bersama istilah “migran,” “krisis,” atau “keamanan.” Pendekatan ini membuat analisis wacana menjadi lebih sistematis karena didukung data kuantitatif sekaligus interpretasi kualitatif (Baker, 2023).

Dalam pengajaran bahasa, linguistik korpus telah mengubah cara materi pembelajaran disusun. Buku ajar modern banyak menggunakan daftar kosakata frekuensi tinggi, pola kolokasi, dan contoh kalimat dari korpus. Pembelajar dapat mempelajari ungkapan yang benar-benar lazim digunakan penutur asli, bukan sekadar bentuk yang secara teori benar. Pendekatan ini dikenal sebagai *data-driven learning*, yaitu pembelajaran melalui eksplorasi data bahasa nyata (Lusta et al., 2023).

Selain itu, hubungan linguistik korpus dengan linguistik modern semakin kuat melalui perkembangan teknologi bahasa. Korpus menjadi fondasi bagi sistem penerjemahan mesin, pengenalan suara, pemeriksa tata bahasa, analisis sentimen, dan model bahasa berbasis Artificial Intelligence. Tanpa data korpus dalam jumlah besar, banyak inovasi teknologi bahasa saat ini tidak dapat berkembang secara optimal.

Linguistik korpus bukanlah cabang yang berdiri terpisah dari linguistik modern, melainkan pendekatan yang memperkuat hampir seluruh bidang kajian bahasa. Ia menyediakan jembatan antara teori dan data, antara penjelasan konseptual dan bukti empiris, serta antara linguistik tradisional dan kebutuhan analisis bahasa di era digital.

KELEBIHAN DAN KETERBATASAN PENDEKATAN KORPUS

Pendekatan korpus telah menjadi salah satu metode penting dalam penelitian bahasa modern karena menempatkan data nyata sebagai dasar analisis. Melalui kumpulan teks atau ujaran yang tersusun secara sistematis, peneliti dapat menelusuri pola bahasa secara lebih objektif dan terukur. Namun, seperti metode ilmiah lainnya, pendekatan korpus memiliki kelebihan sekaligus keterbatasan yang perlu dipahami agar penggunaannya tepat dan proporsional.

Salah satu kelebihan utama pendekatan korpus adalah penggunaan data autentik. Data yang dianalisis berasal dari penggunaan bahasa nyata, seperti percakapan, artikel berita, tulisan akademik, atau komunikasi digital. Hal ini memungkinkan peneliti melihat bagaimana bahasa benar-benar digunakan dalam masyarakat, bukan hanya berdasarkan intuisi atau contoh buatan. Sinclair (1991) menegaskan bahwa banyak pola bahasa hanya dapat ditemukan melalui pengamatan terhadap penggunaan aktual.

Kelebihan berikutnya adalah sifatnya yang objektif dan dapat diverifikasi. Hasil penelitian korpus umumnya lebih transparan karena sumber data, prosedur pencarian, dan hasil frekuensi dapat diperiksa ulang oleh peneliti lain. Dengan demikian, penelitian korpus memiliki tingkat reliabilitas yang lebih tinggi dibandingkan dengan analisis yang hanya bergantung pada intuisi pribadi.

Pendekatan korpus juga unggul dalam mengolah data besar. Dengan bantuan perangkat lunak seperti AntConc atau Sketch Engine, jutaan kata dapat dianalisis dalam waktu singkat. Peneliti dapat menelusuri frekuensi kata, kolokasi, pola gramatikal, serta perubahan bahasa lintas waktu. Kemampuan ini sangat relevan di era digital ketika data bahasa tersedia dalam jumlah besar.

Selain itu, pendekatan korpus mampu menemukan pola yang sulit dideteksi secara intuitif. Misalnya, korpus dapat menunjukkan bahwa frasa *make a decision* jauh lebih umum dibandingkan dengan *do a decision*, atau *strong coffee* lebih lazim daripada *powerful coffee*. Temuan seperti ini penting dalam leksikografi, penerjemahan, dan pengajaran bahasa.

Kelebihan lainnya adalah fleksibilitas metodologis. Pendekatan korpus dapat digunakan secara kuantitatif, misalnya menghitung frekuensi dan distribusi, maupun secara kualitatif, seperti menafsirkan konteks penggunaan kata atau strategi wacana. Oleh karena itu, metode ini dapat diterapkan dalam berbagai bidang, seperti sosiolinguistik, pragmatik, analisis wacana, pengajaran bahasa, hingga teknologi bahasa berbasis *Artificial Intelligence* (McEnery & Hardie, 2019).

Meskipun demikian, pendekatan korpus juga memiliki sejumlah keterbatasan. Pertama, kualitas hasil sangat bergantung pada kualitas korpus yang digunakan. Jika data tidak representatif, terlalu kecil, bias, atau tidak sesuai dengan tujuan penelitian, maka hasil analisis juga berpotensi lemah. Misalnya, meneliti bahasa Indonesia umum hanya dengan data dari media sosial tentu tidak cukup mewakili seluruh ragam bahasa Indonesia.

Keterbatasan kedua adalah bahwa korpus tidak selalu dapat menjelaskan alasan di balik suatu fenomena bahasa. Korpus sangat baik untuk menunjukkan apa yang terjadi dalam data, tetapi tidak selalu menjawab mengapa hal itu terjadi. Untuk memahami motivasi sosial, psikologis, atau kognitif, peneliti sering perlu menggunakan metode tambahan seperti wawancara, observasi, atau eksperimen.

Ketiga, tidak semua konteks sosial terekam dengan lengkap dalam data korpus. Pada beberapa sumber data digital, informasi mengenai usia penutur, latar pendidikan, hubungan sosial, atau situasi komunikasi sering kali tidak tersedia. Padahal, unsur-

unsur tersebut penting dalam penelitian pragmatik dan sosiolinguistik.

Keterbatasan lain adalah kebutuhan terhadap keterampilan teknis. Peneliti perlu memahami cara mengumpulkan data, membersihkan teks, menggunakan perangkat lunak, membaca tabel frekuensi, dan menafsirkan hasil pencarian. Tanpa pemahaman tersebut, data yang besar justru dapat menghasilkan kesimpulan yang keliru.

Selain itu, pendekatan korpus kadang menimbulkan risiko terlalu bergantung pada angka statistik. Frekuensi tinggi tidak selalu berarti paling penting secara linguistik, dan fenomena yang jarang muncul belum tentu tidak signifikan. Oleh karena itu, angka statistik perlu dibaca bersama konteks sosial dan makna penggunaan (Baker, 2010, 2023)

Dengan demikian, pendekatan korpus menawarkan banyak keunggulan: objektif, berbasis data nyata, mampu mengolah data besar, dan fleksibel untuk berbagai tujuan penelitian. Namun, penggunaannya tetap menuntut kehati-hatian dalam memilih data, memahami konteks, serta menafsirkan hasil. Pendekatan korpus akan paling efektif apabila digunakan secara kritis dan, bila perlu, dipadukan dengan metode penelitian lainnya.

KESIMPULAN

Linguistik korpus merupakan pendekatan modern dalam kajian bahasa yang menempatkan data autentik sebagai dasar utama analisis. Melalui korpus, peneliti dapat mengamati bagaimana bahasa benar-benar digunakan dalam berbagai konteks, sehingga hasil penelitian menjadi lebih objektif, terukur, dan relevan. Perkembangan linguistik korpus dari pengumpulan teks manual hingga pemanfaatan teknologi digital menunjukkan bahwa bidang ini terus beradaptasi dengan kebutuhan penelitian bahasa kontemporer.

Pembahasan dalam bab ini menegaskan bahwa konsep korpus tidak sekadar merujuk pada kumpulan teks, tetapi pada data bahasa yang disusun secara sistematis dan representatif untuk tujuan ilmiah. Kehadiran data otentik memungkinkan peneliti memahami pola kosakata, struktur gramatikal, variasi register, dan perubahan bahasa secara lebih akurat. Selain itu, linguistik korpus memiliki hubungan erat dengan berbagai cabang linguistik modern, seperti sosiolinguistik, pragmatik, analisis wacana, pengajaran bahasa, hingga pengembangan teknologi bahasa berbasis *Artificial Intelligence*.

Meskipun menawarkan banyak kelebihan, pendekatan korpus tetap memiliki keterbatasan, terutama terkait representativitas data, konteks sosial, dan kebutuhan keterampilan teknis. Oleh karena itu, penggunaan korpus memerlukan perencanaan metodologis yang matang dan interpretasi yang kritis. Secara keseluruhan, linguistik korpus memberikan kontribusi besar dalam memperkuat penelitian bahasa berbasis bukti serta membuka peluang baru bagi studi kebahasaan di era digital.

DAFTAR PUSTAKA

- Aijmer, K., & Rühlemann, C. (2015). *Corpus pragmatics*. Cambridge University Press.
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- Baker, P. (2023). *Using corpora in discourse analysis*. Bloomsbury Publishing.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics- Investigating language structure and use (2nd ed.)*. Cambridge University Press.
- Chomsky, N. (2014). *Aspects of the Theory of Syntax*. MIT Press.

- Dash, N. S. (2008). *CORPUS LINGUISTICS: AN INTRODUCTION*. Pearson Education India.
- Francis, W. N., & Kucera, H. (1964). A standard corpus of present-day edited American English, for use with digital computers. *Brown University, Providence*, 2.
- Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology*, 25(1), 4–19. <http://hdl.handle.net/10125/44747>
- Harris, Z. S. (1951). *Methods in structural linguistics*. University of Chicago Press.
- Landert, D., Dayter, D., Messerli, T. C., & Locher, M. A. (2023). *Corpus pragmatics*. Cambridge University Press.
- Lin, P., & Adolphs, S. (2023). Corpus linguistics. In *The Routledge Handbook of Applied Linguistics* (pp. 296–308). Routledge.
- Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data-driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, 9(12).
- McEnery, T., & Hardie, A. (2019). *Corpus linguistics: Method, theory and practice (2nd ed.)*. Cambridge University Press.
- Meyer, C. F. (2023). *English corpus linguistics: An introduction*. Cambridge University Press.
- Pérez-Paredes, P., & Curry, N. (2024). Epistemologies of corpus linguistics across disciplines. *Research Methods in Applied Linguistics*, 3(3), 100141.
- Sevilla-Requena, L. (2025). *Determining and Evaluating the Quality of Corpora in the LLM Era*. <https://github.com/soskek/bookcorpus?tab=readme-ov-file>
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Stepanov, V. V. (2022). CORPUS VS NON-CORPUS: MAIN DISTINCTIVE FEATURES. “*Scientific Notes of V. I. Vernadsky Taurida National University*”, Series:

PROFIL PENULIS



Sardian Maharani Asnur

Penulis adalah akademisi, peneliti, dan pengajar di bidang bahasa Inggris, linguistik, serta pendidikan tinggi. Ia memiliki pengalaman panjang dalam pengajaran pada tingkat universitas, khususnya dalam bidang linguistik, penerjemahan, keterampilan berbahasa, dan metodologi penelitian. Fokus keilmuannya mencakup linguistik korpus, psikolinguistik, pengajaran bahasa Inggris sebagai bahasa asing (EFL), serta pengembangan kurikulum berbasis capaian pembelajaran. Sebagai peneliti, Sardian aktif menulis karya ilmiah, artikel akademik, modul pembelajaran, dan book chapter yang berkaitan dengan kajian bahasa dan pendidikan. Minat penelitiannya banyak berorientasi pada pemanfaatan data autentik dalam penelitian bahasa, integrasi teknologi dalam pembelajaran, serta pengembangan model pembelajaran inovatif di perguruan tinggi. Ia juga terlibat dalam berbagai kegiatan akademik seperti penyusunan kurikulum, penjaminan mutu pendidikan, pendampingan penelitian mahasiswa, dan pengembangan bahan ajar. Sebagai pendidik, ia meyakini bahwa bahasa bukan sekadar alat komunikasi, tetapi juga sarana berpikir kritis, pengembangan budaya, dan transformasi sosial. Komitmen tersebut tercermin dalam kontribusinya pada pengembangan literasi akademik, riset kebahasaan, dan peningkatan kualitas pendidikan di Indonesia.

BAB 2

JENIS DAN KARAKTERISTIK KORPUS

Nurhayati S.
Universitas Bina Sarana Informatika, Jakarta
E-mail: nurhayati.nhi@bsi.ac.id

PENDAHULUAN

Dalam dua dekade terakhir, linguistik korpus telah bertransformasi secara radikal dari sekadar metode pengumpulan data menjadi infrastruktur fundamental bagi era kecerdasan buatan (AI) dan *Machine Learning*. Korpus kini dipandang sebagai representasi bahasa yang sistematis dan terstruktur dalam konteks sosialnya, di mana, menurut Flowerdew (2023), kekuatannya terletak pada kemampuan mengungkap pola linguistik yang tidak tertangkap oleh intuisi manusia, seperti frekuensi kolokasi halus dan pergeseran makna melalui observasi empiris pada skala besar. Dalam lanskap teknologi terbaru, Zhu (2023) menekankan bahwa kualitas penalaran AI sangat bergantung pada keragaman data pelatihan, sehingga pemahaman tipologi korpus menjadi kompetensi esensial bagi praktisi teknologi untuk menghindari bias algoritma akibat data yang tidak representatif.

Peran korpus kini meluas menjadi tulang punggung pengembangan *Large Language Models* (LLM), di mana McEnery & Hardie (2021) menegaskan bahwa fungsionalitasnya tidak lagi terbatas pada studi deskriptif, tetapi juga sebagai penyedia data "kebenaran dasar" (*ground truth*) bagi algoritma pemrosesan bahasa alami. Oleh karena itu, validitas dan reliabilitas penelitian sangat bergantung pada pemahaman karakteristik data, apakah bersifat lisan, tulisan, teknis, atau umum, guna menghindari generalisasi yang keliru.

Biber & Egbert (2024) mengingatkan bahwa kesalahan dalam pemilihan jenis korpus, seperti menggunakan data sastra untuk merepresentasikan percakapan sehari-hari, dapat menyebabkan bias signifikan dalam interpretasi hasil analisis.

Dinamika bahasa di ruang digital saat ini menuntut adanya standarisasi arsitektur korpus yang ketat agar data dari berbagai sumber tetap terorganisir dan dapat ditelusuri. Gray, Biber, & Reppen (2024) mencatat bahwa tantangan utama peneliti modern bukan lagi kelangkaan data, melainkan kemampuan menavigasi variasi register dan genre dalam korpus berskala besar. Sebagai upaya merespons urgensi tersebut, bab ini akan membedah secara komprehensif arsitektur korpus berdasarkan standar global, mulai dari dikotomi korpus umum dan spesifik hingga dimensi waktu sinkronis dan diakronis, untuk memberikan landasan teoretis dan praktis yang kuat bagi peneliti dalam memilih atau membangun korpus yang tepat guna.

KORPUS UMUM DAN KORPUS SPESIFIK

Korpus Umum (*General/Reference Corpus*)

Korpus umum, yang sering dikenal sebagai korpus referensi, merupakan kumpulan teks yang disusun secara sistematis untuk merepresentasikan penggunaan bahasa pada skala nasional maupun global dalam kurun waktu tertentu. Fungsi utamanya adalah menjadi standar pembandingan (*benchmark*) bagi para peneliti untuk menguji apakah suatu fenomena bahasa bersifat universal atau terbatas pada kelompok penutur tertentu. Korpus ini memiliki karakteristik profil yang masif dengan keragaman genre yang luas, mulai dari transkripsi lisan hingga dokumen akademik. Seiring perkembangan zaman, Gray, Biber, & Reppen (2024) mencatat adanya adaptasi radikal melalui penyertaan "bahasa digital" (*E-language*) dari blog dan

media sosial guna menjaga relevansi data terhadap pergeseran komunikasi termediiasi komputer kontemporer.

Dalam penyusunannya, korpus umum sangat mengedepankan prinsip keterwakilan (*representativeness*) melalui penyeimbangan proporsi genre untuk mencegah dominasi wacana tertentu. Saat ini, korpus referensi modern telah bertransformasi hingga mencapai skala miliaran token, seperti *Corpus of Contemporary American English* (COCA) yang memungkinkan analisis perubahan linguistik secara *real-time* serta studi diakronis yang sangat mendetail. Volume data yang sangat besar ini bukan hanya sekadar mengejar kuantitas, melainkan kebutuhan teknis untuk menangkap penggunaan kata-kata jarang (*low-frequency words*) serta pola kolokasi spesifik yang sering kali luput dari pemetaan dalam korpus berskala kecil.

Di konteks lokal, pengembangan Korpus Indonesia (KOIN) menjadi langkah strategis untuk memetakan variasi bahasa Indonesia formal dan semiformal di tengah pengaruh bahasa gaul serta serapan asing. Kurniawan (2024) menekankan bahwa integrasi data digital ke dalam korpus nasional Indonesia sangat krusial demi memahami dinamika linguistik generasi saat ini. Kehadiran infrastruktur data yang seimbang ini pada akhirnya menjadi fondasi vital bagi kemajuan teknologi *Natural Language Processing* (NLP) dan *Large Language Models* (LLM), karena tanpa referensi yang komprehensif, model AI berisiko mengalami bias dan kegagalan dalam menangkap nuansa konteks bahasa manusia yang sebenarnya

Tabel 2.1. Profil Korpus Umum Global
(Data Mutakhir 2024-2026)

Nama Korpus	Ukuran (Token/Data)	Komposisi Utama	Peran dalam Linguistik	
BNC 2014	~100 Juta	Lisan, Tulisan, <i>E-language</i>	Standar Inggris Modern	Bahasa Britania
COCA	1 Miliar+	TV, Film, Blog, Jurnal, Berita	Analisis register diakronis	variasi dan
C4 Corpus	750 GB+	<i>Web Crawling (Cleaned)</i>	Dasar model bahasa besar (LLM)	pelatihan
KOIN (Indo)	100 Juta+	Berita, Sastra, Web	Referensi Bahasa Formal	utama Indonesia

Sumber: Diolah dari (McEnery & Hardie, 2021), (Gray, Biber, & Reppen, 2024), dan metadata resmi penyedia korpus.

Korpus Spesifik (*Specialized Corpus*)

Berbeda dengan korpus umum yang mengejar cakupan luas, korpus spesifik (*specialized corpus*) berfokus secara mendalam pada domain, genre, atau kelompok sosial tertentu untuk menghasilkan representasi subbahasa yang akurat. Nesi (2021) menegaskan bahwa dalam konteks *English for Specific Purposes* (ESP), instrumen ini jauh lebih efektif karena mampu menangkap leksikon, konvensi retorik, dan pola gramatikal unik yang sering kali terabaikan dalam korpus umum yang heterogen. Kekuatan utamanya terletak pada presisi terminologis dan kerapatan informasi; studi Chen & Wang (2025) membuktikan bahwa penggunaan korpus hukum spesifik mampu meningkatkan akurasi terjemahan teknis hingga 40%. Contoh nyata terlihat pada kata "*discharge*" yang dalam korpus medis memiliki makna teknis krusial terkait pemulangan pasien, berbeda jauh dengan kolokasi umumnya di korpus besar seperti

COCA. Oleh karena itu, Flowerdew (2023) mewajibkan peneliti wacana profesional, seperti bidang hukum, untuk menggunakan dokumen otentik seperti putusan pengadilan guna membedah struktur argumen yang spesifik secara empiris.

Selain itu, O'Keeffe & McCarthy (2022) menyatakan bahwa korpus spesifik memungkinkan identifikasi "*niche*" linguistik yang mendetail, seperti pola penggunaan kata ganti untuk membangun empati dalam interaksi medis atau penggunaan kalimat pasif dalam laporan ilmiah. Meskipun ukurannya lebih ringkas, tingkat kurasi yang tinggi meminimalisir "kebisingan" data sehingga setiap teks yang masuk tetap relevan dengan kriteria desain yang ditetapkan. Dalam lanskap penelitian tahun 2025-2026, korpus ini menjadi tulang punggung pengembangan *Domain-specific AI*, seperti asisten virtual hukum atau penerjemah teknik, yang sangat bergantung pada akurasi data pelatihan. Vaughan & Clancy (2023) menyimpulkan bahwa detail mikroskopis yang disediakan korpus spesifik tidak hanya krusial bagi analisis profesional, tetapi juga menjadi instrumen vital dalam merancang kurikulum pendidikan berbasis kompetensi yang selaras dengan tuntutan dunia kerja nyata.

KORPUS LISAN DAN KORPUS TULISAN

Perbedaan moda antara bahasa lisan dan tulisan menciptakan karakteristik linguistik yang kontras, sehingga metodologi pengumpulan dan analisisnya memerlukan pendekatan yang berbeda secara fundamental.

Korpus Lisan (*Spoken Corpus*)

Bahasa lisan merupakan instrumen komunikasi manusia yang paling primer, namun secara teknis sangat menantang untuk dikelola dalam bentuk korpus karena sifatnya yang efemer dan sarat fitur kontekstual. Vaughan & Clancy (2023) menyoroti bahwa paradigma korpus lisan modern kini telah bergeser ke

arah multimodalitas, di mana teks transkripsi mentah dianggap tidak lagi memadai tanpa adanya sinkronisasi dengan rekaman audio, video, serta anotasi prosodi yang mendetail. Meskipun teknologi *Automatic Speech Recognition* (ASR) telah mempercepat proses digitalisasi bunyi, O'Keeffe & McCarthy (2022) menegaskan bahwa intervensi manusia tetap esensial untuk menangani fenomena *disfluency* seperti gagap, pengulangan, dan tumpang tindih bicara. Tanpa verifikasi manual yang teliti, fitur-fitur linguistik kritis tersebut berisiko terdistorsi, sehingga menghilangkan nuansa interaksional asli yang sangat penting dalam analisis pragmatik.

Secara empiris, teks lisan menunjukkan karakteristik unik berupa kepadatan leksikal yang rendah, namun memiliki frekuensi kata ganti orang serta verba dinamis yang tinggi jika dibandingkan dengan teks tulis. McEnery & Hardie (2021) memberikan bukti melalui *British National Corpus* (BNC) 2014, di mana penanda pragmatik seperti "you know" atau "I mean" muncul secara dominan sebagai alat manajemen interaksi yang jarang ditemukan dalam ragam tulisan formal. Selain kendala teknis, pembangunan korpus lisan juga dihadapkan pada kompleksitas etika dan privasi yang ketat. Breidina & Egbert (2022) mengingatkan bahwa penggunaan data biometrik berupa suara dan wajah memerlukan izin eksplisit (*informed consent*) guna mematuhi regulasi perlindungan data modern seperti GDPR. Oleh karena itu, penciptaan korpus lisan yang berkualitas tinggi menuntut investasi waktu, biaya, dan ketelitian metodologis yang jauh lebih besar daripada pengembangan korpus tulisan.

Korpus Tulisan dan Era *Computer-Mediated Communication* (CMC)

Evolusi korpus tulisan saat ini telah bertransformasi secara masif ke ranah digital melalui *Computer-Mediated*

Communication (CMC), yang mulai menggeser dominasi teks konvensional seperti buku atau dokumen resmi. Vaughan & Clancy (2023) menyebut fenomena ini sebagai kelahiran genre "bahasa lisan yang tertulis", di mana batasan antara ragam formal dan gaya lisan menjadi kabur akibat pesan spontan yang terekam secara permanen di media sosial. Karakteristik linguistiknya yang menonjolkan elemen nonalfabetis seperti emoji, singkatan, serta sintaksis longgar mencerminkan transformasi bahasa yang bergerak jauh lebih pesat dibandingkan dengan literatur cetak (McCulloch, 2020). Kondisi ini menurut McEnery & Hardie (2021) memicu kemunculan neologisme dan percampuran kode yang masif sebagai bentuk adaptasi terhadap teknologi, sekaligus menuntut peneliti untuk menguasai teknik pra-pemrosesan data "kotor" (*noisy data*) guna menjamin akurasi analisis (Braidina & Egbert, 2022).

Di samping tantangan teknis, pengembangan korpus berbasis CMC juga menghadapi kompleksitas etika dan privasi yang jauh lebih tinggi daripada korpus tradisional. Dalam perdebatan ilmiah periode 2025-2026 mengenai hak cipta dan data pengguna, Biber & Egbert (2024) menekankan krusialnya anonimisasi data otomatis untuk melindungi privasi individu tanpa menghilangkan konteks sosiolinguistik teks yang dianalisis. Upaya proteksi ini menjadi sangat penting karena korpus tulisan di era digital bukan sekadar kumpulan teks internet biasa, melainkan laboratorium linguistik dinamis yang merekam jejak evolusi bahasa manusia secara *real-time*. Dengan demikian, integrasi antara pemrosesan data yang canggih dan kepatuhan terhadap standar etika menjadi kunci utama dalam memanfaatkan korpus CMC sebagai cermin peradaban bahasa modern.

KORPUS SINKRONIS DAN DIAKRONIS

Korpus Sinkronis

Korpus sinkronis dikembangkan untuk memotret realitas bahasa pada satu titik waktu tertentu tanpa melibatkan tinjauan historis, sehingga memungkinkan analisis horizontal terhadap variasi register, dialek, dan gaya bahasa yang muncul secara bersamaan. Instrumen ini memegang peranan krusial dalam leksikografi modern untuk memastikan kamus dan tata bahasa tetap relevan dengan menangkap neologisme serta pergeseran makna kontemporer (McEnery & Hardie, 2021). Melampaui deskripsi struktural, Gries (2023) menekankan efektivitas korpus ini dalam sosiolinguistik komputasional, seperti analisis sentimen dan opini publik, di mana data dari rentang waktu singkat mampu merefleksikan suasana sosial-politik serta diskursus dominan suatu bangsa. Hal ini memungkinkan peneliti mendeteksi tren budaya secara instan melalui frekuensi kata kunci sebagai indikator perilaku kolektif masyarakat terhadap isu nasional tertentu.

Selain itu, korpus sinkronis menjadi fondasi vital dalam pengembangan kecerdasan buatan, di mana Smith (2026) melalui *Linguistic Tech Review* mencatat bahwa model bahasa yang dilatih dengan data terkini (2024-2026) jauh lebih adaptif dalam mengenali istilah teknologi baru dibandingkan dengan model berbasis data usang. Akurasi dalam memahami isu kesehatan global atau kemajuan teknologi ini sangat bergantung pada kemampuan korpus menangkap dinamika bahasa yang frekuensinya dapat melonjak drastis akibat peristiwa global (Biber & Egbert, 2024). Lebih jauh lagi, Gray, Biber, dan Reppen (2024) menunjukkan bahwa korpus sinkronis mampu mengidentifikasi perbedaan gaya komunikasi lintas generasi, seperti antara Gen Z dan milenial, dengan tingkat presisi tinggi. Dengan demikian, korpus ini berfungsi sebagai cermin digital

yang merepresentasikan identitas linguistik dan psikologi sosial masyarakat pada momen spesifik dalam sejarah.

KORPUS DIAKRONIS

Berbeda dengan korpus sinkronis yang menangkap momen statis, korpus diakronis atau historis diibaratkan sebagai rekaman evolusi bahasa secara vertikal yang memungkinkan observasi terhadap perubahan struktur gramatikal, fonologi, hingga leksikon dari berbagai periode waktu. Hilpert (2021) menegaskan bahwa keunggulan utama korpus ini adalah kemampuannya memvalidasi teori perubahan bahasa secara empiris menggunakan data frekuensi objektif guna menggantikan spekulasi teoretis. Dalam ekosistem linguistik komputasional modern, penerapan algoritma distribusi semantik dan pemodelan ruang vektor telah meningkatkan analisis ini ke level baru, sehingga peneliti dapat melacak pergeseran makna kata secara radikal melalui analisis konteks kata di sekitarnya yang dipicu oleh inovasi teknologi, transformasi sosial, maupun peristiwa global masif.

Tabel 2.2. Contoh Pergeseran Makna Berdasarkan Studi Diakronis

Kata	Makna Dahulu (Pre-2000)	Makna Sekarang (Post-2020)	Konteks Perubahan
Awan	Objek meteorologi/kumpulan uap air	Sistem penyimpanan data digital (<i>Cloud</i>)	Revolusi Teknologi Informasi
Daring	Istilah teknis yang jarang digunakan	Metode utama dalam pendidikan dan pekerjaan	Dampak Pandemi COVID-19
Viral	Fenomena medis berkaitan dengan virus	Penyebaran informasi secara masif dan cepat	Ledakan Budaya Media Sosial

Sumber: Analisis diakronis berbasis korpus (Hilpert, 2021); diadaptasi untuk konteks kontemporer).

Data faktual dari berbagai korpus historis menunjukkan bahwa evolusi bahasa senantiasa mengikuti pola sistematis, termasuk fenomena "gramatikalisasi" di mana unit leksikal bermakna konkret bertransformasi menjadi elemen gramatikal yang lebih abstrak (McEnery & Hardie, 2021). Gries (2023) menambahkan bahwa analisis diakronis kini merambah disiplin *Culturomics*, yang memanfaatkan frekuensi kata kunci dalam jutaan buku digital sebagai proksi untuk memahami pergeseran minat budaya dan norma sosiopsikologis peradaban lintas abad. Melalui pelacakan penurunan frekuensi kata, korpus ini memainkan peran vital dalam membedah "ekologi bahasa" guna mengidentifikasi istilah yang mulai usang beserta pemicu sosiolinguistiknya. Dalam lanskap akademik 2025-2026, Biber & Egbert (2024) menegaskan bahwa integrasi korpus diakronis dengan teknik *deep learning* telah memungkinkan ilmuwan memprediksi arah perubahan bahasa di masa depan secara akurat, sehingga memposisikan korpus diakronis bukan sekadar arsip masa lalu, melainkan instrumen prediktif yang menghubungkan sejarah manusia dengan evolusi komunikasinya.

KORPUS BERBAHASA TUNG GAL DAN MULTIBAHASA

Berdasarkan jumlah bahasa yang dikandungnya, korpus dikategorikan untuk melayani tujuan penelitian yang berbeda, mulai dari analisis struktur internal bahasa hingga studi komparatif lintas budaya.

Korpus Monolingual

Korpus monolingual berfungsi sebagai instrumen fundamental dalam penelitian linguistik untuk membedah struktur internal satu bahasa secara mendalam, mulai dari fonologi hingga semantik. Pemanfaatan data yang luas dan

representatif memungkinkan peneliti mengidentifikasi pola penggunaan bahasa alami melalui analisis kolokasi dan konkordansi yang melampaui aturan tata bahasa tradisional (McEnery & Hardie, 2021). Dalam perkembangannya, Sinclair & Carter (2022) mencatat transformasi korpus ini menjadi alat utama analisis wacana kritis untuk mendeteksi bias sistemik, seperti stereotip gender atau diskriminasi tersembunyi, yang sering kali hanya terlihat jelas melalui validasi statistik pada skala masif. Selain itu, korpus ini menjadi bahan baku esensial dalam pengembangan *Large Language Models* (LLM) seperti GPT-4, di mana kurasi data yang bersih dan ketat kini lebih diutamakan untuk mencegah replikasi prasangka sosial selama fase pra-pelatihan (Gray, Biber, & Reppen, 2024).

Lebih lanjut, Flowerdew (2023) menegaskan peran vital korpus monolingual dalam mendokumentasikan bahasa yang kurang terwakili atau dialek yang terancam punah, sehingga berfungsi sebagai instrumen sosiokultural untuk melestarikan warisan linguistik sekaligus membedah ideologi di balik teks. Dalam ranah sosiolinguistik, korpus ini memainkan peran strategis untuk melacak pergeseran norma kesantunan dan gaya bahasa lintas generasi secara *real-time*. Vaughan & Clancy (2023) menjelaskan bahwa melalui analisis statistik data media sosial, peneliti dapat mendeteksi kemunculan slang baru maupun perubahan fungsi gramatikal secara instan. Dengan kemampuannya merekam memori kolektif dan mencerminkan perilaku komunikatif komunitas penutur, korpus monolingual tetap menjadi elemen paling vital dalam linguistik modern untuk memahami evolusi bahasa dalam ekosistem sosialnya.

Korpus Multibahasa

Evolusi korpus multibahasa di era globalisasi dan digitalisasi telah melahirkan dua kategori instrumen penelitian utama, yaitu korpus paralel dan korpus komparabel, yang

melampaui batasan analisis bahasa tunggal. Korpus paralel menyandingkan teks asli dan terjemahannya secara simetris, sementara korpus komparabel menghimpun teks orisinal dengan kemiripan topik dan genre tanpa hubungan terjemahan langsung. Looock (2023) menjelaskan bahwa perbedaan ini mengarahkan korpus paralel pada studi ekuivalensi makna dan korpus komparabel pada analisis gaya bahasa alami. Pemanfaatan korpus paralel berskala masif menjadi sangat krusial bagi kemajuan teknologi *Neural Machine Translation* (NMT) untuk mengenali pola antarbahasa sekaligus mendeteksi *translationese* atau kekakuan hasil terjemahan (McEnery & Hardie, 2021). Di sisi lain, O'Keeffe & McCarthy (2022) menegaskan bahwa korpus komparabel adalah alat terbaik dalam studi lintas budaya karena menyediakan data orisinal penutur asli yang mencerminkan norma komunikatif autentik. Meski pembangunannya menghadapi tantangan penyelarasan (*alignment*) yang rumit, Biber & Egbert (2024) mencatat bahwa tren tahun 2026 yang memanfaatkan *Cross-lingual Word Embeddings* telah memungkinkan penyelarasan otomatis pada tingkat semantik secara lebih cepat dan akurat.

Selain aspek teknis, korpus multibahasa memiliki implikasi sosiolinguistik yang luas, terutama dalam mendukung revitalisasi bahasa daerah yang terancam punah melalui penyediaan sumber daya digital dwibahasa. Vaughan & Clancy (2023) juga menyoroti peran penting korpus ini dalam ranah forensik untuk mendeteksi plagiarisme internasional maupun penipuan lintas bahasa. Sebagai laboratorium global, keberadaan korpus ini memfasilitasi pemahaman mendalam mengenai universalitas dan keunikan cara manusia mengonstruksi realitas melalui lensa bahasa yang beragam. Flowerdew (2023) menyimpulkan bahwa didukung oleh kemajuan komputasi canggih, korpus multibahasa akan terus berfungsi sebagai

infrastruktur data yang menjadi fondasi utama bagi inovasi linguistik dan perkembangan teknologi bahasa di masa depan.

KORPUS PARALEL DAN KORPUS KOMPARABEL

Dalam studi penerjemahan dan linguistik kontrastif, korpus multibahasa menjadi instrumen empiris utama yang terbagi ke dalam dua arsitektur fungsional: korpus paralel dan korpus komparabel. Menurut Looock (2023), pemilihan di antara keduanya sangat bergantung pada tujuan penelitian, apakah ingin mendalami proses ekuivalensi antara teks sumber dan sasaran melalui korpus paralel, atau mengeksplorasi penggunaan bahasa alami yang diproduksi secara orisinal dalam dua budaya berbeda melalui korpus komparabel. Korpus paralel sendiri memerlukan proses penyelarasan kalimat (*sentence alignment*) yang akurat untuk memetakan korespondensi antarbahasa. McEnery & Hardie (2021) mencatat bahwa data ini menjadi "bahan bakar" vital bagi teknologi *Neural Machine Translation* (NMT) seperti Google Translate, dengan *Europarl Corpus* sebagai salah satu standar global yang digunakan untuk melatih kecerdasan buatan dalam mempelajari probabilitas pergeseran sintaksis dan semantik.

Namun, penggunaan korpus paralel sering kali dihadapkan pada tantangan *translationese*, di mana bahasa terjemahan cenderung kaku akibat interferensi struktur bahasa sumber. Untuk mengatasi hal ini, korpus komparabel memainkan peran penting dengan menghimpun teks-teks dalam bahasa berbeda yang memiliki kriteria desain identik, seperti topik dan genre, namun tidak terkait melalui proses penerjemahan. Looock (2023) berargumen bahwa korpus komparabel jauh lebih efektif dalam menangkap gaya bahasa alami karena teks tersebut diproduksi secara orisinal oleh penutur asli untuk komunikasi internal budayanya. Secara metodologis, O'Keeffe & McCarthy (2022) menjelaskan bahwa korpus komparabel berfungsi sebagai

kelompok kontrol untuk mengevaluasi kewajaran leksikal dan gramatikal hasil terjemahan, sehingga mampu menyediakan data kolokasi yang lebih otentik dibandingkan dengan teks terjemahan yang mungkin terdistorsi.

Menatap masa depan, tantangan teknis pengembangan korpus terletak pada aspek skalabilitas dan diversitas genre guna menjangkau domain digital yang lebih cair seperti ulasan produk atau blog. Biber & Egbert (2024) menyoroti tren integrasi dalam *Deep Learning*, di mana korpus komparabel digunakan untuk meningkatkan pemahaman konteks pada model yang awalnya hanya berbasis data paralel. Sinergi antara keduanya memberikan cakrawala baru; korpus paralel memetakan transfer makna antarbahasa, sementara korpus komparabel menyediakan tolok ukur bagaimana bahasa beroperasi secara mandiri dalam ekosistem aslinya. Bagi penulis dan penerjemah profesional, Flowerdew (2023) menegaskan bahwa pemanfaatan kedua jenis korpus ini adalah kunci utama untuk menghasilkan karya yang tidak hanya akurat secara semantik, tetapi juga memiliki resonansi stilistik yang tepat bagi pembaca sasaran.

KRITERIA KUALITAS DAN KEREPRESENTATIFAN KORPUS

Sebuah korpus dianggap berkualitas jika memenuhi standar metodologis yang ketat, bukan hanya karena ukurannya yang besar.

Kerepresentatifan (*Representativeness*)

Dalam metodologi linguistik korpus modern, integritas sebuah korpus tidak lagi diukur hanya melalui kuantitas data, melainkan melalui ketajaman metodologis untuk menjamin validitas ilmiah. Biber & Egbert (2024) menggarisbawahi bahwa kualitas utama bersumber dari kerepresentatifan, di mana sampel teks harus mampu merefleksikan variabilitas linguistik

dari populasi target secara akurat. Gray, Biber, & Reppen (2024) menambahkan bahwa hal ini memerlukan *sampling frame* yang komprehensif untuk mencakup seluruh spektrum genre guna menghindari bias wacana. Implementasi nyata dari prinsip ini terlihat pada *British National Corpus* (BNC) 2014 yang menerapkan stratifikasi sosial secara ketat (McEnery & Hardie, 2021), sementara Flowerdew (2023) menekankan bahwa penggunaan anotasi linguistik seperti *POS tagging* dan lematisasi menjadi faktor pengali nilai korpus yang memungkinkan analisis pola gramatikal yang lebih mendalam.

Memasuki lanskap penelitian tahun 2025-2026, tantangan baru muncul dalam bentuk pembersihan data (*data cleaning*) untuk membedakan tuturan alami manusia dari konten yang dihasilkan oleh kecerdasan buatan (*AI-generated content*). Gries (2023) menegaskan pentingnya normalisasi dan deduplikasi yang disiplin untuk memurnikan data linguistik dari kebisingan digital (*digital noise*) yang berpotensi merusak temuan. Di sisi lain, dimensi etika kini menempati posisi sentral dalam penentuan kualitas korpus profesional, di mana perlindungan data pribadi dan perolehan izin eksplisit (*informed consent*) menjadi syarat mutlak yang tidak dapat ditawar (Vaughan & Clancy, 2023). Dengan mengintegrasikan keunggulan teknis dan kepatuhan etis tersebut, korpus tidak hanya berfungsi sebagai kumpulan teks, tetapi berevolusi menjadi instrumen riset yang kredibel dan dapat dipertanggungjawabkan secara akademik.

ANOTASI DAN METADATA

Kualitas sebuah korpus sangat ditentukan oleh ketersediaan metadata yang menyajikan informasi kontekstual komprehensif, mencakup identitas penutur atau penulis, waktu, lokasi, hingga tujuan komunikatif teks tersebut. Metadata yang kaya menjadi prasyarat utama bagi peneliti untuk melakukan analisis sosiolinguistik mendalam, karena memungkinkan adanya

korelasi antara pola penggunaan bahasa dengan berbagai variabel eksternal penutur. Di samping metadata, anotasi linguistik berperan dalam meningkatkan nilai fungsional korpus secara eksponensial melalui proses pelabelan teknis yang sistematis. Bentuk anotasi utama, seperti *Part-of-Speech (POS) Tagging* untuk klasifikasi kelas kata dan lematisasi untuk mengembalikan kata ke bentuk dasarnya, menjadi fondasi teknis yang memperkuat struktur data korpus (McEnery & Hardie, 2021)

Flowerdew (2023) mencatat bahwa korpus dengan anotasi yang baik memberikan fleksibilitas bagi peneliti untuk mengeksplorasi pola gramatikal yang kompleks, seperti mengidentifikasi struktur kata kerja tertentu yang diikuti oleh objek ganda secara akurat. Dengan demikian, integrasi antara metadata yang detail dan anotasi linguistik yang mendalam bukan sekadar aspek teknis tambahan, melainkan elemen esensial yang memungkinkan korpus menjadi instrumen penelitian yang jauh lebih berdaya guna dalam memetakan perilaku bahasa secara empiris.

KESIMPULAN

Linguistik korpus telah berevolusi dari sekadar metode pengumpulan data menjadi infrastruktur fundamental bagi era Kecerdasan Buatan (AI) dan pengembangan *Large Language Models* (LLM). Tipologi korpus saat ini mencakup dikotomi antara korpus umum yang mengejar keterwakilan bahasa secara luas dan korpus spesifik yang berfokus pada domain tertentu guna meningkatkan akurasi teknis. Selain itu, perbedaan moda antara korpus lisan yang kini mengarah pada multimodalitas dan korpus tulisan yang bertransformasi melalui *Computer-Mediated Communication* (CMC) menuntut pendekatan metodologi serta pertimbangan etika privasi yang lebih ketat. Dimensi waktu juga membagi korpus menjadi tipe sinkronis untuk memotret realitas

bahasa kontemporer dan diakronis untuk melacak evolusi sistematis bahasa serta pergeseran makna lintas zaman.

Integritas sebuah korpus di era modern tidak hanya ditentukan oleh volume datanya, tetapi juga oleh pemenuhan kriteria kualitas yang mencakup kerepresentatifan, keseimbangan genre, serta keandalan anotasi dan metadata. Penggunaan anotasi linguistik seperti *POS tagging* dan lematisasi meningkatkan nilai korpus secara eksponensial dengan memungkinkan analisis pola gramatikal yang kompleks. Menghadapi tantangan tahun 2025-2026, peneliti dituntut untuk mampu melakukan pembersihan data yang disiplin guna memisahkan bahasa alami manusia dari konten buatan mesin (*AI-generated content*) serta mematuhi standar etika internasional. Melalui pemahaman arsitektur korpus yang komprehensif, korpus berfungsi sebagai laboratorium linguistik dinamis dan instrumen prediktif yang menghubungkan sejarah komunikasi manusia dengan inovasi teknologi masa depan.

DAFTAR PUSTAKA

- Biber, D., & Egbert, J. (2024). *Register Variation Online*. Cambridge Online University.
- Breidina, T., & Egbert, J. (2022). *Methodological Triangulation in Corpus Linguistics*. New York: Routledge.
- Chen, L., & Wang, J. (2025). Precision in Legal Translation: A Specialized Corpus Approach. *Journal of Computational Linguistics*.
- Flowerdew, L. (2023). *The Routledge Handbook of Corpus Linguistics (2nd ed.)*. Routledge.
- Gray, B., Biber, D., & Reppen, R. (2024). *Corpus Linguistics for English Teachers: New Tools, New Insights*. Routledge.
- Gries, S. (2023). *Statistics for Linguistics with R: A Practical Introduction (3rd ed.)*. Amsterdam: De Gruyter Mouton.

- Hilpert, M. (2021). *Construction Grammar and Its Applications to English*. Edinburgh: Edinburgh University Press.
- Kurniawan, A. (2024). Mapping E-Language in Indonesia: KOIN Development Report. *Jurnal Linguistik Komputasional Indonesia*.
- Loock, R. (2023). *Translation Quality Assessment: A Corpus-Based Approach*. Cham: Springer Nature.
- McCulloch, G. (2020). *Because Internet: Understanding the New Rules of Language*. Riverhead Books.
- McEnery, T., & Hardie, A. (2021). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- Nesi, H. (2021). *The Handbook of English for Specific Purpose*. Hoboken : NJ: Wiley Blackwell.
- O'Keeffe, A., & McCarthy, M. (2022). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
- Sinclair, J., & Carter, R. (2022). *Trust the text: Language, Corpus and Discourse*. London: Routledge.
- Smith, R. (2026). Synchronic Data in AI Training. *LinguisticTech Review*.
- Vaughan, E., & Clancy, B. (2023). *Data Collection in Spoken Corpus Linguistics*. London: Bloomsbury Academic.
- Zhu, Y. d. (2023). *Foundation Models: A Survey on Large Language Models Training Data*. ArXiv Pre-print.

PROFIL PENULIS



Nurhayati S.

Pengarang adalah seorang dosen pada program studi Sastra Inggris pada Universitas Bina Sarana Informatika. Penelitiannya terutama berfokus pada linguistik terapan dalam bidang pendidikan, analisis wacana, sociolinguistik, dan sintaksis. Bungan rampai ini berisi penjelasan tentang jenis dan karakteristik korpus yang merupakan prasyarat mutlak bagi peneliti bahasa di era digital. Integrasi antara kedalaman teoretis dan kemajuan komputasi telah membuka peluang bagi linguistik korpus untuk menjawab tantangan sosiokultural dan teknologi. Dengan menjaga kriteria kualitas dan kerepresentatifan, korpus akan terus menjadi aset intelektual yang tak ternilai bagi perkembangan sains di masa depan.

BAB 3

TEKNIK PENGUMPULAN DAN PENYUSUNAN KORPUS

Bejo Sutrisno
Universitas Al-Azhar Indonesia, Jakarta
E-mail: bejo.sutrisno@uai.ac.id

PENDAHULUAN

Penelitian linguistik modern semakin bertransformasi berkat kemajuan teknologi dan metode kuantitatif. Salah satu terobosan paling signifikan dalam dekade terakhir adalah munculnya dan berkembangnya Linguistik Korpus (Al-Qurishi et al., 2021; Hadi et al., 2023; Hutson et al., 2022). Linguistik korpus mendefinisikan bahasa sebagai kumpulan besar teks tertulis atau ucapan yang dikumpulkan secara sistematis dan terorganisir, yang kemudian dianalisis menggunakan alat komputasi untuk mengidentifikasi pola, frekuensi, dan distribusi fitur linguistik. Inti dari setiap penelitian linguistik korpus adalah bahwa korpus itu sendiri adalah sebuah dataset bahasa yang representatif dan terstruktur (Hamalainen et al., 2021; Rahgozar, 2020).

Kualitas dan keberhasilan sebuah penelitian linguistik korpus sangat bergantung pada teknik pengumpulan dan penyusunan korpus yang digunakan. Sebuah korpus yang buruk, tidak representatif, atau tidak tersusun dengan baik akan menghasilkan temuan yang bias dan tidak dapat diandalkan. Sebaliknya, korpus yang dirancang dan dibangun dengan cermat dapat mengungkap fenomena linguistik yang sebelumnya sulit atau tidak mungkin diidentifikasi melalui metode penelitian tradisional (Bory, 2020; Bories et al., 2022; Bigoni et al., 2024).

Bab ini akan menggali lebih dalam fondasi penting dari linguistik korpus: teknik pengumpulan dan penyusunan korpus.

Membangun sebuah korpus yang berkualitas tinggi merupakan langkah krusial yang menentukan keberhasilan analisis selanjutnya dan, pada akhirnya, efektivitas aplikasi pembelajarannya (Dwivedi et al., 2023; Varnosfaderani & Forouzanfar, 2024; Yenduri et al., 2024). Kita akan mengeksplorasi berbagai metode dan pertimbangan yang terlibat dalam proses ini, mulai dari sumber data hingga pengelolaan akhir. Tujuannya adalah untuk memberikan pemahaman yang komprehensif bagi para pendidik, peneliti, dan pengembang materi pembelajaran bahasa tentang bagaimana membangun korpus yang dapat diandalkan.

Korpus adalah fondasi dari setiap penelitian atau aplikasi berbasis linguistik korpus. Kualitas korpus secara langsung memengaruhi validitas dan reliabilitas hasil analisis (Singhal et al., 2023; Rashkin et al., 2017; Choudhary et al., 2022). Sebuah korpus yang kurang representatif, tidak bersih, atau tidak terorganisir dengan baik dapat menghasilkan kesimpulan yang menyesatkan sehingga berpotensi merugikan proses pembelajaran bahasa. Oleh sebab itu, pengumpulan dan penyusunan korpus tidak dapat dipandang sebagai tugas yang sederhana. Proses tersebut memerlukan perencanaan yang matang, perhatian yang cermat terhadap detail, serta pemahaman yang memadai mengenai tujuan penyusunan korpus. Bab ini membahas langkah-langkah utama yang diperlukan dalam membangun korpus yang berkualitas dan andal.

Linguistik korpus berakar pada paradigma empirisme linguistik, yang menekankan pentingnya observasi dan bukti nyata dalam memahami bahasa (Hjorland, 2025). Berbeda dengan pendekatan generatif atau teoretis murni yang sering kali bergantung pada intuisi penutur asli atau konstruksi teoretis, linguistik korpus berupaya untuk mendeskripsikan bahasa berdasarkan data yang teramati (Molinari, 2022).

Oleh karena itu, teknik pengumpulan dan penyusunan korpus harus diarahkan untuk memenuhi kebutuhan analitis ini, memastikan bahwa data yang dihasilkan kaya informasi dan sesuai dengan tujuan pembelajaran bahasa (Rejeb et al., 2023; Krister et al., 2021).

PENGUMPULAN DATA TEKS DARI SUMBER DIGITAL

Pengumpulan data teks dari sumber digital merupakan tahap krusial dalam pembangunan korpus linguistik yang representatif dan mutakhir. Keunggulan sumber digital terletak pada skalabilitas, keragaman genre, serta kemudahan akses yang memungkinkan peneliti memperoleh data dalam jumlah besar dengan efisiensi tinggi (Bonatti et al., 2019; Blank, 2009; Zhang et al., 2022). Namun, proses ini menuntut perencanaan metodologis yang matang, termasuk penetapan kriteria seleksi sumber, teknik akuisisi data yang sesuai, serta kepatuhan terhadap aspek legal dan etika penggunaan data (Brundage et al., 2018; Chi & Chen, 2023; Seguin et al., 2022).

Sumber data teks digital sangat melimpah dan beragam, menawarkan peluang besar untuk membangun korpus. Sumber umum meliputi:

- **Internet:** situs web publik: berita online, blog, forum diskusi, situs web pemerintah, dan situs web organisasi. Pengumpulan dapat dilakukan dengan web scraping menggunakan skrip *Python* (dengan *library* seperti *BeautifulSoup* atau *Scrapy*) atau alat yang lebih khusus. Penting untuk memperhatikan hak cipta dan persyaratan penggunaan situs web.
- **E-book dan Dokumen Publik:** Proyek seperti *Project Gutenberg* menyediakan buku-buku yang bebas hak cipta dalam format digital. Dokumen-dokumen PDF atau Word dari sumber resmi juga bisa dimanfaatkan.

- **Repositori Digital:** Jurnal Akademik Online. Banyak jurnal menyediakan akses ke arsip artikel mereka.
- Repositori Tesis dan Disertasi: Universitas sering kali memiliki repositori digital.
- **Sosial Media:** Twitter, Facebook, Reddit. Data dari media sosial dapat memberikan wawasan tentang bahasa informal dan terkini. Pengumpulan data ini sering kali memerlukan API (*Application Programming Interface*) penyedia platform dan perhatian terhadap privasi pengguna serta etika pengumpulan data.

Dengan memanfaatkan beragam sumber digital tersebut, peneliti dapat membangun korpus yang kaya secara linguistik dan kontekstual. Meskipun demikian, kualitas korpus sangat ditentukan oleh ketelitian dalam proses seleksi, dokumentasi metadata, serta kepatuhan terhadap prinsip legalitas dan etika penelitian (Carradini & Swartz, 2023; Setyaningsih et al., 2021). Oleh karena itu, pengumpulan data digital sebaiknya dilakukan secara sistematis, transparan, dan bertanggung jawab agar korpus yang dihasilkan valid, reliabel, serta bermanfaat bagi analisis kebahasaan lanjutan.

TRANSKRIPSI TUTURAN UNTUK KORPUS LISAN

Transkripsi tuturan merupakan tahap fundamental dalam pembangunan korpus lisan karena memungkinkan data ujaran yang bersifat temporer diubah menjadi representasi tekstual yang dapat dianalisis secara sistematis (Mitchell, 2021). Melalui transkripsi yang akurat dan konsisten, peneliti dapat menelaah fenomena kebahasaan secara lebih mendalam, baik dari perspektif fonetik, fonologis, sintaktis, maupun pragmatik. Oleh sebab itu, proses transkripsi tidak hanya bersifat teknis, tetapi juga metodologis, karena kualitas transkripsi akan sangat

menentukan validitas analisis yang dihasilkan (Baker, 2022; Antonio et al., 2020).

Korpus lisan menangkap kekayaan percakapan, pidato, dan bentuk tuturan lainnya. Transkripsi adalah proses mendokumentasikan tuturan secara tertulis yang merupakan langkah krusial untuk analisis kuantitatif dan kualitatif.

- **Format Transkripsi:** Ortografis Standar. Penulisan kata-kata sesuai dengan ejaan standar. Cocok untuk analisis kosakata dan sintaksis.
- **Fonetik/Fonemik:** Menggunakan simbol *International Phonetic Alphabet* (IPA) atau sistem lainnya untuk merekam pelafalan yang tepat. Penting untuk studi fonologi, fonetik, dan variasi dialek.
- **Transkripsi Kuantitatif (CHAT):** Sistem yang sering digunakan dalam analisis percakapan anak, memungkinkan penandaan speaker, jeda, suara non-linguistik, dan informasi temporal.
- **Proses Transkripsi:** Peralatan. Rekaman audio berkualitas tinggi (menggunakan mikrofon eksternal, perekam digital) adalah prasyarat.
- **Perangkat Lunak Transkripsi:** Alat seperti ELAN, Praat, *Express Scribe*, atau bahkan editor teks sederhana dapat digunakan. ELAN dan Praat sangat populer karena kemampuan anotasi multi-tier dan sinkronisasi audio-teks.
- **Panduan Transkripsi:** Membuat panduan transkripsi yang rinci untuk tim transkriptor (jika ada) sangat penting untuk memastikan konsistensi. Ini mencakup bagaimana menangani tumpang tindih ucapan, suara ambient, kata-kata yang tidak jelas, dan penggunaan singkatan atau simbol khusus.
- **Sumber Data Lisan:** Rekaman Percakapan. Percakapan tatap muka, panggilan telepon.
- **Pidato Publik:** Konferensi, kuliah, presentasi.

- **Media Siaran:** Program televisi, radio, podcast (seringkali memerlukan izin).
- **Data Pembelajaran Bahasa:** Rekaman saat pembelajar berbicara atau melakukan tugas komunikasi.

Dengan menerapkan prosedur transkripsi yang sistematis, terstandar, dan didukung perangkat yang memadai, korpus lisan dapat menjadi sumber data yang kaya dan reliabel bagi penelitian kebahasaan (Park & Kim, 2022; Malgaroli et al., 2023; Antonio et al., 2020). Ketelitian dalam proses dokumentasi serta konsistensi dalam penerapan konvensi transkripsi akan meningkatkan keterbandingan data dan memungkinkan analisis lintas studi secara lebih akurat (Buth & Miller-Naudé, 2023; Ndemere, 2020; Connell, 2020). Oleh karena itu, transkripsi tuturan tidak hanya berfungsi sebagai tahap teknis, tetapi juga sebagai fondasi ilmiah yang menentukan kualitas keseluruhan penelitian berbasis korpus lisan.

NORMALISASI DAN PEMBERSIHAN FORMAT DATA

Normalisasi dan pembersihan format data merupakan tahap esensial dalam pengolahan korpus karena kualitas analisis sangat ditentukan oleh kualitas data yang digunakan. Data mentah yang diperoleh dari berbagai sumber digital maupun hasil transkripsi tuturan umumnya mengandung variasi format, inkonsistensi ejaan, serta elemen non-linguistik yang berpotensi mengganggu proses analisis (Zhu, 2022). Oleh karena itu, diperlukan prosedur sistematis untuk menata, menyederhanakan, dan menyeragamkan data sehingga siap digunakan dalam pemodelan linguistik maupun komputasional.

Data mentah, baik teks maupun tuturan yang ditranskrip, sering kali mengandung "kebisingan" atau format yang tidak konsisten. Normalisasi dan pembersihan adalah langkah penting

untuk menyiapkan data untuk analisis (Prasetyo et al., 2023; Purnamawati et al., 2023).

- **Pembersihan Teks:** Menghapus Tag HTML/XML. Jika data diambil dari web, tag markup perlu dihapus. Menghapus Karakter Khusus/Symbol: Simbol yang tidak relevan (misalnya, ikon media sosial, karakter kontrol) harus dihapus atau diganti.
- **Menangani Tanda Baca:** Konsistensi dalam penggunaan tanda baca (misalnya, titik, koma).
- **Menghapus Header/Footer/Sidebar:** Menghilangkan elemen navigasi atau informasi nonkonten dari halaman web.
- **Penanganan Informasi yang Hilang/Rusak:** Menentukan cara menangani bagian teks yang tidak dapat diambil atau rusak.
- **Normalisasi Teks:** Mengubah ke huruf kecil (*Lowercasing*). Umum dilakukan untuk memastikan bahwa "Anjing" dan "anjing" diperlakukan sebagai kata yang sama.
- **Tokenisasi:** Memecah teks menjadi unit-unit diskrit (kata, tanda baca).
- **Lemmatisasi/Stemming:** Lemmatisasi. Mengurangi kata ke bentuk dasarnya (lemma), misalnya, "berlari", "berlari-lari", "berlari" menjadi "lari". Membutuhkan kamus leksikal.
- **Stemming:** Memotong akhiran kata untuk mendapatkan akar kata, misalnya, "berlari", "berlari-lari" menjadi "lari". Lebih sederhana tetapi mungkin menghasilkan akar yang tidak dikenali.
- **Penanganan Angka dan Tanggal:** Menentukan apakah angka dan tanggal akan dipertahankan, dinormalisasi (misalnya, semua angka menjadi ``), atau dihapus.

- **Penanganan Kontraksi dan Singkatan:** Mengembangkan kata menjadi bentuk lengkapnya (misalnya, "nggak" menjadi "tidak") atau konsisten dengan bentuk singkatnya.
- **Normalisasi Tuturan yang Ditranskrip:** Konsistensi Ejaan (jika ortografis). Memastikan ejaan kata yang sama konsisten.
- **Penanganan Suara Non-Linguistik:** Menentukan apakah suara seperti "hmm", "eh", atau tawa akan dihapus, dikodekan, atau dipertahankan.

Melalui proses normalisasi dan pembersihan yang terstruktur, data korpus menjadi lebih konsisten, representatif, dan siap untuk dianalisis secara komputasional maupun linguistik (Satpathy, R., 2025; Malagi, V., 2025; Borge, R., 2025). Tahap ini tidak hanya meningkatkan akurasi hasil analisis, tetapi juga memungkinkan replikasi penelitian dan keterbandingan antarkorpus. Dengan demikian, normalisasi dan pembersihan data berperan sebagai fondasi metodologis yang memastikan bahwa temuan penelitian benar-benar merefleksikan fenomena kebahasaan yang diteliti, bukan artefak dari ketidakteraturan data mentah (Godfrey, 2024).

PENANDAAN METADATA DAN KLASIFIKASI

Setelah data korpus dinormalisasi dan dibersihkan, langkah berikutnya adalah memberikan konteks yang jelas terhadap setiap unit data melalui penandaan metadata dan klasifikasi. Tahap ini sangat penting karena data linguistik tidak hanya dipahami dari bentuk kebahasaannya, tetapi juga dari latar sosial, temporal, dan situasional yang melingkupinya. Dengan metadata yang sistematis dan klasifikasi yang tepat, peneliti dapat melakukan analisis yang lebih terarah, komparatif, dan kontekstual (varios, 2021).

Metadata adalah "data tentang data." Penandaan metadata yang relevan untuk setiap unit dalam korpus sangat penting untuk memungkinkan pemfilteran, analisis yang lebih mendalam, dan pemahaman konteks (Pacheco et al., 2023; Steinmetz, 2013; Ayobami, 2025).

- **Jenis Metadata:** Informasi Sumber: Nama dokumen, URL (jika dari web), nama buku/artikel, penulis, penerbit.
- **Informasi Temporal:** tanggal publikasi, tanggal rekaman. Informasi Geografis: Lokasi pembuatan teks/tuturan, dialek.
- **Informasi Demografis:** Usia, jenis kelamin, tingkat pendidikan (jika relevan dan tersedia).
- **Informasi Genre/Register:** berita, fiksi, akademis, percakapan informal, pidato formal.
- **Informasi Tugas/Konteks:** Untuk korpus pembelajar bahasa, bisa mencakup tingkat kemahiran, tugas yang diberikan, jenis kesalahan yang diantisipasi.
- **Informasi Teknis (untuk data lisan):** Durasi rekaman, kualitas audio, nama partisipan.
- **Metode Penandaan:** Manual. Memasukkan metadata secara manual ke dalam tabel atau file deskriptif.
- **Otomatis (dengan bantuan alat):** Menggunakan skrip untuk mengekstrak informasi dari header dokumen atau URL.
- **Anotasi dalam File Korpus:** Metadata dapat disertakan dalam file teks itu sendiri (misalnya, dalam bentuk komentar di awal file) atau dalam file terpisah yang terkait dengan file data. Format umum termasuk XML, JSON, atau file CSV.
- **Klasifikasi Data:** Berdasarkan Genre: Mengelompokkan data ke dalam kategori yang jelas (misalnya, "Berita", "Blog", "Percakapan").
- **Berdasarkan Penulis/Pembicara:** Mengategorikan data berdasarkan karakteristik demografis atau linguistik.

- **Berdasarkan Topik:** Mengklasifikasikan dokumen berdasarkan subjek utama.

Penandaan metadata dan klasifikasi tidak hanya memperkaya informasi kontekstual, tetapi juga meningkatkan fleksibilitas penggunaan korpus untuk berbagai tujuan penelitian. Dengan struktur metadata yang konsisten, peneliti dapat melakukan penyaringan data secara presisi, membandingkan subkorpus secara sistematis, serta memastikan transparansi dan replikabilitas penelitian (Olu James Mbanugo, 2025; Shin, J., 2025; Vinciguerra, K., 2025). Oleh karena itu, tahap ini menjadi komponen strategis dalam pembangunan korpus yang berkualitas dan siap mendukung analisis linguistik maupun komputasional secara komprehensif.

PENYIMPANAN, PENGELOLAAN, DAN BACKUP DATA

Setelah korpus tersusun, ditandai, dan diklasifikasikan, tahap krusial berikutnya adalah memastikan data tersimpan secara sistematis, mudah diakses, dan terlindungi dari kehilangan atau kerusakan. Dalam praktik penelitian linguistik, termasuk pengembangan korpus untuk aplikasi pembelajaran EFL yang Bapak kerjakan, pengelolaan data yang baik tidak hanya menjamin keberlanjutan penggunaan korpus, tetapi juga mendukung transparansi metodologis dan replikabilitas analisis (Smith, 2022).

Korpus yang telah disusun dan ditandai perlu disimpan dan dikelola dengan baik agar tetap dapat diakses, aman, dan terorganisir.

- **Struktur Direktori:** Buat struktur folder yang logis untuk mengorganisir korpus. Misalnya, berdasarkan genre, tahun, atau subkategori lainnya. Pisahkan data mentah, data yang dibersihkan, dan data yang dianotasi.

- **Format File:** Pilih format file yang umum dan mudah dibaca oleh berbagai perangkat lunak (misalnya, `.txt` untuk teks, `.cha` untuk transkrip CHAT, `.csv` atau `.json` untuk metadata). Hindari format proprietary yang mungkin menjadi usang.
- **Sistem Manajemen Basis Data (untuk korpus besar):** Untuk korpus yang sangat besar, pertimbangkan menggunakan sistem manajemen basis data (DBMS) yang dirancang untuk teks, seperti Apache Solr, Elasticsearch, atau basis data relasional (dengan tabel khusus untuk teks). Ini memungkinkan pencarian yang sangat efisien.
- **Pengelolaan Versi (Version Control):** Gunakan sistem kontrol versi seperti Git untuk melacak perubahan pada data dan metadata. Ini memungkinkan Anda kembali ke versi sebelumnya jika terjadi kesalahan.
- **Keamanan dan Pencadangan (Backup):** Backup Berkala: Lakukan pencadangan korpus secara teratur dan simpan di lokasi yang berbeda (misalnya, hard drive eksternal, layanan cloud).
- **Pengamanan Fisik/Digital:** Lindungi media penyimpanan dari kerusakan fisik atau akses tidak sah.
- **Replikasi Data:** Simpan salinan data di beberapa lokasi geografis yang berbeda untuk ketahanan bencana.
- **Dokumentasi Korpus:** Buat dokumen yang menjelaskan secara rinci tentang korpus yaitu bagaimana data dikumpulkan, kriteria pemilihan, panduan transkripsi, deskripsi metadata, dan batasan-batasannya. Dokumen ini, yang sering disebut "korpus documentation" atau "readme file," sangat penting untuk siapa pun yang akan menggunakan korpus tersebut.

Dengan sistem penyimpanan, pengelolaan, dan pencadangan yang terstruktur, korpus tidak hanya berfungsi

sebagai kumpulan data, tetapi juga sebagai sumber ilmiah yang berkelanjutan, dapat diverifikasi, dan siap digunakan kembali. Praktik manajemen data yang baik akan meningkatkan kredibilitas penelitian, mempermudah integrasi dengan perangkat analisis digital, serta memastikan bahwa upaya pengumpulan data—seperti yang telah Bapak lakukan dalam pengembangan sumber teks untuk pembelajaran—memberikan manfaat jangka panjang bagi penelitian linguistik dan pendidikan bahasa (Dash et al., 2019; Kotani & Yoshimi, 2021).

Proses pengumpulan dan penyusunan korpus adalah investasi waktu dan sumber daya yang signifikan, namun hasilnya adalah fondasi yang tak ternilai untuk berbagai aplikasi dalam pembelajaran bahasa. Dengan menerapkan teknik yang tepat dalam pengumpulan data dari sumber digital dan lisan, melakukan transkripsi yang akurat, membersihkan dan menormalisasi data, menandai metadata yang relevan, serta mengelola data dengan aman, kita dapat membangun korpus yang tidak hanya kaya informasi tetapi juga mudah digunakan dan dianalisis.

Korpus yang berkualitas tinggi memungkinkan para pendidik dan peneliti untuk mengidentifikasi pola penggunaan bahasa yang otentik, mengembangkan materi pengajaran yang relevan, dan memahami tantangan yang dihadapi pembelajar. Langkah-langkah yang diuraikan dalam bab ini merupakan panduan esensial untuk memastikan bahwa fondasi linguistik korpus yang dibangun kuat, kokoh, dan siap mendukung inovasi dalam pendidikan bahasa.

KESIMPULAN

Pembangunan korpus adalah seni dan sains tersendiri dalam linguistik korpus, yang merupakan prasyarat fundamental untuk aplikasi yang kuat dalam pembelajaran bahasa. Bab ini telah menguraikan langkah-langkah krusial yang terlibat dalam teknik

pengumpulan dan penyusunan korpus, mulai dari identifikasi sumber data digital dan lisan, proses transkripsi yang teliti, pembersihan dan normalisasi data yang cermat, hingga penandaan metadata yang informatif dan pengelolaan data yang aman.

Setiap tahapan memiliki tantangan tersendiri, namun dengan perencanaan yang matang, perhatian terhadap detail, dan pemahaman yang kuat tentang tujuan korpus, fondasi yang kokoh dapat dibangun. Korpus yang berkualitas tinggi menyediakan bukti empiris yang tak ternilai tentang bagaimana bahasa benar-benar digunakan, yang sangat esensial untuk mengembangkan materi pembelajaran bahasa yang efektif, memvalidasi teori linguistik terapan, dan mendorong inovasi dalam teknologi pendidikan bahasa. Dengan mengadopsi praktik terbaik yang dibahas, para profesional dapat memastikan bahwa korpus yang mereka bangun tidak hanya berfungsi sebagai kumpulan data, tetapi juga sebagai sumber daya yang dinamis dan kuat untuk memajukan studi dan pengajaran bahasa.

DAFTAR PUSTAKA

- Antonio, P., Maria, B., C., L. B., & Christian, C. (2020). Development of linguistically linked open data resources for collaborative data-intensive research in the language sciences. The MIT Press eBooks. <https://doi.org/10.7551/mitpress/10990.001.0001>
- Ayobami, U. (2025). Automated metadata extraction and correlation techniques for digital evidence analysis in cybercrime investigations. International Journal of Research Publication and Reviews. <https://doi.org/10.55248/gengpi.6.0625.2177>
- Baker, A. (2022). Pronunciation pedagogy: Second language teacher cognition and practice. Digital Archive @ GSU. <https://doi.org/10.57709/1949560>

- Berez-Kroeker, A.L., McDonnell, B., Collister, L.B., & Koller, E. (2022). Data, Data Management, and Reproducible Research in Linguistics: On the Need for The Open Handbook of Linguistic Data Management. The Open Handbook of Linguistic Data Management. <https://doi.org/10.7551/mitpress/12200.003.0005>
- Bigoni, M., Maran, L., & Occhipinti, Z. (2024). Of power, knowledge, and method: The influence of Michel Foucault in accounting history. *Accounting History*. <https://doi.org/10.1177/10323732241243088>
- Borge, R., Johannessen, H., Fostervold, K., & Nielsen, M. (2025). Does Attending Work When Ill Vary Across Different Office Concepts? A Bayesian Analysis of Differences in Sickness Presenteeism using Nationally Representative Data. *Occupational Health Science*, 409-435. <https://doi.org/10.1007/s41542-025-00223-4>
- Bories, A., Plecháč, P., Fabo, P. R., & KG, W. D. G. G. & C. (2022). Computational stylistics in poetry, prose, and drama. *Unknown Journal*. <https://doi.org/10.1515/9783110781502>
- Bory, P. (2020). The internet myth: From the internet imaginary to network ideologies. University of Westminster Press eBooks. <https://doi.org/10.16997/book48>
- Brundage, M., Avin, S., Clark, J., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *Apollo* (University of Cambridge). <https://doi.org/10.17863/cam.22520>
- Carradini, S. & Swartz, J. (2023). Text at Scale: Corpus Analysis in Technical Communication. The WAC Clearinghouse; University Press of Colorado. <https://doi.org/10.37514/tpc-b.2023.2104>

- Chi, Y. & Chen, H. (2023). Investigating Substance Use Via Reddit: Systematic Scoping Review. *Journal of Medical Internet Research*. <https://doi.org/10.2196/48905>
- Choudhary, K., DeCost, B., Chen, C., et al. (2022). Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*. <https://doi.org/10.1038/s41524-022-00734-6>
- Dash, NS., Elumalai, KV., Alqahtani, MSM., & Abumelha, MA. (2019). Proposing a Customised Method for Extratextual Documentative Annotation on Written Text Corpus. *International Journal of English Linguistics*, 9, 99. <https://doi.org/10.5539/ijel.v9n2p99>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary Perspectives on Opportunities, Challenges, and Implications of Generative Conversational AI for Research, practice, and policy. *International Journal of Information Management*. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Godfrey, J. (2024). Prompt and circumstance: Investigating the relationship between college writing and postsecondary policy. *Deep Blue* (University of Michigan). <https://doi.org/10.7302/23889>
- Hadi, MU., Tashi, QA., Shah, A., et al. (2024). Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects. <https://doi.org/10.36227/techrxiv.23589741.v6>
- Hjørland, B. (2025). Semantic Primitives and Compositionality: An Annual Review of Information Science and Technology paper. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.70011>
- Hutson, J., Jeevanjee, T., Graaf, VV., et al. (2022). Artificial Intelligence and the Disruption of Higher Education: Strategies for Integration Across Disciplines. *Creative*

- Education, 13, 3953-3980.
<https://doi.org/10.4236/ce.2022.1312253>
- Kotani, K. & Yoshimi, T. (2021). Prediction of general ESL Proficiency Considering Learners' Dictation Performance. *SHS Web of Conferences*, 102, 01003.
<https://doi.org/10.1051/shsconf/202110201003>
- Krister, L., Jyrki, N., Göteborg, B., et al. (2021). The Swedish framenet. *Natural language processing*.
<https://doi.org/10.1075/nlp.14>
- Malagi, V., M. S., A., Namani, S., Kukreja, S., Vibha, K., & Reka, R. (2025). Innovative Quantum Systems Analysis Through Machine Learning and Quantum Computing. *Advances in Computational Intelligence and Robotics*, 27-52. <https://doi.org/10.4018/979-8-3693-6225-9.ch002>
- Mitchell, P. (2021). Frankenstein's Creature on Film in the Twenty-First Century: Posthuman Monster, Saviour, and Victim Narratives. *Interdisciplinary Literary Studies*.
<https://doi.org/10.5325/intelitestud.23.1.0001>
- Molinari, J. (2022). What Makes Writing academic? *Bloomsbury Publishing Plc eBooks*.
<https://doi.org/10.5040/9781350243958>
- Olu James Mbanugo, Andrea Taylor, & Sweta Sneha (2025). Buttressing the Power of the Entity Relationships Model in Database Structure and Information Visualization: Insights from the Technology Association of Georgia's Digital Health Ecosystem. *World Journal of Advanced Research and Reviews*, 1294-1313.
<https://doi.org/10.30574/wjarr.2025.25.2.0521>
- Pacheco, A., Silva, C. G. D., & Freitas, M. C. V. D. (2023). A Metadata Model for Authenticity in Digital Archival Descriptions. *Archival Science*.
<https://doi.org/10.1007/s10502-023-09422-w>

- Purnamawati, A., Winarto, M.N., & Mailasari, M. (2023). Analisis Sentimen Aplikasi TikTok Menggunakan Metode BM25 dan Improved K-NN Fitur Chi-Square. *Jurnal Komtika (Komputasi dan Informatika)*. <https://doi.org/10.31603/komtika.v7i1.8938>
- Rahgozar, A. & Inkpen, D. (2016). Poetry Chronological Classification: Hafez. *Lecture Notes in Computer Science*, 131-136. https://doi.org/10.1007/978-3-319-34111-8_18
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics*. <https://doi.org/10.18653/v1/d17-1317>
- Rejeb, A., Rejeb, K., Simske, S., & Keogh, J. G. (2023). Exploring blockchain research in supply chain management: A latent dirichlet allocation-driven systematic review. *Information*. <https://doi.org/10.3390/info14100557>
- Satpathy, R. (2025). Exploring Potential Drug Targets in Plasmodium falciparum 3D7 Through Computational Analysis. *Research Journal of Biotechnology*, 103-110. <https://doi.org/10.25303/204rjbt1030110>
- Seguin, J. P., Varghese, D., Anwar, M., Bartindale, T., & Olivier, P. (2022). Co-designing digital platforms for volunteer-led migrant community welfare support. *Designing Interactive Systems Conference*. <https://doi.org/10.1145/3532106.3533544>
- Setyaningsih, N., Larassati, A., & Suryaningtyas, V.W. (2021). Kolokasi leksikon Covid-19 dalam bahasa Indonesia dan bahasa Inggris di ranah cyberspace. *HUMANIKA*. <https://doi.org/10.14710/humanika.v28i2.40600>

- Shin, J., Kim, M., Park, S. (2025). Improving character recognition by using metadata on personal attributes. *Journal of the Korea Society of Computer and Information*, 45-51. <https://doi.org/10.9708/jksci.2025.30.02.045>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., et al. (2023). Large language models encode clinical knowledge. *Nature*. <https://doi.org/10.1038/s41586-023-06291-2>
- Steinmetz, N. (2013). Context-aware semantic analysis of video metadata. publish.UP (University of Potsdam).
- Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, & Fitri Nurapriani. (2023). Analisis sentimen relokasi ibukota nusantara menggunakan algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*. <https://doi.org/10.35134/komtekinfo.v10i1.330>
- varios, A. (2021). Cts spring-cleaning: A critical reflection. *MonTi Monografías de Traducción e Interpretación*. <https://doi.org/10.6035/monti.2021.13>
- Varnosfaderani, S. M. & Forouzanfar, M. (2024). The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering*. <https://doi.org/10.3390/bioengineering11040337>
- Vinciguerra, K. & Marchegiani, L. (2025). knowing when and where improves acoustic classification. A different perspective on auditory contextual information. <https://doi.org/10.36227/techrxiv.174059900.07239200/v1>
- Yenduri, G., Ramalingam, M., Selvi, G. C., et al. (2024). GPT (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*. <https://doi.org/10.1109/access.2024.3389497>
- Zhu, J. (2022). A computational account of selected patterns of linguistic variation and change. *Deep Blue* (University of Michigan). <https://doi.org/10.7302/5900>

PROFIL PENULIS



Bejo Sutrisno

Penulis berasal dari Kecamatan Wangon, Banyumas, Jawa Tengah. Ia meraih gelar S1 dari Universitas Muhammadiyah Prof. Dr. Hamka (UHAMKA) pada tahun 2004, kemudian menyelesaikan studi S2 di Universitas Negeri Jakarta (UNJ) pada tahun 2008. Pada 2015, ia melanjutkan pendidikan Doktor di Sekolah Pascasarjana UNJ dan berhasil lulus pada 2019.

Saat ini, ia merupakan dosen tetap pada Program Magister Linguistik Terapan, Universitas Al-Azhar Indonesia, dengan fokus pada kajian linguistik terapan, pedagogi bahasa, dan kewirausahaan bahasa.

BAB 4

PENANDAAN (ANNOTATION) DALAM LINGUISTIK KORPUS

Akhmad Hairul Umam
Universitas Tanri Abeng, Jakarta
E-mail: ahmad.umam@tau.ac.id

PENDAHULUAN

Perkembangan linguistik korpus telah mengubah cara peneliti bahasa melakukan deskripsi, analisis, dan verifikasi teori kebahasaan. Jika pada pendekatan tradisional analisis sering bertumpu pada intuisi penutur atau contoh buatan, linguistik korpus menekankan penggunaan data bahasa autentik dalam jumlah besar untuk memahami pola kebahasaan secara lebih empiris (Biber, Conrad, & Reppen, 1998; McEnery & Hardie, 2011). Dalam konteks ini, penandaan (annotation) menjadi komponen penting karena memungkinkan data korpus diperkaya dengan informasi linguistik yang tidak tampak langsung pada teks mentah, seperti kelas kata, lema, relasi sintaktis, atau kategori semantik (Gries & Berez, 2017; Xiao, 2008).

Penandaan memberikan nilai tambah pada korpus karena memungkinkan peneliti mengajukan pertanyaan yang lebih beragam dan melakukan analisis secara lebih sistematis. Berbeda dengan *markup* struktural yang berfokus pada struktur dokumen, penandaan bersifat interpretatif karena menambahkan hasil analisis linguistik ke dalam data korpus. Oleh sebab itu, anotasi menjadi komponen penting dalam pemanfaatan korpus untuk penelitian linguistik, pengajaran bahasa, leksikografi, dan *Natural Language Processing/NLP* (Leech, 1997; Wynne, 2005).

Bab ini membahas konsep penandaan dalam linguistik korpus secara sistematis, meliputi pengertian, tujuan, jenis-jenis penandaan, dan contoh/*case study* pada bahasa Inggris dan bahasa Indonesia.

PENGERTIAN PENANDAAN (*ANNOTATION*)

Secara umum, penandaan dalam linguistik korpus adalah proses penambahan informasi linguistik yang bersifat interpretatif ke dalam data bahasa, baik berupa teks tulis maupun tuturan lisan, agar data tersebut dapat dianalisis secara lebih kaya, sistematis, dan terarah. Dalam korpus mentah, banyak informasi linguistik belum tampak secara eksplisit, sehingga penandaan berfungsi untuk membuat struktur dan makna kebahasaan lebih mudah dikenali, dicari, dan diuji secara empiris. Karena itu, anotasi dipandang sebagai salah satu komponen inti dalam pemanfaatan korpus modern, terutama ketika peneliti ingin bergerak dari sekadar pengamatan bentuk permukaan menuju analisis linguistik yang lebih mendalam (Leech, 1997; McEnery & Hardie, 2011).

Informasi yang ditambahkan dalam penandaan dapat berada pada berbagai level. Pada level morfologis, anotasi biasanya mencakup penandaan kelas kata (*part-of-speech*) dan fitur gramatikal seperti waktu (*tense*), jumlah (*number*), jenis kelamin (*gender*), kasus (*case*) dan orang (*person*). Pada level leksikal, penandaan dapat berupa *lemmatization*, yaitu pengelompokan berbagai bentuk kata ke dalam bentuk dasarnya. Pada level sintaktis, penandaan digunakan untuk menandai struktur frasa, relasi dependensi, atau fungsi sintaksis antarkata dalam kalimat. Selain itu, penandaan juga dapat dilakukan pada level semantik, misalnya dengan memberi label makna, kategori semantik, atau relasi makna tertentu yang membantu peneliti menelaah pola penggunaan leksis dan interpretasi konteks (Xiao, 2008; McEnery & Hardie, 2011).

Di luar level tersebut, perkembangan linguistik korpus juga menunjukkan pentingnya penandaan pada level wacana dan pragmatik. Penandaan wacana dapat mencakup penandaan *coreference*, anafora, struktur percakapan, atau *discourse markers*, sedangkan pragmatik dapat digunakan untuk mengidentifikasi fungsi tutur, tindak tutur, dan maksud komunikatif penutur dalam konteks tertentu. Untuk korpus lisan, penandaan bahkan dapat diperluas ke aspek prosodik, seperti intonasi, jeda, tekanan, dan pola nada, karena unsur-unsur tersebut sering berkontribusi pada makna interaksional tuturan. Dengan demikian, penandaan tidak hanya memperkaya korpus sebagai sumber data, tetapi juga memperluas jenis pertanyaan penelitian yang dapat dijawab dalam linguistik teoretis, pengajaran bahasa, leksikografi, analisis wacana, hingga *natural language processing* (Gries, 2017; Archer, Culpeper & Davies, 2008).

FUNGSI DAN TUJUAN

Penandaan dalam korpus linguistik tidak dilakukan sekadar untuk menambahkan label pada data bahasa. Lebih dari itu, penandaan memiliki berbagai fungsi penting yang membantu peneliti memperoleh analisis yang lebih akurat, sistematis, dan dapat dipertanggungjawabkan. Melalui proses ini, korpus tidak lagi hanya menjadi kumpulan teks mentah, tetapi berubah menjadi sumber data yang lebih terstruktur dan siap digunakan untuk berbagai kebutuhan penelitian maupun pengembangan teknologi bahasa. Berikut beberapa fungsi dan tujuan utama penandaan dalam korpus linguistik.

Membuat Pertanyaan Riset Lebih Tajam dan Mudah Dijawab

Tujuan utama penandaan adalah membuka jenis pertanyaan penelitian yang sulit dijawab jika hanya mengandalkan teks mentah. Misalnya, peneliti ingin mengetahui seberapa sering kata *record* dipakai sebagai nomina dibandingkan dengan verba dalam berita. Tanpa penandaan, pencarian “*record*” akan mencampur dua fungsi itu. Dengan POS tagging, peneliti bisa memisahkan *record/NOUN* dan *record/VERB*, sehingga hasil analisis menjadi lebih tepat. Inilah alasan mengapa penandaan sering disebut memperluas “jangkauan pertanyaan” yang bisa dijawab oleh korpus (Xiao, 2008).

Mengurangi Ambiguitas dan Meningkatkan Ketelitian Analisis

Bahasa alami penuh ambiguitas. Penandaan membantu disambiguasi, yaitu membedakan bentuk yang sama tetapi fungsi/analisisnya berbeda. Ini sangat penting untuk menjaga ketelitian hasil penelitian. Fungsi ini juga kuat dalam bahasa Indonesia. Misalnya, bentuk “di” pada “di rumah” berfungsi sebagai preposisi, sedangkan “di-” pada “dimakan” adalah prefiks pasif. Jika tokenisasi dan penandaan tidak konsisten, analisis gramatikal bisa meleset. Penandaan membantu “menertibkan” ambiguitas seperti ini sehingga data lebih stabil untuk analisis lanjutan (Xiao, 2008).

Menstandarkan Data Agar Hasil Riset Bisa Dibandingkan

Penandaan juga bertujuan membuat deskripsi linguistik menjadi lebih seragam, sehingga hasil riset dapat dibandingkan antarkorpus, antargenre, bahkan antarbahasa. Kerangka Universal Dependencies (UD) adalah contoh yang menonjol: UD menyediakan pedoman dan label yang konsisten untuk POS, fitur morfologis, dan relasi dependensi sintaktis lintas bahasa

(Universal Dependencies, n.d.-a). Dengan standardisasi semacam ini, peneliti bisa membandingkan, misalnya, pola *ADJ* + *NOUN* di bahasa Inggris dan bahasa Indonesia dengan kerangka anotasi yang sepadan.

Membuat Korpus “Siap Pakai Ulang” (*Reusability*)

Membuat korpus beranotasi itu mahal, perlu waktu, tenaga, dan keahlian. Namun, “biaya” ini menjadi investasi karena korpus beranotasi dapat dipakai berulang-ulang untuk banyak penelitian. Sekali korpus diberi penandaan yang baik, peneliti lain dapat memanfaatkannya untuk studi berbeda tanpa harus mengulang interpretasi dari awal. Itulah mengapa penandaan sering dipahami sebagai pembangunan infrastruktur data untuk komunitas riset (Xiao, 2008). Dalam konteks UD, *treebank* (korpus teks/kumpulan teks) yang tersedia luas dipakai ulang untuk penelitian linguistik dan juga pengembangan NLP (Universal Dependencies, n.d.-b).

Meningkatkan Transparansi dan Keterulangan (*Replicability*)

Salah satu masalah dalam riset bahasa adalah hasil analisis sulit diulang jika langkah-langkahnya tidak jelas. Penandaan yang terdokumentasi, misalnya melalui pedoman anotasi, definisi label, dan format data standar, membuat proses analisis lebih transparan. Peneliti lain dapat menelusuri mengapa suatu token diberi label tertentu atau mengulang kueri yang sama pada dataset yang sama. Dengan begitu, penelitian menjadi lebih akuntabel dan mudah diverifikasi (Hovy & Lavid, 2010). UD juga mendorong praktik terbuka lewat pedoman dan dokumentasi yang memudahkan evaluasi lintas peneliti (Universal Dependencies, n.d.-a).

Menjadi Fondasi Analisis Tingkat Lanjut dan Aplikasi NLP

Dalam banyak kasus, penandaan dasar adalah “bahan baku” untuk analisis yang lebih tinggi. POS tagging sering menjadi prasyarat untuk parsing sintaktis, ekstraksi entitas, atau analisis relasi gramatikal. Karena itu, penandaan bukan hanya berguna untuk riset linguistik teoretis, tetapi juga penting untuk NLP. UD secara eksplisit dirancang agar anotasi tata bahasa dapat mendukung pemodelan komputasional lintas bahasa (Universal Dependencies, n.d.-a). Dengan kata lain, penandaan menjembatani dua dunia: analisis linguistik yang rapi dan aplikasi komputasional yang terukur.

Menjaga Kualitas Data: Reliabilitas Label dan Ketidakpastian

Karena penandaan bersifat interpretatif, kualitasnya perlu dijaga. Dua anotator bisa saja berbeda pendapat pada kasus ambigu. Di sinilah pentingnya uji kesepakatan, diskusi kasus, dan *adjudication*. Selain itu, pendekatan modern juga menekankan bahwa ketidakpastian bisa dimodelkan, bukan disembunyikan. Passonneau dan Carpenter (2014) menunjukkan manfaat model anotasi yang mempertimbangkan kualitas label (misalnya tingkat keyakinan atau variasi antar-anotator) untuk evaluasi yang lebih realistis. Artinya, anotasi yang baik tidak sekadar “memberi label”, tetapi juga mengelola kualitas label agar hasil riset lebih kuat.

JENIS-JENIS PENANDAAN

Dalam linguistik korpus, penandaan adalah “label tambahan” yang ditempelkan pada teks agar teks tersebut lebih mudah dianalisis. Tanpa penandaan, korpus hanya berupa teks mentah; dengan penandaan, peneliti bisa mencari pola bahasa dengan lebih rapi (misalnya pola kelas kata, struktur kalimat, atau hubungan makna). Penandaan ini bersifat interpretatif

karena melibatkan keputusan analisis linguistik (Xiao, 2008). Berikut jenis-jenis anotasi yang paling umum dipakai.

Penandaan struktural (*structural markup*)

Ini adalah penandaan bagian-bagian teks agar mudah dipotong sebagai unit analisis, misalnya batas dokumen, paragraf, kalimat, judul, atau giliran bicara (untuk korpus lisan). Walau sering dibedakan dari penandaan linguistik, penandaan struktural ini penting sebagai fondasi, karena peneliti biasanya bekerja pada unit seperti “kalimat” atau “tuturan” (Xiao, 2008).

Contoh bahasa Indonesia

Teks mentah:

Bab 1. Pendahuluan

Ini adalah penelitian tentang...

Sesudah ditandai (contoh format sederhana/XML):

```
<doc id="D1" genre="artikel">
  <title>Bab 1. Pendahuluan</title>
  <p>Ini adalah penelitian tentang...</p>
</doc>
```

Penjelasan:

Penandaan ini membantu korpus tahu mana judul, paragraf, atau dokumen. Ini penting untuk pencarian berbasis bagian teks (misalnya, hanya cari di <title> atau di paragraf tertentu). Dalam literatur korpus, ini sering dibedakan dari anotasi linguistik karena lebih bersifat struktural, tetapi tetap fondasional untuk analisis (Xiao, 2005).

POS Tagging (*Part-of-Speech* / Kelas Kata)

Jenis POS *tagging* yang paling populer. Setiap kata diberi label kelas kata seperti NOUN, VERB, ADJ atau label lain sesuai tagset. Contoh sederhana: *makan/VERB*, *rumah/NOUN*. Dengan POS tagging, peneliti bisa menghitung frekuensi verba, mencari pola *ADJ* + *NOUN*, dan membedakan kata yang sama

tetapi memiliki fungsi berbeda (Xiao, 2008). Universal Dependencies (UD) juga menyediakan sistem POS lintas bahasa (UPOS) agar anotasi lebih konsisten antarbahasa (Universal Dependencies, n.d.-a).

Contoh bahasa Inggris

Kalimat: *They record a video.*

POS-tag (gaya UD sederhana):

- They/PRON
- record/VERB
- a/DET
- video/NOUN

Penjelasan:

Setiap kata diberi label kelas kata. Ini membuat pencarian pola jadi mudah, misalnya mencari semua pola VERB + NOUN atau menghitung frekuensi verba tertentu (Xiao, 2005; Universal Dependencies, n.d.-a).

Contoh bahasa Indonesia

Kalimat: “Dia makan nasi”.

POS-tag:

- Dia/PRON
- makan/VERB
- nasi/NOUN

Dengan POS tagging, peneliti bisa menganalisis pola struktur kalimat sederhana atau pola kolokasi berbasis kategori (misalnya kata kerja apa yang sering berpasangan dengan objek makanan).

Lemmatization (Lemma)

Lemmatization menandai bentuk dasar kata. Misalnya dalam bahasa Inggris: *running* → *run*, *books* → *book*. Dalam

bahasa Indonesia: *memakan* → *makan*, *berlari* → *lari* (tergantung aturan analisis). Lemma membantu analisis frekuensi yang lebih adil karena variasi bentuk kata dihitung sebagai satu “keluarga” (Xiao, 2008).

Contoh bahasa Inggris

Teks: *running*, *ran*, *runs*

Lemma:

- *running* → *run*
- *ran* → *run*
- *runs* → *run*

Penjelasan:

Lemma menyatukan bentuk berbeda menjadi satu bentuk dasar, sehingga analisis frekuensi lebih akurat (misalnya “run” dihitung sebagai satu konsep, bukan tiga kata berbeda) (Xiao, 2005).

Contoh bahasa Indonesia

Teks: *memakan*, *dimakan*, *makanlah*

Lemma (contoh sederhana):

- *memakan* → *makan*
- *dimakan* → *makan*
- *makanlah* → *makan*

Penjelasan:

Ini penting untuk bahasa yang kaya imbuhan seperti bahasa Indonesia karena satu lema bisa muncul dalam banyak bentuk (Xiao, 2005).

Fitur Morfologis (*Morphological Features*)

Selain kelas kata, kata juga bisa diberi informasi gramatikal, misalnya *tense*, *number*, *voice*, *person*, dan fitur lain yang sesuai dengan bahasa. Dalam UD, fitur morfologis dibuat lebih

seragam lintas bahasa sehingga mudah dipakai untuk riset dan pemodelan (Universal Dependencies, n.d.-a).

Contoh bahasa Indonesia

Kalimat: *Ikan itu dimakan kucing.*

Contoh penandaan (gaya UD):

- dimakan: VERB, Voice=Pass

Penjelasan:

Selain kelas kata, kata diberi fitur gramatikal, misalnya pasif/aktif, jamak/tunggal, tense, dsb. UD menyediakan daftar fitur yang distandarkan agar bisa dipakai lintas bahasa (Universal Dependencies, n.d.-a).

Penandaan Sintaktis (Parsing: Dependensi Atau Frasa)

Jenis ini menandai struktur kalimat: kata mana yang menjadi subjek, objek, modifier, dan seterusnya. Pada UD, struktur ini sering ditulis dalam bentuk *dependency relations* (misalnya *nsubj*, *obj*), sehingga peneliti dapat menganalisis pola sintaksis secara lebih detail (Universal Dependencies, n.d.-b).

Contoh bahasa Inggris

Kalimat: *She likes coffee.*

Relasi dependensi (ringkas):

- likes = ROOT
- She → *nsubj*(likes)
- coffee → *obj*(likes)

Penjelasan:

Parsing dependensi menunjukkan “siapa melakukan apa kepada siapa”. Ini membantu analisis tata bahasa yang lebih dalam (misalnya pola objek, struktur klausa, dan sebagainya) (Universal Dependencies, n.d.-b).

Contoh bahasa Indonesia

Kalimat: *Budi membeli buku.*

Relasi:

- membeli = ROOT
- Budi → nsubj(membeli)
- buku → obj(membeli)

Penjelasan:

Dengan penandaan ini, peneliti bisa mencari semua kalimat yang polanya *objek = buku* atau melihat variasi verba transitif dalam korpus.

Penandaan semantik

Penandaan semantik menambahkan informasi makna, misalnya kategori medan semantik, *word sense* (makna kata yang mana), atau peran semantik. Ini lebih sulit karena makna sering ambigu dan sangat bergantung pada konteks, sehingga membutuhkan pedoman penandaan yang kuat (Xiao, 2008).

Kalimat: *Saya membeli apel.*

Contoh penandaan semantik sederhana:

- apel → FOOD
- membeli → BUY_EVENT

Penjelasan:

Penandaan semantik memberi label makna/kategori konsep. Ini lebih sulit karena makna sering ambigu dan perlu pedoman yang kuat (Xiao, 2005).

Penandaan Wacana dan Pragmatik

Di level ini, penandaan bisa mencakup *coreference* (kata ganti merujuk ke siapa), hubungan antarbagian teks, *discourse markers*, atau fungsi tutur (misalnya meminta, memerintah, menyindir). Jenis ini berguna untuk analisis percakapan, narasi, dan interaksi sosial, tetapi biasanya lebih kompleks dan mahal dibuat (Xiao, 2008).

Coreference (Rujukan)

Teks: *Ani membeli buku. Dia membacanya semalam.*

Coreference:

- Dia → Ani
- -nya (membacanya) → buku

Penjelasan:

Ini membantu analisis teks yang lebih kompleks, misalnya siapa merujuk ke siapa dalam paragraf naratif atau berita (Xiao, 2005).

PRINSIP-PRINSIP PENANDAAN YANG BAIK

Prinsip penandaan dalam linguistik korpus adalah “aturan main” agar label linguistik yang ditempelkan pada teks konsisten, bisa dipercaya, dan berguna untuk analisis. Karena penandaan bersifat *interpretatif* (ada keputusan analisis di balik setiap label), kualitas penandaan sangat ditentukan oleh cara kita merancang skema, melatih anotator, dan melakukan kontrol kualitas (Xiao, 2008).

Jelas Tujuan dan Level Penandanya

Sebelum memberi label, tentukan dulu kebutuhan riset: apakah butuh POS (kelas kata), lemma, fitur morfologi, sintaksis (dependensi), semantik, atau wacana. Semakin tinggi level anotasi, semakin kompleks aturan dan semakin besar biaya/risikonya (Xiao, 2008).

Contoh: Jika tujuan hanya menghitung pola *ADJ + NOUN*, cukup POS tagging; tidak harus parsing dependensi.

Skema Penandaan Harus Eksplisit dan Stabil

Skema berarti daftar label dan definisinya, termasuk unit yang diberi label (token, frasa, kalimat). Skema yang stabil mencegah anotator memakai label “sesuka hati” dan membuat hasil bisa dibandingkan antarpencarian (Xiao, 2008).

Menggunakan skema standar seperti Universal Dependencies membantu konsistensi lintas bahasa (Universal Dependencies, n.d.-a).

Contoh: Putuskan untuk memakai UPOS (NOUN, VERB, ADJ) dan konsisten menggunakannya di seluruh korpus.

Guidelines Harus Operasional (Ada Aturan Keputusan + Contoh)

Guidelines yang baik tidak hanya “definisi label”, tetapi juga:

- Aturan untuk kasus ambigu,
- Prioritas jika ada dua label mungkin,
- Contoh benar/salah.

Guidelines penting karena penandaan adalah keputusan analitis yang perlu ditulis agar transparan dan bisa diulang (Hovy & Lavid, 2010; Xiao, 2008).

Contoh (Bahasa Indonesia):

- “di” sebagai preposisi lokasi → ADP (*di rumah*)
- “di-” prefiks pasif → bagian VERB (*dimakan*), bukan ADP

Konsistensi Antar Anotator Harus Diuji (IAA) dan Perbedaan Diselesaikan (*Adjudication*)

Korpus yang baik tidak cukup “kelihatan rapi”; ia harus konsisten. Karena itu biasanya dilakukan:

- Penandaan ganda pada sebagian data,
- Uji *inter-annotator agreement* (IAA),
- *Adjudication* untuk menentukan label final.

Ini membuat kualitas penandaan terukur dan membantu memperbaiki pedoman bila banyak konflik (Hovy & Lavid, 2010).

Contoh: Dua anotator sering berbeda pada label INTJ vs ADJ (mis. “mantap!"). Maka pedoman ditambah aturan: jika berdiri

sendiri sebagai respons → INTJ; jika menerangkan nomina → ADJ.

Kelola Ketidakpastian: Tidak Semua Kasus Harus “Pasti”

Sebagian data memang ambigu. Pendekatan modern menekankan pentingnya memodelkan atau mencatat ketidakpastian (misalnya melalui label alternatif, confidence, atau catatan kasus) agar evaluasi lebih realistis dan tidak memaksa keputusan yang rapuh (Passonneau & Carpenter, 2014).

Contoh: Pada kata yang bisa NOUN/VERB, anotator boleh memberi catatan “ambiguous” untuk dibahas saat *adjudication*.

Quality Control dan Dokumentasi Versi (Versioning)

Penandaan harus disertai:

- Audit sampel berkala,
- Pelacakan versi skema/guidelines,
- Metadata (alat, tanggal, anotator, sumber data),
- Format rilis yang jelas.

Dokumentasi memastikan korpus bisa dipakai ulang dan analisis bisa direplikasi (Hovy & Lavid, 2010; Xiao, 2008).

Contoh: “Korpus v1.1 memperbaiki aturan tokenisasi ‘di-’ dan memperbarui label sesuai UD versi terbaru.”

Gunakan Pendekatan Semi-Otomatis Bila Skalanya Besar

Untuk korpus besar, praktik umum adalah semiotomatis: model memberi label awal, manusia mengoreksi. Ini menyeimbangkan efisiensi dan akurasi—dan disarankan dalam diskusi pengembangan korpus beranotasi (Xiao, 2008).

KESIMPULAN

Penandaan dalam linguistik korpus adalah proses pemberian informasi linguistik tambahan pada data bahasa agar korpus

menjadi lebih kaya, terstruktur, dan dapat dianalisis secara lebih mendalam. Penandaan berbeda dari markup struktural karena bersifat interpretatif; ia melibatkan keputusan analitis yang memerlukan skema, pedoman, dan kontrol kualitas.

Jenis penandaan mencakup POS *tagging*, *lemmatization*, fitur morfologis, sintaktis, semantik, hingga wacana. Di antara semuanya, POS *tagging* merupakan bentuk anotasi paling dasar dan paling luas digunakan, sekaligus fondasi bagi analisis tingkat lanjut. Contoh *British National Corpus* (BNC) pada bahasa Inggris memperlihatkan bagaimana penandaan skala besar memungkinkan penelitian linguistik yang sistematis, sedangkan contoh *Universal Dependencies* (UD) *Indonesian-Google Semantic Dependencies* (GSD) menunjukkan pentingnya standardisasi lintas bahasa sekaligus penyesuaian terhadap kekhasan bahasa Indonesia.

Bab ini juga menegaskan bahwa kualitas penandaan sangat menentukan validitas hasil penelitian dan performa aplikasi *Natural Language Processing* (NLP). Karena itu, praktik anotasi yang baik harus menggabungkan guideline yang jelas, pelatihan anotator, evaluasi reliabilitas, dan strategi semi-otomatis yang efisien. Dengan demikian, anotasi bukan sekadar tahap teknis, melainkan fondasi epistemologis dan metodologis dalam linguistik korpus modern.

DAFTAR PUSTAKA

- Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In A. Lüdeling & M. Kytö (Eds.) (Ed.), *Corpus linguistics: An international handbook* (pp. 613–642). Mouton de Gruyter.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
- BNC. (2009). *Linguistic Annotation of Texts* (“tagging”).

- Oxford University Computing Services / BNC Documentation.
<https://www.natcorp.ox.ac.uk/corpus/creating.xml?ID=annotation>
- Dependencies, U. (n.d.). *Universal Dependencies*.
<https://universaldependencies.org/>
- Dependencies, U. (n.d.). *Introduction*.
<https://universaldependencies.org/introduction>
- Gries, S. T. (2017). Linguistic annotation in/for corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 13(1), 1–16.
- Gries, S. T., & Berez, A. L. (2017). Handbook of Linguistic Annotation. In *Handbook of Linguistic Annotation*. Springer Science & Business Media. <https://doi.org/10.1007/978-94-024-0881-2>
- Hovy, E., & Lavid, J. (2010). Towards a ‘ Science ’ of Corpus Annotation : A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22(1), 25.
- Leech, G. (1997). Introducing corpus annotation. In *Corpus annotation: Linguistic information from computer text corpora* (pp. 1–18).
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Passonneau, R. J., & Carpenter, B. (2014). The benefits of a model of annotation. *Transaction of the Association for Computational Linguistics*, 2, 187–195.
https://doi.org/10.1162/tacl_a_00185
- Wynne, M. (Ed.). (2005). *Developing linguistic corpora: A guide to good practice*. Oxbow Books.
- Xiao, R. (2008). Corpus annotation. In *Using corpora in contrastive and translation studies*. Lancaster University.
https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBL/S/chapters/A04.pdf?utm_source=chatgpt.com

PROFIL PENULIS



Akhmad Hairul Umam

Penulis adalah dosen tetap pada Program Studi Ilmu Komunikasi, Fakultas Manajemen dan Kepemimpinan, Universitas Tanri Abeng, Jakarta. Ia lahir di Sumenep pada 6 Januari 1982 dan telah menekuni bidang pengajaran Bahasa Inggris serta kajian linguistik terapan sejak awal karier akademiknya. Penulis meraih gelar Sarjana (S1) di bidang Pendidikan

Bahasa Inggris dari UIN Syarif Hidayatullah Jakarta pada tahun 2004, kemudian melanjutkan pendidikan Magister (S2) dalam bidang Linguistik Terapan di Universitas Negeri Jakarta dan lulus pada tahun 2010. Gelar Doktor (S3) dalam bidang Linguistik Terapan Bahasa Inggris diselesaikan pada tahun 2024 di Universitas Katolik Atma Jaya, Jakarta, dengan fokus pada praktik-praktik translingual dalam pembelajaran bahasa Inggris yang dilakukan secara jarak jauh.

Dalam perjalanan akademiknya, penulis telah menerima berbagai hibah kompetitif nasional seperti Penelitian Dosen Pemula (PDP) dan Penelitian Disertasi Doktor (PDD) yang didanai oleh Kementerian Riset, Teknologi, dan Pendidikan Tinggi serta Kemendikbudristek. Pada tahun 2025, penulis kembali mendapatkan kepercayaan sebagai penerima Hibah Program Kemitraan Masyarakat (PKM) dari Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat (DRTPM), Kemendikbudristek. Dalam program ini, ia mengusung kegiatan bertema “Pemberdayaan Kemitraan Masyarakat sebagai Strategi *Green Economy* terhadap Pengelolaan Sampah Berbasis IPTEK pada Bank Sampah Tri Alam Lestari di Jakarta Selatan.

BAB 5

ALAT DAN PERANGKAT LUNAK LINGUISTIK KORPUS

Dwi Indarti
Universitas Bina Sarana Informatika, Jakarta
E-mail: dwi.diw@bsi.ac.id

PENDAHULUAN

Perkembangan linguistik korpus dalam dua dekade terakhir tidak dapat dipisahkan dari kemajuan teknologi komputasi dan perangkat lunak analisis bahasa. Linguistik korpus, yang pada dasarnya merupakan pendekatan empiris dalam studi bahasa berbasis data autentik, semakin bergantung pada alat dan perangkat lunak untuk mengelola, menganalisis, dan memvisualisasikan data bahasa dalam skala besar. Dalam konteks ini, perangkat lunak linguistik korpus menjadi komponen utama yang memungkinkan peneliti untuk mengidentifikasi pola linguistik secara sistematis dan efisien.

Sejumlah penelitian terbaru menegaskan bahwa integrasi antara keahlian linguistik dan teknologi komputasi telah menjadi ciri utama penelitian bahasa modern. Misalnya, Wan Jusoh, et al. (2024) menyatakan bahwa penggunaan alat komputasional memungkinkan peneliti untuk “mengidentifikasi pola, frekuensi, tren, dan variasi bahasa secara lebih efisien dalam korpus besar”. Hal ini menunjukkan bahwa tanpa bantuan perangkat lunak, analisis linguistik korpus modern akan menjadi sangat terbatas, terutama ketika berhadapan dengan big data bahasa.

Selain itu, perkembangan Natural Language Processing (NLP) juga berkontribusi secara signifikan terhadap peningkatan kemampuan alat linguistik korpus. Gorozhanov, et al. (2024) menekankan bahwa prosedur NLP memungkinkan analisis teks

yang lebih kompleks, termasuk analisis semantik dan diskursif dalam korpus sastra dan non-sastra. Dengan demikian, batas antara linguistik korpus dan linguistik komputasional semakin kabur, menciptakan pendekatan interdisipliner yang kuat.

Perangkat lunak seperti AntConc, #LancsBox, dan Wmatrix telah menjadi alat utama dalam analisis korpus. AntConc masih menjadi salah satu alat yang paling banyak digunakan karena kemudahan penggunaannya, sementara #LancsBox menawarkan fitur visualisasi yang lebih canggih. Hal ini menunjukkan adanya evolusi dari alat sederhana berbasis frekuensi menuju sistem analitik yang lebih kompleks dan interaktif.

Lebih lanjut, dalam konteks penelitian bahasa pembelajar (*learner corpus*), perangkat lunak memainkan peran penting dalam menganalisis kompleksitas sintaksis, kekayaan leksikal, dan pola argumentasi. Kyle (2021) menegaskan bahwa penggunaan alat NLP dalam korpus pembelajar memungkinkan analisis yang lebih mendalam terhadap struktur bahasa yang digunakan oleh pelajar.

Dengan demikian, bab ini bertujuan untuk membahas berbagai alat dan perangkat lunak dalam linguistik korpus, mulai dari perangkat dasar hingga sistem berbasis NLP dan machine learning. Pembahasan ini diharapkan dapat memberikan pemahaman komprehensif mengenai fungsi, kelebihan, dan keterbatasan masing-masing alat, serta relevansinya dalam penelitian linguistik kontemporer.

KLASIFIKASI ALAT LINGUISTIK KORPUS

Alat linguistik korpus dapat diklasifikasikan berdasarkan fungsi, kompleksitas, dan tujuan penggunaannya. Secara umum, alat tersebut dapat dibagi menjadi tiga kategori utama: (1) alat konkordansi (*concordancers*), (2) alat analisis statistik, dan (3) alat berbasis NLP dan *machine learning*.

Alat konkordansi merupakan jenis perangkat lunak paling dasar dalam linguistik korpus. Fungsinya adalah menampilkan kata atau frasa dalam konteks (*Key Word in Context/KWIC*). Contoh populer dari alat ini adalah *AntConc* dan *WordSmith Tools*. Alat ini memungkinkan peneliti untuk melihat pola penggunaan kata dalam konteks nyata, sehingga sangat berguna dalam analisis leksikal dan kolokasi. *WordSmith Tools*, misalnya, menyediakan fitur pencarian pola bahasa yang luas dan telah lama digunakan dalam penelitian linguistik.

Kategori kedua adalah alat analisis statistik, yang digunakan untuk mengukur frekuensi, distribusi, dan hubungan antar unsur bahasa. Dalam studi corpus-based terbaru, penggunaan alat statistik memungkinkan peneliti untuk membandingkan data dari berbagai korpus secara lebih akurat. Sebagai contoh, penelitian oleh Hiipala & Bateman (2022) menunjukkan bahwa ukuran statistik berbasis korpus dapat digunakan untuk membedakan antara infleksi dan derivasi dalam berbagai bahasa.

Kategori ketiga adalah alat berbasis NLP dan machine learning, yang merupakan perkembangan terbaru dalam linguistik korpus. Alat ini tidak hanya menganalisis data, tetapi juga mampu melakukan prediksi, klasifikasi, dan pemodelan bahasa. Gorozhanov, et al. (2024). menunjukkan bahwa integrasi NLP memungkinkan analisis teks yang lebih kompleks, termasuk analisis semantik dan pragmatic.

Selain klasifikasi berdasarkan fungsi, alat linguistik korpus juga dapat dibedakan berdasarkan aksesibilitasnya, yaitu *open-source* dan *proprietary software*. Alat open-source seperti *AntConc* dan *LanCSBox* memberikan fleksibilitas tinggi bagi peneliti, sementara software berbayar biasanya menawarkan fitur yang lebih lengkap.

Dengan berkembangnya teknologi, batas antara kategori-kategori ini semakin kabur. Banyak alat modern yang

menggabungkan berbagai fungsi dalam satu platform, seperti analisis frekuensi, visualisasi data, dan pemodelan statistik. Oleh karena itu, pemahaman terhadap klasifikasi alat menjadi penting untuk memilih perangkat yang sesuai dengan kebutuhan penelitian.

PERANGKAT LUNAK POPULER DALAM LINGUISTIK KORPUS

Berbagai perangkat lunak telah dikembangkan untuk mendukung penelitian linguistik korpus. Beberapa di antaranya telah menjadi standar dalam bidang ini karena kemudahan penggunaan dan fitur yang komprehensif.

AntConc merupakan salah satu perangkat lunak yang paling banyak digunakan dalam linguistik korpus, terutama dalam konteks pendidikan dan penelitian dasar. Menurut Staples & Anthony (2023), *AntConc* banyak digunakan dalam pengajaran EFL karena kemampuannya dalam menampilkan pola leksikogramatikal secara jelas. Keunggulan utama *AntConc* adalah antarmuka yang sederhana dan tidak memerlukan keahlian pemrograman. Selain *AntConc*, *#LancsBox* merupakan alat yang semakin populer karena menyediakan fitur visualisasi data yang canggih. *Wmatrix* juga merupakan alat penting yang digunakan untuk analisis semantik dan klasifikasi teks. Dalam penelitian tentang sinonim, *Wmatrix* digunakan untuk menganalisis kolokasi dan perbedaan makna secara sistematis. Alat ini sangat berguna dalam studi semantik korpus.

Selain itu, *Leipzig Corpus Miner* merupakan platform yang mengintegrasikan analisis kuantitatif dan kualitatif dalam satu sistem. Platform ini memungkinkan peneliti untuk melakukan analisis teks secara otomatis dengan dukungan antarmuka grafis dan bahasa pemrograman *R*.

Perangkat lunak lainnya termasuk *Sketch Engine*, yang dikenal dengan fitur word sketch dan analisis kolokasi yang

kuat. Meskipun tidak disebutkan dalam sumber di atas, alat ini широко digunakan dalam penelitian linguistik modern.

Secara keseluruhan, pemilihan perangkat lunak sangat bergantung pada tujuan penelitian, jenis data, dan tingkat keahlian pengguna. Peneliti pemula cenderung menggunakan AntConc, sementara peneliti lanjutan mungkin lebih memilih alat berbasis NLP atau platform yang lebih kompleks.

INTEGRASI NLP DALAM LINGUISTIK KORPUS

Integrasi Natural Language Processing (NLP) dalam linguistik korpus telah membawa perubahan signifikan dalam cara peneliti menganalisis bahasa. NLP memungkinkan analisis otomatis terhadap struktur bahasa, termasuk sintaksis, semantik, dan pragmatik.

Kyle (2021) menekankan bahwa penggunaan NLP dalam learner corpus memungkinkan analisis yang lebih mendalam terhadap struktur bahasa pelajar, termasuk kesalahan gramatikal dan variasi leksikal. Hal ini membuka peluang untuk penelitian yang lebih kompleks dan berbasis data besar. Selain itu, penelitian oleh Perkins & Roe (2024) menunjukkan bahwa analisis korpus dapat digunakan untuk mengkaji kebijakan akademik terkait AI, menunjukkan bahwa linguistik korpus dapat diterapkan dalam berbagai bidang, termasuk pendidikan dan kebijakan. Integrasi NLP juga memungkinkan penggunaan teknik machine learning, seperti clustering dan classification, dalam analisis korpus. Hal ini memungkinkan peneliti untuk mengidentifikasi pola yang tidak terlihat secara manual. Dengan demikian, NLP tidak hanya memperluas kemampuan analisis linguistik korpus, tetapi juga membuka peluang baru untuk penelitian interdisipliner.

TANTANGAN DAN KETERBATASAN ALAT LINGUISTIK KORPUS

Meskipun alat linguistik korpus telah berkembang pesat dan menawarkan berbagai keunggulan dalam analisis bahasa berbasis data, penggunaannya tidak terlepas dari sejumlah tantangan dan keterbatasan yang signifikan. Perkembangan teknologi digital, khususnya dalam bidang *Natural Language Processing* (NLP) dan machine learning, memang telah memperluas kemungkinan analisis linguistik secara kuantitatif dan empiris. Namun demikian, berbagai kendala metodologis, teknis, serta konseptual masih menjadi isu penting yang perlu dipertimbangkan secara kritis oleh para peneliti. Oleh karena itu, pemahaman yang komprehensif mengenai keterbatasan ini sangat penting agar hasil penelitian berbasis korpus dapat diinterpretasikan secara tepat dan tidak menimbulkan bias atau kesimpulan yang keliru.

Salah satu tantangan utama dalam penggunaan alat linguistik korpus adalah keterbatasan dalam menangani keragaman bahasa, khususnya bahasa dengan karakteristik morfologi yang kompleks. Sebagian besar perangkat lunak korpus, seperti AntConc, pada awalnya dikembangkan dengan fokus pada bahasa Inggris yang relatif memiliki struktur morfologi yang lebih sederhana dibandingkan dengan bahasa-bahasa lain. Hal ini menyebabkan performa alat tersebut menjadi kurang optimal ketika digunakan untuk menganalisis bahasa yang memiliki sistem morfologi yang kaya dan kompleks, seperti bahasa aglutinatif (misalnya bahasa Korea, Turki, atau Indonesia dalam beberapa konteks), serta bahasa dengan sistem infleksi yang tinggi.

Korpus yang ideal seharusnya mencerminkan variasi bahasa yang luas, mencakup berbagai genre, register, konteks sosial, serta latar belakang penutur. Namun, dalam praktiknya, mengumpulkan data yang representatif bukanlah hal yang

mudah. Proses pengumpulan data sering kali dibatasi oleh faktor aksesibilitas, hak cipta, serta keterbatasan sumber daya. Selain itu, peneliti juga harus mempertimbangkan keseimbangan data (*balance*) dalam korpus, misalnya proporsi antara teks lisan dan tulisan, atau antara berbagai jenis teks seperti akademik, jurnalistik, dan percakapan sehari-hari.

Tantangan berikutnya berkaitan dengan kebutuhan akan keahlian teknis dalam penggunaan alat linguistik korpus. Seiring dengan perkembangan teknologi, banyak alat korpus modern yang mengintegrasikan teknik NLP dan machine learning. Meskipun hal ini membuka peluang analisis yang lebih kompleks, seperti analisis sentimen, *topic modeling*, atau *word embeddings*, penggunaan teknik tersebut sering kali memerlukan pemahaman tentang pemrograman dan statistik. Bagi peneliti linguistik tradisional yang tidak memiliki latar belakang teknis, hal ini dapat menjadi hambatan yang signifikan.

Keterbatasan lainnya adalah masalah aksesibilitas dan lisensi data. Banyak korpus besar yang berkualitas tinggi, seperti korpus akademik atau korpus berbasis jurnal, tidak tersedia secara bebas karena alasan hak cipta. Hal ini membatasi akses bagi peneliti, terutama di negara berkembang atau institusi dengan sumber daya terbatas. Akibatnya, peneliti sering kali harus menggunakan korpus yang tersedia secara gratis tetapi mungkin memiliki kualitas yang lebih rendah atau cakupan yang terbatas.

Selain itu, isu etika juga menjadi perhatian dalam penggunaan korpus, terutama yang berasal dari data daring seperti media sosial. Pengumpulan dan penggunaan data dari platform digital menimbulkan pertanyaan tentang privasi, persetujuan pengguna, dan penggunaan data secara bertanggung jawab. Peneliti harus memastikan bahwa data yang digunakan tidak melanggar hak individu dan sesuai dengan standar etika penelitian.

Tantangan lain yang perlu diperhatikan adalah keterbatasan alat dalam menangkap aspek pragmatik dan konteks sosial bahasa. Linguistik korpus cenderung berfokus pada pola penggunaan bahasa yang dapat diukur secara kuantitatif, tetapi sering kali kurang mampu menangkap makna implisit, ironi, atau nuansa budaya yang terkandung dalam bahasa. Hal ini menjadi keterbatasan penting, terutama dalam analisis wacana atau studi pragmatik.

Terakhir, terdapat pula isu terkait dengan standarisasi dalam penggunaan alat dan metode korpus. Berbagai alat korpus memiliki fitur, format data, dan metode analisis yang berbeda-beda. Hal ini dapat menyulitkan perbandingan hasil penelitian antardisiplin. Kurangnya standarisasi juga dapat menyebabkan inkonsistensi dalam metodologi penelitian, sehingga memengaruhi replikasi dan validitas hasil.

PANDUAN PRAKTIS: PENGGUNAAN ANTCONC (STEP-BY-STEP)

Analisis Korpus Menggunakan AntConc

AntConc adalah perangkat lunak gratis yang sangat populer untuk analisis korpus. Berikut langkah-langkah penggunaannya:

Langkah 1: Persiapan Data

- Siapkan file teks (.txt)
- Pastikan encoding UTF-8
- Gabungkan beberapa file jika diperlukan

Langkah 2: Membuka Korpus

1. Buka AntConc
2. Klik **File** → **Open Files**
3. Pilih file teks yang akan dianalisis

Langkah 3: Analisis Concordance (KWIC)

1. Klik tab **Concordance**
2. Masukkan kata kunci (misalnya: *education*)
3. Klik **Start**

Output: daftar kalimat dengan kata tersebut dalam konteks

Langkah 4: Analisis Word List

1. Klik tab **Word List**
2. Klik **Start**

Output: daftar frekuensi kata

Langkah 5: Collocation Analysis

1. Klik tab **Collocates**
2. Masukkan kata target
3. Klik **Start**

Output: kata-kata yang sering muncul bersama.

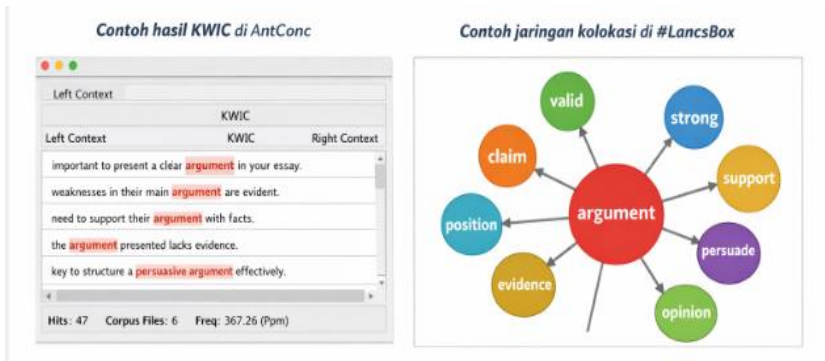
Langkah 6: N-gram Analysis

1. Klik tab **Clusters/N-Grams**
2. Tentukan panjang (misal 2–4 kata)
3. Klik **Start**

Output: frasa umum dalam korpus

Contoh hasil analisis:

- “Make decision”
- “Take action”
- “Global warming”



Gambar 5.1. Contoh tampilan hasil analisis di antconc dan #LancsBox

PANDUAN PRAKTIS: PENGGUNAAN #LANCSBOX (STEP-BY-STEP)

Analisis Korpus dengan #LancsBox

Langkah 1: Import Data

1. Buka #LancsBox
2. Pilih **Corpora** → **Add Corpus**
3. Import file teks

Langkah 2: GraphColl (Kolokasi Visual)

1. Pilih **GraphColl**
2. Masukkan kata target (misalnya: *language*)
3. Klik **Run**

Output: grafik jaringan kata

Langkah 3: KWIC Concordance

1. Pilih **KWIC**
2. Masukkan kata
3. Klik **Search**

Output: konteks penggunaan kata

Langkah 4: Frequency Analysis

1. Pilih **Words**
2. Klik **Calculate**

Output: frekuensi kata

Langkah 5: Dispersion Plot

1. Pilih **Dispersion**
2. Masukkan kata target

Output: distribusi kata dalam teks

Keunggulan utama:

- Visualisasi jaringan kata
- Analisis berbasis graf
- Lebih intuitif untuk penelitian wacana

KESIMPULAN

Berdasarkan pembahasan yang telah dipaparkan, dapat disimpulkan bahwa perkembangan linguistik korpus sangat erat kaitannya dengan kemajuan teknologi komputasi yang memungkinkan analisis bahasa dilakukan secara lebih sistematis, objektif, dan berbasis data dalam skala besar. Dari segi klasifikasi, alat linguistik korpus dapat dibedakan menjadi tiga kategori utama, yaitu alat konkordansi, alat analisis statistik, serta alat berbasis NLP dan machine learning. Masing-masing kategori memiliki fungsi dan keunggulan tersendiri. Alat konkordansi, misalnya, sangat efektif untuk menampilkan kata dalam konteks (KWIC), sehingga membantu peneliti memahami penggunaan kata dalam konteks nyata. Sementara itu, alat analisis statistik memungkinkan pengukuran frekuensi dan distribusi bahasa secara kuantitatif. Di sisi lain, alat berbasis

NLP menawarkan kemampuan analisis yang lebih kompleks, seperti pemodelan bahasa dan klasifikasi teks.

Perkembangan perangkat lunak dalam linguistik korpus juga menunjukkan adanya pergeseran dari alat yang sederhana menuju sistem yang lebih canggih dan terintegrasi. AntConc, sebagai salah satu alat yang paling populer, menawarkan kemudahan penggunaan dan sangat cocok bagi peneliti pemula. Sebaliknya, #LancsBox menyediakan fitur visualisasi yang lebih kompleks, seperti jaringan kolokasi, yang memungkinkan peneliti untuk memahami hubungan semantik secara lebih mendalam. Selain itu, perangkat seperti Leipzig Corpus Miner dan Sketch Engine menunjukkan bahwa analisis korpus kini semakin terintegrasi dengan teknik pemrograman dan analisis data lanjutan.

Integrasi NLP dalam linguistik korpus menjadi salah satu perkembangan paling signifikan dalam bidang ini. Teknologi ini memungkinkan analisis bahasa yang lebih mendalam, tidak hanya pada level leksikal dan sintaksis, tetapi juga pada level semantik dan pragmatik. Dengan bantuan NLP, peneliti dapat melakukan analisis otomatis terhadap data korpus dalam jumlah besar, termasuk mendeteksi pola yang tidak mudah terlihat secara manual. Selain itu, penggunaan machine learning dalam analisis korpus membuka peluang baru untuk penelitian interdisipliner yang menggabungkan linguistik, ilmu komputer, dan data science. Namun demikian, meskipun alat linguistik korpus menawarkan berbagai keunggulan, terdapat sejumlah tantangan dan keterbatasan yang perlu diperhatikan. Salah satu tantangan utama adalah keterbatasan alat dalam menangani bahasa dengan struktur morfologi yang kompleks. Banyak perangkat lunak yang dikembangkan dengan fokus pada bahasa Inggris, sehingga kurang optimal ketika digunakan untuk bahasa lain dengan karakteristik yang berbeda. Selain itu, kualitas hasil analisis sangat bergantung pada kualitas dan representativitas

korpus yang digunakan. Korpus yang tidak seimbang atau tidak representatif dapat menghasilkan temuan yang bias dan kurang valid.

Dari sisi metodologis, linguistik korpus juga memiliki keterbatasan dalam menangkap aspek pragmatik dan konteks sosial bahasa. Meskipun alat korpus sangat efektif dalam mengidentifikasi pola yang dapat diukur secara kuantitatif, aspek makna implisit, ironi, dan nuansa budaya sering kali sulit dianalisis secara otomatis. Oleh karena itu, pendekatan korpus sebaiknya dikombinasikan dengan pendekatan kualitatif untuk menghasilkan analisis yang lebih komprehensif. Selain itu, penting untuk dicatat bahwa penggunaan alat linguistik korpus tidak dapat menggantikan peran peneliti sepenuhnya. Alat tersebut hanya berfungsi sebagai sarana untuk memproses dan menganalisis data, sementara interpretasi hasil tetap memerlukan keahlian dan pemahaman teoritis dari peneliti. Dengan kata lain, keberhasilan penelitian korpus tidak hanya ditentukan oleh kecanggihan alat yang digunakan, tetapi juga oleh kemampuan peneliti dalam merancang penelitian, memilih data yang tepat, serta menafsirkan hasil secara kritis.

Secara keseluruhan, dapat disimpulkan bahwa alat dan perangkat lunak linguistik korpus memainkan peran yang sangat penting dalam penelitian bahasa modern. Perkembangan teknologi telah memperluas cakupan dan kedalaman analisis yang dapat dilakukan, sekaligus membuka peluang baru untuk penelitian interdisipliner. Namun demikian, penggunaan alat tersebut harus dilakukan secara kritis dengan mempertimbangkan keterbatasan yang ada. Ke depan, perkembangan linguistik korpus diperkirakan akan semakin terintegrasi dengan teknologi kecerdasan buatan dan analisis big data. Oleh karena itu, peneliti linguistik perlu terus mengembangkan kompetensi digital dan metodologis agar dapat memanfaatkan potensi alat-alat tersebut secara optimal.

DAFTAR PUSTAKA

- Gorozhanov, A., Guyesnova, I., Stepanova, D. (2024). Natural language processing and fiction text: Basis for corpus research. *Journal of Language Studies, Semiotics, and Semantics*, 15(1), 195–210. <https://doi.org/10.22363/2313-2299-2024-15-1-195-210>
- Hiipala, T., & Bateman, J. (2022). Semiotically-grounded distant viewing of diagrams: Insights from two multimodal corpora. *Digital Scholarship in the Humanities*, 37(2), 405–425. <https://doi.org/https://doi.org/10.1093/llc/fqab063>
- Kyle, K. (2021). Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, 7(1), 1–16. <https://doi.org/https://doi.org/10.1075/iklcr.00019.int>
- Perkins, M., & Roe, J. (2024). Decoding academic integrity policies: A corpus linguistics investigation of AI and other technological threats. *High Edu Policy*, 37, 633–653. <https://doi.org/https://doi.org/10.1057/s41307-023-00323-2>
- Staples, S., & Anthony, L. (2023). Expositions: Corpus linguistics in EFL language teaching: Insights from research and practice. *Japan Association for Language Teaching Journal*, 45(2), 246–363. <https://doi.org/http://doi.org/10.37546/JALTJJ45.2-4>
- Wan Jusoh, W., Jobar, N., Yusoff, M., & Mahat, H. (2024). Exploring corpus linguistics via computational tool analysis: Key finding review. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(2), 1052–1062. <https://doi.org/https://doi.org/10.11591/ijeecs.v34.i2.pp1052-1062>

PROFIL PENULIS



Dwi Indarti

Penulis berdomisili di Kabupaten Tangerang, Banten, tetapi sebagian besar aktivitasnya berlangsung di Jakarta. Saat ini, ia sedang menempuh pendidikan doktoral dengan konsentrasi Linguistik Terapan Bahasa Inggris di Universitas Katolik Indonesia Atma Jaya. Bidang penelitiannya mencakup linguistik terapan, sastra, penerjemahan, dan linguistik korpus. Ia telah mempublikasikan artikel penelitian pada berbagai jurnal nasional terakreditasi, seperti *Studies in English Language and Education*, *Journal of Languages and Language Teaching*, *Lingua Cultura*, dan jurnal lainnya.

BAB 6

ANALISIS FREKUENSI DAN DISTRIBUSI KORPUS

Stevanus Ngenget
Universitas Katolik De La Salle Manado
E-mail: sngenget@unikadelasalle.ac.id

PENDAHULUAN

Linguistik korpus berkembang menjadi salah satu pendekatan paling berpengaruh dalam penelitian bahasa modern karena kemampuannya menghadirkan bukti empiris dari penggunaan bahasa nyata. Jika pada masa sebelumnya analisis bahasa banyak bertumpu pada intuisi peneliti atau contoh-contoh terbatas, pendekatan korpus memungkinkan analisis dilakukan terhadap jutaan kata yang berasal dari teks autentik. Perubahan ini membawa dampak besar terhadap cara peneliti memahami struktur bahasa, pola penggunaan kata, hingga variasi bahasa dalam masyarakat. Dalam konteks tersebut, analisis frekuensi dan distribusi menjadi titik awal yang sangat penting. Peneliti biasanya memulai eksplorasi korpus dengan melihat kata apa yang paling sering muncul, bagaimana penyebarannya dalam berbagai teks, dan dalam konteks apa kata tersebut digunakan. Walaupun tampak sederhana, analisis frekuensi sebenarnya mampu membuka banyak temuan linguistik yang mendalam.

Frekuensi kata sering kali mencerminkan kepentingan komunikatif suatu unsur bahasa. Kata-kata yang paling sering digunakan biasanya berkaitan dengan kebutuhan utama penutur dalam menyampaikan makna. Oleh sebab itu, penelitian berbasis frekuensi tidak hanya digunakan dalam linguistik murni, tetapi juga dalam pengajaran bahasa, penyusunan kamus, analisis media, penerjemahan, hingga pengembangan teknologi bahasa

berbasis kecerdasan buatan. Di era digital saat ini, jumlah data bahasa meningkat secara eksponensial melalui media sosial, portal berita daring, jurnal elektronik, hingga komunikasi berbasis aplikasi pesan instan. Kondisi tersebut membuat pendekatan korpus semakin relevan karena peneliti membutuhkan metode yang mampu mengolah data bahasa dalam jumlah besar secara sistematis. Analisis frekuensi membantu peneliti menyederhanakan kompleksitas data tersebut menjadi pola-pola yang dapat diinterpretasikan.

Penelitian Xu dan Zhang (2021) menunjukkan bahwa frekuensi kata masih menjadi dasar penting dalam ekstraksi kata kunci otomatis pada sistem NLP modern. Sementara itu, Omarova et al. (2025) memperlihatkan bagaimana distribusi frekuensi kata dalam bahasa Kazakh dapat mengungkap karakteristik tipologis dan gramatikal bahasa tersebut. Temuan-temuan semacam ini memperlihatkan bahwa analisis frekuensi tidak lagi sekadar statistik linguistik sederhana, melainkan bagian penting dari kajian bahasa kontemporer. Selain memberikan gambaran kuantitatif, analisis frekuensi juga membantu peneliti memahami perubahan bahasa dari waktu ke waktu. Kata tertentu dapat mengalami peningkatan atau penurunan frekuensi akibat perubahan sosial, politik, budaya, maupun perkembangan teknologi. Misalnya, kata seperti *digital transformation*, *AI*, dan *ChatGPT* mengalami lonjakan penggunaan dalam korpus media global sejak perkembangan teknologi kecerdasan buatan generatif.

Bab ini membahas secara mendalam konsep analisis frekuensi dan distribusi dalam linguistik korpus dengan memadukan landasan teoretis, contoh praktis, dan hasil penelitian terbaru. Pembahasan difokuskan pada penghitungan frekuensi kata, kolokasi, keyness, distribusi kata dan frasa, interpretasi hasil frekuensi, serta implikasi analisis frekuensi dalam penelitian bahasa modern.

KONSEP DASAR FREKUENSI DALAM KAJIAN KORPUS

Dalam linguistik korpus, frekuensi dipandang sebagai salah satu indikator paling penting dalam memahami perilaku bahasa . Pendekatan berbasis frekuensi berangkat dari asumsi bahwa unsur linguistik yang paling sering muncul cenderung memiliki fungsi sentral dalam komunikasi. Oleh karena itu, analisis frekuensi bukan sekadar proses statistik, melainkan sarana untuk memahami bagaimana bahasa digunakan dalam konteks sosial dan komunikatif yang nyata.

Hejazi (2022) menjelaskan bahwa pendekatan korpus memungkinkan peneliti mengidentifikasi pola penggunaan bahasa yang tidak mudah diamati melalui introspeksi. Pendekatan ini memperlihatkan bahwa bahasa memiliki kecenderungan repetitif dan sistematis. Dengan bantuan perangkat lunak korpus, jutaan kata dapat dianalisis untuk menemukan pola distribusi, hubungan leksikal, dan variasi penggunaan bahasa. Dalam kajian linguistik modern, frekuensi juga berkaitan dengan teori probabilistik bahasa. Kata atau struktur yang memiliki frekuensi tinggi dianggap lebih mudah diproses oleh otak manusia karena sering terekspos dalam pengalaman linguistik sehari-hari. Oleh sebab itu, frekuensi kata memiliki hubungan erat dengan pemerolehan bahasa, pemrosesan bahasa, dan kemampuan membaca.

Penelitian (Pereira et al., 2022) menunjukkan bahwa kata dengan frekuensi tinggi menghasilkan waktu fiksasi mata yang lebih singkat dibandingkan dengan kata dengan frekuensi rendah. Hal ini membuktikan bahwa otak manusia memproses kata yang familiar secara lebih cepat dan efisien. Dalam konteks ini, frekuensi menjadi indikator penting dalam psikolinguistik dan penelitian membaca. Selain itu, frekuensi juga menjadi dasar penting dalam pengembangan teknologi bahasa berbasis kecerdasan buatan. Banyak algoritma NLP seperti TF-IDF,

language modelling, text classification, dan keyword extraction memanfaatkan distribusi frekuensi sebagai dasar pemrosesan teks. Xu & Zhang (2021) menegaskan bahwa kombinasi frekuensi dan asosiasi kata menghasilkan performa ekstraksi kata kunci yang lebih akurat dibandingkan dengan metode tradisional.

Dalam praktik penelitian, analisis frekuensi biasanya dilakukan terhadap beberapa jenis unit linguistik, antara lain:

1. Kata tunggal (*single words*)
2. Frasa atau n-grams
3. Lemma
4. Bentuk gramatikal
5. Struktur sintaksis tertentu
6. Kelas kata

Analisis terhadap berbagai unit tersebut memungkinkan peneliti memperoleh gambaran yang lebih komprehensif mengenai struktur dan pola penggunaan bahasa.

PERANGKAT LUNAK ANALISIS FREKUENSI

Perkembangan linguistik korpus sangat dipengaruhi oleh perkembangan perangkat lunak analisis teks. Saat ini terdapat berbagai perangkat lunak yang digunakan untuk melakukan penghitungan frekuensi dan distribusi kata secara otomatis. Beberapa perangkat lunak yang paling umum digunakan antara lain: AntConc, Sketch Engine, LancsBox, WordSmith Tools, LexiTurk, dan Voyant Tools.

AntConc merupakan salah satu perangkat lunak paling populer karena bersifat gratis dan mudah digunakan. Program ini memungkinkan pengguna membuat daftar frekuensi kata (*word list*), konkordansi, kolokasi, dan klaster. Sketch Engine menyediakan fitur analisis yang lebih kompleks seperti *word sketch*, thesaurus otomatis, dan keyword extraction. Program ini

banyak digunakan dalam penelitian akademik skala besar dan leksikografi.

Sementara itu, Başaran (2022) memperkenalkan LexiTürk sebagai alat analisis frekuensi untuk buku ajar bahasa Turki. Program tersebut dirancang untuk membantu peneliti dan pengembang materi ajar mengidentifikasi distribusi kosakata dalam buku teks bahasa.

Penggunaan perangkat lunak memberikan beberapa keuntungan utama:

1. Mempercepat analisis data dalam jumlah besar
2. Mengurangi kesalahan manual
3. Memungkinkan visualisasi data
4. Mendukung analisis statistik lanjutan
5. Mempermudah replikasi penelitian

Namun demikian, hasil analisis perangkat lunak tetap memerlukan interpretasi linguistik yang cermat. Perangkat lunak hanya menghasilkan data statistik, sedangkan makna linguistik tetap harus ditafsirkan oleh peneliti.

TEKNIK IDENTIFIKASI DAN PENGUKURAN FREKUENSI

Penghitungan frekuensi kata merupakan langkah pertama yang hampir selalu dilakukan dalam penelitian linguistik korpus (Chiarcos et al., 2022). Melalui analisis ini, peneliti dapat mengetahui kata atau unsur linguistik apa yang paling dominan dalam suatu korpus. Informasi tersebut menjadi dasar untuk analisis lanjutan seperti kolokasi, keyword analysis, discourse analysis, hingga text mining. Secara sederhana, frekuensi kata menunjukkan jumlah kemunculan suatu kata dalam korpus tertentu. Akan tetapi, dalam praktik penelitian, konsep ini jauh lebih kompleks dibandingkan sekadar menghitung jumlah kata. Peneliti perlu mempertimbangkan ukuran korpus, jenis teks,

variasi genre, hingga proses normalisasi data agar hasil analisis dapat dibandingkan secara valid.

Dalam linguistik korpus dikenal dua jenis frekuensi utama, yaitu frekuensi absolut dan frekuensi relatif. Frekuensi absolut menunjukkan jumlah kemunculan langsung suatu kata. Sebagai contoh, jika kata *research* muncul sebanyak 1.245 kali dalam korpus, maka angka tersebut merupakan frekuensi absolutnya. Sebaliknya, frekuensi relatif menunjukkan jumlah kemunculan yang telah dinormalisasi terhadap ukuran korpus. Normalisasi penting dilakukan karena ukuran korpus dapat berbeda secara signifikan. Sebuah kata yang muncul 500 kali dalam korpus kecil mungkin sebenarnya lebih dominan dibandingkan kata yang muncul 1.000 kali dalam korpus besar.

Selain itu, analisis frekuensi juga melibatkan konsep *type* dan *token*. Token merujuk pada keseluruhan kata dalam korpus, sedangkan *type* merujuk pada jumlah kata unik. Hubungan antara keduanya dikenal sebagai *Type-Token Ratio* (TTR). Semakin tinggi TTR, semakin tinggi pula keragaman leksikal suatu teks. Dalam praktiknya, analisis frekuensi sering digunakan untuk mengidentifikasi kosakata inti dalam suatu bidang ilmu. Penelitian Prapobratanakul (2024), misalnya, menunjukkan bahwa artikel penelitian material science memiliki dominasi kosakata akademik tertentu yang secara konsisten muncul dalam jurnal ilmiah bidang tersebut. Temuan tersebut membantu mahasiswa dan peneliti memahami kosakata yang paling relevan dalam disiplin ilmu tertentu.

Analisis frekuensi juga sangat penting dalam pengajaran bahasa. Başaran (2022) menjelaskan bahwa buku ajar bahasa asing sebaiknya dikembangkan berdasarkan kosakata dengan frekuensi tinggi agar pembelajar memperoleh akses lebih cepat terhadap kata-kata yang paling sering digunakan dalam komunikasi. Pendekatan ini memungkinkan proses pembelajaran menjadi lebih efektif dan kontekstual.

Dalam bidang psikolinguistik, frekuensi kata berkaitan erat dengan proses pemahaman bacaan. Pereira, Costa, dan Guerreiro (2022) menemukan bahwa kata dengan frekuensi tinggi diproses lebih cepat oleh pembaca dibandingkan kata dengan frekuensi rendah. Melalui teknologi eye-tracking, mereka menunjukkan bahwa kata yang jarang digunakan menghasilkan waktu fiksasi mata yang lebih panjang dan lebih banyak regresi saat membaca. Fakta tersebut menunjukkan bahwa frekuensi bukan hanya konsep statistik, tetapi juga berkaitan langsung dengan proses kognitif manusia dalam memahami bahasa. Kata yang lebih sering terekspos dalam pengalaman linguistik cenderung lebih mudah dikenali dan diproses oleh otak.

Dalam era digital, penghitungan frekuensi kini dilakukan menggunakan berbagai perangkat lunak korpus seperti AntConc, Sketch Engine, LancsBox, dan WordSmith Tools. Program-program tersebut mampu menganalisis jutaan kata dalam waktu singkat. Peneliti dapat menghasilkan daftar frekuensi, kolokasi, klaster, dan concordance lines secara otomatis.

PENGHITUNGAN FREKUENSI KATA

Penghitungan frekuensi kata merupakan teknik dasar dalam linguistik korpus yang digunakan untuk mengetahui jumlah kemunculan suatu kata, frasa, atau bentuk linguistik tertentu dalam korpus. Analisis ini memungkinkan peneliti mengidentifikasi unsur bahasa yang dominan, kosakata inti (*core vocabulary*), dan karakteristik register suatu teks. Secara umum, frekuensi dibedakan menjadi frekuensi absolut dan frekuensi relatif. Frekuensi absolut menunjukkan jumlah kemunculan langsung suatu kata dalam korpus, sedangkan frekuensi relatif menunjukkan jumlah kemunculan yang telah dinormalisasi berdasarkan ukuran korpus, misalnya per satu juta

kata. Normalisasi penting dilakukan agar perbandingan antara korpus yang berbeda ukuran tetap valid.

Dalam analisis korpus modern, konsep *type* dan *token* juga sangat penting. *Token* merujuk pada seluruh jumlah kata dalam korpus, sedangkan *type* merujuk pada jumlah kata unik. Perbandingan keduanya dikenal sebagai *Type-Token Ratio* (TTR) yang digunakan untuk mengukur keragaman leksikal suatu teks.

Sebagai ilustrasi, Tabel 6.1 menunjukkan contoh hasil analisis frekuensi pada korpus artikel akademik.

Tabel 6.1. Contoh Frekuensi Kata dalam Korpus Akademik

No	Kata	Frekuensi Absolut	Frekuensi Relatif
1	research	1.245	450/million
2	data	1.180	426/million
3	analysis	1.032	372/million
4	learning	924	333/million
5	students	876	315/million

Sumber: Simulasi Analisis AntConc (2026)

Tabel tersebut menunjukkan bahwa kata seperti *research*, *data*, dan *analysis* memiliki frekuensi tinggi karena merupakan unsur utama dalam register akademik. Frekuensi tersebut merepresentasikan fokus utama diskursus ilmiah pada penelitian, pengolahan data, dan analisis. Başaran (2022) menjelaskan bahwa analisis frekuensi sangat penting dalam pengembangan buku ajar bahasa karena kosakata dengan frekuensi tinggi lebih relevan untuk diprioritaskan dalam pembelajaran bahasa asing. Pembelajaran berbasis frekuensi memungkinkan pembelajar memperoleh akses lebih cepat terhadap kata-kata yang paling sering digunakan dalam komunikasi.

Selain dalam pengajaran bahasa, frekuensi kata juga memiliki implikasi penting dalam psikolinguistik. Pereira,

Costa, dan Guerreiro (2022) menunjukkan bahwa kata dengan frekuensi tinggi diproses lebih cepat oleh pembaca dibandingkan kata dengan frekuensi rendah. Dalam penelitian eye-tracking mereka, kata yang jarang muncul memerlukan durasi fiksasi mata yang lebih lama dan menghasilkan lebih banyak regresi mata selama membaca. Dalam linguistik korpus digital, penghitungan frekuensi kini tidak hanya digunakan untuk analisis statistik sederhana, tetapi juga untuk representasi semantik dan integrasi linked linguistic data. Chiarcos et al. (2022) melalui model OntoLex-FrAC mengembangkan representasi frekuensi, kolokasi, dan informasi berbasis korpus dalam format interoperabel yang mendukung pengembangan teknologi bahasa dan leksikografi digital.

VISUALISASI FREKUENSI KATA MENGGUNAKAN WORD CLOUD

Selain disajikan dalam bentuk tabel dan statistik numerik, data frekuensi dalam linguistik korpus juga dapat divisualisasikan menggunakan *word cloud*. Word cloud merupakan representasi visual yang menampilkan kata berdasarkan tingkat frekuensi kemunculannya dalam korpus. Semakin sering suatu kata muncul, semakin besar ukuran kata tersebut dalam visualisasi.

Dalam penelitian bahasa digital dan media sosial, word cloud banyak digunakan untuk membantu peneliti mengidentifikasi tema dominan secara cepat dan intuitif. Visualisasi ini sangat membantu terutama pada korpus berukuran besar karena mampu menyederhanakan kompleksitas data tekstual menjadi representasi visual yang mudah dipahami.

Saat ini, visualisasi word cloud dapat dibuat menggunakan bahasa pemrograman Python dengan bantuan pustaka seperti wordcloud, matplotlib, dan pandas. Pendekatan ini memungkinkan peneliti melakukan integrasi antara analisis

frekuensi otomatis dan visualisasi data dalam satu alur kerja penelitian. Sebagai contoh, Ngenget (2025) menggunakan visualisasi word cloud untuk menganalisis persepsi publik terhadap Tomohon International Flower Festival (TIFF) berdasarkan data digital. Visualisasi tersebut membantu menunjukkan dominasi kata-kata tertentu yang berkaitan dengan makna sosial, ekonomi, dan digital dalam diskursus publik mengenai festival tersebut.

Gambar 6.1. Contoh Word Cloud

Kolokasi merujuk pada kecenderungan dua atau lebih kata muncul bersama secara signifikan dalam korpus (Meisenbacher et al., 2024). Dalam linguistik korpus, kolokasi dianggap penting karena bahasa tidak digunakan secara acak, melainkan membentuk pola hubungan leksikal tertentu. Sebagai contoh, kata *strong* lebih sering berkolokasi dengan *argument* atau *evidence*, sedangkan kata *powerful* lebih lazim berkolokasi dengan *engine* atau *computer*. Meskipun kedua kata tersebut memiliki makna yang mirip, distribusi kolokasionalnya berbeda.

Analisis kolokasi biasanya menggunakan ukuran statistik seperti:

1. Mutual Information (MI)
2. t-score
3. Log-likelihood
4. Dice coefficient

Berikut ilustrasi hasil analisis kolokasi kata *innovation* dalam korpus pendidikan tinggi.

Tabel 6.2. Contoh Kolokasi Kata “Innovation”

No	Kolokasi	MI Score
1	digital innovation	8.5
2	educational innovation	7.9
3	innovation strategy	7.2
4	technological innovation	6.8

Sumber: Simulasi Sketch Engine (2026)

Tabel tersebut menunjukkan bahwa kata *innovation* memiliki hubungan kuat dengan bidang pendidikan dan teknologi. Hal ini mengindikasikan bahwa inovasi dalam konteks akademik sering dikaitkan dengan transformasi digital dan strategi institusional. Kolokasi juga penting dalam pengajaran bahasa. Başaran (2022) menjelaskan bahwa pembelajar bahasa asing sering mengalami kesulitan memahami kombinasi kata alami karena mereka cenderung menerjemahkan kata secara literal. Oleh sebab itu, pendekatan berbasis korpus membantu pembelajar memahami pola penggunaan bahasa yang lebih autentik. Selain itu, kolokasi banyak digunakan dalam text mining dan NLP. Xu dan Zhang (2021) menemukan bahwa kombinasi frekuensi kata dan asosiasi kolokasional menghasilkan performa ekstraksi kata kunci yang lebih baik dibandingkan dengan pendekatan berbasis frekuensi semata.

IDENTIFIKASI KEKHASAN LEKSIKAL DALAM KORPUS

Keyness merupakan ukuran statistik yang digunakan untuk menentukan kata-kata yang secara signifikan lebih sering muncul dalam suatu korpus dibandingkan korpus referensi (Anthony et al., 2023; Hejazi, 2022). Kata-kata tersebut disebut sebagai *keywords*. Analisis *keyness* umumnya menggunakan uji statistik seperti log-likelihood dan chi-square. Kata dengan nilai *keyness* tinggi dianggap sebagai representasi khas suatu domain, genre, atau register. Sebagai ilustrasi, dalam korpus berita politik, kata seperti *government*, *election*, dan *policy* biasanya memiliki nilai *keyness* yang lebih tinggi dibandingkan dengan korpus umum.

Tabel 6.3. Contoh Analisis Keyness pada Korpus Politik

No	Kata	Keyness Score
1	election	356.4
2	government	341.2
3	policy	298.7
4	parliament	264.5

Sumber: Simulasi Analisis Keyness (2026)

Analisis *keyness* sangat penting dalam analisis wacana kritis karena dapat membantu peneliti mengidentifikasi fokus ideologis dan tema dominan suatu teks. Prapobtanakul (2024) menggunakan analisis frekuensi dan *keyness* untuk mengidentifikasi kosakata akademik yang dominan dalam artikel penelitian material science. Penelitian tersebut menemukan bahwa sekitar 75,97% kosakata artikel ilmiah material science tercakup dalam *New Academic Word List* (NAWL), menunjukkan pentingnya daftar kosakata akademik dalam memahami teks ilmiah.

PENYEBARAN KATA DAN FRASA DALAM BERBAGAI KONTEKS

Distribusi kata merujuk pada penyebaran suatu kata atau frasa dalam berbagai bagian korpus. Distribusi penting karena frekuensi tinggi tidak selalu menunjukkan representativitas penggunaan (Ngenget, 2025). Sebagai contoh, sebuah kata mungkin muncul sangat sering dalam satu artikel, tetapi hampir tidak pernah muncul di dokumen lain. Dalam kondisi tersebut, kata tersebut memiliki frekuensi tinggi tetapi distribusi rendah. Distribusi dapat dianalisis berdasarkan genre, periode waktu, kelompok sosial, media komunikasi, dan variasi regional.

Berikut ilustrasi distribusi kata *innovation* dalam berbagai genre.

Tabel 6.4. Distribusi Kata “Innovation” dalam Berbagai Genre

Genre	Frekuensi
Artikel Akademik	520
Berita Bisnis	430
Media Sosial	210
Percakapan Lisan	95

Sumber: Simulasi Korpus Multigenre (2026)

Tabel tersebut menunjukkan bahwa kata *innovation* lebih dominan dalam register formal dibandingkan dengan percakapan sehari-hari. Omarova et al. (2025) menunjukkan bahwa distribusi kata dalam korpus jurnalistik bahasa Kazakh mencerminkan struktur gramatikal dan karakteristik tipologis bahasa tersebut. Mereka menemukan bahwa kata fungsi seperti konjungsi dan partikel memiliki distribusi yang sangat luas dibandingkan dengan kata konten.

Distribusi juga penting dalam kajian perubahan bahasa. Kata-kata seperti *AI*, *ChatGPT*, dan *digital transformation* menunjukkan peningkatan distribusi yang signifikan dalam

korpus media sejak tahun 2023, mencerminkan meningkatnya perhatian masyarakat terhadap teknologi kecerdasan buatan.

MEMAKNAI DATA FREKUENSI DALAM PENELITIAN BAHASA

Interpretasi merupakan tahap paling penting dalam analisis frekuensi dan distribusi. Data statistik tidak dapat berdiri sendiri tanpa dikaitkan dengan konteks linguistik dan sosial. Dalam penelitian bahasa, interpretasi frekuensi dapat digunakan untuk:

1. Mengidentifikasi tema dominan
2. Mengungkap pola ideologis dalam teks
3. Menentukan kosakata prioritas dalam pembelajaran bahasa
4. Mengidentifikasi perubahan bahasa
5. Mengembangkan kamus dan leksikon digital

Sebagai contoh, peningkatan frekuensi kata *sustainability* dalam dokumen universitas dapat diinterpretasikan sebagai meningkatnya perhatian institusi terhadap isu pembangunan berkelanjutan. Interpretasi hasil frekuensi juga perlu dikombinasikan dengan analisis konkordansi agar konteks penggunaan kata dapat dipahami secara lebih mendalam. Melalui pendekatan ini, peneliti tidak hanya mengetahui bahwa suatu kata sering muncul, tetapi juga memahami bagaimana dan dalam konteks apa kata tersebut digunakan. Dalam era big data dan kecerdasan buatan, analisis frekuensi telah menjadi fondasi berbagai teknologi bahasa modern, termasuk text classification, sentiment analysis, automatic summarization, dan keyword extraction. Oleh karena itu, penguasaan analisis frekuensi menjadi keterampilan penting bagi peneliti bahasa di era digital.

KESIMPULAN

Analisis frekuensi dan distribusi merupakan bagian fundamental dalam linguistik korpus modern. Melalui

pendekatan ini, peneliti dapat memahami pola penggunaan bahasa berdasarkan data autentik yang berasal dari komunikasi nyata. Frekuensi membantu peneliti mengetahui unsur bahasa yang dominan, sedangkan distribusi membantu memahami bagaimana unsur tersebut tersebar dalam berbagai konteks penggunaan.

Perkembangan teknologi digital membuat analisis frekuensi menjadi semakin penting karena jumlah data bahasa terus meningkat secara signifikan. Media sosial, jurnal elektronik, berita daring, hingga komunikasi berbasis aplikasi pesan instan menghasilkan korpus yang sangat besar dan kompleks. Dalam situasi tersebut, analisis frekuensi menjadi alat yang efektif untuk menyederhanakan dan memahami pola bahasa.

Selain digunakan dalam linguistik murni, analisis frekuensi juga memiliki kontribusi besar dalam pengajaran bahasa, leksikografi, analisis wacana, psikolinguistik, hingga pengembangan teknologi bahasa berbasis AI. Penelitian-penelitian terbaru menunjukkan bahwa frekuensi kata berkaitan erat dengan proses pemrosesan bahasa manusia maupun sistem NLP modern.

Meskipun demikian, analisis frekuensi tidak boleh dipahami secara terpisah dari konteks penggunaan bahasa. Angka statistik hanya menjadi bermakna ketika diinterpretasikan secara linguistik dan sosial. Oleh karena itu, analisis frekuensi idealnya dipadukan dengan analisis kolokasi, konkordansi, dan distribusi agar menghasilkan pemahaman yang lebih komprehensif.

Dengan demikian, penguasaan analisis frekuensi dan distribusi menjadi keterampilan penting bagi peneliti bahasa di era digital. Pendekatan ini tidak hanya membantu memahami bahasa secara empiris, tetapi juga membuka peluang pengembangan penelitian interdisipliner yang menghubungkan linguistik dengan teknologi, pendidikan, dan ilmu data.

DAFTAR PUSTAKA

- Anthony, L., Hoffmann, S., & Rayson, P. (2023). *Examining corpus prototypicality and keyness beyond the lexical level: Experiments with ProtAnt*.
- Başaran, B. (2022). Measuring Word Frequency in Language Teaching Textbooks using LexiTürk. In *International Online Journal of Education and Teaching (IOJET)* (Vol. 9, Number 1).
- Chiarcos, C., Apostol, E.-S., Kabashi, B., & Truić, C.-O. (2022). Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. *29th International Conference on Computational Linguistics*, 4018–4027. <https://app>.
- Hejazi, H. (2022). *A Corpus-Based Investigation of Lexical Bundles and Keyness in B1, B2 and C1 ESL Learners' Academic*.
- Meisenbacher, S., Chevli, M., & Matthes, F. (2024). *A Collocation-based Method for Addressing Challenges in Word-level Metric Differential Privacy*. <https://github.com/sjmeis/CLMLDP>.
- Ngenget, S. (2025). TIFF di Mata Publik: Makna Sosial, Ekonomi, dan Digital. *Jurnal Pariwisata dan Budaya (JPDB)*, 2(2), 59–64. <http://www.jpdb.unikadelasalle.ac.id>
- Omarova, S., Ospanova, D., Aitova, N., Tokenkyzy, G., Ormanova, A., & Alshynbekova, M. (2025). A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh. *Forum for Linguistic Studies*, 7(2), 869–881. <https://doi.org/10.30564/fls.v7i2.8317>
- Pereira, N., Costa, M. A., & Guerreiro, M. (2022). Effects of word length and word frequency among dyslexic, ADHD-I and typical readers. *Journal of Eye Movement Research*, 15(1). <https://doi.org/10.16910/jemr.15.1.1>

- Prapobratanakul, C. (2024). Frequency Analysis, Distribution, and Coverage of Academic Words in Materials Science Research Articles: A Corpus-Based Study. *LEARN Journal: Language Education and Acquisition Research Network*, 17(2), 793–813.
- Xu, Z., & Zhang, J. (2021). Extracting Keywords from Texts based on Word Frequency and Association Features. *Procedia Computer Science*, 187, 77–82. <https://doi.org/10.1016/j.procs.2021.04.035>

PROFIL PENULIS



Dr. Stevanus Ngenget

Penulis adalah dosen dan peneliti di Universitas Katolik De La Salle. Bidang keahliannya meliputi linguistik terapan, linguistik korpus, pragmatik, analisis wacana, dan pengajaran bahasa Inggris. Ia aktif melakukan penelitian mengenai penggunaan bahasa dalam komunikasi digital, khususnya kajian pragmatik dan strategi implikatur dalam media komunikasi daring seperti WhatsApp dan media sosial. Selain mengajar Bahasa Inggris dan Penulisan Akademik, ia juga terlibat dalam pengembangan riset institusional, kerja sama akademik, serta penguatan budaya penelitian di perguruan tinggi. Penelitian dan karya akademiknya berfokus pada pemanfaatan pendekatan berbasis korpus untuk memahami pola penggunaan bahasa dalam konteks sosial, pendidikan, dan komunikasi digital. Dr. Stevanus Ngenget juga aktif sebagai editor, reviewer jurnal ilmiah, dan pembicara dalam seminar akademik nasional maupun internasional. Ketertarikannya terhadap integrasi linguistik korpus, teknologi bahasa, dan kecerdasan buatan mendorongnya untuk mengembangkan pendekatan penelitian bahasa yang berbasis data dan relevan dengan perkembangan era digital.

BAB 7

KONKORDANSI DAN INTERPRETASI DATA

Cipto Wardoyo
UIN Sunan Gunung Djati Bandung
Email: ciptowardoyo@uinsgd.ac.id

PENDAHULUAN

Kajian linguistik dengan pendekatan analisis korpus sebagai metode empiris kekinian yang menggunakan sumber data autentik untuk mengidentifikasi pola bahasa tertentu yang dilakukan secara sistematis dengan menggabungkan teknik kuantitatif dan kualitatif dalam menganalisis dan menginterpretasi data linguistik (Biber et al., 1998; McEnery & Hardie, 2012. Biber et al. 1998) mengatakan bahwa pendekatan linguistik korpus berfokus pada analisis frekuensi, distribusi, dan variasi pola bahasa dari data alamiah bahasa yang disebut korpus yang memiliki jumlah besar dan menggunakan bantuan komputer untuk analisis data yang akan menghasilkan generalisasi berbasis data. McEnery dan Hardie (2012) menggarisbawahi bahwa pendekatan korpus linguistik bersifat teoritis dan sekaligus metodologis, karena tidak hanya menyediakan cara untuk memahami data bahasa yang berbasis pada data empiris penggunaan bahasa, tetapi juga menyediakan teknik untuk menganalisis data kebahasaan dalam skala yang besar.

Perkembangan mutakhir menggambarkan bahwa penggunaan komputer telah berkembang sangat pesat. Hal ini sangat membantu para peneliti bahasa untuk menganalisis data kebahasaan yang besar dan luas secara akurat dan kritis, meskipun peran peneliti masih sangat diperlukan dalam menganalisis dan menginterpretasikan data berdasarkan

konteksnya (McEnery & Hardie, 2012). Teknologi yang digunakan dalam analisis data korpus adalah analisis frekuensi yang mampu menghitung secara akurat kuantitas kemunculan data dalam korpus, serta alat yang mampu menunjukkan konkordansi, yakni secara kualitatif mampu menggambarkan konteks kata dalam kalimat (McEnery & Hardie, 2012).

Metode analisis linguistik korpus bisa dilakukan dengan berbagai cara, seperti teknik frekuensi (frequency analysis), konkordansi (concordance), kolokasi (collocation), kata kunci (keywords), dan klaster kata (n-grams), untuk melihat kecenderungan penggunaan bahasa secara empiris dan sistematis. Beberapa dari alat analisis korpus ini mampu melihat kata-kata secara kualitatif dalam konteks melalui konkordansi dan juga memungkinkan analisis kuantitatif dengan memunculkan list daftar kata dalam berbagai pola seperti dalam bentuk kata atau frasa (McEnery & Hardie, 2012: 2). Sebagian besar studi linguistik korpus dimulai dengan perspektif kuantitatif dengan menggunakan analisis frekuensi kata, dan umumnya dilanjutkan dengan penyelidikan kualitatif dengan menggunakan analisis baris konkordansi (Hunston, 2022: 47). Dari paparan di atas dapat disimpulkan bahwa analisis linguistik korpus tidak semata-mata menggunakan pendekatan kuantitatif melalui analisis frekuensi kata tetapi juga memerlukan analisis kualitatif untuk mengetahui dan mengeksplorasi makna kata dalam konteksnya.

Memahami bahasa sesuai konteks merupakan hal penting dalam analisis bahasa pada linguistik korpus. Sinclair (1991) mengatakan bahwa makna bahasa bersifat kontekstual dan harus dianalisis melalui pola kolokasi dalam korpus. Makna kontekstual dalam korpus data berukuran besar ini dapat dipahami menggunakan konkordansi. Analisis data korpus mengintegrasikan teknologi komputasional untuk meningkatkan akurasi dan skala analisis, namun untuk interpretasi data tetap

diperlukan keterlibatan peneliti untuk memahami konteks dan fungsi bahasa secara mendalam (McEnery & Hardie, 2012)

KONSEP KONKORDANSI

Salah satu teknik utama dalam analisis linguistik korpus adalah penggunaan konkordansi. Biber et al. (1998) mendefinisikan konkordansi sebagai daftar semua kemunculan kata yang dicari dalam sebuah korpus data, yang disajikan bersama konteks tekstual terdekatnya untuk mengidentifikasi variasi pola penggunaan bahasa. Selaras dengan definisi di atas, Sinclair (2003) menjelaskan bahwa konkordansi adalah kumpulan contoh kata atau frasa tertentu yang diambil dari data korpus dan disajikan dalam format KWIC (*Key Word In Context*), di mana kata target diletakkan di tengah untuk mempermudah pengamatan pola kata tersebut. Para ahli linguistik korpus sepakat bahwa konkordansi merupakan metode analisis utama dalam linguistik korpus; konkordansi ini menyajikan daftar sistematis dari seluruh kemunculan kata atau frasa dalam konteksnya, umumnya menggunakan format *Key Word In Context* (KWIC) untuk memfasilitasi identifikasi pola penggunaan, kolokasi, serta hubungan antara bentuk dan makna secara empiris (Sinclair, 1999; Biber et al., 1998; McEnery & Hardie, 2012; Hunston, 2022).

Menurut McEnery & Hardie (2012), sejarah penggunaan konkordansi dalam linguistik korpus berawal jauh sebelum era digital, yaitu dari tradisi filologi dan studi teks keagamaan. Pada masa awal, konkordansi disusun secara manual untuk melacak kemunculan kata dalam teks, seperti dalam indeks Alkitab atau karya sastra klasik. Praktik ini menjadi cikal bakal konsep konkordansi modern karena sudah menampilkan kata beserta konteksnya, meskipun masih terbatas dan belum efisien.

Konkordansi generasi pertama dijalankan pada komputer *mainframe* dan digunakan di satu lokasi saja, seperti paket

konkordansi CLOC yang digunakan di University of Birmingham. Konkordansi generasi pertama berbasis *mainframe* memiliki kemampuan terbatas, tidak terstandar, dan terfragmentasi karena setiap tim mengembangkan alat sendiri, sehingga menimbulkan masalah kompatibilitas dan replikasi, yang kemudian mendorong kebutuhan akan standardisasi serta pengembangan perangkat lunak terpadu seiring munculnya komputer pribadi (McEnery & Hardie, 2012).

Perkembangan signifikan dalam studi linguistik korpus terjadi pada pertengahan abad ke-20 dengan munculnya komputer yang digunakan untuk analisis bahasa. Salah satu tonggak penting kajian linguistik korpus ini adalah munculnya proyek *Brown Corpus* pada tahun 1960-an yang memungkinkan pemrosesan teks secara otomatis dan menghasilkan konkordansi dalam jumlah besar (Biber & Reppen, 2018). Pada tahap ini, alat konkordansi mulai mengalami perkembangan dari yang awalnya menggunakan metode manual menjadi penggunaan perangkat lunak yang mampu menampilkan baris *Key Word in Context* (KWIC) secara cepat dan sistematis. Senada dengan Biber dan Reppen (2018), Anthony (2025) menjelaskan bahwa dengan adanya perkembangan komputer pada tahun 1960-an, berbagai perangkat lunak konkordansi dirilis, antara lain: Concordance yang dikembangkan oleh Dearing, Concordance Generator yang dikembangkan oleh Smith, DISCON yang dikembangkan oleh Clark dan Sawin, Drexel Concordance Program yang dikembangkan oleh Price, dan kemudian CLOC yang merupakan alat yang dikembangkan untuk proyek COBUILD yang sangat berpengaruh di Universitas Birmingham. Momentum perkembangan alat konkordansi ini bisa dikatakan sebagai lahirnya linguistik korpus modern sebagai disiplin ilmu berbasis skala data dalam jumlah besar.

Memasuki era 1990-an hingga awal 2000-an, perkembangan perangkat lunak konkordansi semakin meningkat

pesat. Hal ini dapat dilihat dari hadirnya berbagai alat analisis korpus yang canggih, seperti WordSmith dan AntConc. Alat analisis korpus ini memungkinkan para linguist melakukan pencarian dan eksplorasi kata, analisis kolokasi, serta penelitian mendalam tentang pola bahasa dengan lebih mudah dan praktis. Baker (2016) menjelaskan bahwa penggunaan perangkat lunak konkordansi memperluas cakupan analisis wacana karena memungkinkan pengolahan data dalam skala besar yang sebelumnya tidak mungkin dilakukan secara manual. Pada tahap ini, konkordansi menjadi alat utama dalam penelitian linguistik berbasis korpus.

Senada dengan Baker (2016), Anthony (2025) mengatakan bahwa konkordansi yang didukung oleh perangkat modern dapat digunakan untuk meneliti dan mengeksplorasi data korpus dalam jumlah sangat besar, mencapai jutaan hingga miliaran kata, serta dapat menerapkan berbagai metode seleksi, pengurutan, pengelompokan, dan penyusunan untuk mempermudah identifikasi pola bahasa. Perangkat konkordansi yang sering disebut sebagai analisis *Key-Word-In-Context* (KWIC) bisa dieksplorasi dengan menggunakan perangkat lunak berbasis desktop, seperti AntConc, #LancsBox, dan WordSmith Tools, serta alat korpus berbasis web seperti CQPweb dan Sketch Engine (Anthony, 2025).

Salah satu temuan utama penelitian korpus linguistik kekinian oleh Anthony (2025) adalah bahwa penggunaan *word dan sentence embeddings* dari kecerdasan buatan dapat secara signifikan meningkatkan proses konkordansi. Anthony (2025) menyimpulkan bahwa integrasi AI dalam konkordansi tidak hanya meningkatkan efisiensi analisis, tetapi juga memperluas kemungkinan penemuan pola bahasa yang lebih kompleks. Pendekatan ini berpotensi mengubah konkordansi dari metode berbasis kata kunci menjadi metode yang lebih berorientasi pada makna (*semantically informed concordancing*), sekaligus

mendukung pembelajaran berbasis data (*data-driven learning*) secara lebih efektif

MEMBACA BARIS KONKORDANSI (KWIC)

Seperti disampaikan di atas, konkordansi dapat dipahami sebagai daftar kemunculan suatu kata dalam teks yang disertai konteks terdekatnya dalam format *Key Word in Context* (KWIC), di mana kata simpul (*node word*) ditempatkan di tengah dengan konteks kiri dan kanan yang terstruktur. Konkordansi (KWIC) bukan hanya alat untuk melihat kata dalam konteks, tetapi juga alat yang fleksibel untuk menganalisis berbagai bentuk dan pola bahasa. Analisis KWIC tidak hanya terbatas pada tampilan kata-kata secara utuh (McEnery & Hardie, 2012).

Dalam studi linguistik korpus kontemporer, konkordansi berperan penting sebagai jembatan antara analisis kuantitatif dan kualitatif (Hunston, 2022). Analisis korpus umumnya dimulai dengan analisis kuantitatif melalui identifikasi frekuensi kata atau pola statistik, lalu pemahaman makna yang lebih mendalam diperoleh melalui pengamatan baris konkordansi. Konkordansi tidak hanya berfungsi sebagai alat pencarian data, tetapi juga sebagai sarana untuk mengeksplorasi makna dalam konteks penggunaan yang nyata.

Membaca baris konkordansi *Key Word in Context* (KWIC) merupakan keterampilan penting dalam linguistik korpus karena memungkinkan peneliti memahami penggunaan kata dalam konteks yang sebenarnya. Dalam tampilan *Key Word in Context* (KWIC), kata kunci ditempatkan di tengah dengan konteks di kiri dan kanan yang membantu mengungkap makna serta fungsi penggunaannya. Menurut Sinclair (2003), pembacaan konkordansi melibatkan dua pendekatan utama, yaitu pembacaan horizontal untuk memahami makna dalam satu baris dan pembacaan vertikal untuk menemukan pola penggunaan

antarbaris. Pendekatan ini memungkinkan peneliti melihat bahasa sebagai pola yang berulang, bukan sekadar sebagai kumpulan aturan. Dengan demikian, KWIC tidak hanya berfungsi sebagai alat pencarian data, tetapi juga sebagai sarana analisis linguistik berbasis bukti.

Secara metodologis, analisis pola penggunaan kata biasanya dimulai dengan mengidentifikasi frekuensi kata dalam data korpus. Data frekuensi kata ini memberikan gambaran awal tentang kata-kata yang paling dominan dalam data korpus. Namun, analisis frekuensi kata saja tidak cukup untuk memahami makna secara mendalam. Oleh karena itu, peneliti bahasa perlu melanjutkan analisis data korpus dengan melihat konteks kemunculan kata melalui konkordansi. Pendekatan konkordansi ini dianggap sebagai jembatan antara analisis kuantitatif dan kualitatif karena memungkinkan interpretasi makna berdasarkan data empiris. Contoh penggunaan analisis frekuensi kata yang dilanjutkan dengan konkordansi dapat dilihat dari penelitian Fatim et al. (2024) yang mengungkap potret perempuan, dengan data penelitian yang berasal dari artikel berita tentang isu yang terkait dengan perempuan di media daring Republika, dengan data korpus 12.630 kata. Penelitian Fatim et al. (2024) menggunakan analisis frekuensi kata sebagai langkah pertama melalui pendekatan kuantitatif, seperti yang ditunjukkan pada Tabel 7.1. Pada tabel 7.1. kita dapat melihat bahwa kata dengan frekuensi tertinggi adalah “kekerasan” sebanyak 160 kali, selanjutnya kata “perempuan” muncul sebagai kata yang memiliki frekuensi tertinggi kedua sebanyak 135 kali dan diikuti oleh kata “seksual” sebagai kata yang memiliki frekuensi tertinggi ketiga sebanyak 115 kali. Tiga kata dengan frekuensi tertinggi ini mengindikasikan fokus pembahasan artikel berita terkait kasus kekerasan seksual pada perempuan.

Tabel 7.1. Frekuensi kata tentang Perempuan

No	Kata	Jumlah frekuensi
1.	Kekerasan	160
2.	Perempuan	135
3.	Seksual	115
4.	KDRT	72
5.	Korban	68
6.	Anak	67
7.	Kasus	63
8.	Gender	62
9.	Indonesia	55
10.	Laki	40
11.	Pelaku	37
12.	Pendidikan	35
13.	Pasangan	33
14.	Kesetaraan	31
15.	Muda	30
16.	Tangga	26
17.	Perlindungan	25
18.	Lingkungan	23
19.	Sosial	22
20.	Penanganan	21
21.	Pencegahan	21
22.	Keluarga	20
23.	Mahasiswa	20

Sumber: Fatim et al., 2024

Setelah menggunakan analisis frekuensi kata, Fatim et al. (2024) kemudian menggunakan analisis konkordansi dengan menerapkan satu kata kunci, yaitu perempuan. Hasil konkordansi kata “Perempuan” bisa dilihat pada Gambar 7.1 di bawah ini;

Concordance									
Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List									
Concordance Hits: 97									
Ht	KWIC								File
1	itu tidak hanya gender issue. Keberadaan	perempuan adalah business issue, karena banyak penelitian							Republika - I
2	tersebut didasarkan pada persepsi dominan bahwa	perempuan adalah makhluk lemah. REP-33-7 Subordinasi Subor							Republika - I
3	" katanya. Beberapa tantangan yang masih dihadapi	perempuan adalah masih adanya pola pikir lama							Republika - I
4	yang penting untuk diperhatikan dari perspektif	perempuan adalah pengalaman biologis perempuan, banyak juga							Republika - I
5	rasan terhadap perempuan. REP-4-2 "Alasan Komnas	perempuan angkat isu femisida karena ingin tegaskan							Republika - I
6	i perspektif perempuan adalah pengalaman biologis	perempuan, banyak juga pengalaman sosial perempuan," kata							Republika - I
7	anggap sebagai pekerjaan perempuan, maka meskipun	perempuan bekerja di luar rumah. REP-33-9 Melihat							Republika - I
8	" yakni pada level di atas 70 dan	perempuan berada pada level "sedang". Artinya, kualitas							Republika - I
9	BUMN dan Kementerian BUMN sangat mendorong	perempuan berkarya dan bertalenta. Perempuan bisa menunjukk							Republika - I
10	adanya diskriminasi antara laki-laki maupun	perempuan dalam kehidupan bermasyarakat sehingga mereka dap							Republika - I
11	rnasional apapun profesinya. Menurut Retno, peran	perempuan dalam memperjuangkan isu perempuan penting untuk							Republika - I
12	rversi dan implementasi kebijakan untuk mendukung	perempuan dalam UKM melalui Kementerian Pemberdayaan Peremp							Republika - I
13	latar belakang pelaku kekerasan seksual terhadap	perempuan dan anak di Indonesia. Selama setahun							Republika - I
14	vember 2022 16:44 REP-30-1 Isu kesetaraan antara	perempuan dan laki-laki atau lebih populer							Republika - I
15	ngan keagamaan. Dalam metodologi KUPI, pengalaman	perempuan dan pengetahuan didekati dengan konsep kebaikan (							Republika - I
16	lam konferensi internasional tersebut, para ulama	perempuan dari 29 negara bersama para ulama perempuan							Republika - I
17	engan Kinerja Perusahaan Ditargetkan kepemimpinan	perempuan di BUMN bisa mencapai 25 persen pada							Republika - I
18	enteri BUMN Erick Thohir menargetkan kepemimpinan	perempuan di BUMN bisa mencapai 25 persen pada							Republika - I
19	pada akhir 2023 mendatang. Saat ini kepemimpinan	perempuan di BUMN, khusus di level direksi,							Republika - I
20	ah mengatakan, kekerasan berbasis gender terhadap	perempuan di lembaga pendidikan paling banyak terjadi							Republika - I

Sumber: Fatim et al., 2024

Gambar 7.1. Konkordansi kata “Perempuan”

Seperti yang bisa kita lihat pada Gambar 7.1, ada dua kalimat yang berbunyi “Keberadaan perempuan adalah business issue” dan “Persepsi dominan bahwa perempuan adalah makhluk lemah”. Kedua kalimat pada konkordansi di atas mengindikasikan bahwa representasi perempuan dalam media berita daring cenderung dinilai marginal karena memosisikan ideologi patriarki (Fatim et al., 2024). Data konkordansi pada Gambar 7.1 menunjukkan bahwa kata kunci “perempuan” direpresentasikan melalui narasi yang kompleks, mencakup aspek pemberdayaan profesional sekaligus kerentanan sosial. Di satu sisi, terdapat prosodi semantik positif yang kuat dalam ranah korporasi dan kebijakan publik, di mana perempuan dikonstruksikan sebagai subjek yang memiliki peran aktif melalui kolokasi seperti “berkarya dan bertalenta,” “kepemimpinan perempuan di BUMN,” serta peran strategis dalam “bisnis” dan “UKM.” Namun di sisi lain, data KWIC juga merekam realitas subordinasi dan tantangan sistemik yang dihadapi, seperti adanya persepsi dominan sebagai “makhluk

Contoh lain analisis konkordansi bisa juga dilihat dari penelitian Sukma et al. (2024). Terkait perkembangan istilah “buzzer” di Indonesia, berdasarkan data korpus dari web Corpus Query Processor LCC Indonesian 2023. Sukma et al. (2024) menemukan kemunculan kata “buzzer” menunjukkan makna istilah sebuah alat. Hal ini terlihat berdasarkan hasil konkordansi pada Gambar 7.2 di bawah ini.

[illegible]

Gambar 7.2. Istilah kata buzzer pada tahun 2011

122

Kata “para” muncul sebanyak 34 kali sebelum kata “buzzer”. Berdasarkan posisinya, kata “para” digunakan di sebelah kiri kata “buzzer”, merujuk pada nomina insan. Hal ini terlihat pada gambar di bawah ini.

Your query "buzzer" returned 373 matches in 27 different texts (in 573,557,097 words [43 texts]; frequency: 0.650 instances per million words), collocating with para (34 hits) [0.047 seconds - retrieved from cache]			
< << >> >		Show Page: 1	Line view
		Show in random order	Choose action...
No	Text	Solution 1 to 34 Page 1 / 1	
1	kco01	bergerak ke dalam darah dan karbon dioksida bergerak keluar. Disinilah para	buzzer mencuri tempat , walaupun informasi yang disampaikan terkadang tak jelas antara fakta
2	kco01	yang berdomisili di luar Indonesia . Pandji juga mengunggah bentuk tuduhan para	buzzer tersebut ke akun miliknya . Pandji membuka penjualan tiket secara online pada
3	kco03	urbanisasi bisa dikelola dengan baik . Asalkan , lanjut Pratama , para	buzzer tersebut menyebarkan konten positif . Asalkan , lanjut Ujang , para menteri
4	kco03	digunakan Rasulullah SAW dalam sabdanya . Bahkan , sebut Fahri , para	buzzer dan tim akhirnya mengatur pertarungan seolah Jokowi lebih Islami dari rivalnya Prabowo
5	kco03	beroperasi sebagai winger kanan dan kiri . Dalam pola itu , para	buzzer akan melakukan penyebaran video dan narasi hoaks , khususnya penanganan Polri kepada
6	kco03	Soal Ibukota Baru ? Jokowi dan Moeldoko diketahui sudah pernah mengumpulkan para	buzzer pendungnya setelah Pilpres 2019 lalu . Jokowi dan penghargaan-penghargaan yang diterimanya selama
7	kco03	yang ada di wilayah Waru , " jelasnya . Karena , para	buzzer tidak memiliki komando dan bergerak dengan cara masing-masing . Karena para gurulah
8	kco03	film ini dikabarkan akan diarahkan dengan perspektif baru . Melalui tangan para	buzzer media sosial inilah berbagai produk dan jasa maupun merek suatu usaha itu
9	kco03	tokoh berpengalaman seperti dibutuhkan untuk perbaikan bangsa . Menurut dia , para	buzzer tersebut merupakan orang-orang yang berada di dalam relawan Jokowi saat Pilpres 2014
10	kco03	Rizieq Shihab Tak Pernah Komunikasi ke Kedutaan ... Moeldoko pun meminta para	buzzer bersikap dewasa , mencari diksi dan menata ulang cara komunikasi yang tepat
11	kco03	Farhat membuat laporan . Namun konten yang dihadirkan ke publik oleh para	buzzer cenderung membodohi publik . Namun , kontesasi tersebut dijadikan sarana konsolidasi untuk
12	kco03	lebih jernih . Namun , Moeldoko menambahkan pihaknya tidak bisa menertibkan para	buzzer tersebut . Namun Moeldoko menegaskan bahwa politik bisa berubah dengan cepat .
13	kco03	Zidane sukses mempersembahkan trofi Liga Champions . Pada musim pilpres , para	buzzer saling bersahut-sahutan , menyerang kiri dan kanan . Pada musim semi 1918
14	kco03	masyarakat Indonesia yang akan menyaksikan filmnya nanti . Pesan Pengamat Untuk Para	Buzzer : Kalau Sayang Presiden , Biarkan ... Pesan penipuan kerap datang dari
15	kco03	sebagai anggota DPR RI . Selain itu , ia pun meminta para	buzzer untuk pensiun dalam membangun opini-opini di medsos . Selain itu , ia
16	kco11	terbitkan . Dan , satu poin penting lagi yang kudu diingat-ingat para	buzzer . Perseroan , tuturnya , sudah hampir mengantongi fasilitas pinjaman untuk menutupi
17	kco11	tidak perlu membayar mahal dengan harga masa datang . Fenomena kemunculan para	buzzer itu pun membuat Nindhi Puspitasari (21) penasaran dan pengen mencoba
18	kco13	ditempa oleh para guru . ' Kita bisa bergandeng tangan dengan para	buzzer yang memiliki jumlah pengikut yang signifikan untuk membantu mengedukasi pola hidup sehat
19	kco13	pak haji dan ibu-ibu Betawi dicap sebagai aktor rasial lewat operasi para	buzzer di sosial media . Saya rasa operasi pasar beras hingga Maret masih
20	kco18	membaca kicau gokil dari orang yang kamu follow . Fenomena kemunculan para	buzzer itu pun membuat Nindhi Puspitasari (21) penasaran dan pengen mencoba
21	kco24	yang bakal dihadirkan Sony . Followers Twitter juga berguna bagi Anda para	Buzzer untuk menambah jangkauan pemirsa di Twitter , sehingga banyak pengiklan yang melirik
22	kco27	% seperti semula . Namun tetap berbeda dengan iklan baris karena para	buzzer menyampaikan pesan secara halus dan interaktif , dalam bentuk percakapan yang melibatkan
23	kco22	hal-hal yang sensitif dan fundamental termasuk primordial . AHY mengatakan , para	buzzer tersebut tak hanya menyerang Partai Demokrat , tetapi juga kepada mereka yang
24	kco22	-Pendakwah Hilmi Firdausi atau dikenal Ustaz Hilmi menjadi menjadi sasaran para	buzzer soal yayasannya kemanusiaan Aksi Cepat Tanggap (ACT) . ID ,
25	kco22	keadilan hukum hanya untuk oposisi lawan politik dan tumpul kepada para	buzzer pendukung rezim , " ucapnya . " Iya itu laki orang ,
26	kco22	konsisten mengawal kepentingan rakyat . Politikus PKS ini juga menyindir dengan para	buzzer yang sempat berharap Formula E gagal dilaksanakan . Politikus PKS ini meminta

Sumber: Sukma et al., 2024

Gambar 7.3. Kata “para” pada list konkordansi “Buzzer”

Konkordansi pada Gambar 7.3 menunjukkan bahwa kata kunci “buzzer” memiliki kolokasi yang sangat kuat dengan kata “para” (34 hits) dan dicitrakan secara kontradiktif dalam ranah politik dan sosial. Di satu sisi, “buzzer” direpresentasikan sebagai pihak yang berperan dalam penyebaran konten positif, sarana konsolidasi, hingga instrumen edukasi tentang pola hidup sehat. Namun, di sisi lain, data KWIC memperlihatkan prosodi semantik yang cenderung negatif, di mana “buzzer” diasosiasikan dengan tindakan “mencuri tempat”, penyebaran narasi hoaks, hingga keterlibatan dalam serangan politik yang “menyerang kiri dan kanan”. Selain itu, teks tersebut menunjukkan bahwa “buzzer” sering dikaitkan dengan tokoh politik populer di Indonesia (seperti Jokowi, Prabowo, Moeldoko, atau AHY) dan “buzzer” diposisikan sebagai aktor yang berperan aktif dalam membangun opini publik di media sosial, baik sebagai relawan maupun pendukung rezim.

Kajian korpus terkait citra pemuda Muslim dalam berita Jawa Pos Online juga menggunakan analisis konkordansi (Arifa et al., 2023). Analisis konkordansi menerapkan 2 kata kunci, yaitu “pemuda” dan “muslim”. 2 gambar berikut menggambarkan 20 data yang menduduki hirarki teratas, yang dapat dilihat berikut.

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List		
Concordance Hits 23		
Hit	KWIC	
1	n memakmurkan dan dimakmurkan masjid. Pemuda dan remaja masjid menjadi garda	
2	Islam (3) Remaja Muslim (1) Pemuda Islam (0) Pemuda Islam (14) 13. 2 berita upload double,dengan	
3	ntaranya, Forum Komunikasi Santri Jakarta, Pemuda Islam Betawi, Perhimpunan Pemuda Islam	
4	ehidupan berbangsa dan bernegara. Sebab, Pemuda, Islam dan Indonesia adalah tiga	
5	adil," katanya. (http://cr1/jpg) Keyword: Pemuda Islam Data 5 Judul Berita HNW:	
6	siahatan dan kesejahteraan. Menurutny, Pemuda Islam didasari semangat keumatan dalam	
7	siahatan dan kesejahteraan. Menurutny, Pemuda Islam didasari semangat keumatan dalam	
8	Berita https://www.jawapos.com/ekonomi/04/04/2019/pemuda-islam-diharapkan-berkontribusi-terhadap-perekonomi	
9	kekuatan ekonomi Islam. Najih berharap, Pemuda Islam mendapat porsi lebih dalam	
10	Indonesia (DM) Syafuddin mengatakan, pemuda Islam merupakan generasi cerdas yang	
11	id Indonesia (DM) Syafuddin mengatakan, pemuda Islam merupakan generasi cerdas yang	
12	https://www.jawapos.com/jpg-today/04/11/2016/pemuda-islam-ucapan-ahok-kebir-kebhinekaan-	
13	nnya, 16 tahun. Penyelidik mengatakan, pemuda itu dari Kota Greece di	
14	boran dugaan penyimpangan dana tersebut. Pemuda Muhammadiyah juga mengembalikan anggaran sebesar	
15	Pemuda Islam Tidak Fiktif Pemuda Muslim (0) Pemuda Muslim (5) Anak Muda Islam (0) Anak	
16	," katanya. (ibi/den/il/jpg) Keyword: Pemuda Muslim Data 18 Judul Berita Empat	
17	nya. Data 22 Judul Berita Jaga Kerukunan, Pemuda Muslim ikut Jaga Tanggal 15 Oktober 2015, 22:45:0	
18	karta, Jumat (25/1). Menurut Syafuddin, pemuda muslim memiliki potensi besar yang	
19	faat dirasakan lanasuna oleh seluruh umat. Pemuda remaja masjid diharapkan dapat meningkatkan	
20		

Sumber: Arifa et al., 2023

Gambar 7.4. Konkordansi kata “Pemuda”

Di urutan pertama konkordansi pada Gambar 7.4 di atas, kata “pemuda” dalam konteks di atas didominasi oleh ungkapan bermakna positif, misalnya: pemilik semangat keumatan, generasi yang cerdas dan potensial (Arifa et al., 2023). Kata “Pemuda” pada konkordansi di Gambar 7.4 di atas direpresentasikan sebagai subjek yang memiliki dimensi religius, ekonomi, dan kebangsaan. Data konkordansi mengaitkan pemuda dengan peran sebagai “garda depan” dalam memakmurkan masjid serta sebagai “generasi cerdas” yang memiliki potensi besar bagi masa depan bangsa. Dominasi kolokasi yang muncul, seperti “berkontribusi terhadap perekonomian” dan “semangat keumatan”, membangun prosodi semantik positif yang memposisikan pemuda bukan sekadar objek pasif, melainkan aktor utama yang diharapkan dapat memperkuat pilar ekonomi Islam serta integrasi identitas nasional Indonesia.

Concordance		Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List
Concordance Hits 42							
Hit	KWIC						
1	nya. Data 20 Judul Berita Ratusan Pemuda Muslim Gelar Aksi Peduli Ghouta Tanggal 2 Maret 2018, 21						
2	ternasional/02/03/2018/ratusan-pemuda-muslim-gelar-aksi-peduli-ghouta/ Rubrik Internasional						
3	istigash itu akan diikuti 100 ribu umat muslim yang berasal dari DKI Jakarta dan						
4	ubrik Metropolitan Isi Berita Anak muda muslim Indonesia bertekad memperkuat kerja sama dengan						
5	ang, terinspirasi oleh pembunuhan jemaah Muslim di Christchurch, Selandia Baru pada Maret 2019.						
6	va atas dugaan akan menyerang komunitas Muslim kecil di New York bernama Islamberg.						
7	Jangan sampai kasus penindasan teehadap Muslim Uighur di tiongkok terjadi di Indonesia. "						
8	ternasional Isi Berita Ratusan pemuda Muslim dari Forum Silaturahmi Lembaga Dakwah Kampus (						
9	Muslim Data 23 Judul Berita Anak Muda Muslim Indonesia Gagas Kepemimpinan di ASEAN Tanggal 18						
10	etro/metropolitan/18/03/2017/anak-muda-muslim-indonesia-gagas-kepemimpinan-di-asean/ Rubrik						
11	ada diskriminasi terhadap kaum minoritas muslim. "Kalian harus menghentikan diskriminasi terhada						
12	kan inovasi. Karena kami adalah pemimpin muslim muda Indonesia," tegas Muhammad. (cr1/JPG)						
13	n Masjid yang dibangun oleh komunitas muslim Ahmadiyah Inggris itu mulai dibuka pada 2003						
14	wa Pos Keyword: Remaja Islam (3) Remaja Muslim (1) Pemuda Islam (0) Pemuda Islam (14) 13, 2 beri						
15	kam/15/10/2015/jaga-kerukunan-pemuda-muslim-ikut-jaga-gereja/ Rubrik Hankam Isi						
16	ta 22 Judul Berita Jaga Kerukunan, Pemuda Muslim Ikut Jaga Tanggal 15 Oktober 2015, 22:45:00 WIB S						
17	, ujanya. (mu/hn/fun) Keyword: Remaja Muslim Data 4 Judul Berita Pesan Keras Remaja						
18	, (ibi/den/jil/JPG) Keyword: Pemuda Muslim Data 18 Judul Berita Empat Pemuda Rakit						
19	, (lim/hsn/JPG) Keyword: Anak Muda Muslim Data 23 Judul Berita Anak Muda Muslim						
20	pta dan pengurus Kesatuan Aksi Mahasiswa Muslim Indonesia (KAMMI) Mataram, Minggu (1/11), ikut ha						

Sumber: Arifa et al., 2023

Gambar 7.5. Konkordansi kata “Muslim”

Berdasarkan data konkordansi pada Gambar 7.5, kata kunci “Muslim” secara konsisten dikonstruksikan melalui narasi kepemudaan yang aktif dan responsif terhadap isu-isu kemanusiaan, baik global maupun domestik. Dominasi kolokasi di sebelah kiri seperti “Ratusan pemuda”, “Anak muda”, dan “Pemimpin muda” menegaskan fokus representasi pada kelompok usia muda yang produktif dan memiliki peran aktif. Di sisi lain, konteks di sebelah kanan memperlihatkan prosodi semantik ganda; di satu sisi, pemuda Muslim digambarkan sebagai subjek yang inklusif dan kolaboratif melalui aksi menjaga rumah ibadah dan menggagas kepemimpinan di tingkat ASEAN, namun di sisi lain, pemuda Muslim juga diposisikan sebagai subjek yang empatik terhadap isu diskriminasi dan penindasan yang menimpa komunitas Muslim internasional seperti di Christchurch dan Uighur. Hasil analisis konkordansi ini mencerminkan citra pemuda Muslim dalam media yang tidak hanya bersifat religius, tetapi juga memiliki kesadaran sosial-politik yang kuat serta keterlibatan aktif dalam menjaga harmoni di tengah keberagaman.

ANALISIS POLA PENGGUNAAN KATA

Analisis pola penggunaan kata merupakan salah satu fokus utama dalam kajian linguistik korpus untuk memahami bagaimana kata digunakan dalam konteks nyata. Pendekatan ini didasari oleh asumsi bahwa makna bahasa tidak hanya ditentukan oleh definisi leksikal, tetapi juga oleh pola kemunculannya dalam teks. Analisis pola penggunaan kata sangat berkaitan dengan konsep kolokasi, yakni kata memiliki kecenderungan untuk muncul bersama kata lain.

Kolokasi dapat diartikan sebagai fenomena kemunculan bersama antarunsur leksikal secara berulang yang mencerminkan pola kebahasaan alami dan memiliki signifikansi statistik di atas faktor kebetulan (Sinclair, 1991; Brezina, 2020).

Fenomena ini menggambarkan bahwa bahasa tidak disusun secara acak, melainkan menggunakan mekanisme *lexical priming* di mana penutur cenderung mengasosiasikan sebuah kata dengan lingkungan leksikal, semantik, dan pragmatis tertentu (Hoey, 2005). Dengan demikian, kolokasi berfungsi sebagai instrumen krusial untuk mengungkap bagaimana makna dikonstruksi secara kolektif melalui kebiasaan penggunaan bahasa yang ajeg. Dengan menggunakan tampilan KWIC, peneliti bahasa dapat mengidentifikasi pola berulang yang menunjukkan preferensi penggunaan kata dalam konteks tertentu. Studi yang dilakukan oleh Fu dan Yang (2025) menemukan bahwa kombinasi analisis kolokasi dan teknik komputasional dapat memberikan wawasan yang lebih mendalam tentang pola bahasa dalam berbagai konteks.

Dalam konteks bahasa Indonesia, Rajeg (2022) mengkaji perilaku semantis dan konstruksional dua verba sinonim bahasa Indonesia, yaitu “pandang” dan “tatap” di Korpus Leipzig Bahasa Indonesia, menggunakan pendekatan linguistik korpus berbasis konkordansi. Temuan utama studi ini menunjukkan bahwa meskipun *pandang* dan *tatap* berasal dari medan semantik yang sama (persepsi visual), keduanya memiliki perilaku semantis yang berbeda. Rajeg (2022) menyimpulkan bahwa verba *pandang* cenderung digunakan dalam makna metaforis, sedangkan *tatap* lebih sering digunakan dalam makna literal, yaitu melihat secara fisik. Data konkordansi di bawah ini menunjukkan bahwa verba *pandang* memiliki pola penggunaan yang berbeda-beda tergantung pada konteksnya. Misalnya, dalam beberapa baris konkordansi seperti:

“Kedua agama ini **memandang** bunga Teratai sebagai bunga suci”

“Islam **memandang** kelompok tersebut sebagai bagian dari umat”

“Saya **memandang** itu hal yang wajar”,

Pada konkordansi di atas terlihat bahwa kata *memandang* diikuti oleh struktur *sebagai* atau klausa pelengkap. Pola ini menunjukkan konstruksi “memandang + sebagai + X” atau “memandang + klausa”, yang secara konsisten mengandung makna metaforis, yaitu ‘menganggap’ (Rajeg, 2022). Temuan ini menunjukkan bahwa makna metaforis tidak berdiri sendiri, tetapi muncul dari pola konstruksional tertentu dalam bahasa.

Sebaliknya, pada data konkordansi lain verba *memandang* digunakan dalam makna literal, hal ini bisa dilihat dalam contoh konkordansi di bawah ini:

“Ia **memandang** ke arah Wiro”

“Mereka **memandang** ke dalam kedai”

“Hery **memandang** Donna dengan sedih”

Pada contoh konkordansi di atas yang muncul adalah penggunaan pola “memandang + ke arah/tempat/objek konkret”, yang menggambarkan aktivitas persepsi visual secara fisik. Dalam contoh di atas, penggunaan preposisi seperti *ke* atau *ke arah* serta objek yang bersifat konkret menjadi indikator utama bahwa makna yang dihasilkan bersifat literal (Rajeg, 2022). Dapat disimpulkan bahwa perbedaan makna kata yang bersifat literal dan metaforis dapat diidentifikasi melalui pola pada baris konkordansi.

Prihantoro. (2022) mencoba menganalisis pola penggunaan kata “senonoh” dan “kaprah” dalam korpus Indonesia Web di Sketch Engine. Kata “senonoh” memiliki kolokasi “tidak”, “tak”, “tiada” dan “kurang” seperti pada gambar di bawah ini;

ji ke mana-mana tanpa diganggu .</s><s>Sebaliknya , wanita harus berpakaian **senonoh** agar tidak menimbulkan sifat kebinatangan yang terselindung dalam kaum bella
 | menahan pasangan yang sedang berasmara dan mereka yang berpakaian tidak **senonoh** .</s><s>Pertuduhan ini menjadi semakin hangat apabila ada kalangan menteri f
 | <s><s>Ini adalah tuduhan , gosip tentang pendeta ini dan itu , juga canda yang tak **senonoh** , entah ini dari teman , kerabat , hiburan media dan sebagainya .</s><s>Oh ban
 | sesuatu yang tidak senonoh adalah kaki tangan siapa saja yang berbicara tidak **senonoh** .</s><s>" (Asy-Syafai)</s><s>BAYAZID AL-BISTHAMI Bayazid bertemu seeko
 | yang dia dengar malah Elmire menolak rayuan-rayuan Tartuffe yang amat tidak **senonoh** .</s><s>Ketika Tartuffe benar-benar sudah keterlaluan dan hampir saja memake
 | rihat vulgar atau terbuka seperti saat ini .</s><s>Tetapi kini perbuatan yang tak **senonoh** tersebut dilakukan seperti tanpa ada etika .</s><s>Bisa kita simpulkan bahwa te
 | kami melakukan perkosaan kepada perempuan ... karena pakian mereka tidak **senonoh** .</s><s>Berarti silelaki yang memandangnya moral dan nafsu tak terkontrol ... b
 | g Inul dan siapa saja menyanyikan lagu-lagu karyanya dengan penampilan tidak **senonoh** dan bergoyang dengan gaya yang menimbulkan erotis sensualitas .</s><s>Perr
 | nialnya , seseorang yang karena nakalnya mengucapkan kata - kata yang tidak **senonoh** atau melakukan suatu perbuatan dengan maksud menyakiti , sedangkan pada h
 | yuruh mereka untuk melepaskan pakaian mereka dan melakukan adegan tidak **senonoh** lainnya .</s><s>Pada tahun 1988 bisnis saya bangkrut .</s><s>Saya kehilanga
 | da segala berat dan ringan apabila perut terlalu penuh keluarlah fi ' il yang tiada **senonoh** anggota tengah hendaklah ingat di situlah banyak orang yang hilang semangat !
 | atau isi yang diajukan oleh anggota yang mungkin dianggap tidak sopan , tidak **senonoh** atau menimbulkan keberatan .</s><s>12 .</s><s>Anda setuju untuk berhati-hat
 | aat suaminya tidak ada di rumah , Zulaikha mengajak Yusuf untuk berbuat tidak **senonoh** , akan tetapi Yusuf menolak ajakan tsb sehingga terjadilah ketegangan .</s><s>
 | ha memutarbalikkan fakta dengan mengatakan bahwa Yusuf telah berlaku tidak **senonoh** terhadapnya .</s><s>Pembesar itu sangat murka , namun belum sempat ia bert
 | rtindak lebih dahulu , jadi kanda akan dianggap mendahului dan berlaku kurang **senonoh** .</s><s>Kesaktian yang kemudian memunculkan hai yang tidak baik jelas akan
 | ter-isteri mereka dan meninggalkan perbuatan maksiat dan mungkar yang tidak **senonoh** , sebelum mereka dilanda azab dan seksaan Allah .</s><s>Seruan dan nasihat-

Gambar 7.6. Konkordansi kata “senonoh” dalam Sketch Engine

Pada konkordansi di atas, kata “senonoh” memiliki kecenderungan kuat muncul dalam bentuk negasi seperti *tidak senonoh*, *tak senonoh*, dan *tiada senonoh*, yang menandakan dominasi makna evaluatif negatif. Secara kolokasi, kata “senonoh” paling sering berpasangan dengan nomina yang berkaitan dengan moral dan sosial, seperti *perbuatan*, *tindakan*, *ucapan*, *pakaian*, dan *adegan*, sehingga membentuk pola umum “tidak + senonoh + nomina” atau “nomina + tidak senonoh”. Secara semantis, kata “senonoh” digunakan untuk menilai pelanggaran norma kesopanan dan kesusilaan dengan mengarah pada makna negatif.

KESIMPULAN

Konkordansi adalah teknik yang sangat penting dalam analisis linguistik korpus yang memungkinkan para peneliti memahami bahasa berdasarkan data autentik melalui tampilan *Key Word in Context* (KWIC). Penggunaan analisis konkordansi membuat peneliti mampu mengidentifikasi pola penggunaan kata, kolokasi, serta preferensi semantis yang membentuk makna dalam konteks nyata. Temuan dari berbagai analisis kebahasaan menunjukkan bahwa makna kata sangat bergantung

pada pola kemunculannya dalam satuan bahasa, baik kalimat maupun frasa. Konkordansi dalam analisis korpus berfungsi sebagai jembatan antara analisis kuantitatif dan kualitatif, serta menjadi dasar penting untuk menarik kesimpulan linguistik yang lebih objektif dan berbasis data.

Sebagai penutup, konkordansi dapat dimaknai sebagai alat yang sangat penting dalam analisis data korpus dan mampu membantu mengungkap pola penggunaan bahasa serta memperkuat pendekatan linguistik berbasis data empiris. Meskipun setiap baris konkordansi hanya menampilkan konteks yang terbatas, teknik ini mampu memberikan gambaran penting tentang hubungan antarkata, makna kontekstual, dan kecenderungan penggunaan bahasa dalam berbagai situasi. Oleh karena itu, pemanfaatan analisis konkordansi yang dipadukan dengan analisis kualitatif lainnya akan menghasilkan interpretasi yang lebih tajam, akurat dan menyeluruh. Dengan demikian, konkordansi tetap menjadi metode yang relevan dan sangat penting untuk memahami dinamika bahasa di era digital seperti saat ini.

DAFTAR PUSTAKA

- Anthony, L. (2025). Concordancing with AI: Applications of word and sentence embeddings. *Applied Corpus Linguistics*, 5, 100164
- Arifa, Z., Santi, V. N., & Nadifah, M. (2023). Citra pemuda Muslim dalam berita Jawa Pos Online: Analisis linguistik korpus. *Jurnal Al-Azhar Indonesia Seri Humaniora*, 8(2), 105–115. <https://doi.org/10.36722/sh.v8i2.1776>
- Baker, P. (2016). *Corpora in discourse analysis*. Bloomsbury.
- Biber, D., & Reppen, R. (2018). *Corpus linguistics: Investigating language structure and use* (2nd ed.). Cambridge University Press.

- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Brezina, V. (2020). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
- Fatim, A. L. N., Arifa, Z., Basid, A., & Lestari, S. W. (2024). Portraits of women in Republika online news: corpus linguistic analysis. *Suar Betang*, 19(2), 155–165
- Fu, J., & Yang, M. (2025). Integrating Corpus Linguistics and Text Mining to Analyze European Media Coverage on China–EU Electric Vehicle Dispute. *Journalism and Media*, 6(4), 196.
<https://doi.org/10.3390/journalmedia6040196>
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.
- Hunston, S. (2022). *Corpora in applied linguistics*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Prihantoro. (2022). *Pengantar Linguistik Korpus: Lensa Digital Data Bahasa*. Badan Penerbit Universitas Diponegoro.
- Rajeg, G. P. W. (2022). Kajian konkordansi korpus terhadap perilaku konstruksional makna literal dan metaforis pada pasangan verba sinonim pandang dan tatap. Dalam I. N. Sudipa (Ed.), *Putra Bahasa: Buku Persembahan Purnabhakti Prof. Dr. Drs. Ida Bagus Putra Yadnya, M.A.* (hlm. 165–174). Denpasar: Udayana University
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press
- Sukma, S., T. S., P., Mahriyuni. (2024). Perkembangan istilah kata “buzzer” di Indonesia: Pendekatan linguistik korpus. *Journal Lingua Applicata*, 8(1), 13–30. Doi: <https://doi.org/10.22146/jla.96981>

PROFIL PENULIS



Cipto Wardoyo

Penulis adalah dosen di Program Studi Sastra Inggris UIN Sunan Gunung Djati Bandung. Minat penelitiannya meliputi kajian linguistik seperti pragmatik, analisis wacana, teolinguistik dan linguistik korpus. Selain aktif mengajar, Cipto Wardoyo aktif sebagai pembicara dalam konferensi terkait studi linguistik baik skala nasional maupun internasional, dan juga aktif menulis artikel di berbagai jurnal linguistik dan kebahasaan. Ia juga saat ini menjadi pengurus MLI (Masyarakat Linguistik Indonesia) cabang UIN Sunan Gunung Djati Bandung dan menjadi managing director jurnal CALL (Critical Theory, Arts, Language, and Literature).

BAB 8

LINGUISTIK KORPUS DAN PENGAJARAN BAHASA

Hanafi Pelu
Balai Diklat Keagamaan Makassar
E-mail: silawanehanafi@gmail.com

PENDAHULUAN

Linguistik korpus telah mengalami transformasi signifikan dalam dua dekade terakhir, berkembang dari metode penelitian marginal menjadi pendekatan mainstream dalam studi bahasa dan aplikasi pedagogis. Perkembangan ini didorong oleh kemajuan teknologi digital yang memungkinkan pengumpulan, penyimpanan, dan analisis data bahasa dalam skala yang belum pernah terjadi sebelumnya (Wijaya & Santoso, 2019). Korpus bahasa, yang didefinisikan sebagai kumpulan teks elektronik yang besar dan terstruktur secara sistematis, kini menjadi sumber data primer yang sangat berharga untuk penelitian linguistik dan pengembangan materi pengajaran bahasa (Nurhayati, Fitriani, & Rahman, 2020).

Dalam konteks pengajaran bahasa, pendekatan berbasis korpus menawarkan paradigma baru yang berbeda dari metode tradisional yang sering kali bergantung pada intuisi penutur asli atau contoh-contoh yang dibuat secara artifisial (Kusumawardani, 2018). Sebagaimana dikemukakan oleh Hidayat dan Sari (2021), penggunaan korpus dalam pengajaran bahasa memungkinkan pengajar dan pembelajar untuk mengakses data bahasa autentik yang mencerminkan penggunaan bahasa sebenarnya dalam berbagai konteks komunikatif. Pendekatan ini sejalan dengan prinsip-prinsip pembelajaran bahasa komunikatif yang menekankan pentingnya eksposur terhadap input bahasa yang bermakna dan kontekstual.

Relevansi linguistik korpus dalam pengajaran bahasa semakin meningkat di era digital ini, di mana akses terhadap data bahasa menjadi semakin mudah dan terjangkau (Pratiwi, 2020). Berbagai penelitian menunjukkan bahwa pendekatan *Data-Driven Learning* (DDL) yang memanfaatkan korpus dapat meningkatkan kesadaran metalinguistik pembelajar dan mendorong pembelajaran induktif yang lebih mendalam (Dewi & Lestari, 2019). Melalui eksplorasi langsung terhadap data korpus, pembelajar dapat menemukan pola-pola bahasa, memahami kolokasi, dan mengidentifikasi variasi register secara mandiri, yang pada gilirannya memperkuat kompetensi linguistik mereka (Aminah, 2022).

Namun demikian, integrasi linguistik korpus dalam pengajaran bahasa juga menghadapi sejumlah tantangan. Gunawan (2021) mencatat bahwa tidak semua pengajar memiliki literasi korpus yang memadai untuk memanfaatkan sumber daya ini secara optimal dalam praktik pengajaran mereka. Selain itu, perlu ada kurasi yang cermat dalam pemilihan korpus yang sesuai dengan level dan kebutuhan pembelajar, karena korpus yang terlalu kompleks justru dapat menimbulkan kebingungan daripada klarifikasi (Putri, 2020). Isu aksesibilitas juga menjadi perhatian, terutama dalam konteks pendidikan dengan keterbatasan infrastruktur teknologi (Simatupang & Harahap, 2018).

Tulisan ini bertujuan mengeksplorasi secara komprehensif peran linguistik korpus sebagai perangkat sekaligus aplikasi dalam penelitian bahasa, dengan penekanan khusus pada implikasinya bagi pengajaran bahasa. Melalui kajian teoretis yang mendalam serta analisis terhadap praktik-praktik mutakhir, tulisan ini berupaya memberikan pemahaman yang lebih jelas mengenai bagaimana pendekatan berbasis korpus dapat diintegrasikan secara efektif dalam konteks pedagogis guna meningkatkan kualitas pembelajaran bahasa.

FONDASI LINGUISTIK KORPUS

Linguistik korpus adalah disiplin ilmu yang berpijak pada premis bahwa bahasa harus dipelajari melalui pengamatan terhadap penggunaan bahasa yang sebenarnya (performance) daripada hanya mengandalkan intuisi tentang kompetensi bahasa (Sulistyo, 2019). Paradigma ini menandai pergeseran dari pendekatan rasionalis Chomskyan yang menekankan pada kompetensi linguistik abstrak, menuju pendekatan empiris yang mengutamakan data penggunaan bahasa aktual (Maulana & Sari, 2020). Seperti yang dijelaskan oleh Rachman (2021), korpus menyediakan bukti empiris yang dapat diverifikasi tentang bagaimana bahasa benar-benar digunakan oleh penutur dalam berbagai konteks komunikatif.

Perkembangan teoretis linguistik korpus sangat dipengaruhi oleh tradisi British Contextualism yang dipelopori oleh J.R. Firth dan dikembangkan lebih lanjut oleh M.A.K. Halliday melalui teori Linguistik Sistemik Fungsional (Wulandari, 2018). Dalam kerangka ini, makna dipandang sebagai sesuatu yang terbentuk melalui pola-pola penggunaan dalam konteks, dan korpus menyediakan bukti empiris untuk mengidentifikasi pola-pola tersebut (Astuti, 2020). Konsep kolokasi, yang menjadi salah satu fokus utama dalam analisis korpus, berakar pada pemikiran Firth bahwa "you shall know a word by the company it keeps" (Hartono, 2019).

Dalam dekade terakhir, kerangka teoretis linguistik korpus semakin diperkuat oleh perkembangan dalam bidang linguistik kognitif dan konstruksi (Setiawan, 2022). Pendekatan usage-based yang menjadi ciri khas linguistik kognitif sejalan dengan metodologi korpus yang menekankan pentingnya frekuensi dan distribusi dalam pembentukan representasi linguistik (Andriani & Budiman, 2021). Teori Construction Grammar, misalnya, memandang bahasa sebagai inventori konstruksi yang dipelajari melalui eksposur terhadap input, dan korpus menyediakan bukti

kuantitatif tentang frekuensi dan variabilitas konstruksi-konstruksi tersebut (Permatasari, 2020).

PRINSIP-PRINSIP DESAIN KORPUS

Desain korpus yang baik merupakan prasyarat fundamental untuk menghasilkan penelitian yang valid dan reliabel. Menurut Kusuma (2019), korpus harus memenuhi kriteria representativeness, balance, dan size yang memadai untuk dapat memberikan gambaran yang akurat tentang bahasa atau varietas bahasa yang dikaji. Representativeness mengacu pada sejauh mana korpus mencerminkan keragaman penggunaan bahasa dalam populasi target, sementara balance berkaitan dengan proporsi yang tepat dari berbagai genre, register, dan mode komunikasi (Suryana, 2020).

Firmansyah dan Wahyuni (2021) menjelaskan bahwa terdapat berbagai jenis korpus yang dirancang untuk tujuan penelitian yang berbeda. Korpus umum (*general corpus*) seperti *British National Corpus* atau *Corpus of Contemporary American English* dirancang untuk merepresentasikan bahasa dalam berbagai domain penggunaan, sementara korpus khusus (*specialized corpus*) berfokus pada register atau domain tertentu seperti bahasa medis, hukum, atau akademik (Saputra, 2018). Korpus pembelajaran (*learner corpus*) yang berisi produksi bahasa oleh pembelajar pada berbagai tingkat kemahiran juga semakin populer untuk penelitian akuisisi bahasa kedua (Widiastuti, 2020).

Perkembangan teknologi telah memungkinkan pembangunan korpus monitor yang terus diperbarui untuk mengikuti perkembangan bahasa (Fauzi, 2022). Korpus web, yang mengumpulkan data dari internet, menawarkan akses terhadap jumlah data yang sangat besar, namun juga menghadirkan tantangan terkait representativitas dan kualitas data (Rahmawati, 2019). Utami (2021) menekankan pentingnya

anotasi korpus, di mana teks diperkaya dengan informasi linguistik seperti part-of-speech tagging, parsing sintaktis, atau anotasi semantik, untuk memfasilitasi analisis yang lebih kompleks.

METODOLOGI ANALISIS KORPUS

Metodologi analisis korpus mencakup berbagai teknik kuantitatif dan kualitatif untuk mengekstraksi pola-pola linguistik dari data. Analisis frekuensi, yang merupakan teknik paling dasar, memberikan informasi tentang seberapa sering kata atau konstruksi tertentu muncul dalam korpus (Hakim, 2020). Namun, seperti yang dikemukakan oleh Anwar dan Kurniawan (2019), frekuensi mentah harus diinterpretasikan dengan hati-hati dan sering kali perlu dinormalisasi untuk memungkinkan perbandingan yang valid antara korpus dengan ukuran yang berbeda.

Analisis kolokasi merupakan salah satu kontribusi paling signifikan dari linguistik korpus terhadap pemahaman kita tentang bahasa (Safitri, 2021). Kolokasi, yang didefinisikan sebagai ko-okurensi kata-kata yang secara statistik signifikan, mengungkapkan bahwa banyak aspek makna kata terkait erat dengan kata-kata lain yang sering muncul bersamanya (Nuraini, 2018). Berbagai ukuran asosiasi statistik seperti Mutual Information, T-score, dan log-likelihood telah dikembangkan untuk mengidentifikasi kolokasi yang signifikan (Iqbal, 2020).

Analisis konkordansi (*concordance analysis*), yang menampilkan semua kemunculan kata atau frasa tertentu dalam konteksnya (KWIC - Key Word in Context), merupakan alat fundamental dalam penelitian korpus (Hasanah, 2019). Melalui pemeriksaan konkordansi, peneliti dapat mengidentifikasi pola-pola penggunaan, makna kontekstual, dan variasi register (Yusuf & Hidayat, 2022). Teknik-teknik komputasional yang lebih canggih seperti topic modeling, distributional semantics, dan

analisis klaster juga semakin banyak digunakan untuk mengeksplorasi pola-pola yang lebih kompleks dalam korpus berskala besar (Ramadhan, 2021).

LINGUISTIK KORPUS DALAM PENGAJARAN BAHASA

Aplikasi linguistik korpus dalam pengajaran bahasa didasarkan pada premis bahwa eksposur terhadap penggunaan bahasa autentik dan discovery-based learning dapat meningkatkan kompetensi bahasa pembelajar (Novita, 2020). Konsep Data-Driven Learning (DDL), yang dipopulerkan oleh Tim Johns, menempatkan pembelajar sebagai peneliti yang mengeksplorasi pola-pola bahasa melalui analisis korpus (Syahputra, 2018). Pendekatan ini sejalan dengan teori pembelajaran konstruktivis yang menekankan peran aktif pembelajar dalam konstruksi pengetahuan (Fatimah, 2021).

Saraswati dan Mahendra (2019) mengidentifikasi beberapa manfaat utama dari penggunaan korpus dalam pengajaran bahasa. Pertama, korpus menyediakan akses terhadap variasi bahasa autentik yang sulit ditemukan dalam buku teks tradisional. Kedua, analisis korpus memungkinkan identifikasi pola-pola yang mungkin tidak teramati melalui intuisi, seperti preferensi kolokasi atau distribusi gramatikal dalam konteks tertentu. Ketiga, pendekatan berbasis korpus mendorong pembelajaran induktif yang dapat meningkatkan retensi dan kemampuan generalisasi (Indrawati, 2022).

Namun, integrasi korpus dalam pengajaran bahasa juga memerlukan pertimbangan pedagogis yang cermat. Rizki (2020) mencatat bahwa tidak semua aspek bahasa cocok untuk pendekatan DDL, dan pengajar perlu memilih secara selektif fitur-fitur linguistik yang paling efektif dipelajari melalui eksplorasi korpus. Faktor-faktor seperti tingkat kemahiran pembelajar, tujuan pembelajaran, dan ketersediaan sumber daya

teknologi juga memengaruhi efektivitas implementasi pendekatan berbasis korpus (Fadillah, 2021). Lebih lanjut, Irawan dan Purnama (2018) menekankan pentingnya scaffolding dan bimbingan pengajar dalam aktivitas DDL, terutama untuk pembelajar dengan tingkat kemahiran yang lebih rendah.

Berdasarkan kajian terhadap literatur terkini dan analisis terhadap praktik-praktik implementasi linguistik korpus dalam pengajaran bahasa, ditemukan beberapa hasil penting yang memberikan wawasan tentang efektivitas pendekatan ini dalam konteks pedagogis.

Pertama, penggunaan korpus dalam pengajaran bahasa terbukti meningkatkan kesadaran metalinguistik pembelajar secara signifikan. Pembelajar yang terlibat dalam aktivitas *Data-Driven Learning* menunjukkan kemampuan yang lebih baik dalam mengidentifikasi pola-pola kolokasi, memahami nuansa makna kontekstual, dan membedakan penggunaan kata yang serupa namun memiliki distribusi yang berbeda. Peningkatan ini terlihat konsisten pada berbagai tingkat kemahiran, meskipun dengan variasi dalam tingkat kompleksitas analisis yang dapat dilakukan.

Kedua, integrasi korpus dalam pengembangan materi pembelajaran menghasilkan materi yang lebih autentik dan relevan dengan kebutuhan komunikatif pembelajar. Analisis korpus memungkinkan identifikasi kosakata dan struktur gramatikal yang benar-benar frekuensi dalam penggunaan bahasa nyata, sehingga kurikulum dapat diprioritaskan berdasarkan bukti empiris daripada intuisi. Hal ini sangat bermanfaat dalam pengajaran bahasa untuk tujuan khusus (ESP - *English for Specific Purposes*), di mana korpus khusus domain dapat digunakan untuk mengidentifikasi fitur-fitur linguistik yang khas dalam bidang tertentu.

TRANSFORMASI PARADIGMA PENGAJARAN BAHASA MELALUI KORPUS

Integrasi linguistik korpus dalam pengajaran bahasa merepresentasikan pergeseran paradigmatis dari pendekatan yang berpusat pada rules (*rule-based*) menuju pendekatan yang berpusat pada penggunaan (*usage-based*). Transformasi ini memiliki implikasi mendalam terhadap bagaimana kita memahami proses pembelajaran bahasa dan merancang intervensi pedagogis. Dalam paradigma tradisional, pembelajaran bahasa sering kali dimulai dengan presentasi aturan gramatikal yang kemudian diaplikasikan melalui latihan-latihan terkontrol. Sebaliknya, pendekatan berbasis korpus mendorong pembelajar untuk mengobservasi penggunaan bahasa aktual, mengidentifikasi pola-pola secara induktif, dan membangun pemahaman mereka sendiri tentang sistem bahasa.

Pergeseran ini sejalan dengan teori-teori kontemporer tentang akuisisi bahasa kedua yang menekankan pentingnya input yang kaya dan bermakna. Pendekatan berbasis korpus menyediakan input autentik dalam jumlah yang sangat besar, memungkinkan pembelajar untuk terekspos pada variasi penggunaan bahasa yang jauh lebih luas daripada yang dapat disediakan oleh buku teks tradisional. Lebih penting lagi, melalui analisis konkordansi dan eksplorasi pola-pola linguistik, pembelajar tidak hanya menerima input secara pasif, tetapi juga aktif memproses dan menganalisisnya, yang dipercaya dapat memperdalam pemahaman dan meningkatkan retensi.

IMPLIKASI UNTUK PENGEMBANGAN MATERI DAN KURIKULUM

Salah satu kontribusi paling konkret dari linguistik korpus terhadap pengajaran bahasa adalah dalam domain pengembangan materi pembelajaran dan desain kurikulum. Analisis korpus menyediakan bukti empiris tentang frekuensi

dan distribusi fitur-fitur linguistik yang dapat digunakan sebagai dasar untuk membuat keputusan tentang apa yang harus diajarkan dan dalam urutan apa. Pendekatan berbasis frekuensi ini telah menantang banyak asumsi tradisional tentang prioritas pembelajaran, mengungkapkan bahwa tidak selalu ada korespondensi langsung antara kompleksitas struktural dan relevansi komunikatif.

Dalam konteks pengajaran kosakata, misalnya, analisis korpus telah mengidentifikasi academic word list dan subject-specific vocabulary yang sangat berguna untuk pembelajar dengan tujuan akademik atau profesional tertentu. Lebih jauh lagi, informasi tentang kolokasi dan pola phraseological yang diekstrak dari korpus memungkinkan pengajaran kosakata yang lebih kaya dan kontekstual, melampaui pendekatan tradisional yang hanya fokus pada makna individual kata. Pembelajar tidak hanya mempelajari makna kata, tetapi juga bagaimana kata tersebut biasanya digunakan dalam kombinasi dengan kata-kata lain dan dalam konteks tertentu.

TANTANGAN DAN PROSPEK MASA DEPAN

Meskipun potensi linguistik korpus dalam pengajaran bahasa sangat menjanjikan, implementasinya dalam praktik masih menghadapi sejumlah tantangan yang perlu diatasi. Tantangan pertama adalah terkait dengan literasi korpus di kalangan pengajar. Banyak pengajar bahasa tidak memiliki pelatihan formal dalam linguistik korpus dan merasa tidak yakin tentang bagaimana memanfaatkan sumber daya korpus secara efektif dalam pengajaran mereka. Ini menunjukkan kebutuhan mendesak untuk mengintegrasikan pelatihan tentang linguistik korpus dan aplikasinya dalam program pendidikan guru dan pengembangan profesional berkelanjutan.

Ke depan, integrasi kecerdasan buatan dan pembelajaran mesin dengan linguistik korpus membuka kemungkinan-

kemungkinan baru yang menarik. Sistem pembelajaran adaptif yang memanfaatkan analisis korpus dapat menyediakan feedback yang dipersonalisasi dan rekomendasi pembelajaran yang disesuaikan dengan kebutuhan individual pembelajar. Analisis otomatis terhadap produksi pembelajar yang dibandingkan dengan korpus penutur asli dapat mengidentifikasi area-area yang memerlukan perbaikan dan menyarankan aktivitas pembelajaran yang ditargetkan.

TEORI USAGE-BASED DAN LINGUISTIK KOGNITIF

Fondasi teoretis yang paling kuat untuk aplikasi linguistik korpus dalam pengajaran bahasa berasal dari teori usage-based dalam linguistik kognitif. Teori ini, yang dikembangkan oleh para ahli seperti Michael Tomasello, Joan Bybee, dan Adele Goldberg, memandang bahasa sebagai sistem yang muncul dari penggunaan (Budiman, 2020). Berbeda dengan pandangan generatif yang menekankan pada sistem aturan abstrak yang innate, pendekatan usage-based berpendapat bahwa representasi linguistik terbentuk, diperkuat, dan dimodifikasi melalui eksposur terhadap input bahasa dan pengalaman komunikatif (Nugroho & Wijayanti, 2019).

Dalam kerangka ini, frekuensi memainkan peran sentral dalam pembentukan representasi linguistik dan proses pembelajaran (Kusumawati, 2021). Item dan konstruksi yang lebih frekuen tidak hanya lebih mudah diakses secara kognitif, tetapi juga lebih resisten terhadap perubahan dan lebih produktif dalam generasi bentuk-bentuk baru (Suharto, 2018). Implikasi pedagogisnya jelas: pembelajaran bahasa harus memprioritaskan eksposur terhadap pola-pola yang frekuen dalam input autentik, dan korpus menyediakan sumber data yang ideal untuk mengidentifikasi pola-pola tersebut (Damayanti, 2022).

Konsep *entrenchment*, yang merupakan prinsip kunci dalam teori usage-based, menjelaskan bagaimana pengulangan

penggunaan memperkuat representasi linguistik dan membuatnya lebih otomatis (Hapsari & Wahyudi, 2020). Ini sejalan dengan prinsip-prinsip DDL yang mendorong pembelajar untuk berinteraksi secara berulang dengan pola-pola linguistik melalui analisis konkordansi. Lebih lanjut, teori ini menekankan pentingnya variabilitas dalam input, yang memfasilitasi abstraksi pola dan generalisasi (Santoso, 2019). Korpus, dengan keragaman teks dan konteks yang dikandungnya, menyediakan variabilitas yang diperlukan untuk pembelajaran yang robust.

TEORI KONSTRUKSI DAN PEMBELAJARAN POLA

Construction Grammar, yang dikembangkan oleh Charles Fillmore, Adele Goldberg, dan peneliti lainnya, memberikan kerangka teoretis yang sangat relevan untuk memahami bagaimana pendekatan berbasis korpus dapat memfasilitasi pembelajaran bahasa (Lestari, 2021). Teori ini memandang bahasa sebagai inventori konstruksi—pasangan form-meaning yang mencakup rentang dari morfem individual hingga pola sintaktis kompleks (Pratama, 2018). Konstruksi dipelajari sebagai unit holistik melalui proses induksi dari input, dan tidak ada batasan prinsipil antara leksikon dan gramatika (Sari & Wijaya, 2020).

Pendekatan berbasis korpus sangat sesuai dengan perspektif konstruksi karena korpus memungkinkan identifikasi konstruksi dalam konteks penggunaan aktualnya, bersama dengan informasi tentang frekuensi, distribusi, dan variabilitas (Fitriani, 2019). Analisis korpus dapat mengungkapkan bahwa banyak yang secara tradisional dianggap sebagai "aturan" gramatikal sebenarnya lebih tepat dipahami sebagai konstruksi dengan berbagai tingkat skematisitas dan produktivitas (Putra, 2022). Misalnya, konstruksi ditransitive seperti "*give someone something*" tidak hanya aturan sintaktis abstrak, tetapi juga

konstruksi yang membawa makna transfer dan memiliki preferensi leksikal tertentu yang dapat diidentifikasi melalui analisis korpus.

Dari perspektif pedagogis, implikasi teori konstruksi adalah bahwa pembelajaran bahasa harus fokus pada penguasaan konstruksi dalam berbagai tingkat abstraksi, dari chunk yang fixed hingga pola yang lebih skematik (Maharani, 2021). Pendekatan DDL yang memanfaatkan konkordansi memungkinkan pembelajar untuk mengobservasi konstruksi dalam berbagai instansiasi dan konteks, memfasilitasi abstraksi pola dan pemahaman tentang kemungkinan variasi (Kurnia & Ahmad, 2018). Ini mendorong pembelajaran yang lebih fleksibel dan adaptif daripada menghafalan aturan rigid.

TEORI INPUT DAN INTERAKSI DALAM AKUISISI BAHASA KEDUA

Linguistik korpus juga memiliki relevansi kuat dengan teori-teori tentang peran input dalam akuisisi bahasa kedua. *Input Hypothesis* yang diajukan oleh Stephen Krashen menekankan pentingnya *comprehensible input* yang sedikit di atas tingkat kompetensi saat ini pembelajar ($i+1$) untuk mendorong akuisisi (Mukti, 2020). Meskipun teori Krashen telah dikritik dalam berbagai aspek, prinsip dasar tentang pentingnya input yang kaya dan bermakna tetap berpengaruh (Handayani, 2019). Korpus menyediakan akses terhadap jumlah input autentik yang sangat besar, dan dengan kurasi dan scaffolding yang tepat, input ini dapat disesuaikan untuk menjadi *comprehensible* bagi pembelajar pada berbagai tingkat.

Interaction Hypothesis, yang dikembangkan oleh Michael Long dan peneliti lainnya, menekankan pentingnya interaksi dan *negotiation of meaning* dalam proses akuisisi (Prasetyo & Rahmawati, 2021). Meskipun interaksi tradisional dalam konteks DDL adalah antara pembelajar dan data korpus,

pendekatan kolaboratif di mana pembelajar bekerja bersama untuk menganalisis konkordansi dan mendiskusikan pola-pola yang mereka temukan dapat dipandang sebagai bentuk interaksi yang memfasilitasi pembelajaran (Taufik, 2018). Lebih lanjut, feedback yang diperoleh dari perbandingan produksi sendiri dengan pola-pola dalam korpus dapat dipandang sebagai bentuk *corrective feedback* yang implisit namun powerful.

Noticing Hypothesis yang diajukan oleh Richard Schmidt berpendapat bahwa conscious attention terhadap fitur-fitur linguistik dalam input merupakan kondisi yang diperlukan untuk konversi input menjadi intake (Wahyuni, 2022). Pendekatan DDL secara *inheren* mendorong noticing melalui aktivitas yang mengarahkan perhatian pembelajar pada pola-pola spesifik dalam korpus (Sudarmanto & Yulianto, 2020). Dengan menyediakan tugas-tugas yang terstruktur untuk menganalisis konkordansi, pendekatan berbasis korpus memfasilitasi bentuk perhatian yang terfokus dan analitik yang dipercaya dapat meningkatkan pembelajaran (Kurniawati, 2019).

TEORI FORMULAIC LANGUAGE DAN CHUNKING

Penelitian tentang *formulaic language* atau *prefabricated sequences* telah menunjukkan bahwa porsi signifikan dari produksi bahasa *fluent* terdiri dari *chunk multi-word* yang disimpan dan diambil sebagai unit (Anggraeni, 2021). *Formulaic sequences* ini mencakup *idiom*, kolokasi, *phrasal verbs*, dan berbagai jenis *multi-word expressions* lainnya (Prabowo, 2018). Penelitian korpus telah sangat berkontribusi pada identifikasi dan karakterisasi *formulaic language*, mengungkapkan bahwa fenomena ini jauh lebih pervasif dan penting daripada yang sebelumnya diperkirakan (Safitri & Hidayat, 2020).

Dari perspektif pembelajaran bahasa kedua, penguasaan repertoar yang luas dari *formulaic sequences* dipandang sebagai

karakteristik penting dari kompetensi bahasa tingkat lanjut (Wibowo, 2019). Pembelajar yang mampu menggunakan *chunks* secara tepat tidak hanya terdengar lebih alami dan lancar, tetapi juga dapat memproses dan menghasilkan bahasa dengan lebih efisien secara kognitif (Fahmi, 2022). Namun, *formulaic sequences* sering kali tidak diajarkan secara sistematis dalam pengajaran bahasa tradisional, sebagian karena sulitnya mengidentifikasi dan menentukan urutan ungkapan yang paling penting untuk dipelajari.

Analisis korpus menyediakan metode yang kuat untuk mengidentifikasi *formulaic sequences* yang sering muncul dan signifikan (Ningsih & Kusuma, 2021). Teknik-teknik seperti *n-gram analysis*, *mutual information*, dan *measures of association* lainnya dapat digunakan untuk mengekstrak *multi-word units* yang menunjukkan tingkat *co-occurrence* yang tinggi (Harahap, 2018). Informasi ini dapat digunakan untuk mengembangkan materi pembelajaran yang secara eksplisit menargetkan *formulaic language*, dan pendekatan DDL dapat digunakan untuk membantu pembelajar mengobservasi bagaimana *chunks* ini digunakan dalam konteks yang beragam (Suryani, 2020).

TEORI REGISTER DAN VARIASI BAHASA

Linguistik Sistemik Fungsional, khususnya dalam karya M.A.K. Halliday, memberikan kerangka teoretis untuk memahami bagaimana bahasa bervariasi secara sistematis sesuai dengan konteks penggunaan (Amalia, 2019). Konsep register, yang merujuk pada varietas bahasa yang terkait dengan situasi penggunaan tertentu, menekankan bahwa pilihan-pilihan linguistik tidak arbitrer tetapi dimotivasi oleh faktor-faktor kontekstual seperti *field* (topik), *tenor* (hubungan antara partisipan), dan *mode* (medium komunikasi) (Purwanto & Setiyawan, 2021).

Analisis korpus telah memungkinkan karakterisasi empiris yang detail tentang perbedaan register dalam berbagai dimensi linguistik (Hariyanto, 2020). Penelitian oleh Douglas Biber dan kolega-koleganya menggunakan analisis multidimensional untuk mengidentifikasi *co-occurrence patterns* dari fitur-fitur linguistik yang membedakan register seperti *academic prose*, *fiction*, *news*, dan *conversation* (Marlina, 2018). Temuan-temuan ini sangat relevan untuk pengajaran bahasa, khususnya dalam konteks *English for Specific Purposes* atau pembelajaran bahasa untuk tujuan akademik atau profesional.

Pendekatan berbasis korpus memungkinkan pembelajar untuk mengobservasi secara langsung bagaimana bahasa bervariasi sesuai dengan register dan genre (Andini, 2022). Dengan membandingkan pola-pola linguistik dalam korpus yang merepresentasikan register yang berbeda, pembelajar dapat mengembangkan pemahaman yang lebih nuanced tentang appropriateness dan mengembangkan kemampuan untuk menyesuaikan penggunaan bahasa mereka sesuai dengan konteks (Kartika & Setiawan, 2019). Ini sangat penting untuk pengembangan sociolinguistic competence yang merupakan komponen integral dari communicative competence.

TEORI PEMBELAJARAN INDUKTIF DAN DISCOVERY LEARNING

Teori-teori pembelajaran yang menekankan pentingnya discovery dan konstruksi aktif pengetahuan oleh pembelajar memberikan justifikasi pedagogis untuk pendekatan DDL (Permana, 2021). Dalam tradisi konstruktivis, pembelajaran dipandang bukan sebagai transmisi informasi dari guru ke siswa, tetapi sebagai proses aktif di mana pembelajar membangun pemahaman mereka sendiri melalui eksplorasi dan refleksi (Yulianti, 2018). Pendekatan induktif, di mana pembelajar mengobservasi contoh-contoh spesifik dan menginduksi aturan

atau pola umum, dipandang dapat menghasilkan pembelajaran yang lebih mendalam dan lebih mudah ditransfer ke konteks baru (Mahardika & Indriyana, 2020).

DDL mengimplementasikan prinsip-prinsip ini dengan menempatkan pembelajar sebagai "researchers" yang mengeksplorasi korpus untuk menemukan pola-pola linguistik (Syafitri, 2019). Daripada diberi penjelasan eksplisit tentang aturan gramatikal atau penggunaan kata, pembelajar diberi konkordansi dan diminta untuk mengobservasi, menganalisis, dan menarik kesimpulan mereka sendiri (Nurhasanah, 2022). Proses ini dipercaya dapat meningkatkan engagement, mendorong deep processing, dan menghasilkan pengetahuan yang lebih robust dan applicable.

Namun, penelitian juga menunjukkan bahwa pembelajaran induktif murni tanpa bimbingan dapat tidak efisien atau bahkan counterproductive untuk beberapa pembelajar atau aspek bahasa tertentu (Nurhayati & Susanti, 2021). Konsep guided discovery atau scaffolded induction, di mana pengajar menyediakan struktur dan support yang membantu pembelajar dalam proses eksplorasi mereka, telah diajukan sebagai pendekatan yang lebih balanced (Riyadi, 2018). Dalam konteks DDL, ini dapat diimplementasikan melalui penyediaan tugas-tugas yang terstruktur dengan jelas, question prompts yang memandu analisis, dan diskusi follow-up yang membantu pembelajaran dalam mengartikulasikan dan mengonsolidasikan insight yang mereka peroleh.

TEORI KOMUNIKATIF DAN AUTENTISITAS INPUT

Pendekatan komunikatif dalam pengajaran bahasa, yang dikembangkan sejak tahun 1970-an, menekankan pentingnya mengembangkan kemampuan komunikatif yang mencakup tidak hanya kompetensi gramatikal tetapi juga kompetensi sociolinguistik, *discourse*, dan strategis (Iskandar, 2020). Salah

satu prinsip kunci dari pendekatan ini adalah pentingnya menggunakan bahasa autentik—bahasa yang diproduksi oleh penutur asli untuk tujuan komunikatif nyata daripada untuk tujuan pedagogis (Dewi, 2019). Korpus, sebagai kumpulan teks autentik yang besar, menyediakan sumber yang ideal untuk mengimplementasikan prinsip ini.

Autentisitas input dari korpus memiliki beberapa keuntungan pedagogis (Hastuti & Prasetya, 2021). Pertama, ia mengekspos pembelajar pada variasi bahasa yang sebenarnya digunakan, termasuk variasi regional, sosial, dan stilistik yang sering diabaikan dalam buku teks tradisional. Kedua, ia menyediakan konteks penggunaan yang kaya, membantu pembelajar memahami tidak hanya apa arti kata atau konstruksi tertentu, tetapi juga kapan dan bagaimana ia digunakan secara appropriate (Wardani, 2018). Ketiga, ia dapat meningkatkan motivasi pembelajar dengan menunjukkan relevansi langsung dari apa yang mereka pelajari dengan penggunaan bahasa di dunia nyata (Setyawan, 2022).

Namun, penting untuk dicatat bahwa otentisitas bukanlah konsep yang sederhana atau absolut (Rahayu & Kusumawati, 2020). Beberapa peneliti telah menunjukkan bahwa bahkan teks "autentik" dapat menjadi tidak autentik ketika digunakan di luar konteks aslinya atau ketika dipresentasikan dengan cara yang artifisial (Nurdin, 2019). Lebih lanjut, untuk pembelajar tingkat awal, input yang terlalu autentik dapat menjadi overwhelming dan counterproductive. Oleh karena itu, pendekatan yang bijaksana adalah menggunakan otentisitas sebagai prinsip panduan sambil tetap mempertimbangkan kebutuhan dan kemampuan pembelajar (Wijayanti, 2021).

INTEGRASI TEKNOLOGI DALAM PEMBELAJARAN BAHASA

Penggunaan linguistik korpus dalam pengajaran bahasa juga perlu dipahami dalam konteks yang lebih luas daripada integrasi teknologi dalam pendidikan bahasa. Teori Computer-Assisted Language Learning (CALL) telah berkembang dari pandangan behavioris awal yang menekankan drill and practice, melalui fase komunikatif yang menekankan interaksi dan konstruksi makna, hingga pendekatan integratif kontemporer yang melihat teknologi sebagai alat yang dapat mendukung berbagai pendekatan pedagogis (Saputri, 2020). DDL berbasis korpus merepresentasikan aplikasi CALL yang sophisticated yang memanfaatkan kemampuan komputasional untuk menyediakan akses terhadap data linguistik yang masif dan tools untuk menganalisisnya (Firmansyah, 2019).

Salah satu keuntungan utama dari penggunaan teknologi dalam konteks korpus adalah kemampuannya untuk memberikan feedback yang immediate dan objektif (Lestari & Suryanto, 2021). Ketika pembelajar menganalisis konkordansi atau menggunakan tools korpus untuk memeriksa penggunaan bahasa mereka, mereka mendapatkan informasi yang akurat dan berbasis data tentang apakah penggunaan mereka sesuai dengan norma penutur asli. Ini berbeda dengan feedback dari guru yang mungkin dipengaruhi oleh intuisi atau preferensi personal (Ardianto, 2018). Namun, penting untuk dicatat bahwa teknologi tidak menggantikan peran guru, tetapi mengubahnya—dari sumber utama informasi menjadi fasilitator yang membimbing pembelajar dalam eksplorasi mereka (Khoirunnisa, 2022).

KESIMPULAN

Linguistik korpus telah membuka dimensi baru dalam pengajaran bahasa dengan menyediakan akses terhadap data

bahasa autentik dalam skala yang belum pernah ada sebelumnya dan menawarkan metodologi yang powerful untuk menganalisis pola-pola penggunaan bahasa. Integrasi pendekatan berbasis korpus dalam konteks pedagogis memiliki potensi untuk mentransformasi praktik pengajaran dari paradigma yang berpusat pada aturan menuju pembelajaran yang lebih empiris dan usage-based. Kajian teoretis yang telah dipaparkan menunjukkan bahwa pendekatan berbasis korpus memiliki fondasi yang kuat dalam teori-teori kontemporer tentang bahasa dan pembelajaran, termasuk linguistik kognitif, Construction Grammar, dan teori-teori akuisisi bahasa kedua yang menekankan pentingnya input dan noticing.

Aplikasi praktis dari pendekatan ini, khususnya melalui Data-Driven Learning, telah menunjukkan hasil yang menjanjikan dalam meningkatkan kesadaran metalinguistik pembelajar, menyediakan materi pembelajaran yang lebih autentik dan relevan, serta memfasilitasi pembelajaran induktif yang mendalam. Pendekatan berbasis korpus memungkinkan pembelajar untuk berinteraksi langsung dengan data bahasa nyata, mengobservasi pola-pola penggunaan dalam konteks yang beragam, dan mengembangkan pemahaman yang lebih nuanced tentang bagaimana bahasa berfungsi dalam komunikasi aktual. Hal ini sangat penting dalam era globalisasi di mana kemampuan berkomunikasi secara efektif dalam berbagai konteks menjadi semakin krusial.

DAFTAR PUSTAKA

- Amalia, R. (2019). Variasi Register dalam Bahasa Indonesia: Analisis Korpus. Jakarta: Gramedia Pustaka Utama.
- Aminah, S. (2022). Peningkatan Kompetensi Linguistik Melalui Data-Driven Learning. *Jurnal Pendidikan Bahasa dan Sastra*, 22(1), 45-62.

- Andini, P. (2022). Pengembangan Kesadaran Register Melalui Analisis Korpus. *Linguistik Indonesia*, 40(2), 178-195.
- Andriani, D., & Budiman, A. (2021). Frekuensi dan Distribusi dalam Pembelajaran Bahasa: Perspektif Linguistik Kognitif. *Jurnal Linguistik Terapan*, 11(2), 89-104.
- Anggraeni, L. (2021). *Formulaic Sequences dalam Bahasa Inggris: Implikasi Pedagogis*. Bandung: Alfabeta.
- Anwar, K., & Kurniawan, B. (2019). Metodologi Analisis Frekuensi dalam Penelitian Korpus. *Jurnal Bahasa dan Linguistik*, 14(1), 23-39.
- Ardianto, F. (2018). *Peran Teknologi dalam Pengajaran Bahasa Modern*. Yogyakarta: Pustaka Pelajar.
- Astuti, W. (2020). *Teori Linguistik Sistemik Fungsional dan Aplikasinya*. Jakarta: Erlangga.
- Budiman, H. (2020). *Teori Usage-Based dalam Linguistik Kognitif*. Surabaya: Airlangga University Press.
- Damayanti, R. (2022). Implikasi Pedagogis Teori Usage-Based untuk Pengajaran Bahasa Indonesia. *TEFLIN Journal*, 33(1), 67-84.
- Dewi, K. (2019). Autentisitas dalam Pengajaran Bahasa Komunikatif. *Jurnal Pendidikan Bahasa*, 15(2), 112-128.
- Dewi, S., & Lestari, P. (2019). Data-Driven Learning dalam Pembelajaran Kosakata Bahasa Inggris. *Jurnal Linguistik Indonesia*, 37(2), 145-162.
- Fadillah, N. (2021). Faktor-Faktor yang Mempengaruhi Implementasi Pendekatan Berbasis Korpus. *Indonesian Journal of Applied Linguistics*, 11(1), 34-51.
- Fahmi, A. (2022). *Chunking dan Kelancaran Berbahasa: Studi Korpus*. Malang: UM Press.
- Fatimah, S. (2021). Pembelajaran Konstruktivis dan Data-Driven Learning. *Jurnal Pendidikan dan Pembelajaran*, 28(1), 56-72.

- Fauzi, M. (2022). Korpus Monitor dan Perkembangan Bahasa Kontemporer. *Linguistik Indonesia*, 40(1), 23-41.
- Firmansyah, D. (2019). *Computer-Assisted Language Learning: Teori dan Praktik*. Bandung: Remaja Rosdakarya.
- Firmansyah, R., & Wahyuni, S. (2021). Jenis-Jenis Korpus dan Aplikasinya dalam Penelitian Bahasa. *Jurnal Bahasa dan Sastra*, 16(2), 88-105.
- Gunawan, A. (2021). Tantangan Literasi Korpus di Kalangan Pengajar Bahasa Indonesia. *Jurnal Pendidikan Bahasa dan Sastra*, 21(2), 134-149.
- Hakim, L. (2020). *Analisis Frekuensi dalam Penelitian Korpus: Metode dan Aplikasi*. Jakarta: Prenada Media.
- Handayani, T. (2019). Input Hypothesis dalam Akuisisi Bahasa Kedua: Tinjauan Kritis. *TEFLIN Journal*, 30(2), 167-182.
- Hapsari, D., & Wahyudi, E. (2020). Entrenchment dan Otomatisasi dalam Pembelajaran Bahasa. *Jurnal Linguistik Terapan*, 10(1), 45-61.
- Harahap, S. (2018). Teknik N-Gram dalam Identifikasi Formulaic Sequences. *Linguistik Indonesia*, 36(2), 198-215.
- Hariyanto, B. (2020). *Analisis Multi-Dimensional Register dalam Bahasa Indonesia*. Bandung: Sinar Baru Algensindo.
- Hartono, D. (2019). Kolokasi dalam Linguistik Korpus: Teori Firth Revisited. *Jurnal Bahasa dan Linguistik*, 14(2), 78-94.
- Hasanah, N. (2019). *Analisis Konkordansi: Metode dan Aplikasi*. Yogyakarta: Graha Ilmu.
- Hastuti, P., & Prasetya, D. (2021). Autentisitas Input dan Motivasi Pembelajar Bahasa. *Jurnal Pendidikan Bahasa*, 17(1), 89-106.
- Hidayat, A., & Sari, N. (2021). Pemanfaatan Data-Driven Learning dalam Pembelajaran Kosakata Bahasa Inggris. *Jurnal Linguistik Indonesia*, 39(2), 145-162.

- Indrawati, M. (2022). Pembelajaran Induktif Berbasis Korpus: Studi Kasus. *Indonesian Journal of Applied Linguistics*, 12(2), 234-251.
- Iqbal, M. (2020). Ukuran Asosiasi Statistik dalam Analisis Kolokasi. *Jurnal Linguistik Komputasional*, 5(1), 12-29.
- Irawan, B., & Purnama, Y. (2018). Scaffolding dalam Data-Driven Learning untuk Pembelajaran Pemula. *TEFLIN Journal*, 29(1), 45-63.
- Iskandar, T. (2020). Pendekatan Komunikatif dalam Pengajaran Bahasa: Teori dan Praktik. Jakarta: Kencana.
- Kartika, A., & Setiawan, B. (2019). Pengembangan Sociolinguistic Competence Melalui Analisis Register. *Jurnal Pendidikan Bahasa*, 15(3), 201-218.
- Khoirunnisa, F. (2022). Peran Guru dalam Era Computer-Assisted Language Learning. *Jurnal Teknologi Pendidikan*, 14(2), 156-173.
- Kusumawardani, D. (2018). Kritik terhadap Metode Tradisional dalam Pengajaran Bahasa. *Linguistik Indonesia*, 36(1), 34-51.
- Kusuma, W. (2019). Prinsip-Prinsip Desain Korpus Bahasa. Malang: State University of Malang Press.
- Kusumawati, E. (2021). Frekuensi dan Pembelajaran Bahasa: Perspektif Kognitif. *Jurnal Linguistik Terapan*, 11(1), 23-40.
- Kurnia, R., & Ahmad, F. (2018). Pembelajaran Konstruksi Gramatikal Berbasis Korpus. *Indonesian Journal of Applied Linguistics*, 8(2), 178-195.
- Kurniawati, S. (2019). Noticing dan Pembelajaran Bahasa Kedua. *TEFLIN Journal*, 30(1), 89-106.
- Lestari, D. (2021). Construction Grammar dan Pengajaran Bahasa. Yogyakarta: Pustaka Pelajar.

- Lestari, R., & Suryanto, A. (2021). Feedback dalam Pembelajaran Bahasa Berbasis Teknologi. *Jurnal Teknologi Pendidikan*, 13(1), 67-84.
- Mahardika, P., & Indriyana, S. (2020). Pembelajaran Induktif vs Deduktif: Studi Komparatif. *Jurnal Pendidikan dan Pembelajaran*, 27(2), 134-151.
- Maharani, I. (2021). Pembelajaran Konstruksi dalam Perspektif Pedagogis. *Linguistik Indonesia*, 39(1), 56-73.
- Marlina, L. (2018). Analisis Register dalam Korpus Bahasa Indonesia. Bandung: Alfabeta.
- Maulana, H., & Sari, D. (2020). Pergeseran Paradigma dalam Linguistik: Dari Chomsky ke Korpus. *Jurnal Bahasa dan Linguistik*, 15(2), 112-129.
- Mukti, T. (2020). Input Hypothesis Krashen: Relevansi untuk Pembelajaran Bahasa Kontemporer. *TEFLIN Journal*, 31(2), 145-162.
- Nuraini, F. (2018). Kolokasi dalam Bahasa Indonesia: Analisis Korpus. Jakarta: Gramedia Pustaka Utama.
- Nurhayati, E. (2021). Guided Discovery dalam Data-Driven Learning. *Jurnal Pendidikan Bahasa*, 17(2), 178-195.
- Nurhayati, S., Fitriani, D., & Rahman, A. (2020). Korpus Bahasa Indonesia: Pengembangan dan Aplikasi. Bandung: Simbiosis Rekatama Media.
- Nurhayati, T., & Susanti, R. (2021). Pembelajaran Induktif: Manfaat dan Keterbatasan. *Indonesian Journal of Applied Linguistics*, 11(2), 89-106.
- Nurdin, A. (2019). Autentisitas dalam Konteks Pedagogis: Perspektif Kritis. *Jurnal Pendidikan Bahasa dan Sastra*, 19(1), 45-62.
- Nurhasanah, L. (2022). Discovery Learning dalam Pengajaran Bahasa Berbasis Korpus. Malang: UM Press.

- Ningsih, W., & Kusuma, D. (2021). Identifikasi Formulaic Sequences Menggunakan Analisis Korpus. *Jurnal Linguistik Komputasional*, 6(2), 67-84.
- Novita, R. (2020). *Data-Driven Learning dan Kompetensi Bahasa*. Yogyakarta: Graha Ilmu.
- Nugroho, A., & Wijayanti, E. (2019). Teori Usage-Based: Implikasi untuk Pengajaran Bahasa. *Linguistik Indonesia*, 37(1), 23-41.
- Permana, D. (2021). *Konstruktivisme dalam Pembelajaran Bahasa*. Jakarta: Rajawali Pers.
- Permatasari, A. (2020). *Construction Grammar: Teori dan Aplikasi dalam Analisis Korpus*. Bandung: Alfabeta.
- Prabowo, H. (2018). *Multi-Word Expressions dalam Bahasa Inggris*. Surabaya: Airlangga University Press.
- Prasetyo, B., & Rahmawati, D. (2021). Interaction Hypothesis dalam Pembelajaran Bahasa Kedua. *TEFLIN Journal*, 32(1), 78-95.
- Pratama, Y. (2018). Konstruksi dalam Linguistik: Perspektif Form-Meaning Pairing. *Linguistik Indonesia*, 36(2), 156-173.
- Pratiwi, L. (2020). Aksesibilitas Data Bahasa di Era Digital. *Jurnal Teknologi Pendidikan*, 12(2), 89-106.
- Purwanto, E., & Setiyawan, A. (2021). *Teori Register Halliday dan Aplikasinya*. Jakarta: Erlangga.
- Putri, A. (2020). Kurasi Korpus untuk Pembelajaran Bahasa: Prinsip dan Praktik. *Indonesian Journal of Applied Linguistics*, 10(1), 45-62.
- Putra, R. (2022). Analisis Konstruksi Gramatikal Menggunakan Korpus. *Jurnal Linguistik Terapan*, 12(1), 34-51.
- Rachman, F. (2021). *Bukti Empiris dalam Linguistik Korpus*. Bandung: Sinar Baru Algensindo.

- Rahayu, S., & Kusumawati, L. (2020). Autentisitas sebagai Konsep Relasional dalam Pengajaran Bahasa. *Jurnal Pendidikan Bahasa*, 16(2), 123-140.
- Ramadhan, A. (2021). *Teknik Komputasional dalam Analisis Korpus*. Jakarta: Prenada Media.
- Rahmawati, N. (2019). Korpus Web: Peluang dan Tantangan. *Jurnal Linguistik Komputasional*, 4(1), 23-40.
- Riyadi, S. (2018). *Scaffolded Induction dalam Pembelajaran Bahasa*. Yogyakarta: Pustaka Pelajar.
- Rizki, M. (2020). Pemilihan Fitur Linguistik untuk Data-Driven Learning. *Indonesian Journal of Applied Linguistics*, 10(2), 167-184.
- Safitri, A. (2021). *Analisis Kolokasi dalam Linguistik Korpus*. Malang: UM Press.
- Safitri, L., & Hidayat, R. (2020). Formulaic Language dalam Kompetensi Bahasa. *Linguistik Indonesia*, 38(2), 189-206.
- Santoso, B. (2019). Variabilitas Input dalam Pembelajaran Bahasa. *TEFLIN Journal*, 30(1), 56-73.
- Saputra, D. (2018). *Korpus Khusus: Desain dan Aplikasi*. Bandung: Remaja Rosdakarya.
- Saputri, M. (2020). Evolusi Computer-Assisted Language Learning. *Jurnal Teknologi Pendidikan*, 12(1), 34-51.
- Saraswati, N., & Mahendra, I. (2019). Manfaat Korpus dalam Pengajaran Bahasa. *Jurnal Pendidikan Bahasa dan Sastra*, 19(2), 112-129.
- Sari, P., & Wijaya, K. (2020). Leksikon dan Gramatika dalam Construction Grammar. *Linguistik Indonesia*, 38(1), 67-84.
- Setiawan, A. (2022). *Linguistik Kognitif dan Konstruksi*. Jakarta: Kencana.
- Setyawan, R. (2022). Motivasi Pembelajar dalam Pendekatan Berbasis Autentisitas. *Jurnal Pendidikan Bahasa*, 18(1), 45-62.

- Sudarmanto, A., & Yulianto, B. (2020). Noticing dalam Data-Driven Learning. *Indonesian Journal of Applied Linguistics*, 10(2), 123-140.
- Suharto, G. (2018). Frekuensi dan Produktivitas dalam Linguistik Kognitif. *Linguistik Indonesia*, 36(1), 89-106.
- Sulistyo, G. H. (2019). *Linguistik Korpus dalam Pengajaran Bahasa Inggris*. Malang: State University of Malang Press.
- Suryana, A. (2020). *Balance dan Representativeness dalam Desain Korpus*. Bandung: Alfabeta.
- Suryani, D. (2020). *Formulaic Language dalam Materi Pembelajaran*. Yogyakarta: Graha Ilmu.
- Syafitri, R. (2019). Pembelajar sebagai Peneliti dalam Data-Driven Learning. *Jurnal Pendidikan dan Pembelajaran*, 26(2), 156-173.
- Syahputra, I. (2018). *Data-Driven Learning: Konsep dan Implementasi*. Jakarta: Rajawali Pers.
- Simatupang, R., & Harahap, M. (2018). Aksesibilitas Teknologi dalam Pendidikan Bahasa di Indonesia. *Jurnal Teknologi Pendidikan*, 10(2), 89-106.
- Taufik, M. (2018). Interaksi Kolaboratif dalam Pembelajaran Berbasis Korpus. *Indonesian Journal of Applied Linguistics*, 8(1), 67-84.
- Utami, S. (2021). *Anotasi Korpus: Teori dan Praktik*. Bandung: Sinar Baru Algensindo.
- Wahyuni, E. (2022). Noticing Hypothesis Schmidt dalam Pembelajaran Bahasa Kedua. *TEFLIN Journal*, 33(1), 112-129.
- Wardani, K. (2018). Konteks Penggunaan dalam Pengajaran Kosakata. *Jurnal Pendidikan Bahasa*, 14(2), 178-195.
- Wibowo, A. (2019). *Formulaic Sequences dan Kompetensi Lanjut*. *Linguistik Indonesia*, 37(2), 201-218.
- Widiastuti, P. (2020). *Korpus Pembelajaran: Aplikasi dalam Penelitian Akuisisi Bahasa Kedua*. Malang: UM Press.

- Wijaya, M., & Santoso, D. (2019). Perkembangan Linguistik Korpus di Indonesia. *Linguistik Indonesia*, 37(1), 12-29.
- Wijayanti, A. (2021). Autentisitas dalam Pembelajaran Bahasa: Pendekatan Pragmatis. *Jurnal Pendidikan Bahasa dan Sastra*, 21(1), 89-106.
- Wulandari, T. (2018). *British Contextualism dan Linguistik Sistemik Fungsional*. Jakarta: Gramedia Pustaka Utama.
- Yulianti, S. (2018). *Konstruktivisme dalam Pendidikan Bahasa*. Yogyakarta: Pustaka Pelajar.
- Yusuf, H., & Hidayat, A. (2022). Konkordansi dan Analisis Pola Bahasa. *Jurnal Linguistik Komputasional*, 7(1), 45-62.

PROFIL PENULIS



Hanafi Pelu

Lahir dan besar di Desa Wakal, Kecamatan Leihitu, Kabupaten Maluku Tengah, Provinsi Maluku. Menyelesaikan pendidikan SD Negeri 1 Wakal, SMP Negeri Hitu dan SMA Negeri Hila. Menyelesaikan pendidikan sarjana di Universitas Negeri Makassar. Pernah menjadi guru di Madrasah Tsanawiyah Negeri Masohi, Madrasah Aliyah Negeri 2 Ambon dan SUPM Negeri Ambon. Pernah bertugas di Balai Diklat Keagamaan Ambon dan sekarang bertugas di Balai Diklat Keagamaan Makassar. Pernah menulis di Jurnal Waiheru Balai Diklat Keagamaan Ambon, Jurnal Baruga Balai Diklat Keagamaan Makassar, Merit International Journal, Jurnal Andragogi Pusdiklat, Jurnal Universitas Iqro Buru, dan pernah menulis buku.

BAB 9

LINGUISTIK KORPUS DALAM LEKSIKOGRAFI

Loso Judijanto
IPOSS, Jakarta
E-mail: losojudijantobumn@gmail.com

PENDAHULUAN

Linguistik korpus telah merevolusi praktik leksikografi modern dengan menggeser dasar penyusunan kamus dari intuisi ahli menuju bukti empiris yang terukur. Dalam konteks ini, korpus bahasa berfungsi sebagai himpunan data autentik yang merekam penggunaan bahasa nyata dalam berbagai situasi komunikasi. Bagi leksikografi, ketersediaan data yang luas dan terstruktur memungkinkan perumusan entri kamus yang lebih akurat, representatif, dan kontekstual. Pendekatan berbasis korpus tidak hanya memperkaya pemahaman tentang frekuensi dan distribusi kata, tetapi juga membuka akses terhadap variasi makna, pola kolokasi, serta pergeseran penggunaan leksikal dalam masyarakat bahasa. Oleh karena itu, integrasi linguistik korpus dalam leksikografi tidak lagi dipandang sebagai inovasi tambahan, melainkan sebagai fondasi metodologis yang menentukan kualitas dan relevansi produk leksikografis (Hashimoto & Nelson, 2025).

Perkembangan teknologi komputasi bahasa turut mempercepat transformasi ini dengan menyediakan perangkat analisis yang mampu mengekstraksi pola leksikal secara sistematis. Dalam praktik leksikografi tradisional, pemilihan lema dan penentuan makna sering bergantung pada pengamatan terbatas terhadap teks tertentu. Sebaliknya, linguistik korpus memungkinkan pengumpulan data dari berbagai sumber yang mencerminkan keragaman penggunaan bahasa, termasuk ragam

formal, informal, akademik, hingga media digital (Nuraisyah et al., 2023).

LANDASAN OPERASIONAL LINGUISTIK KORPUS DALAM LEKSIKOGRAFI

Pemanfaatan linguistik korpus dalam leksikografi menghadirkan pendekatan empiris yang menempatkan data kebahasaan nyata sebagai fondasi penyusunan kamus. Orientasi ini menggeser praktik intuitif menuju prosedur berbasis bukti. Proses pengumpulan, pemilahan, dan analisis data bahasa membentuk sistem kerja leksikografi modern yang menekankan akurasi penggunaan, variasi makna, dan representasi pemakaian autentik dalam masyarakat bahasa (Glogar, 2025).

Desain Korpus untuk Tujuan Leksikografi

Desain korpus untuk kebutuhan leksikografi menuntut perencanaan sistematis yang mempertimbangkan tujuan deskriptif kamus, jenis pengguna, serta ruang lingkup bahasa yang akan direpresentasikan. Proses desain tidak sekadar mengumpulkan teks dalam jumlah besar, melainkan membangun kumpulan data yang secara metodologis mampu mencerminkan variasi pemakaian bahasa secara nyata. Struktur korpus dirancang melalui seleksi genre, sumber, medium komunikasi, serta periode waktu penggunaan bahasa. Pendekatan ini memastikan bahwa setiap lema yang dihasilkan dalam kamus memiliki dasar empiris yang kuat dan dapat ditelusuri pada konteks pemakaiannya. Representasi data menjadi aspek utama karena korpus yang tidak dirancang secara terarah akan menghasilkan bias distribusi kosakata. Praktik leksikografi modern menempatkan korpus sebagai sumber utama untuk mengidentifikasi kecenderungan frekuensi, pola penggunaan, serta perubahan makna. Oleh karena itu, desain korpus tidak hanya bersifat teknis, tetapi juga epistemologis,

karena menentukan bagaimana bahasa dipahami dan direpresentasikan dalam kamus (Hashimoto & Nelson, 2025).

Desain korpus yang berorientasi pada leksikografi juga memperhatikan prinsip keberagaman sumber bahasa sehingga variasi sosial, regional, dan fungsional dapat terwakili secara proporsional. Praktik ini memastikan bahwa kamus tidak hanya mencerminkan bahasa standar, tetapi juga dinamika penggunaan aktual dalam masyarakat. Strategi *sampling* menjadi mekanisme utama untuk menjaga keseimbangan data sehingga setiap kategori teks memiliki kontribusi yang terukur terhadap keseluruhan korpus. Pendekatan berbasis bukti ini menghasilkan proses seleksi lema yang lebih objektif serta memungkinkan pemetaan hubungan leksikal secara sistematis (Hashimoto & Nelson, 2025).

Kriteria Representativitas dan Keseimbangan Korpus Leksikografis

Representativitas dalam korpus leksikografis merujuk pada kemampuan data bahasa untuk mencerminkan pola penggunaan bahasa secara luas dan beragam. Prinsip ini memastikan bahwa frekuensi dan distribusi kosakata yang dihasilkan dari analisis korpus tidak bersifat parsial. Korpus yang representatif memperlihatkan distribusi teks yang mencakup berbagai domain komunikasi seperti akademik, jurnalistik, percakapan, dan media digital. Keseimbangan korpus berperan sebagai mekanisme pengendalian terhadap dominasi jenis teks tertentu yang dapat menggeser gambaran umum penggunaan bahasa. Pengukuran representativitas dilakukan melalui evaluasi distribusi genre, frekuensi bentuk leksikal, serta keberagaman konteks penggunaan (Glogar, 2025).

Prinsip keseimbangan korpus juga berkaitan dengan dinamika perubahan bahasa yang terjadi dalam masyarakat. Representativitas tidak bersifat statis, melainkan menuntut

pembaruan data secara berkala agar kamus tetap relevan dengan perkembangan penggunaan bahasa. Pendekatan ini menempatkan korpus sebagai sistem terbuka yang berkembang mengikuti perubahan sosial dan teknologi komunikasi. Korpus leksikografis yang seimbang memungkinkan identifikasi makna dominan serta variasi penggunaan kata secara lebih akurat. Pengelolaan distribusi data menjadi landasan dalam menentukan frekuensi lema, penyusunan definisi, serta pemilihan contoh penggunaan. Representativitas korpus tidak ditentukan oleh ukuran semata, tetapi oleh distribusi data yang mampu mencerminkan variasi penggunaan bahasa dalam praktik komunikasi nyata (Glogar, 2025).

Teknik Ekstraksi Lema dan Kandidat Entri

Ekstraksi lema dalam linguistik korpus merupakan proses identifikasi bentuk leksikal yang layak menjadi entri kamus berdasarkan data penggunaan aktual. Teknik ini memanfaatkan analisis frekuensi, pemrosesan morfologis, serta algoritma pengelompokan kata untuk menemukan kandidat lema secara sistematis. Pendekatan berbasis korpus memungkinkan leksikografer menghindari seleksi intuitif karena setiap kandidat entri didasarkan pada bukti distribusi dalam data bahasa. Proses ekstraksi melibatkan normalisasi bentuk kata, pengelompokan varian morfologis, serta identifikasi satuan leksikal multikata. Metode komputasional mempermudah pemetaan hubungan antara bentuk dasar dan variasi penggunaan sehingga struktur lema dapat dibangun secara konsisten (Bakhtiyorovna, 2025).

Pengembangan teknik ekstraksi lema memperlihatkan pergeseran dari pendekatan manual menuju sistem otomatis berbasis komputasi linguistik. Analisis distribusi kata memberikan informasi mengenai produktivitas leksikal serta kecenderungan pembentukan makna baru. Data korpus berfungsi sebagai sumber empiris untuk menentukan batasan

lema dan relasi semantik antar-entri. Pendekatan ini menghasilkan struktur kamus yang lebih sistematis dan transparan karena setiap keputusan leksikografis memiliki dasar bukti penggunaan. Ekstraksi lema berbasis korpus memungkinkan penyusunan kamus yang mencerminkan frekuensi dan variasi bentuk kata secara objektif melalui analisis data bahasa autentik (Bakhtiyorovna, 2025).

Analisis Frekuensi dan Distribusi Leksikal

Analisis frekuensi dalam linguistik korpus berfungsi sebagai mekanisme utama untuk mengidentifikasi tingkat produktivitas kosakata dalam penggunaan bahasa. Frekuensi kemunculan kata memberikan gambaran mengenai prioritas penyusunan entri serta tingkat relevansi bentuk leksikal dalam komunikasi. Distribusi leksikal memperlihatkan bagaimana suatu kata digunakan dalam berbagai konteks dan domain komunikasi. Pendekatan statistik dalam linguistik korpus menghasilkan pemetaan kosakata yang bersifat kuantitatif sekaligus deskriptif. Data frekuensi menjadi dasar dalam menentukan urutan penyajian makna dalam entri kamus (Wahyuningtyas & Kesuma, 2024).

Distribusi leksikal memberikan wawasan mengenai fungsi pragmatik dan semantik suatu kata dalam praktik komunikasi. Informasi mengenai konteks kemunculan kata memungkinkan penyusunan definisi yang mencerminkan penggunaan aktual. Pendekatan ini menempatkan frekuensi sebagai indikator relevansi leksikal dalam masyarakat bahasa. Analisis berbasis korpus memperlihatkan hubungan antara frekuensi, makna, dan variasi penggunaan secara sistematis. Proses ini memperkuat validitas deskripsi leksikal dalam kamus modern. Frekuensi dan distribusi leksikal dalam korpus menyediakan dasar empiris untuk menentukan relevansi makna dan prioritas penyajian entri dalam kamus (Wahyuningtyas & Kesuma, 2024).

Identifikasi Pola Kolokasi untuk Penyusunan Entri

Kolokasi merupakan hubungan kemunculan bersama antara kata dalam konteks tertentu yang mencerminkan pola penggunaan bahasa alami. Identifikasi kolokasi dalam linguistik korpus memberikan informasi penting mengenai preferensi kombinasi leksikal yang tidak selalu dapat dijelaskan melalui aturan gramatikal. Analisis kolokasi membantu leksikografer memahami makna kontekstual kata serta batasan penggunaannya. Informasi ini menjadi dasar dalam penyusunan definisi, contoh pemakaian, serta informasi frasa dalam entri kamus. Metode statistika digunakan untuk mengukur kekuatan hubungan antar kata sehingga pola kolokasi dapat diidentifikasi secara objektif. Analisis ini juga memungkinkan identifikasi makna idiomatik serta variasi semantik yang muncul dalam konteks tertentu (Nuraisyah et al., 2023).

PEMANFAATAN DATA KORPUS DALAM PENYUSUNAN ENTRI KAMUS

Penentuan Makna Berdasarkan Bukti Korpus

Penentuan makna dalam penyusunan entri kamus berbasis linguistik korpus bertumpu pada pengamatan sistematis terhadap pemakaian kata dalam konteks nyata. Data korpus menyediakan ribuan hingga jutaan contoh penggunaan yang menampilkan variasi semantik secara empiris, sehingga definisi leksikal tidak lagi bergantung pada intuisi penyusun kamus semata. Analisis distribusi konteks memungkinkan pengidentifikasian makna dominan, makna turunan, serta pergeseran makna dalam penggunaan aktual. Frekuensi kemunculan bentuk leksikal dalam konteks tertentu menjadi indikator penting dalam menetapkan prioritas makna utama. Pengamatan terhadap lingkungan leksikal yang konsisten memperlihatkan hubungan semantik yang stabil dan dapat digunakan sebagai dasar

deskripsi makna. Pendekatan berbasis bukti juga memfasilitasi identifikasi nuansa makna yang sering luput dalam pendekatan tradisional. Proses analisis melibatkan pemilahan konteks berdasarkan domain penggunaan, fungsi komunikatif, serta hubungan sintagmatik yang muncul secara berulang. Representasi makna yang dihasilkan menjadi lebih deskriptif daripada preskriptif karena berlandaskan fakta penggunaan. Penyusunan definisi tidak sekadar menjelaskan arti kata, melainkan menggambarkan pola pemakaian yang aktual dalam komunikasi. Data korpus juga memungkinkan pemetaan perubahan makna secara diakronis ketika korpus mencakup data lintas waktu (Goncharenko, 2023).

Pendalaman makna berbasis korpus memperlihatkan bahwa makna leksikal terbentuk melalui pola penggunaan yang berulang dalam komunitas tutur. Analisis konkordansi mengungkap kecenderungan semantik melalui pengamatan posisi kata dalam struktur kalimat dan asosiasi dengan kata lain. Pemodelan distribusional memungkinkan pemetaan hubungan makna melalui kedekatan konteks pemakaian. Deskripsi makna kemudian dirumuskan melalui generalisasi pola penggunaan yang konsisten. Pendekatan ini membantu mengurangi subjektivitas dalam interpretasi makna serta meningkatkan reliabilitas deskripsi leksikografis. Penerapan metode statistik sederhana seperti penghitungan frekuensi relatif memberikan gambaran mengenai dominasi makna tertentu dalam praktik bahasa. Integrasi analisis kualitatif dan kuantitatif memperkaya proses interpretasi data. Penyusunan definisi berbasis bukti juga memperkuat fungsi kamus sebagai dokumentasi praktik bahasa aktual (Goncharenko, 2023).

Pemodelan Polisemi dan Variasi Makna

Fenomena polisemi menjadi salah satu tantangan utama dalam leksikografi karena satu bentuk leksikal sering

merepresentasikan berbagai makna yang saling berkaitan. Linguistik korpus menyediakan kerangka analitis yang memungkinkan pemodelan polisemi secara sistematis melalui pengamatan distribusi konteks penggunaan. Data korpus memperlihatkan bahwa variasi makna tidak muncul secara acak, melainkan mengikuti pola penggunaan yang konsisten dalam lingkungan leksikal tertentu. Identifikasi makna dilakukan melalui pengelompokan konteks berdasarkan kesamaan fungsi semantik. Analisis ini menghasilkan pemetaan hubungan antar makna yang bersifat bertingkat, mulai dari makna inti hingga makna turunan. Pendekatan berbasis korpus membantu membedakan polisemi dari homonimi melalui pengamatan kontinuitas konteks penggunaan. Variasi makna dapat dilacak melalui perubahan asosiasi leksikal yang menyertai suatu kata dalam konteks berbeda. Metode klasterisasi konteks memungkinkan pengelompokan penggunaan berdasarkan kemiripan distribusi (Turnip et al., 2025).

Pemanfaatan data korpus dalam pemodelan polisemi juga memperlihatkan bahwa variasi makna dipengaruhi oleh faktor pragmatik dan sosial. Konteks komunikasi memainkan peran penting dalam pembentukan interpretasi makna tertentu. Analisis penggunaan dalam berbagai register mengungkap bahwa makna dapat berubah sesuai dengan domain komunikasi. Pengamatan terhadap frekuensi distribusi makna membantu menentukan urutan penyajian dalam entri kamus. Makna yang paling sering muncul ditempatkan sebagai makna utama, sedangkan makna yang lebih jarang disajikan sebagai turunan. Pendekatan berbasis bukti memungkinkan penyusunan struktur entri yang mencerminkan realitas penggunaan bahasa (Turnip et al., 2025).

Penyusunan Contoh Pemakaian Otentik

Contoh pemakaian dalam entri kamus berfungsi sebagai representasi konkret dari penggunaan kata dalam konteks komunikatif. Linguistik korpus menyediakan sumber data autentik yang memungkinkan penyusunan contoh penggunaan berbasis bukti empiris. Contoh yang diambil dari korpus mencerminkan variasi struktur sintaksis, fungsi pragmatik, serta nuansa makna dalam komunikasi nyata. Penggunaan data autentik meningkatkan kejelasan interpretasi makna karena pengguna kamus dapat melihat bagaimana kata digunakan dalam situasi sebenarnya. Proses pemilihan contoh melibatkan seleksi konteks yang representatif, jelas, dan informatif. Korpus memungkinkan identifikasi pola penggunaan yang paling umum sehingga contoh yang dipilih mencerminkan praktik bahasa yang dominan. Representasi konteks yang beragam juga memperlihatkan fleksibilitas penggunaan kata dalam berbagai situasi. Penyusunan contoh berbasis korpus memperkuat hubungan antara definisi dan praktik penggunaan. Pendekatan ini meningkatkan kualitas pedagogis kamus karena pengguna memperoleh gambaran nyata tentang fungsi kata dalam komunikasi (Hasbially, 2026).

Kualitas contoh pemakaian berbasis korpus ditentukan oleh relevansi konteks terhadap makna yang dijelaskan dalam entri. Seleksi contoh mempertimbangkan kejelasan struktur, keterbacaan, serta kesesuaian dengan makna yang direpresentasikan. Data korpus memungkinkan penyusunan contoh yang menunjukkan variasi penggunaan tanpa kehilangan kealamian. Contoh autentik memperlihatkan hubungan antara makna leksikal dan fungsi pragmatik dalam komunikasi. Representasi penggunaan nyata memperkaya deskripsi leksikal dengan dimensi kontekstual. Penggunaan contoh berbasis korpus meningkatkan akurasi dan kejelasan entri kamus karena contoh mencerminkan praktik bahasa aktual (Hasbially, 2026).

Analisis Register dan Variasi Penggunaan Leksikal

Analisis register dalam leksikografi berbasis korpus berfokus pada identifikasi perbedaan penggunaan leksikal berdasarkan konteks sosial, situasional, dan fungsional bahasa. Data korpus memungkinkan pengamatan sistematis terhadap variasi penggunaan kata dalam domain komunikasi yang berbeda seperti akademik, jurnalistik, percakapan informal, maupun teks administratif. Distribusi kata dalam berbagai jenis teks memperlihatkan kecenderungan pemakaian yang tidak seragam. Variasi tersebut menjadi indikator penting dalam penyusunan entri kamus karena menunjukkan batasan penggunaan yang relevan bagi pengguna bahasa. Representasi informasi register membantu menjelaskan apakah suatu kata bersifat formal, teknis, atau kolokial. Data distribusional memberikan dasar empiris dalam menentukan label register pada entri leksikografis. Informasi mengenai domain penggunaan memperkaya deskripsi makna dengan dimensi pragmatik. Pengamatan terhadap pola pemakaian juga memperlihatkan hubungan antara pilihan leksikal dan fungsi komunikasi. Penyusunan entri yang mempertimbangkan variasi register menghasilkan kamus yang lebih representatif terhadap praktik bahasa aktual. Analisis register berbasis korpus meningkatkan presisi deskripsi leksikal karena informasi penggunaan didasarkan pada bukti empiris (Herlina & Hendrokumoro, 2024).

Pendalaman analisis register melalui linguistik korpus memperlihatkan bahwa variasi penggunaan leksikal dipengaruhi oleh tujuan komunikasi dan karakteristik komunitas tutur. Perbedaan distribusi kata dalam korpus spesifik domain menunjukkan bahwa pilihan leksikal mencerminkan norma komunikatif tertentu. Pengamatan terhadap pola penggunaan memungkinkan identifikasi kecenderungan semantik yang khas pada setiap register. Deskripsi leksikografis kemudian

merumuskan variasi penggunaan sebagai bagian integral dari makna. Informasi register tidak hanya menjelaskan konteks pemakaian, tetapi juga memberikan panduan interpretasi makna dalam situasi tertentu. Representasi variasi penggunaan membantu menghindari generalisasi berlebihan dalam definisi leksikal. Analisis berbasis korpus memberikan kerangka metodologis yang lebih akurat dalam mengidentifikasi variasi penggunaan daripada pendekatan introspektif (Herlina & Hendrokumoro, 2024).

Integrasi Informasi Gramatikal Berbasis Korpus

Integrasi informasi gramatikal dalam entri kamus berbasis linguistik korpus melibatkan deskripsi pola sintaksis yang menyertai penggunaan kata dalam konteks nyata. Data korpus memungkinkan pengamatan langsung terhadap struktur gramatikal yang paling sering muncul bersama suatu leksem. Pola kolokasi gramatikal memberikan informasi mengenai kategori kata, fungsi sintaktis, serta konstruksi yang dominan. Representasi informasi gramatikal berbasis bukti empiris meningkatkan keakuratan deskripsi leksikografis. Penyusunan entri tidak hanya memuat definisi makna, tetapi juga pola penggunaan struktural yang relevan. Pendekatan ini membantu pengguna memahami cara penggunaan kata dalam struktur kalimat yang benar. Analisis korpus memperlihatkan bahwa pilihan struktur gramatikal tidak bersifat acak, melainkan mengikuti kecenderungan distribusional tertentu. Data frekuensi konstruksi sintaktis menjadi dasar dalam menentukan pola penggunaan utama. Integrasi informasi gramatikal berbasis korpus menghasilkan deskripsi lebih komprehensif mengenai perilaku leksikal. Representasi pola gramatikal membantu menjelaskan hubungan antara bentuk dan fungsi dalam penggunaan bahasa (Desfiona et al., 2025).

Pengembangan leksikografi modern menempatkan informasi gramatikal sebagai komponen integral dalam representasi makna leksikal. Linguistik korpus menyediakan kerangka metodologis untuk mengidentifikasi pola struktur melalui analisis konkordansi dan distribusi sintaktis. Integrasi data struktural memperlihatkan bahwa perilaku gramatikal merupakan bagian dari makna penggunaan. Penyusunan entri berbasis korpus memungkinkan penyajian informasi mengenai konstruksi yang paling produktif. Integrasi informasi gramatikal berbasis korpus menghasilkan deskripsi leksikal yang lebih komprehensif karena pola struktur diidentifikasi melalui penggunaan nyata (Desfiona et al., 2025).

INOVASI LEKSIKOGRAFI BERBASIS LINGUISTIK KORPUS

Otomatisasi Penyusunan Entri dengan Teknologi Korpus

Otomatisasi dalam penyusunan entri kamus merepresentasikan perubahan metodologis yang mendasar dalam praktik leksikografi modern. Proses identifikasi lema, pengelompokan bentuk turunan, serta pemetaan distribusi makna kini dapat dilakukan melalui sistem komputasional yang memproses korpus dalam skala besar secara efisien dan konsisten. Teknologi pemrosesan bahasa alami berperan penting dalam mengekstraksi kandidat entri melalui analisis frekuensi, asosiasi leksikal, serta struktur sintagmatik yang berulang dalam korpus. Keunggulan utama otomatisasi terletak pada kemampuannya mengurangi bias subjektif, mempercepat proses kompilasi, dan menyediakan dasar empiris yang transparan bagi setiap keputusan leksikografis. Integrasi algoritma klasifikasi semantik memungkinkan sistem mengenali hubungan makna antarlema secara sistematis, sehingga struktur entri kamus dapat disusun secara lebih koheren (Siregar et al., 2025).

Pemanfaatan teknologi korpus dalam otomatisasi penyusunan entri tidak hanya berkaitan dengan efisiensi, tetapi juga menyangkut validitas deskripsi makna yang dihasilkan. Sistem komputasional mampu menampilkan bukti penggunaan autentik dalam jumlah besar yang memperlihatkan variasi konteks secara sistematis. Analisis distribusi semantik berbasis korpus memungkinkan identifikasi makna dominan, makna perifer, serta pergeseran penggunaan dalam berbagai domain komunikasi. Hasil pengolahan data tersebut memberikan landasan objektif bagi perumusan definisi leksikal yang mencerminkan praktik bahasa aktual, bukan hanya norma preskriptif. Proses ini menunjukkan bahwa inovasi teknologi tidak menggantikan interpretasi linguistik, melainkan memperluas ruang analisis yang dapat diakses oleh leksikografer (Siregar et al., 2025).

Pengembangan Kamus Elektronik Berbasis Data Korpus

Kamus elektronik berbasis korpus merepresentasikan evolusi format leksikografi dari media statis menuju sistem informasi linguistik yang dinamis dan interaktif. Integrasi data korpus memungkinkan kamus digital tidak hanya menyajikan definisi, tetapi juga menyediakan bukti penggunaan nyata dalam berbagai konteks komunikasi. Pengguna memperoleh akses langsung terhadap contoh autentik yang memperlihatkan variasi makna, kolokasi, serta pola sintaksis yang menyertai suatu lema. Struktur kamus elektronik memungkinkan penyajian informasi leksikal secara berlapis, mencakup frekuensi penggunaan, distribusi register, dan relasi semantik antarkata. Teknologi pencarian berbasis korpus memberikan fleksibilitas bagi pengguna untuk menelusuri pola bahasa berdasarkan kebutuhan spesifik, sehingga kamus berfungsi sebagai alat eksplorasi linguistik yang adaptif. Pengembangan kamus digital berbasis data empiris juga meningkatkan relevansi leksikografi dalam

lingkungan komunikasi yang terus berubah (Febrina et al., 2024).

Perkembangan kamus elektronik berbasis korpus menunjukkan bahwa teknologi digital memperluas fungsi leksikografi sebagai sumber pengetahuan linguistik yang kontekstual. Integrasi visualisasi data dalam kamus digital memperlihatkan distribusi frekuensi dan hubungan kolokasional secara intuitif, sehingga pemahaman makna menjadi lebih komprehensif. Sistem digital berbasis korpus juga memungkinkan personalisasi akses informasi sesuai kebutuhan pengguna, misalnya berdasarkan domain, tingkat formalitas, atau jenis teks. Inovasi ini memperlihatkan bahwa leksikografi modern bergerak menuju model representasi bahasa yang berbasis data dan berorientasi pada pengguna. Kamus digital berbasis korpus menghadirkan paradigma baru leksikografi karena makna tidak lagi disajikan sebagai definisi statis, melainkan sebagai representasi penggunaan bahasa yang dapat diamati secara langsung dalam data autentik (Febrina et al., 2024).

Pemutakhiran Leksikon melalui Pemantauan Korpus Dinamis

Pemantauan korpus dinamis memungkinkan leksikografi mengikuti perubahan bahasa secara real time melalui pengumpulan data yang berkelanjutan. Pendekatan ini menempatkan bahasa sebagai sistem yang selalu berkembang, sehingga leksikon harus diperbarui secara sistematis untuk mempertahankan relevansinya. Korpus dinamis menghimpun data dari berbagai sumber komunikasi kontemporer, termasuk media digital, teks jurnalistik, dan wacana publik, sehingga variasi penggunaan bahasa dapat diamati secara luas. Analisis longitudinal terhadap data korpus memungkinkan identifikasi kemunculan kata baru, pergeseran makna, serta perubahan

frekuensi penggunaan dalam rentang waktu tertentu. Proses pemutakhiran leksikon tidak lagi bersifat periodik, melainkan berkelanjutan sesuai dinamika data bahasa yang tersedia. Pendekatan ini meningkatkan sensitivitas leksikografi terhadap inovasi linguistik yang muncul dalam praktik komunikasi nyata. Pemantauan berbasis korpus dinamis juga memungkinkan evaluasi keberlangsungan suatu bentuk leksikal melalui analisis stabilitas distribusinya dalam data. Transformasi metodologis ini menunjukkan bahwa leksikografi modern tidak hanya mendeskripsikan bahasa, tetapi juga merekam evolusi penggunaannya secara sistematis (Ashari et al., 2024).

Pemanfaatan korpus dinamis dalam pemutakhiran leksikon memperlihatkan bahwa perubahan bahasa dapat didokumentasikan secara empiris melalui observasi penggunaan aktual. Sistem pemantauan otomatis mampu mendeteksi pola kemunculan bentuk leksikal baru dan mengukur tingkat penyebarannya dalam berbagai domain komunikasi. Data tersebut memberikan dasar objektif bagi keputusan leksikografis mengenai kelayakan suatu bentuk untuk dimasukkan dalam kamus. Pendekatan berbasis bukti ini meningkatkan transparansi proses pemutakhiran leksikon dan mengurangi ketergantungan pada intuisi individual (Ashari et al., 2024).

Evaluasi Kualitas Entri Leksikografis Berbasis Bukti Korpus

Evaluasi kualitas entri kamus berbasis korpus menempatkan penggunaan bahasa nyata sebagai tolok ukur utama dalam menilai ketepatan deskripsi leksikal. Pendekatan ini berangkat dari asumsi bahwa validitas suatu definisi tidak dapat dilepaskan dari bukti empirik yang menunjukkan bagaimana kata digunakan dalam berbagai konteks komunikasi. Analisis kualitas entri dilakukan melalui pemeriksaan kesesuaian antara definisi, contoh penggunaan, serta pola distribusi leksikal yang tercatat

dalam korpus. Representativitas data menjadi aspek penting karena kualitas deskripsi makna bergantung pada keragaman konteks penggunaan yang dianalisis. Evaluasi berbasis korpus juga mencakup pengujian konsistensi antar-entri melalui analisis hubungan semantik dan pola kolokasional yang berulang dalam data bahasa. Proses ini memungkinkan leksikografer mengidentifikasi ketidaksesuaian antara definisi normatif dan praktik penggunaan aktual. Pendekatan empiris dalam evaluasi entri memperkuat objektivitas leksikografi karena setiap keputusan deskriptif dapat ditelusuri kembali pada bukti penggunaan nyata. Transformasi ini menegaskan bahwa kualitas kamus tidak hanya ditentukan oleh kelengkapan informasi, tetapi juga oleh akurasi representasi penggunaan bahasa dalam data yang dapat diverifikasi (Nuratri et al., 2025).

Evaluasi kualitas entri berbasis bukti korpus juga memperluas pemahaman mengenai hubungan antara makna leksikal dan konteks penggunaannya. Proses verifikasi deskripsi makna dilakukan melalui analisis kesesuaian antara definisi dan distribusi konteks dalam data korpus, sehingga makna tidak dipahami sebagai entitas abstrak, melainkan sebagai pola penggunaan yang dapat diamati. Evaluasi leksikografis berbasis korpus memperkuat validitas deskripsi makna karena kesesuaian antara definisi dan pola penggunaan dapat diuji secara empiris melalui distribusi konteks dalam data bahasa (Nuratri et al., 2025).

Arah Pengembangan Leksikografi Berbasis Korpus di Masa Depan

Perkembangan leksikografi berbasis korpus menunjukkan kecenderungan menuju integrasi yang semakin kuat antara teknologi komputasional dan analisis linguistik empiris. Arah pengembangan masa depan menekankan penggunaan kecerdasan buatan dalam menginterpretasi pola makna yang

kompleks dari data bahasa berskala besar. Sistem analisis semantik berbasis pembelajaran mesin memungkinkan pengenalan hubungan makna yang lebih halus daripada metode statistika konvensional. Integrasi teknologi ini memperluas kemampuan leksikografi dalam merepresentasikan variasi makna, perubahan penggunaan, serta hubungan konseptual antarlema. Pemanfaatan *big data* memungkinkan kamus berfungsi sebagai sistem representasi bahasa yang adaptif dan terus diperbarui sesuai perubahan praktik komunikasi. Transformasi ini menegaskan bahwa masa depan leksikografi terletak pada kemampuan mengintegrasikan analisis linguistik, teknologi komputasional, dan data penggunaan nyata secara berkelanjutan (Almos et al., 2023).

Arah pengembangan leksikografi berbasis korpus juga mencerminkan pergeseran paradigma dari deskripsi statis menuju pemodelan bahasa sebagai sistem dinamis yang dapat diprediksi melalui data. Integrasi teknologi kecerdasan buatan memungkinkan sistem leksikografi memproses variasi penggunaan bahasa dalam skala yang tidak dapat dicapai melalui metode manual. Pendekatan berbasis *big data* mendukung penyusunan kamus yang responsif terhadap perubahan sosial, budaya, dan teknologi yang memengaruhi praktik komunikasi. Pengembangan antarmuka interaktif memungkinkan pengguna berinteraksi langsung dengan data korpus untuk eksplorasi makna kata secara nyata (Almos et al., 2023).

KESIMPULAN

Integrasi linguistik korpus dalam leksikografi menandai transformasi metodologis yang mendasar dalam cara kosakata bahasa dideskripsikan dan direpresentasikan. Melalui desain korpus yang representatif, analisis frekuensi, identifikasi kolokasi, serta pengamatan konteks penggunaan, penyusunan

entri kamus beralih dari pendekatan intuitif menuju deskripsi berbasis bukti empiris. Pemanfaatan data korpus memungkinkan penentuan makna yang lebih akurat, pemodelan variasi semantik yang lebih sistematis, serta penyajian contoh penggunaan yang autentik. Selain itu, perkembangan teknologi komputasi memperluas fungsi korpus dalam otomatisasi penyusunan entri, pembaruan leksikon, dan evaluasi kualitas deskripsi leksikal. Integrasi teknologi komputasional memungkinkan pengelolaan data berskala besar sehingga desain korpus kini melibatkan pertimbangan metadata, anotasi linguistik, serta kompatibilitas perangkat lunak analisis.

Kolokasi berfungsi sebagai jembatan antara makna leksikal dan praktik komunikasi. Analisis kolokasi berbasis korpus memungkinkan leksikografi menggambarkan makna kata melalui hubungan penggunaannya dalam konteks nyata. Pemanfaatan data korpus menghadirkan perubahan metodologis dalam praktik leksikografi modern. Bukti penggunaan nyata menggantikan intuisi sebagai dasar analisis makna, struktur, dan distribusi leksikal. Pendekatan berbasis data memungkinkan deskripsi bahasa yang lebih akurat, adaptif terhadap perubahan, serta mampu merepresentasikan dinamika pemakaian kata dalam berbagai konteks sosial dan komunikatif.

Leksikografi kontemporer mengalami transformasi signifikan melalui pemanfaatan data bahasa berskala besar yang dihimpun secara sistematis. Praktik penyusunan kamus tidak lagi bertumpu pada intuisi semata, melainkan pada bukti empirik penggunaan bahasa nyata. Integrasi teknologi komputasional dengan analisis korpus membuka kemungkinan inovasi metodologis yang memperluas akurasi deskripsi leksikal sekaligus meningkatkan relevansi kamus dalam dinamika komunikasi modern.

Hubungan antara linguistik korpus dan leksikografi membentuk kerangka kerja yang menempatkan bahasa sebagai

fenomena dinamis yang dapat diamati melalui data nyata. Dengan demikian, leksikografi berbasis korpus tidak hanya meningkatkan akurasi kamus sebagai produk ilmiah, tetapi juga memperkuat relevansinya sebagai representasi penggunaan bahasa dalam masyarakat. Pendekatan ini menunjukkan bahwa pemahaman kosakata tidak dapat dipisahkan dari bukti penggunaan aktual, sehingga linguistik korpus menjadi fondasi epistemologis yang menentukan arah perkembangan leksikografi modern.

DAFTAR PUSTAKA

- Almos, R., Pramono, P., Seswita, S., Asdaqul Asma, R., & Putri, N. O. (2023). Linguistik korpus: Sarana dan media pembelajaran pada mata kuliah leksikologi dan leksikografi di perguruan tinggi. *Lectura: Jurnal Pendidikan*, 14(1). <https://doi.org/https://doi.org/10.31849/lectura.v14i1.11705>
- Ashari, M., Suryadarma, Y., & Sofian, A. A. (2024). CORPUS-BASED DESIGN OF THE “SOCIAL MEDIA RESEARCH DICTIONARY” AT UNIDA GONTOR. *Tadris Al-‘Arabiyyah: Jurnal Pendidikan Bahasa Arab Dan Kebahasaaraban* 4, 4(2), 186–198. <https://doi.org/https://doi.org/10.15575/ta.v4i2.46655>
- Bakhtiyorovna, E. D. (2025). The Role of Corpus Linguistics in Modern Lexicography. *International Journal of Artificial Intelligence*, 5(11), 1179–1182.
- Desfiona, E., Hardiasari, S. A. P., Qorizki, D., Awaliyah, S., & Baehaqie, I. (2025). Analisis Komponen Makna Kata Bermakna Berbicara dalam Bahasa Indonesia dan Bahasa Jawa. *Jurnal Ilmiah KORPUS*, 9(2), 264–283. <https://doi.org/10.33369/jik.v9i2.38692>
- Febrina, R., Suhandano, & Sutrisno, A. (2024). Indonesian Women in Kamus Besar Bahasa Indonesia (KBBI) (1988–2018): A Lexicographic Corpus. *Journal Arbitrer*, 11(3),

- 297–313. <https://doi.org/10.25077/ar.11.3.297-313.2024>
- Glogar, O. (2025). Representative Corpus of the Czech Legal Language: A Tool for Jurisprudence. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 38(5), 1681–1703. <https://doi.org/10.1007/s11196-025-10254-3>
- Goncharenko, I. G. (2023). Using language corpora in building learners' combinatorial dictionaries. *Philological Class*, 28(2), 55–68.
- Hasbiallah, Z. H. (2026). Kolokasi dan Prosodi Semantik pada Kata Capek, Lelah dan Penat Berbasis Linguistik Korpus. *Kajian Linguistik Dan Sastra*, 5(1), 130–144. <https://doi.org/10.22437/kalistra.v5i1.51419>
- Hashimoto, B., & Nelson, K. (2025). Recent trends in corpus design and reporting: A methodological synthesis. *Research in Corpus Linguistics*, 12(1), 59–88. <https://doi.org/https://doi.org/10.32714/ricl.12.01.03>
- Herlina, D., & Hendrokumoro, H. (2024). Variasi dan Komponen Makna Verba Pewarta pada Korpus Berita Daring. *Humanis*, 28(1). <https://doi.org/https://doi.org/10.24843/JH.2024.v28.i01.p01>
- Nuraisyah, D., Sumarni, S., Nor Azizah, R., & Syahid, A. (2023). Menyelidiki pola kolokasi dalam terjemahan bahasa Inggris dan bahasa Indonesia dari The Little Prince: Studi korpus. *Jurnal Pendidikan Bahasa*, 15(2). <https://doi.org/10.37630/jpb.v15i2.2923>
- Nuratri, R. K., Kusumaningtyas, D. P. S., Rahayu, K. I. S., Hasel, C. A., & Sandyaningrum, J. N. (2025). The Application of Mandarin Industry Vocabulary Corpus Pages to Improve Translators' Competence in the Industrial World. *ABISATYA: Journal of Community Engagement*, 3(2), 9–19.

- Siregar, A. L., Manik, C. S., Dalimunthe, E. S., Amaliyah, R., Mangunsong, S. E., & Puteri, A. (2025). Perkembangan Leksikografi Digital di Indonesia: Analisis Literatur Inovasi dan Tantangan. *Jurnal Pendidikan Tambusai*, 9(3), 37467–37471. <https://doi.org/10.31004/jptam.v9i3.34537>
- Turnip, P. F., Chaniago, F. H., Sinaga, S. R. D., & Sari, C. J. (2025). Permasalahan Ketidakjelasan Definisi dalam Penyusunan Entri Kamus Bahasa Indonesia. *Jurnal Pendidikan Tambusai*, 9(3), 36706–36710. <https://doi.org/10.31004/jptam.v9i3.34235>
- Wahyuningtyas, D., & Kesuma, T. M. J. (2024). Pemanfaatan linguistik korpus dalam menentukan kata berfrekuensi tinggi pada buku Sahabatku Indonesia BIPA 1. *Jurnal Bahasa Indonesia Bagi Penutur Asing*.

PROFIL PENULIS



Loso Judijanto

Penulis adalah peneliti yang bekerja pada lembaga penelitian IPOSS Jakarta. Penulis dilahirkan di Magetan pada tanggal 19 Januari 1971. Penulis menamatkan pendidikan *Master of Statistics* di *the University of New South Wales*, Sydney, Australia pada tahun 1998 dengan dukungan beasiswa ADCOS (*Australian Development Cooperation Scholarship*) dari Australia. Sebelumnya penulis menyelesaikan Magister Manajemen di Universitas Indonesia pada tahun 1995 dengan dukungan beasiswa dari Bank Internasional Indonesia. Pendidikan sarjana diselesaikan di Institut Pertanian Bogor pada Jurusan Statistika – FMIPA pada tahun 1993 dengan dukungan beasiswa dari KPS-Pertamina. Penulis menamatkan Pendidikan dasar hingga SMA di Maospati, Sepanjang karirnya, Penulis pernah ditugaskan untuk menjadi anggota Dewan Komisaris dan/atau Komite Audit pada beberapa perusahaan/lembaga yang bergerak di berbagai sektor antara lain pengelolaan pelabuhan laut, telekomunikasi seluler, perbankan, pengembangan infrastruktur, sekuritas, pembiayaan infrastruktur, perkebunan, pertambangan batu bara, properti dan rekreasi, dan pengelolaan dana perkebunan. Penulis memiliki minat dalam riset di bidang kebijakan publik, ekonomi, keuangan, *human capital*, dan *corporate governance*.

BAB 10

LINGUISTIK KORPUS DALAM KAJIAN WACANA

Siska Eka Syafitri
Universitas Samudra, Langsa
E-mail: siskaekasyafitri@unsam.ac.id

PENDAHULUAN

Linguistik korpus merupakan pendekatan dalam kajian bahasa yang memanfaatkan kumpulan data autentik (korpus) untuk menganalisis pola penggunaan bahasa secara empiris. Dalam konteks kajian wacana, linguistik korpus menawarkan keunggulan berupa analisis berbasis data besar (large-scale data) yang memungkinkan peneliti mengidentifikasi kecenderungan bahasa yang tidak selalu tampak melalui analisis manual. Kajian wacana sendiri berfokus pada bagaimana bahasa digunakan dalam konteks sosial, politik, dan budaya untuk membentuk makna. Oleh karena itu, integrasi linguistik korpus dalam kajian wacana memberikan pendekatan yang lebih objektif, sistematis, dan dapat direplikasi dalam memahami praktik diskursif di berbagai konteks.

IDENTIFIKASI TEMA DAN TOPIK

Dalam perspektif SFL, identifikasi tema tidak hanya berkaitan dengan topik umum, tetapi juga dengan struktur *Theme-Rheme* dalam klausa. Tema berfungsi sebagai titik awal pesan, sedangkan rheme mengembangkan informasi baru. Halliday dan Matthiessen menegaskan bahwa tema merupakan titik awal pesan dalam sebuah klausa (Halliday & M.I.M., 2014). Dengan demikian, analisis tema dalam wacana tidak hanya melihat apa yang dibicarakan, tetapi juga bagaimana informasi diorganisasikan untuk membangun makna.

Melalui linguistik korpus, identifikasi tema dan topik dapat dilakukan secara lebih sistematis dengan menelusuri pola kemunculan unsur leksikal dalam teks. Kata atau frasa yang muncul secara berulang dalam korpus mencerminkan fokus utama wacana, sementara distribusinya dalam berbagai konteks membantu mengungkap bagaimana suatu topik dikembangkan. Dalam konteks ini, pendekatan berbasis korpus memberikan dasar empiris yang kuat karena analisis korpus mampu menghadirkan gambaran sistematis mengenai pola leksikal dan representasi tema dalam teks (Khojasteh & Mukundan, 2025), sehingga penentuan tema tidak hanya bergantung pada intuisi peneliti.

Selain itu, analisis korpus juga memungkinkan peneliti memahami bagaimana unsur bahasa tertentu berkontribusi dalam membangun tema secara lebih kontekstual. Penggunaan unsur leksikal yang berulang dalam berbagai situasi menunjukkan bahwa makna tidak hanya ditentukan oleh topik besar, tetapi juga oleh cara bahasa digunakan dalam praktik nyata. Hal ini menunjukkan bahwa pola penggunaan unsur bahasa dalam korpus mencerminkan distribusi makna yang autentik dalam wacana (Kumar et al., 2025), sehingga tema dapat dipahami sebagai hasil dari interaksi antara pilihan bahasa dan konteks penggunaannya.

Dalam kaitannya dengan teori Appraisal, identifikasi tema juga mencakup dimensi evaluatif dalam teks. Martin dan White menekankan bahwa evaluasi dalam bahasa berkaitan dengan bagaimana sikap dan penilaian dinegosiasikan dalam teks (Martin & White, 2005). Dengan demikian, analisis tema tidak hanya menjawab pertanyaan “apa yang dibicarakan”, tetapi juga “bagaimana topik tersebut dinilai atau diposisikan” dalam wacana, termasuk kecenderungan sikap yang terkandung di dalamnya.

Lebih lanjut, identifikasi tema dan topik melalui pendekatan korpus ini menjadi landasan awal untuk analisis yang lebih mendalam terhadap pola bahasa dalam diskursus. Setelah tema utama dan fokus wacana terpetakan, analisis dapat dilanjutkan dengan bagaimana tema tersebut direalisasikan melalui pilihan leksikogramatikal yang berulang, seperti pola transitivitas dan strategi evaluatif. Dengan demikian, analisis tema berfungsi sebagai pintu masuk untuk memahami pola bahasa yang lebih luas dalam wacana, sehingga kedua aspek ini saling melengkapi dalam mengungkap konstruksi makna dan ideologi.

ANALISIS POLA BAHASA DALAM DISKURUS

Dalam kerangka SFL, pola bahasa merupakan realisasi dari pilihan sistemik dalam bahasa yang membentuk makna. Hal ini juga menegaskan bahwa sebuah makna akan dibentuk melalui pola pilihan leksikogramatikal dalam bahasa (Halliday & M.I.M., 2014). Oleh karena itu, analisis pola bahasa menjadi kunci untuk memahami bagaimana wacana dibangun, karena setiap pilihan bahasa mencerminkan cara penutur atau penulis merepresentasikan pengalaman dan realitas sosial.

Dalam perspektif Appraisal, pola evaluatif dianalisis melalui tiga subsistem utama, yaitu *attitude*, *engagement*, dan *graduation*. Ketiga aspek ini memungkinkan peneliti melihat bagaimana sikap, posisi suara, dan intensitas penilaian direalisasikan dalam teks. Martin dan White menekankan bahwa evaluasi dalam bahasa mencakup ekspresi perasaan, penilaian terhadap perilaku, serta apresiasi terhadap objek atau fenomena (Martin & White, 2005). Dengan bantuan linguistik korpus, pola-pola ini dapat diidentifikasi secara lebih sistematis melalui frekuensi, kolokasi, dan konkordansi, sehingga analisis menjadi lebih berbasis data.

Tabel 10.1. Analisis Pola Bahasa Berbasis SFL dan Appraisal

No	Kutipan Data	Transitivitas (Ideasional)	Attitude	Engagement	Graduation	Interpretasi Wacana
1	“Pemerintah berhasil menangani krisis dengan cepat”	Material (Actor: pemerintah; Process: berhasil menangani; Goal: krisis)	Judgement (+) kapasitas	Monogloss	Force (intensifier: “berhasil”, “cepat”)	Pemerintah direpresentasikan positif dan kompeten
2	“Kebijakan ini sangat merugikan masyarakat kecil”	Relational (Carrier: kebijakan; Attribute: merugikan)	Appreciation (-)	Monogloss	Force (high: “sangat”)	Kebijakan dibingkai negatif dengan dampak kuat
3	“Beberapa pihak menilai program tersebut kurang efektif”	Verbal (Sayer: beberapa pihak; Verbiage: penilaian)	Judgement (-) kapasitas	Heterogloss (atribusi: “menilai”)	Focus (softening: “kurang”)	Kritik disampaikan secara tidak langsung
4	“Masyarakat tampak kecewa terhadap pelayanan publik”	Mental (Senser: masyarakat; Phenomenon: pelayanan)	Affect (-)	Monogloss	Force (implicit)	Emosi negatif masyarakat ditonjolkan

		publik)				
5	“Program ini cukup membantu, meskipun belum optimal”	Relational + concessive	Appreciation (+/-)	Heterogloss (kontras)	Focus (softening: “cukup”)	Evaluasi ambivalen
6	“Diduga terjadi penyalahgunaan dana oleh oknum tertentu”	Material (pasif)	Judgement (-) propriety	Heterogloss (modalitas: “diduga”)	Focus (blur)	Strategi kehati-hatian

Tabel di atas menunjukkan bahwa pola bahasa dalam diskursus dapat dianalisis secara sistematis dengan mengintegrasikan metafungsi ideasional dan interpersonal dalam kerangka SFL serta sistem evaluasi dalam Appraisal. Melalui analisis transitivitas, terlihat bagaimana realitas sosial dikonstruksi dengan menempatkan aktor tertentu sebagai pelaku atau sasaran tindakan. Hal ini sejalan dengan pandangan bahwa sistem transitivitas digunakan untuk merepresentasikan pengalaman manusia dalam berbagai jenis proses (Almurashi, 2018).

Sementara itu, dimensi *attitude*, *engagement*, dan *graduation* memperlihatkan bagaimana sikap, posisi suara, dan tingkat intensitas evaluasi direalisasikan dalam teks, sehingga mengungkap kecenderungan ideologis yang mendasari wacana. Pola-pola ini tidak muncul secara acak, melainkan merupakan hasil dari pilihan leksikogramatikal yang berulang. Dengan demikian, pola evaluasi dalam bahasa dapat diidentifikasi secara sistematis dalam berbagai teks (Martin & White, 2005), sehingga analisis menjadi lebih terstruktur dan dapat dipertanggungjawabkan.

Dengan menggabungkan kerangka SFL dan linguistik korpus, analisis pola bahasa tidak hanya menjelaskan bagaimana makna dibentuk, tetapi juga bagaimana makna tersebut digunakan untuk membangun sikap, posisi sosial, dan ideologi dalam wacana. Pendekatan ini memberikan pemahaman yang lebih komprehensif terhadap hubungan antara bahasa dan realitas sosial, sekaligus menjadi jembatan menuju analisis wacana yang lebih kritis.

KORPUS DALAM ANALISIS WACANA KRITIS

Pendekatan *corpus-assisted discourse studies* (CADS) menggabungkan linguistik korpus dengan Analisis Wacana Kritis (AWK) untuk menghasilkan analisis yang lebih sistematis

dan berbasis data. Dalam perspektif SFL, wacana dipandang sebagai praktik sosial yang tidak terlepas dari relasi kekuasaan dan ideologi yang direalisasikan melalui bahasa. Dengan demikian, bahasa tidak hanya berfungsi sebagai alat komunikasi, tetapi juga sebagai sarana untuk membentuk dan merepresentasikan realitas sosial.

Halliday menekankan bahwa bahasa berfungsi dalam konteks sosial tertentu dan selalu terkait dengan makna yang lebih luas. Bahasa berkembang sesuai dengan fungsi yang dijalankannya dalam kehidupan sosial manusia (Syafitri et al., 2023). Pernyataan ini menegaskan bahwa setiap pilihan bahasa dalam teks mencerminkan posisi sosial, kepentingan, dan ideologi tertentu, sehingga analisis wacana tidak dapat dipisahkan dari konteks sosial yang melatarbelakanginya.

Dalam kerangka ini, teori appraisal dari J.R. Martin dan P.R.R. White memberikan kontribusi penting dalam mengungkap dimensi evaluatif dalam teks. Melalui analisis *attitude*, *engagement*, dan *graduation*, peneliti dapat melihat bagaimana suatu kelompok direpresentasikan secara positif atau negatif melalui pilihan leksikal. Dengan dukungan data korpus, pola evaluatif tersebut dapat diidentifikasi secara lebih sistematis, sehingga analisis ideologi dalam wacana menjadi lebih terukur dan tidak semata-mata bergantung pada interpretasi subjektif.

Lebih lanjut, pendekatan linguistik korpus menunjukkan bahwa pola bahasa dalam wacana bersifat terstruktur dan dapat ditelusuri melalui distribusi serta frekuensi kemunculan unsur leksikal. Dalam konteks ini, analisis korpus membantu memperlihatkan bahwa penggunaan bahasa tidak acak, melainkan mengikuti kecenderungan tertentu yang berulang dalam wacana. Sejalan dengan itu, analisis korpus mampu mengungkap pola penggunaan bahasa yang sistematis dalam

berbagai konteks wacana (Shubha, 2021), sehingga memberikan dasar empiris dalam memahami bagaimana makna dikonstruksi.

Selain itu, variasi penggunaan bahasa dalam korpus juga menunjukkan bahwa makna tidak bersifat tetap, melainkan berkembang sesuai dengan konteks sosial dan tujuan komunikasi. Perbedaan distribusi dan penggunaan unsur bahasa mencerminkan bagaimana penutur atau penulis menyesuaikan pilihan bahasanya untuk mencapai tujuan tertentu. Hal ini menegaskan bahwa variasi pola bahasa mencerminkan perbedaan konteks dan fungsi dalam proses pembentukan makna (Khojasteh & Mukundan, 2025), sehingga analisis wacana perlu mempertimbangkan dinamika konteks tersebut.

Dalam ranah wacana kritis, pendekatan ini menjadi sangat penting karena memungkinkan pengungkapan relasi kekuasaan yang tersembunyi dalam pola bahasa. Penggunaan bahasa tertentu tidak hanya merepresentasikan realitas, tetapi juga dapat memperkuat atau menantang struktur sosial yang ada, termasuk relasi gender dan dominasi kelompok tertentu. Oleh karena itu, analisis linguistik korpus dapat mengungkap hubungan antara bahasa, kekuasaan, dan posisi ideologis dalam wacana (Mae et al., 2025).

Dengan demikian, integrasi linguistik korpus dalam Analisis Wacana Kritis memungkinkan peneliti untuk mengidentifikasi pola ideologis secara lebih objektif dan terukur. Pendekatan ini tidak hanya memperkuat validitas analisis, tetapi juga memperluas pemahaman tentang bagaimana bahasa berfungsi sebagai alat untuk membangun dan merepresentasikan realitas sosial dalam berbagai konteks wacana.

Setelah mengidentifikasi pola ideologis dalam wacana melalui pendekatan korpus, langkah selanjutnya adalah melakukan perbandingan antarkelompok atau media untuk melihat bagaimana makna dan ideologi direalisasikan secara berbeda. Pendekatan ini membantu mengungkap variasi representasi yang dipengaruhi oleh latar sosial, budaya, dan institusional. Dalam perspektif SFL, perbedaan tersebut berkaitan dengan konteks situasi (*field*, *tenor*, dan *mode*) yang membentuk pilihan bahasa dalam teks. Halliday dan Matthiessen menegaskan bahwa terdapat hubungan sistematis antara konteks sosial dan organisasi bahasa (Halliday & M.I.M., 2014), sehingga variasi wacana mencerminkan perbedaan kondisi sosial yang melatarbelakanginya.

Dalam praktiknya, analisis berbasis korpus menunjukkan bahwa satu isu yang sama dapat direpresentasikan secara berbeda antarmedia melalui pilihan leksikal dan strategi diskursif yang digunakan. Perbedaan ini tidak bersifat kebetulan, melainkan berkaitan dengan posisi ideologis yang diambil oleh masing-masing pihak. Sebagaimana terlihat dalam kajian wacana media, representasi suatu isu dapat dibingkai secara kontras melalui bahasa yang digunakan, sehingga analisis korpus mampu mengungkap perbedaan representasi wacana yang dipengaruhi oleh posisi ideologis (Yoshimoto, 2025).

Melalui linguistik korpus, perbandingan wacana dilakukan dengan menganalisis frekuensi kata, pola kolokasi, dan distribusi struktur bahasa dalam korpus yang berbeda. Pendekatan ini memungkinkan peneliti mengidentifikasi pola representasi yang khas dalam masing-masing wacana serta memahami bagaimana suatu isu dikonstruksi secara berbeda. Dalam konteks ini, analisis terhadap berbagai jenis teks juga menunjukkan bahwa terdapat variasi representasi yang

sistematis sesuai dengan jenis teks dan konteks penggunaannya (Jennings et al., 2025), sehingga memperkuat pentingnya pendekatan komparatif dalam kajian wacana.

Selain itu, dalam kerangka Appraisal, perbandingan wacana dapat difokuskan pada variasi evaluasi dalam teks. Martin dan White menegaskan bahwa evaluasi merupakan aspek penting dalam membangun dan menegosiasikan hubungan sosial dalam wacana (Martin & White, 2005). Oleh karena itu, perbedaan dalam penggunaan *attitude*, *engagement*, dan *graduation* menunjukkan bagaimana suatu isu dinilai dan diposisikan secara berbeda. Untuk memperkuat hasil analisis, pendekatan ini juga dapat dikombinasikan dengan berbagai metode analisis, sehingga penggunaan triangulasi metodologis dapat meningkatkan kekuatan dan keandalan analisis berbasis korpus (Gillings & Jaworska, 2025).

Secara keseluruhan, perbandingan wacana antarkelompok atau media tidak hanya mengungkap variasi penggunaan bahasa, tetapi juga memperlihatkan bagaimana realitas sosial dibangun melalui pola diskursif yang berbeda. Dengan mengintegrasikan linguistik korpus, SFL, dan teori Appraisal, analisis menjadi lebih komprehensif dan berbasis data. Pendekatan ini memungkinkan peneliti memahami secara lebih mendalam dinamika wacana, termasuk bagaimana ideologi dan relasi kekuasaan direpresentasikan dalam berbagai konteks sosial.

Sebagai ilustrasi penerapan linguistik korpus dalam kajian wacana, bagian ini menyajikan contoh analisis terhadap wacana media yang sempat viral, yaitu pemberitaan mengenai kebijakan pemerintah terkait kenaikan harga bahan bakar minyak (BBM) di berbagai media online Indonesia. Isu ini menjadi sorotan publik karena berdampak langsung pada kehidupan masyarakat, sehingga memunculkan beragam representasi wacana di media (Kompas.com, 2022; Detik.com, 2022). Untuk memahami variasi tersebut secara lebih sistematis, analisis tidak hanya

dilakukan pada satu kutipan, tetapi juga dengan mengklasifikasikan data korpus dari berbagai sumber.

Salah satu contoh kutipan yang banyak muncul dalam pemberitaan adalah: *“Pemerintah terpaksa menaikkan harga BBM demi menjaga stabilitas ekonomi nasional.”* (Kompas.com, 2022). Dalam perspektif SFL, klausa ini menunjukkan proses material dengan “pemerintah” sebagai aktor. Penggunaan kata “terpaksa” membangun makna bahwa kebijakan tersebut dilakukan dalam kondisi tertentu yang tidak sepenuhnya diinginkan. Dengan demikian, pilihan bahasa digunakan untuk membingkai tindakan sebagai bentuk rasionalisasi kebijakan, sehingga berfungsi untuk membangun legitimasi di mata publik.

Dari sudut pandang Appraisal, frasa “demi menjaga stabilitas ekonomi nasional” mengandung evaluasi positif yang menempatkan kebijakan sebagai sesuatu yang bernilai baik. Hal ini menunjukkan bahwa evaluasi dalam wacana berperan dalam mengarahkan persepsi pembaca terhadap suatu isu (Martin & White, 2005). Namun, representasi tersebut tidak bersifat tunggal. Dalam media lain atau ruang publik seperti media sosial, muncul wacana alternatif seperti *“Kenaikan harga BBM semakin membebani masyarakat kecil”* (Detik.com, 2022) yang justru menekankan dampak negatif kebijakan terhadap masyarakat. Perbedaan ini menunjukkan adanya variasi posisi ideologis dalam merepresentasikan isu yang sama.

Untuk melihat variasi tersebut secara lebih komprehensif, data wacana dapat diklasifikasikan dalam bentuk korpus berdasarkan sumber dan kecenderungan bahasanya. Klasifikasi ini membantu mengidentifikasi pola leksikal, jenis evaluasi, serta posisi ideologis yang muncul dalam masing-masing kelompok data, sebagaimana ditunjukkan pada tabel 10.2 berikut:

Tabel 10.2. Perbandingan Representasi Wacana Kenaikan Harga BBM di Berbagai Media

No	Sumber Korpus	Jenis Media	Kutipan Data	Tema	Pola Leksikal Dominan	Evaluasi (Appraisal)	Posisi Ideologis
1	Kompas.com (2022)	Media online nasional	“Pemerintah terpaksa menaikkan harga BBM demi menjaga stabilitas ekonomi nasional.”	Kebijakan ekonomi	terpaksa, menjaga, stabilitas	Appreciation (+), Engagement (defensif)	Mendukung / legitimasi kebijakan
2	Detik.com (2022)	Media online nasional	“Kenaikan harga BBM semakin membebani masyarakat kecil.”	Dampak sosial	kenaikan, membebani, masyarakat kecil	Affect (-), Judgement (-implisit)	Kritis/ menyoroti dampak
3	Media sosial (Twitter/X)	Media sosial	“Harga naik lagi, rakyat makin susah hidup.”	Keluhan publik	naik, susah, rakyat	Affect (-) kuat	Penolakan / resistensi

4	Media sosial (Facebook)	Media sosial	“Kebijakan ini tidak berpihak pada rakyat kecil.”	Kritik kebijakan	tidak berpihak, rakyat kecil	Judgement (-) propriety	Kritis terhadap pemerintah
5	Portal berita alternatif	Media online independen	“Langkah pemerintah dinilai terburu-buru dan kurang transparan.”	Kritik kebijakan	dinilai, terburu-buru, kurang transparan	Judgement (-), Engagement (atribusi)	Kritis / evaluatif

Melalui klasifikasi tersebut, terlihat bahwa setiap jenis media memiliki kecenderungan representasi yang berbeda. Media arus utama cenderung menggunakan bahasa yang lebih netral atau mendukung, sementara media sosial dan portal alternatif lebih menonjolkan evaluasi kritis. Pola ini menunjukkan bahwa pola leksikal yang berulang dalam korpus dapat mengungkap kecenderungan ideologis dalam wacana, sehingga analisis tidak hanya berhenti pada level deskriptif, tetapi juga mampu mengungkap relasi kekuasaan dan perspektif yang mendasarinya.

Dengan demikian, integrasi analisis SFL, Appraisal, dan linguistik korpus dalam studi kasus ini memperlihatkan bagaimana wacana politik dan media dibangun melalui pilihan bahasa yang strategis. Setiap variasi bahasa mencerminkan cara yang berbeda dalam merepresentasikan realitas sosial, sehingga analisis korpus menjadi alat penting untuk memahami konstruksi makna dan dinamika ideologi dalam wacana kontemporer.

KESIMPULAN

Secara keseluruhan, kajian linguistik korpus dalam analisis wacana menunjukkan bahwa bahasa bukan sekadar alat komunikasi, melainkan sarana strategis dalam membangun makna, merepresentasikan realitas, dan menegosiasikan ideologi. Melalui integrasi pendekatan SFL dan teori Appraisal, analisis berbasis korpus memungkinkan pengungkapan pola bahasa secara sistematis sekaligus mendalam, baik dari aspek representasi pengalaman maupun evaluasi sosial. Dengan demikian, linguistik korpus menjadi jembatan antara analisis kuantitatif dan interpretasi kritis, tidak hanya memperkuat validitas penelitian, tetapi juga membuka ruang pemahaman yang lebih luas pada beragam konteks dinamika wacana.

DAFTAR PUSTAKA

- Almurashi, W. (2018). *An Introduction to Halliday's Systemic Functional Linguistics*. May 2016. <https://doi.org/10.5296/jsel.v4i1.9423>
- Gillings, M., & Jaworska, S. (2025). How humans and machines identify discourse topics: A methodological triangulation. *Applied Corpus Linguistics*, 5(1), 100121. <https://doi.org/10.1016/j.acorp.2025.100121>
- Halliday, M. A. K. & M., & M.I.M. (2014). *Introduction to Functional Grammar Revised by Christian M.I.M Matthiessen*. Routledge.
- Jennings, B., Powell, D., Jaworska, S., & Joseph, H. (2025). Adolescent reading experience, independent choices and curriculum materials. *Applied Corpus Linguistics*, 5(1), 100124. <https://doi.org/10.1016/j.acorp.2025.100124>
- Khojasteh, L., & Mukundan, J. (2025). *Research Quarterly: A Systematic Review of Corpus-Based Methodologies in Textbook Analyses and Evaluation*. 47, 175–195.
- Kumar, P., Chauhan, S., & Ghosh, S. (2025). *Integrating Corpus Insights in Representation of Phrasal Verbs for ESL Learners: A Case Study of NCERT English Textbooks*. 49(Since 2009), 96–107. <https://doi.org/10.31901/24566322.2025/49.03.1369>
- Mae, J., Cruz, A., & Guzman, M. L. P. De. (2025). *A corpus linguistic analysis of gendered language and power dynamics in early 20th-century movie dialogues*. 2025(13), 123–132.
- Martin, J. R., & White, P. R. R. (2005). *The Language of Evaluation*. Palgrave.
- Shubha, N. S. (2021). *EXPLORATION OF PHRASAL VERBS IN ELT TEXTBOOKS: A CORPUS-BASED ANALYSIS OF LOWER SECONDARY*. 2(4).
- Syafitri, S. E., Sinar, T. S., Mulyadi, M., & Lubis, M. (2023).

“How Do the Brochures Speak?” Systemic Functional Linguistics Analysis in Online Parenting Brochures. *Language Literacy: Journal of Linguistics, Literature, and Language Teaching*, 7(2), 337–345. <https://doi.org/10.30743/ll.v7i2.7053>

Yoshimoto, K. (2025). Anywhere but here : Discourses and representations surrounding same-sex marriage in Japanese newspapers. *Applied Corpus Linguistics*, 5(1), 100120. <https://doi.org/10.1016/j.acorp.2025.100120>

PROFIL PENULIS

Siska Eka Syafitri



Penulis merupakan dosen pada Program Studi Pendidikan Bahasa Inggris di Universitas Samudera (UNSAM) yang memiliki fokus kajian dalam bidang ilmu linguistik, khususnya kajian wacana dan Linguistik Sistemik Fungsional (*Systemic Functional Linguistics/SFL*). Minat akademiknya berpusat pada analisis penggunaan bahasa dalam konteks sosial dan pendidikan, terutama dalam memahami bagaimana makna dibangun melalui interaksi bahasa di dalam kelas. Hal ini tercermin dalam disertasinya yang berjudul “*Wacana Kelas pada Anak Usia Dini: Analisis Linguistik Sistemik Fungsional*” yang mengkaji dinamika komunikasi antara guru dan peserta didik melalui perspektif metafungsi bahasa. Dalam perjalanan akademiknya, Siska Eka Syafitri telah menghasilkan berbagai karya ilmiah yang dipublikasikan dalam jurnal nasional maupun internasional, serta berkontribusi dalam penulisan book chapter yang membahas linguistik, pendidikan bahasa, dan analisis wacana. Ia juga aktif dalam mengembangkan pemikiran terkait penerapan SFL dalam konteks pembelajaran bahasa Inggris, khususnya untuk meningkatkan kualitas interaksi pedagogis di kelas. Selain sebagai pengajar dan peneliti, ia terbuka terhadap kolaborasi akademik, diskusi ilmiah, dan pengembangan riset bersama.

BAB 11

LINGUISTIK KORPUS DALAM SOCIOLINGUISTICS

Yanuarius Yanu Dharmawan
Universitas Bandar Lampung, Bandar Lampung
E-mail: yanu@ubl.ac.id

PENDAHULUAN

Sosiolinguistik memandang bahasa sebagai praktik sosial, bukan sekadar sistem tanda yang berdiri sendiri. Bahasa digunakan untuk berinteraksi, menegosiasikan makna, membangun identitas, serta menandai kedekatan, jarak, dan posisi sosial dalam komunitas tutur tertentu (Holmes & Wilson, 2022; Prasatyo et al., 2025; Rahmawati & Lestari, 2024; Schneider et al., 2025). Karena itu, variasi bahasa bukan penyimpangan dari norma, melainkan bagian wajar dari kehidupan berbahasa. Perbedaan pengucapan, kosakata, pola gramatikal, dan gaya tutur dapat mencerminkan wilayah, usia, gender, kelas sosial, jaringan sosial, hingga orientasi ideologis penutur (Meyerhoff, 2018; Tagliamonte, 2016).

Perkembangan sosiolinguistik menunjukkan pergeseran dari penelitian berbasis observasi lapangan dan wawancara menuju pemanfaatan data yang lebih besar dan beragam. Tradisi variasiis menekankan pentingnya data autentik untuk menjelaskan hubungan antara bentuk bahasa dan variabel sosial secara empiris (Földes, 2023; Gardner & Kostadinova, 2024). Namun, pendekatan ini sering terkendala oleh jumlah data yang terbatas, transkripsi yang memakan waktu, serta sulitnya membandingkan pola bahasa lintas komunitas dan periode. Dalam konteks inilah linguistik korpus memberi kontribusi penting, karena memungkinkan analisis atas data bahasa yang sistematis, terdokumentasi, dan dapat ditelusuri kembali dalam

skala yang lebih luas (Haitong, 2025; Pérez-Paredes & Curry, 2024; Puspitasari et al., 2024).

Linguistik korpus pada dasarnya menggunakan kumpulan data bahasa autentik, baik tulis, lisan, maupun multimodal, sebagai dasar analisis kebahasaan (Friginal, 2024; Puspitasari et al., 2024). Bagi sosiolinguistik, kekuatan pendekatan ini terletak pada kemampuannya menangkap pola penggunaan bahasa yang berulang, seperti frekuensi, distribusi, kolokasi, konkordansi, dan perbedaan antarsubkorpus sosial (Biber et al., 1998). Dengan demikian, variasi bentuk sapaan, partikel, pronomina, atau konstruksi evaluatif dapat diamati secara lebih sistematis antarwilayah, antargenerasi, dan antarkelompok sosial.

Pendekatan korpus juga penting karena menjembatani analisis kuantitatif dan interpretasi sosial. Data statistik tidak berhenti pada angka, tetapi menjadi dasar untuk menjelaskan makna sosial suatu bentuk bahasa dalam konteks tertentu (Paquot & Gries, 2020). Sebuah kata atau konstruksi dapat menandai solidaritas, formalitas, keakraban, atau resistensi bergantung pada penutur, situasi, dan komunitasnya (Johnstone & Andrus, 2024). Karena itu, analisis korpus dalam sosiolinguistik tidak hanya menanyakan apa yang sering muncul, tetapi juga mengapa pola itu muncul dan apa makna sosialnya.

Dalam penelitian bahasa kontemporer, relevansi linguistik korpus semakin besar karena meluasnya data digital, seperti teks berita, media sosial, forum daring, komentar pengguna, podcast, dan arsip digital (Friginal, 2024; Zappavigna, 2018). Sumber-sumber ini memungkinkan peneliti menelaah bahasa sebagai praktik sosial yang berlangsung dalam ruang publik dan komunitas virtual (Kusuma et al., 2023). Identitas sosial pun tidak lagi hanya terkait wilayah, usia, atau kelas, tetapi juga afiliasi digital, gaya partisipasi, dan posisi ideologis yang dinegosiasikan melalui bahasa (András & Balázs, 2025;

Dharmawan et al., 2023; Georgakopoulou et al., 2023; Zappavigna, 2018).

Meski demikian, penggunaan korpus dalam sosiolinguistik menghadapi tantangan metodologis. Representativitas tetap menjadi persoalan utama karena korpus selalu merupakan hasil seleksi, bukan cerminan total praktik berbahasa dalam masyarakat (Brookes et al., 2022; Meilinda et al., 2024). Dalam korpus digital, metadata sosial seperti usia, gender, kelas, dan lokasi sering tidak lengkap, sementara penggunaan bahasa juga dipengaruhi oleh norma platform dan teknologi yang digunakan (Zappavigna, 2018). Di samping itu, kajian sosiolinguistik mutakhir menolak pandangan yang terlalu esensialis tentang identitas sosial. Karena itu, data korpus perlu dibaca secara hati-hati dengan mempertimbangkan konteks, posisi partisipan, dan perpotongan berbagai identitas sosial (Puspitasari et al., 2024; Putri, 2025).

Secara metodologis, kekuatan linguistik korpus bagi sosiolinguistik terletak pada tiga hal: kemampuannya menyediakan data autentik dalam jumlah besar, memungkinkan replikasi dan transparansi analisis, serta menggabungkan pembacaan kuantitatif dan kualitatif (Friginal, 2024; Paquot & Gries, 2020; Pérez-Paredes & Curry, 2024; Puupponen et al., 2025; Suhaili et al., 2024). Karena itu, pendekatan ini sangat berguna untuk menelaah variasi regional, perbedaan antargenerasi, gaya bahasa berbasis gender, stratifikasi sosial, konstruksi identitas, hingga representasi kelompok sosial dalam media dan wacana publik (Bednarek et al., 2024; Suhaili et al., 2024; Umam et al., 2025; Wearne & Riedy, 2024).

Korpus lisan juga memegang peranan penting karena banyak gejala sosiolinguistik paling jelas muncul dalam tuturan, seperti pengucapan, jeda, tawa, overlap, partikel, dan code-switching (Cahyanti & Dharmawan, 2025; Friginal, 2024; Hidayati et al., 2024; Ramadani & Dharmawan, 2025). Selain

itu, korpus diakronis memungkinkan peneliti menelusuri perubahan bahasa, baik dalam frekuensi, fungsi, maupun makna sosial suatu bentuk dari waktu ke waktu (Haitong, 2025; Hilpert, 2013; Pitzl-Hagin, 2024; Rohmah, 2024). Dengan demikian, linguistik korpus menjadi perangkat penting dalam sociolinguistik modern karena membantu membaca bahasa bukan hanya sebagai struktur, tetapi juga sebagai jejak sosial yang merekam identitas, relasi kuasa, solidaritas, dan perubahan budaya (Holmes & Wilson, 2022; Li et al., 2024).

Bab ini membahas peran linguistik korpus dalam sociolinguistik melalui lima fokus utama, yaitu variasi bahasa berdasarkan wilayah; variasi berdasarkan gender, usia, dan latar sosial; korpus lisan sebagai sumber analisis variasi; perubahan bahasa di komunitas; serta hubungan antara korpus dan identitas sosial. Melalui pembahasan ini, korpus dipahami bukan sekadar kumpulan data bahasa, melainkan medium empiris untuk memahami dinamika sosial dalam penggunaan bahasa.

VARIASI BAHASA BERDASARKAN WILAYAH

Variasi bahasa berdasarkan wilayah merupakan salah satu bidang utama dalam sociolinguistik. Bahasa tidak tersebar secara seragam, melainkan memperlihatkan perbedaan dalam pengucapan, kosakata, struktur gramatikal, dan strategi pragmatik sesuai dengan lokasi geografis serta sejarah sosial penuturnya. Karena itu, variasi regional tidak hanya menandai asal daerah, tetapi juga afiliasi budaya, kebanggaan lokal, dan kadang posisi sosial tertentu (Ghifara et al., 2025; Sofa & Febrianti, 2025).

Secara tradisional, variasi wilayah dikaji melalui dialektologi, terutama lewat survei lapangan, wawancara, dan atlas dialek. Pendekatan ini penting untuk memetakan persebaran bentuk bahasa, tetapi sering terbatas pada jumlah informan yang kecil dan belum sepenuhnya menangkap

dinamika penggunaan bahasa dalam konteks sehari-hari (Ghifara et al., 2025; Kamila et al., 2025). Linguistik korpus memperluas pendekatan tersebut dengan menyediakan data yang lebih besar, lebih sistematis, dan lebih kontekstual. Melalui korpus regional, peneliti dapat memanfaatkan data tulis maupun lisan, seperti media lokal, transkrip wawancara, forum daring, dan percakapan sehari-hari, lalu membandingkan ciri bahasa antardaerah secara lebih terukur (Dance, 2025; Soyusiawaty et al., 2025).

Keunggulan pendekatan korpus terletak pada kemampuannya mengaitkan bentuk bahasa dengan frekuensi dan konteks pemakaian. Peneliti tidak hanya mengetahui bahwa suatu bentuk muncul di wilayah tertentu, tetapi juga seberapa sering digunakan, dalam genre apa, dan dalam lingkungan leksikal seperti apa (Agustina et al., 2025; Suhartatik et al., 2026). Pada tingkat leksikal, perbedaan regional dapat terlihat pada nama benda, istilah kuliner, bentuk sapaan, atau ungkapan emosional. Pada tingkat fonologis, variasi tampak pada perbedaan pelafalan dan intonasi, yang biasanya memerlukan korpus lisan untuk dianalisis dengan baik (Azhar, 2025; Mardatillah & Nurfadilah, 2025). Sementara itu, pada tingkat morfosintaksis, korpus memungkinkan deteksi pola-pola yang lebih halus, seperti susunan frasa, pilihan partikel, dan penggunaan pronomina yang khas wilayah tertentu (Di Caro & Samo, 2025).

Pendekatan korpus juga menunjukkan bahwa wilayah geografis bukan sekadar latar tempat, melainkan ruang sosial yang dipengaruhi mobilitas, urbanisasi, media, dan kontak bahasa. Dalam masyarakat modern, bentuk regional dapat menyebar melampaui batas geografis asalnya, bahkan berkembang menjadi ragam supraregional atau ragam digital (Sadigzade, 2025). Dalam konteks bahasa Indonesia, hal ini sangat relevan karena bahasa Indonesia digunakan di tengah

keragaman bahasa daerah yang kuat. Akibatnya, ragam bahasa Indonesia di berbagai wilayah memperlihatkan pengaruh lokal dalam kosakata, intonasi, struktur, dan partikel, baik dalam komunikasi luring maupun daring (Rohmah, 2024; Zamrizal et al., 2025).

Selain mendeskripsikan perbedaan bentuk, korpus juga membantu menjelaskan relasi sosial yang menyertainya. Variasi regional dapat beririsan dengan prestise, stigma, kelas sosial, pusat media, dan stereotip budaya (Johnstone & Andrus, 2024; Li et al., 2024). Untuk menganalisis hal itu, peneliti biasanya menggunakan frekuensi, keyword analysis, kolokasi, konkordansi, dan perbandingan subkorpus. Dalam penelitian yang lebih mutakhir, geolinguistik korpus memungkinkan pemetaan persebaran bentuk secara spasial, sedangkan kajian kontak bahasa membantu menelusuri alih kode, campur kode, dan bentuk hibrida di wilayah multibahasa (Agustina et al., 2025; Ali et al., 2025; Grieve, 2016; Haitong, 2025; Mambu, 2025).

Di era digital, dimensi wilayah tidak hilang, tetapi berubah bentuk. Ciri regional tetap hadir dalam media sosial melalui pilihan kosakata, sapaan, gaya bercanda, atau representasi pengucapan, sekaligus menyebar lebih cepat ke ruang nasional (András & Balázs, 2025; Manurung et al., 2025; Zappavigna, 2018). Meski demikian, analisis korpus berbasis wilayah tetap harus berhati-hati terhadap persoalan metodologis, seperti penentuan kategori wilayah, kelengkapan metadata lokasi, ketimpangan distribusi data, dan pengaruh genre atau platform terhadap hasil analisis (Paquot & Gries, 2020).

Dengan demikian, linguistik korpus memberi dimensi baru pada studi variasi bahasa berdasarkan wilayah. Jika dialektologi tradisional terutama memetakan di mana suatu bentuk ditemukan, pendekatan korpus memungkinkan peneliti menjelaskan seberapa sering bentuk itu digunakan, dalam

konteks apa, dan makna sosial apa yang melekat padanya. Melalui korpus, wilayah dipahami bukan hanya sebagai ruang geografis, tetapi juga sebagai ruang sosial tempat identitas, sejarah, mobilitas, dan perubahan bahasa saling berkelindan.

VARIASI BERDASARKAN GENDER, USIA, DAN SOSIAL

Selain wilayah, variasi bahasa juga berkaitan erat dengan gender, usia, pendidikan, pekerjaan, dan posisi sosial (Dharmawan, 2023). Dalam sociolinguistik, faktor-faktor ini dipahami sebagai variabel yang memengaruhi pilihan bahasa, gaya bertutur, strategi interaksi, dan nilai sosial yang melekat pada bentuk kebahasaan tertentu (Holmes & Wilson, 2022; Meyerhoff, 2018). Karena itu, bahasa tidak pernah sepenuhnya netral; pilihan kata, sapaan, formalitas, humor, dan cara menyatakan sikap sering mencerminkan posisi sosial penutur.

Kajian sociolinguistik klasik telah lama menunjukkan hubungan antara variasi bahasa, stratifikasi sosial, dan penyebaran perubahan bahasa (Sofa & Febrianti, 2025). Namun, linguistik korpus memperluas analisis ini melalui data yang lebih besar, beragam, dan dapat dibandingkan secara sistematis. Dengan metadata yang memadai, seperti usia, jenis kelamin, pendidikan, pekerjaan, dan jenis interaksi, peneliti dapat membandingkan kecenderungan penggunaan bahasa antarkelompok secara lebih terukur (Baker, 2010; Biber et al., 1998).

Dalam kajian gender, korpus dapat digunakan untuk meneliti pilihan leksikal, intensifier, hedge, tag question, bentuk evaluatif, dan strategi kesantunan. Namun, penelitian mutakhir cenderung menolak generalisasi sederhana tentang “bahasa laki-laki” dan “bahasa perempuan” (Coates, 2015; Litosseliti, 2014). Gender kini lebih dipahami sebagai sesuatu yang dinegosiasikan melalui bahasa, bukan kategori biologis yang otomatis

menentukan gaya tutur (Li, 2025). Karena itu, analisis korpus perlu membaca data secara kontekstual agar tidak memperkuat stereotip (Baker & Levon, 2015).

Variasi berdasarkan usia juga penting karena penutur muda sering menjadi sumber inovasi linguistik, terutama dalam ranah informal dan digital. Mereka cenderung lebih cepat mengadopsi slang, singkatan, ejaan kreatif, dan ekspresi pragmatik baru, sedangkan penutur yang lebih tua cenderung mempertahankan bentuk yang lebih mapan (Tagliamonte, 2016). Melalui korpus, peneliti dapat membandingkan pilihan bahasa antargenerasi dan melihat arah perubahan bahasa melalui distribusi kosakata, partikel, struktur kalimat, dan pola interaksi (Tagliamonte & Denis, 2008).

Kelas sosial dan faktor sosial ekonomi juga berpengaruh terhadap variasi bahasa, terutama dalam kaitannya dengan pendidikan, profesi, prestise, dan akses terhadap norma bahasa standar (Manurung et al., 2025; Travis & Gan, 2025). Perbedaan ini dapat terlihat dalam pilihan leksikal, kompleksitas sintaktis, formalitas, dan variasi pragmatik (Holmes & Wilson, 2022). Namun, faktor-faktor ini tidak bekerja secara terpisah. Variasi bahasa biasanya terbentuk melalui perpotongan gender, usia, wilayah, etnisitas, profesi, dan jaringan sosial (Li, 2025; Schneider et al., 2025).

Dalam konteks digital, variasi sosial menjadi semakin kompleks karena bahasa di media sosial juga dipengaruhi oleh konvensi platform dan performa diri, bukan semata-mata identitas offline (Zappavigna, 2018). Meski demikian, korpus digital tetap sangat berguna karena menyediakan jejak penggunaan bahasa yang kaya untuk menganalisis bagaimana identitas sosial dibentuk dan dinegosiasikan. Dengan demikian, linguistik korpus memberi dasar empiris yang lebih kuat untuk menelaah variasi bahasa berdasarkan gender, usia, dan sosial secara lebih dinamis.

KORPUS LISAN UNTUK ANALISIS VARIASI

Dalam sosiolinguistik, korpus lisan sangat penting karena banyak variasi bahasa paling jelas muncul dalam tuturan sehari-hari. Intonasi, jeda, tawa, overlap, interjeksi, partikel wacana, dan peralihan kode sering tidak tampak utuh dalam teks tulis, padahal unsur-unsur itu penting untuk membaca identitas sosial dan relasi interpersonal (Clift, 2016).

Korpus lisan biasanya dibangun dari rekaman percakapan natural, wawancara, diskusi, pidato, atau interaksi institusional yang kemudian ditranskripsikan agar dapat dianalisis secara sistematis. Dalam banyak kasus, korpus ini juga dilengkapi dengan anotasi tambahan, seperti identitas pembicara, situasi tutur, jeda, tumpang tindih, atau ciri prosodik sesuai tujuan penelitian (Schneider et al., 2025). Karena itu, tingkat detail transkripsi menjadi persoalan metodologis penting: terlalu sederhana dapat menghilangkan ciri sosial, terlalu rinci dapat menyulitkan pengolahan data.

Keunggulan utama korpus lisan terletak pada kemampuannya menangkap bahasa sebagaimana digunakan secara aktual. Melalui data lisan, peneliti dapat mengamati bentuk nonstandar, variasi pelafalan, pengelolaan giliran bicara, strategi keakraban, ketidaksantunan, dan unsur pragmatik yang jarang muncul dalam teks tulis formal (Adolphs & Carter, 2013). Pada tingkat leksikal, korpus lisan dapat menunjukkan kata, frasa, dan formula interaksional yang khas kelompok tertentu. Pada tingkat fonologis, korpus ini memungkinkan analisis realisasi bunyi, reduksi, ritme, dan intonasi sebagai penanda sosial maupun regional (Schneider et al., 2025). Pada tingkat pragmatik, korpus lisan memperlihatkan cara penutur mengelola persetujuan, penolakan, interupsi, humor, dan kesopanan.

Korpus lisan juga sangat berguna untuk meneliti fenomena yang sulit ditangkap dalam korpus tulis, seperti code-switching,

backchanneling, repairs, dan negotiated meaning, khususnya dalam masyarakat multilingual (Gardner & Kostadinova, 2024). Namun, penyusunannya menghadapi tantangan etis dan teknis, seperti izin partisipan, kualitas rekaman, proses transkripsi yang memakan waktu, dan observer's paradox (Labov, 1973). Teknologi transkripsi semiotomatis dan forced alignment memang membantu, tetapi tetap memerlukan verifikasi manusia agar interpretasi sosial tidak keliru (Haitong, 2025). Karena itu, korpus lisan tetap menjadi sumber yang sangat penting untuk memahami variasi bahasa dalam praktik komunikasi yang hidup.

ANALISIS PERUBAHAN BAHASA DI KOMUNITAS

Bahasa selalu berubah, dan perubahan itu berlangsung melalui praktik sosial, interaksi antargenerasi, mobilitas, media, dan kontak budaya. Dalam sociolinguistik, perubahan bahasa tidak dipahami semata-mata sebagai perubahan struktur, tetapi sebagai proses sosial yang terjadi dalam komunitas tertentu (Labov, 1973). Linguistik korpus penting di sini karena memungkinkan perubahan tersebut ditelusuri secara empiris.

Melalui korpus diakronis atau korpus yang mencakup rentang waktu berbeda, peneliti dapat mengamati perubahan frekuensi, distribusi, fungsi, dan makna suatu bentuk bahasa. Analisis ini dapat memperlihatkan munculnya ekspresi baru, menurunnya bentuk lama, bergesernya kolokasi, atau berubahnya nilai pragmatik suatu konstruksi (Hilpert & Gries, 2016). Dengan demikian, perubahan bahasa tidak lagi hanya didasarkan pada intuisi, tetapi pada jejak penggunaan yang nyata.

Perubahan bahasa sering bermula dari komunitas tertentu, seperti remaja perkotaan, kelompok profesional, komunitas migran, atau komunitas daring (Ali et al., 2023; Dharmawan et al., 2023; Kusuma et al., 2023). Melalui korpus lokal atau

korpus komunitas, peneliti dapat melihat apakah inovasi bahasa bersifat sementara, terbatas, atau mulai menyebar ke masyarakat yang lebih luas (Androutsopoulos, 2014; Tagliamonte, 2016). Korpus juga memungkinkan penggabungan kuantifikasi dan interpretasi sosial: peningkatan suatu bentuk dapat dihubungkan dengan perubahan media, budaya populer, norma kedekatan, atau relasi antargenerasi (Zappavigna, 2018).

Penyebaran perubahan bahasa juga sangat dipengaruhi oleh jaringan sosial. Bentuk baru cenderung menyebar lebih cepat dalam jaringan yang padat dan aktif berinteraksi, baik secara tatap muka maupun digital (Manurung et al., 2025). Meski begitu, data digital harus dibaca secara hati-hati karena tidak semua komunitas memiliki representasi yang sama, dan bahasa di platform daring juga dibentuk oleh norma platform dan budaya partisipasi digital (Boyd, 2014).

Selain pada kosakata, perubahan bahasa juga dapat terjadi pada konstruksi, pragmatik, dan gaya. Bentuk yang sama dapat mengalami perluasan fungsi atau perubahan asosiasi sosial. Dalam konteks bahasa Indonesia yang multibahasa, pendekatan korpus sangat relevan untuk meneliti interaksi antara bahasa daerah, bahasa nasional, dan bahasa global yang terus memunculkan inovasi leksikal, campur kode, dan penanda identitas baru (Hilpert, 2013). Dengan demikian, linguistik korpus membantu menjelaskan bahwa komunitas sosial merupakan penggerak utama perubahan bahasa.

KORPUS DAN IDENTITAS SOSIAL

Bahasa bukan hanya sarana menyampaikan informasi, tetapi juga cara menampilkan identitas, afiliasi, nilai, dan posisi sosial dalam interaksi. Dalam sociolinguistik kontemporer, identitas dipahami sebagai sesuatu yang terus dibentuk dan dinegosiasikan melalui praktik kebahasaan, bukan sebagai atribut tetap yang sudah ada sebelumnya (De Fina et al., 2006).

Dalam kerangka ini, bahasa menjadi medium utama pembentukan identitas sosial.

Linguistik korpus memungkinkan proses itu diteliti secara empiris. Melalui korpus, peneliti dapat mengamati bentuk-bentuk linguistik yang berulang dan diasosiasikan dengan kelompok tertentu, baik pada tingkat leksikal, gramatikal, pragmatik, maupun gaya. Slang tertentu dapat menandai generasi muda, istilah teknis dapat menandai profesi, dan gaya penulisan tertentu di media sosial dapat menunjukkan orientasi budaya atau ideologis (Baker & Levon, 2015; Zappavigna, 2018). Makna identitas itu muncul bukan dari bentuk itu sendiri, tetapi dari pola penggunaan dan nilai sosial yang diberikan komunitas terhadapnya.

Analisis identitas melalui korpus biasanya dimulai dari pola statistik, lalu dilanjutkan dengan pembacaan konteks. Dengan keyword analysis, kolokasi, dan konkordansi, peneliti dapat melihat bagaimana kelompok tertentu membangun solidaritas, otoritas, humor, resistensi, atau jarak sosial melalui bahasa (Baker, 2023). Pendekatan ini penting karena identitas sosial selalu berlapis: seseorang dapat sekaligus menampilkan identitas regional, generasional, profesional, dan ideologis bergantung pada situasi (Coupland, 2007).

Dalam masyarakat digital, hubungan antara bahasa dan identitas menjadi semakin kompleks. Platform digital memungkinkan penutur membangun persona secara sadar melalui emoji, singkatan, tagar, ejaan, humor, dan campuran kode, tetapi pada saat yang sama juga membatasi bentuk ekspresi melalui norma platform itu sendiri (Barton & Lee, 2025). Korpus digital karena itu sangat berguna untuk meneliti identitas kontemporer, meskipun persoalan etika tetap harus diperhatikan.

Korpus juga menunjukkan bahwa identitas sering dibangun melalui oposisi: penutur tidak hanya menandai siapa dirinya,

tetapi juga membedakan diri dari kelompok lain (Ali et al., 2025). Dalam wacana publik dan media, pilihan leksikal tertentu dapat memperkuat ideologi dan representasi sosial tertentu (Baker et al., 2013). Bahkan pola-pola kecil, seperti pilihan pronomina, tingkat keterusterangan, atau partikel tertentu, dapat menjadi penanda identitas yang penting. Dengan demikian, linguistik korpus memberi dasar yang kuat untuk memahami identitas sebagai praktik sosial yang dijalankan melalui bahasa.

KESIMPULAN

Linguistik korpus telah memperluas sekaligus memperkuat metodologi sociolinguistik. Melalui korpus, variasi bahasa berdasarkan wilayah, gender, usia, dan latar sosial dapat diuji secara lebih sistematis melalui frekuensi, distribusi, kolokasi, dan konteks penggunaan, bukan hanya dideskripsikan secara umum. Korpus dengan demikian menjembatani bukti linguistik dan interpretasi sosial.

Bab ini menunjukkan bahwa korpus tidak hanya berguna untuk analisis teknis, tetapi juga untuk membaca dinamika sosial dalam bahasa: bagaimana ciri regional dipertahankan atau berubah, bagaimana variasi antargenerasi muncul, bagaimana kelas sosial memengaruhi pilihan bentuk, bagaimana korpus lisan menangkap praktik tutur yang hidup, dan bagaimana identitas dinegosiasikan melalui bahasa. Namun, pendekatan ini tetap menuntut kehati-hatian terhadap representativitas data, kualitas metadata, bias platform, dan persoalan etika (Paquot & Gries, 2020).

Pada akhirnya, sociolinguistik dan linguistik korpus merupakan dua pendekatan yang saling melengkapi. Sociolinguistik memberi kerangka teoretis untuk memahami makna sosial bahasa, sedangkan linguistik korpus menyediakan alat untuk mengidentifikasi dan menguji pola itu dalam data autentik. Sinergi keduanya membuat penelitian bahasa lebih

kuat secara empiris, lebih tajam secara teoretis, dan lebih dekat dengan kenyataan bahwa bahasa selalu hidup di dalam masyarakat.

DAFTAR PUSTAKA

- Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. Routledge.
<https://doi.org/10.4324/9780203526149>
- Agustina, Y., Nazri, M. A., & Setiawan, B. (2025). Mapping the landscape of corpus-based pedagogy in English education: global vs. Indonesian researcher perspectives. *Journal of Educational Studies*, 3(3), 82–94.
- Ali, H. V., Hidayati, D., & Dharmawan, Y. Y. (2023). A study of presupposition and conversational implicatures in the comic strip of “The Diary of a Wimpy Kid.” *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, 9(2), 1353–1364.
<https://doi.org/https://doi.org/10.30605/onoma.v9i2.3034>
- Ali, H. V., Prasatyo, B. A., & Dharmawan, Y. Y. (2025). Navigating language barriers: Translingual discrimination and resilience among Indonesian migrant workers in Kuwait. *Indonesian Journal of Applied Linguistics*, 14(3), 496–511. <https://doi.org/10.17509/ijal.v14i3.73641>
- András, I., & Balázs, G. (2025). Representation, virtuality, and cognition. Rethinking language in the digital attention economy. *Magyar Nyelv*, 149(5), 637–648.
<https://doi.org/10.38143/Nyr.2025.5.637>
- Androutsopoulos, J. (2014). *Mediatization and sociolinguistic change*. De Gruyter.
<https://books.google.co.id/books?id=z73oBQAAQBAJ>
- Azhar, I. N. (2025). Toward a refined understanding of cluster reduction: Phonotactic innovation and elision in Bangkalanese Madurese . *Prosodi*, 19(2), 307–324.
<https://doi.org/10.21107/prosodi.v19i2.31915>

- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
<https://books.google.co.id/books?id=ywvPwAEACAAJ>
- Baker, P. (2023). *Using corpora in discourse analysis*. Bloomsbury Publishing.
<https://books.google.co.id/books?id=caa-EAAAQBAJ>
- Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
<https://doi.org/10.1177/0957926508088962>
- Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.
[https://doi.org/DOI: 10.1017/CBO9780511920103](https://doi.org/DOI:10.1017/CBO9780511920103)
- Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse and Communication*, 9(2), 221–236.
<https://doi.org/10.1177/1750481314568542>
- Barton, D., & Lee, C. (2025). *Language online: Investigating digital texts and practices*. Routledge.
<https://doi.org/10.4324/9781003265276>
- Bednarek, M., Schweinberger, M., & Lee, K. K. H. (2024). Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory*, 20(3), 539–566. <https://doi.org/10.1515/cllt-2023-0104>
- Biber, D. (1994). Representativeness in corpus design. In A. Zampolli, N. Calzolari, & M. Palmer (Eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*

- (pp. 377–407). Springer Netherlands.
https://doi.org/10.1007/978-0-585-35958-8_20
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
<https://books.google.co.id/books?id=vMfLCgAAQBAJ>
- Boyd, D. (2014). *It's complicated: The social lives of networked teens*. Yale University Press.
<https://books.google.co.id/books?id=p9u8AgAAQBAJ>
- Brookes, G., Atkins, S., & Harvey, K. (2022). Corpus linguistics and health communication: using corpora to examine the representation of health and illness. In A. O’Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (2nd ed.). Routledge.
- Bullock, B. E., & Toribio, A. J. (2009). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.
- Cahyanti, A. D., & Dharmawan, Y. Y. (2025). Reframing EFL classrooms students’ perspectives on translanguaging as a pedagogical strategy in Indonesian senior high school. *Indonesian Journal of Teaching and Learning (INTEL)*, 4(3), 197–206. <https://doi.org/10.56855/intel.v4i3.1673>
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge University Press.
<https://books.google.co.id/books?id=XnzangEACAAJ>
- Clift, R. (2016). Conversation analysis. In *Cambridge Textbooks in Linguistics*. Cambridge University Press.
<https://doi.org/10.1017/9781139022767>
- Coates, J. (2015). *Women, men, and language: A sociolinguistic account of gender differences in language* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315645612>

- Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.
<https://books.google.co.id/books?id=oJE462b0kv4C>
- Curzan, A. (2014). *Fixing English: Prescriptivism and language history*. Cambridge University Press.
<https://doi.org/10.1017/CBO9781139107327>
- Dance, W. (2025). Corpus linguistics and social media. In the *Reference Module in Social Sciences*. Elsevier.
<https://doi.org/10.1016/B978-0-323-95504-1.00479-8>
- De Fina, A., Schiffrin, D., & Bamberg, M. (2006). Discourse and identity. In *Studies in Interactional Sociolinguistics*. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511584459>
- Dharmawan, Y. Y., Ali, H. V., Prasatyo, B. A., & Amaliah, A. (2023). Comparing The Argumentative Essay Formats of Indonesian and Korean Students Using the Toulmin Model. *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, 9(2), 1099–1114.
<https://doi.org/https://doi.org/10.30605/onoma.v9i2.2922>
- Dharmawan, Yanuarius Yanu. (2023). Professional identity of faculty members: a case of multilingual context in the periphery. In R. J. Dickey & H. K. Lee (Eds.), *AsiaTEFL Proceedings 2023: Papers from the 21st AsiaTEFL Conference* (pp. 1083–1096). AsiaTEFL.
- Di Caro, V. N., & Samo, G. (2025). Quantifying trends of indefiniteness strategies in bilectal speakers of Sicilian and Italian. *Isogloss Open Journal of Romance Linguistics*, 11(6), 1–23. <https://doi.org/10.5565/rev/isogloss.488>
- Földes, C. (2023). Lexical variation in focus: A problem outline. *Acta Facultatis Philosophicae Universitatis Ostravienis Studia Germanistica*, 2023(33), 5–22.
<https://doi.org/10.15452/StudiaGermanistica.2023.33.0001>

- Friginal, E. (2024). Beyond expectations: (Applied) corpus linguistics and a framework for the study of spoken professional talk. *Iberica*, 2024(47), 43–66. <https://doi.org/10.17398/2340-2784.47.43>
- Gardner, M. H., & Kostadinova, V. (2024). Gettin' sociolinguistic data remotely: comparing vernacularity during online remote versus in-person sociolinguistic interviews. *Linguistics Vanguard*, 10, 417–429. <https://doi.org/10.1515/lingvan-2022-0069>
- Georgakopoulou, A., Giaxoglou, K., & Patron, S. (2023). *Small stories research: Tales, tellings, and tellers across contexts* (1st ed.). Routledge. <https://doi.org/10.4324/9781003253563>
- Ghifara, Z., Nurbaiti, S., & Rosidin, O. (2025). Analisis kontrastif ekspresi makian bahasa Jawa dialek Solo dan Malang dalam peristiwa tutur mahasiswa. *Kopula: Jurnal Bahasa, Sastra, Dan Pendidikan*, 7(2), 437–449. <https://doi.org/10.29303/kopula.v7i2.7103>
- Gries, S. Th., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. In *Handbook of Linguistic Annotation* (pp. 379–409). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_15
- Grieve, J. (2016). Regional variation in written American English. In *Studies in English Language*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139506137>
- Haitong, P. (2025). The role of corpus linguistics in contemporary linguistics research and translation studies. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Seriya 2. Yazykoznanie*, 24(1), 95–106. <https://doi.org/10.15688/jvolsu2.2025.1.8>
- Hidayati, D., Dharmawan, Y. Y., Andika Prasatyo, B., & Luciana. (2024). Exploring how translanguaging and

- multimodal learning improve EFL students' enjoyment and proficiency. *Journal on English as a Foreign Language*, 14(2), 446–475. <https://doi.org/10.23971/jeftl.v14i2.8012>
- Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge University Press. <https://books.google.co.id/books?id=dYXOzQEACAAJ>
- Hilpert, M., & Gries, S. Th. (2016). Quantitative approaches to diachronic corpus linguistics. In *The Cambridge Handbook of English Historical Linguistics* (pp. 36–53). Cambridge University Press. <https://doi.org/10.1017/CBO9781139600231.003>
- Holmes, J., & Wilson, N. (2022). *An introduction to sociolinguistics*. Routledge.
- Johnstone, B., & Andrus, J. (2024). *Discourse analysis* (4th ed.). Wiley-Blackwell.
- Kamila, N., Akhyaruddin, & Ningsih, A. G. (2025). Bentuk–bentuk kesantunan dalam bahasa Melayu Jambi. *Jurnal Basicedu*, 9(4), 1167–1175. <https://doi.org/10.31004/basicedu.v9i4.10443>
- Kusuma, E. D., Dharmawan, Y. Y., Oktariani S., A. I., & Mawuntu, S. C. (2023). Analysis of Translingual Practices: A Case Study of Indonesia Teenagers' Instagram Users. *Dialectical Literature and Educational Journal*, 8(2), 45–57. <https://doi.org/https://doi.org/10.51714/dlejpancasakti.v8i2.101.pp.45-57>
- Labov, W. (1973). *Sociolinguistic patterns*. University of Pennsylvania Press.
- Li, S., Halim, H. A., & Mamat, R. (2024). Unravelling Yu Guangzhong's identity through his poetic reflections: A corpus-based investigation. *Theory and Practice in*

- Language Studies*, 14(2), 561–569.
<https://doi.org/10.17507/tpls.1402.29>
- Li, Y. (2025). Gendered language in telephone conversations: a corpus-based study of lexical, grammatical, and pragmatic variation. *Humanities and Social Sciences Communications*, 13(1), 23. <https://doi.org/10.1057/s41599-025-06267-9>
- Litosseliti, Li. (2014). *Gender and language: Theory and practice*. Routledge.
<https://doi.org/10.4324/9780203784792>
- Mambu, J. E. (2025). Corpus-informed and critical approaches to addressing sustainable development goals (SDGS). *LLT Journal: A Journal on Language and Language Teaching*, 28(2), 587–613. <https://doi.org/10.24071/llt.v28i2.12023>
- Manurung, E. P. N., Manik, J. L. K., Sibuea, G. S., & Surip, M. (2025). Dinamika bahasa Indonesia di era digital: Antara perkembangan linguistik dan tantangan preservasi. *Morfologi: Jurnal Ilmu Pendidikan, Bahasa, Sastra Dan Budaya*, 3(2), 248–257.
<https://doi.org/10.61132/morfologi.v3i2.1551>
- Mardatillah, R., & Nurfadilah, N. (2025). Keberagaman morfologi dialek Bahasa Makassar di kabupaten Bantaeng. *DIKBASTRA: Jurnal Pendidikan Bahasa Dan Sastra*, 8(1), 10–19. <https://doi.org/10.22437/dikbastra.v8i1.43562>
- Meilinda, N., Sarwono, B. K., & Salamah, U. (2024). Domestication theory in action: Indonesian twitter corpus analysis during distance learning. *International Journal of Evaluation and Research in Education*, 13(3), 1321–1334.
<https://doi.org/10.11591/ijere.v13i3.27149>
- Meyerhoff, M. (2018). *Introducing sociolinguistics* (3rd ed.). Routledge. <https://doi.org/10.4324/9780429507922>
- Paquot, M., & Gries, S. Th. (2020). *A practical handbook of corpus linguistics*. Springer.

- Pérez-Paredes, P., & Curry, N. (2024). Epistemologies of corpus linguistics across disciplines. *Research Methods in Applied Linguistics*, 3(3). <https://doi.org/10.1016/j.rmal.2024.100141>
- Pitzl-Hagin, M. L. (2024). Emergent pragmatic conventions in spoken ELF corpus data: Micro-diachronic analysis of inclusive vs. exclusive multilingual practices. *Corpus Pragmatics*, 8(3), 175–200. <https://doi.org/10.1007/s41701-024-00166-1>
- Prasaty, B. A., Wiryani, R. S., Listiana, T. A., Siagian, C. E. M., Dharmawan, Y. Y., & Manara, C. (2025). Students' views on language diversity and heritage language maintenance in the Indonesian context. *IAFOR Journal of Education: Language Learning in Education*, 13(1), 195–2018. <https://doi.org/10.22492/ije.13.1.08>
- Puspitasari, D. A., Karlina, Y., Hernina, Kurniawan, Sutejo, & Danardana, A. S. (2024). Language choices and digital identity of high school student text messages in the new capital city of Indonesia: implication for language education. *International Journal of Language Education*, 8(1), 162–184. <https://doi.org/10.26858/ijole.v8i1.63833>
- Putri, D. R. (2025). Discourse analysis: A literature study. *Majapahit Journal of English Studies*, 2(2), 145–153.
- Puupponen, A., Hodge, G., Anible, B., Salonen, J., Wainio, T., Keränen, J., Hernández, D., & Jantunen, T. (2025). Opening up corpus FinSL: Enriching corpus analysis with linguistic ethnography in a study of constructed action. *Linguistics*, 63(1), 277–316. <https://doi.org/10.1515/ling-2023-0196>
- Rahmawati, A., & Lestari, P. M. (2024). Pendakwah berbahasa Jawa ngapak dalam ceramah ustadzah mumpuni pada channel youtube: Kajian sosiolinguistik. In *Bahasa dan Sastra* (Vol. 10, Number 2). Pendidikan. <https://e-journal.my.id/onoma>

- Ramadani, L., & Dharmawan, Y. Y. (2025). English day program and its impact on students' speaking proficiency: A socio-linguistic exploration. *Indonesian Journal of Teaching and Learning*, 4(2), 78–87. <https://doi.org/10.56855/intel.v4i2.1396>
- Rohmah, R. A. (2024). Dinamika penggunaan bahasa Indonesia di media sosial: analisis literatur. *Journals of Indonesian Multidisciplinary Research*, 3(1), 71–77.
- Sadigzade, Z. (2025). Dialect diversity and social change: New approaches in sociolinguistics. *Journal of Azerbaijan Language and Education Studies*, 2(3), 91–105. <https://doi.org/10.69760/jales.2025002017>
- Schneider, J., Fokkema, T., Matias, A. R., Stojčić, S., Ugrina, D., & Vera-Larrucea, C. (2025). Identities: Urban belonging and intercultural relations. In M. Crul, J. Schneider, & F. Lelie (Eds.), *The European Second Generation Compared: Does the Integration Context Matter?* (1st ed., pp. 285–340). Routledge. <https://doi.org/10.4324/9781003706083-8>
- Sofa, A. R., & Febrianti, A. (2025). Dialektologi bahasa Arab: Analisis perbedaan linguistik berdasarkan kajian pustaka. *Lencana: Jurnal Inovasi Ilmu Pendidikan*, 3(2), 76–87. <https://doi.org/10.55606/lencana.v3i2.5018>
- Soyusiawaty, D., Fadlil, A., & Sunardi, S. (2025). Pengembangan korpus bahasa minang pada spell error corpus for minang language (SPEML). *TEKNOSI*, 11(1). <https://doi.org/10.25077/TEKNOSI.v11i01.2025.17-26>
- Suhaili, A., Suganda, D., Darmayanti, N., & Yuliawati, S. (2024). A corpus-assisted critical discourse analysis of the Iranian people in the corpus of Donald Trump's social media texts. *Theory and Practice in Language Studies*, 14(1), 88–97. <https://doi.org/10.17507/tpls.1401.10>
- Suhartatik, Siswanto, W., & Roekhan. (2026). Enhancing vocabulary acquisition and language learning through

- corpus linguistic approaches: the role of onomastics in teaching batik Madura's cultural identity. *Journal of Teaching and Learning*, 20(1), 337–1. <https://doi.org/10.22329/jtl.v19i5.9608>
- Tagliamonte, S. A. (2016). *Teen talk: The language of adolescents*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139583800>
- Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1), 3–34. <https://doi.org/10.1215/00031283-2008-001>
- Travis, C. E., & Gan, Q. (2025). The intersection of ethnicity and social class in language variation and change. *Language Variation and Change*. <https://doi.org/10.1017/S0954394525100537>
- Umam, K., Ridlo, U., & Alek. (2025). Pendekatan dalam penelitian bahasa dan sastra. *Journal of Literature Review*, 1(2), 292–303. <https://doi.org/10.63822/g1zdp64>
- Wearne, S., & Riedy, C. (2024). Whose “place” is it? Using corpus-based techniques to sketch place-based sustainability discourses in public and academic forums. *Sustainability Science*, 19(3), 883–904. <https://doi.org/10.1007/s11625-024-01466-w>
- Zamrizal, N. A., Erlangga, B., Yuniati, I., Yuniarti, N., Julianti, S., Ningsih, D. A., Lorenza, A., Pajri, R., Kurniati, N., Putri, T. A., Sari, N. P., Manulang, J. W., & Utami, P. R. (2025). Ragam bahasa Indonesia dan pedoman umum ejaan bahasa Indonesia (PUEBI). *Lateralisasi*, 13(2), 678–688. <http://jurnal.umb.ac.id/index.php/lateralisasi>
- Zappavigna, M. (2018). *Searchable talk: Hashtags and social media metadiscourse*. Bloomsbury.

PROFIL PENULIS



Yanuarius Yanu Dharmawan

Penulis adalah dosen dan peneliti di bidang linguistik. Ia memiliki latar belakang akademik lintas bidang yang memperkaya pendekatannya dalam pendidikan dan bahasa, dengan kualifikasi Doktor Linguistik Terapan Bahasa Inggris, Magister Manajemen, Magister Humaniora dalam Linguistik

Bahasa Inggris, serta Sarjana Sastra dalam bidang sastra Inggris. Minat keilmuannya mencakup fonologi, morfologi, semantik, sintaksis, pengajaran bahasa berbasis teknologi, dan kebijakan pendidikan. Karya dan perhatiannya berkembang pada berbagai bidang, seperti *translanguaging*, pembelajaran multimodal, penguatan pelafalan melalui media digital, interaksi kelas EFL, literasi akademik dosen untuk publikasi internasional, keberagaman bahasa, serta pemeliharaan bahasa warisan dalam konteks Indonesia. Ia juga terlibat dalam pengembangan pembelajaran berbasis korpus, *blended learning*, dan pemanfaatan media digital untuk mendukung proses belajar yang lebih efektif, adaptif, dan manusiawi. Melalui kerja ilmiah, pedagogis, dan pengembangan pendidikan yang dijalaninya, ia menempatkan bahasa bukan hanya sebagai alat akademik, tetapi juga sebagai ruang dialog, etika, dan pembentukan martabat manusia. Latar belakangnya yang memadukan linguistik, pendidikan, dan manajemen memperlihatkan komitmennya untuk menghadirkan pembelajaran yang tidak hanya unggul secara akademik, tetapi juga relevan bagi perkembangan manusia dan kehidupan bersama.

BAB 12

LINGUISTIK KORPUS DALAM TERJEMAHAN DAN INTERPRETASI

Ria Saraswati
Universitas Esa Unggul, Jakarta
E-mail: ria.saraswati@esaunggul.ac.id

PENDAHULUAN

Perkembangan teknologi digital dalam beberapa dekade terakhir telah memengaruhi cara para peneliti mempelajari bahasa sekaligus mengubah praktik penerjemahan dalam konteks profesional. Salah satu pendekatan yang semakin mendapat perhatian dalam kajian linguistik terapan adalah linguistik korpus, yaitu pendekatan yang memanfaatkan kumpulan teks digital yang disusun secara sistematis dan dianalisis dengan bantuan perangkat komputasional. Melalui korpus, peneliti dapat mengamati penggunaan bahasa berdasarkan data autentik yang merepresentasikan praktik komunikasi nyata dalam berbagai konteks. Pendekatan ini memberikan peluang untuk memahami fenomena kebahasaan secara lebih empiris dibandingkan metode analisis yang hanya bertumpu pada intuisi linguistik. Dalam bidang penerjemahan, linguistik korpus memungkinkan analisis hubungan antara teks sumber dan teks sasaran dilakukan secara lebih sistematis serta membuka ruang bagi penelitian yang berbasis bukti (*evidence-based research*). Dengan demikian, korpus tidak hanya berfungsi sebagai sumber data linguistik, tetapi juga sebagai instrumen metodologis yang penting dalam perkembangan studi penerjemahan modern (McEnery & Hardie, 2012; G. Wang & Xin, 2024).

Dalam konteks studi penerjemahan kontemporer, linguistik korpus memberikan kontribusi yang signifikan dalam mengungkap pola dan kecenderungan bahasa yang muncul dalam teks terjemahan. Dengan memanfaatkan perangkat analisis korpus seperti *concordance* dan analisis frekuensi, peneliti dapat menelusuri bagaimana kata, frasa, maupun struktur tertentu digunakan secara berulang dalam sejumlah besar teks terjemahan. Analisis ini memungkinkan identifikasi berbagai kecenderungan linguistik yang sering disebut sebagai *translation universals*, seperti simplifikasi, eksplisitasi, dan normalisasi bahasa. Selain memperkaya pengembangan teori penerjemahan, pemanfaatan korpus juga memiliki implikasi praktis dalam kegiatan penerjemahan profesional maupun pendidikan penerjemah. Korpus dapat digunakan sebagai sumber rujukan untuk memeriksa penggunaan istilah, kolokasi, dan struktur kalimat dalam bahasa sasaran sehingga membantu menghasilkan terjemahan yang lebih akurat dan sesuai dengan konvensi bahasa. Di sisi lain, dalam proses pembelajaran, korpus memungkinkan mahasiswa mempelajari bahasa melalui data autentik dan mengembangkan kemampuan analisis terhadap pola kebahasaan dalam teks terjemahan (Baker & Saldanha, 2020; Kwok et al., 2025).

KORPUS PARALEL DAN KORPUS KOMPARATIF

Dalam kajian linguistik korpus yang berkaitan dengan penerjemahan, dua jenis korpus yang sering digunakan adalah korpus paralel dan korpus komparatif. Korpus paralel merujuk pada kumpulan teks yang terdiri atas teks sumber dan teks terjemahannya yang disusun secara berpasangan sehingga memungkinkan analisis langsung terhadap hubungan antara kedua teks tersebut. Setiap bagian teks dalam bahasa sumber biasanya disejajarkan dengan segmen yang setara dalam bahasa sasaran sehingga peneliti dapat menelusuri bagaimana makna,

struktur, maupun pilihan leksikal dialihkan selama proses penerjemahan. Melalui pendekatan ini, peneliti dapat mengamati strategi penerjemahan secara empiris berdasarkan data linguistik yang nyata. Selain itu, korpus paralel juga memiliki peran penting dalam pengembangan teknologi bahasa, terutama dalam bidang penerjemahan mesin dan pemrosesan bahasa alami, karena data bilingual yang tersedia memungkinkan sistem komputasional mempelajari hubungan semantik dan struktural antara dua bahasa secara lebih sistematis. Dengan menganalisis sejumlah besar pasangan teks, peneliti dapat mengidentifikasi kecenderungan tertentu dalam praktik penerjemahan yang mungkin tidak terlihat melalui analisis terhadap teks individual (B. Li et al., 2025; Mikeleni & Oliver, 2024; Sulistyo et al., 2025; Yang et al., 2025).

Berbeda dengan korpus paralel, korpus komparatif merupakan kumpulan teks yang dibandingkan berdasarkan karakteristik tertentu tanpa adanya hubungan langsung antara teks sumber dan teks terjemahan. Dalam korpus komparatif, peneliti biasanya membandingkan dua kelompok teks dalam bahasa yang sama, misalnya antara teks asli yang ditulis oleh penutur asli dan teks terjemahan yang berasal dari bahasa lain. Pendekatan ini memungkinkan analisis terhadap perbedaan gaya bahasa, struktur sintaksis, maupun frekuensi penggunaan kosakata antara teks terjemahan dan teks yang ditulis secara langsung dalam bahasa sasaran. Dengan melakukan perbandingan tersebut, peneliti dapat mengidentifikasi ciri khas bahasa terjemahan yang sering disebut sebagai *translation universals*, seperti kecenderungan penggunaan struktur yang lebih sederhana atau penjelasan makna yang lebih eksplisit dibandingkan dengan teks asli. Analisis korpus komparatif juga membantu memahami bagaimana teks terjemahan berfungsi dalam sistem bahasa sasaran serta bagaimana praktik penerjemahan dipengaruhi oleh perkembangan teknologi dan

perubahan konteks komunikasi global (Ji et al., 2025; Lapshinova-Koltunski et al., 2022; B. Li et al., 2025; Tianqi, 2024).

IDENTIFIKASI POLA TERJEMAHAN

Identifikasi pola terjemahan merupakan salah satu pendekatan penting dalam penelitian penerjemahan berbasis korpus. Melalui analisis terhadap kumpulan teks yang besar dan terstruktur, peneliti dapat mengamati bagaimana unsur leksikal, struktur sintaksis, maupun strategi pragmatik digunakan dalam teks terjemahan. Pendekatan ini memungkinkan peneliti mengidentifikasi kecenderungan linguistik yang muncul secara berulang sehingga memberikan gambaran mengenai karakteristik bahasa terjemahan. Dengan bantuan perangkat analisis korpus seperti *concordance tools*, analisis frekuensi, serta analisis kolokasi, peneliti dapat menelusuri pola penggunaan kata atau struktur tertentu dalam sejumlah besar teks. Hasil analisis tersebut membantu menjelaskan bagaimana keputusan penerjemahan tidak hanya dipengaruhi oleh struktur bahasa sumber, tetapi juga oleh norma penggunaan bahasa dalam bahasa sasaran serta konteks komunikasi yang melatarbelakangi proses penerjemahan. Dengan demikian, pendekatan berbasis korpus memungkinkan penelitian penerjemahan dilakukan secara lebih objektif karena didasarkan pada bukti linguistik yang dapat diamati secara sistematis (Biber et al., 1998; Granger & Lefer, 2023).

Salah satu pola yang sering ditemukan dalam penelitian penerjemahan adalah fenomena *simplification*, yaitu kecenderungan teks terjemahan menggunakan struktur bahasa yang lebih sederhana dibandingkan dengan teks sumber. Penelitian Afzaal et al. (2025), misalnya, menunjukkan bahwa teks terjemahan cenderung menggunakan kalimat yang lebih pendek serta struktur sintaksis yang lebih sederhana. Selain

simplifikasi, penelitian berbasis korpus juga mengidentifikasi fenomena *explicitation*, yaitu kecenderungan penerjemah menyatakan informasi yang sebelumnya bersifat implisit dalam teks sumber secara lebih jelas dalam bahasa sasaran. Dalam penelitian Bao (2025), penerjemah ditemukan sering menambahkan penanda hubungan logis seperti *therefore*, *however*, atau *in other words* untuk memperjelas hubungan antargagasan dalam teks akademik. Selain itu, pola lain yang juga sering ditemukan adalah *normalization* dan *interference*. Normalisasi menunjukkan kecenderungan penerjemah menggunakan bentuk bahasa yang lebih konvensional dalam bahasa sasaran, sedangkan interferensi terjadi ketika struktur bahasa sumber masih memengaruhi bentuk bahasa dalam teks sasaran, terutama pada pasangan bahasa yang memiliki perbedaan tipologis yang signifikan (Aleksandrovna & Zubayda, 2026; Mahadi et al., 2010).

PENERJEMAHAN BERBASIS DATA KORPUS

Penerjemahan berbasis data korpus merupakan pendekatan yang menempatkan korpus sebagai sumber rujukan utama dalam proses pengambilan keputusan penerjemahan. Dalam pendekatan ini, penerjemah tidak hanya mengandalkan intuisi linguistik atau pengalaman pribadi, tetapi juga memanfaatkan data bahasa autentik yang tersedia dalam korpus digital. Melalui analisis frekuensi kata, kolokasi, dan struktur sintaksis dalam sejumlah besar teks, penerjemah dapat mengidentifikasi pola penggunaan bahasa yang lazim dalam bahasa sasaran. Pendekatan ini membantu memastikan bahwa hasil terjemahan tidak hanya akurat secara makna, tetapi juga sesuai dengan konvensi linguistik yang berlaku dalam komunitas pengguna bahasa sasaran. Selain itu, pendekatan berbasis korpus juga memungkinkan analisis penerjemahan dilakukan secara lebih objektif karena memadukan analisis kuantitatif dengan

interpretasi kualitatif terhadap data linguistik yang tersedia, sehingga memberikan dasar empiris yang lebih kuat dalam studi penerjemahan (Ding, 2024; Saldanha & O'Brien, 2014).

Dalam praktik profesional, penerjemahan berbasis korpus sering dikombinasikan dengan teknologi penerjemahan berbantuan komputer seperti *translation memory* dan *computer-assisted translation tools*. Teknologi tersebut memungkinkan penerjemah mengakses korpus paralel serta memanfaatkan terjemahan sebelumnya sebagai referensi selama proses penerjemahan berlangsung. Dengan memanfaatkan basis data bilingual yang besar, penerjemah dapat menjaga konsistensi terminologi sekaligus meningkatkan efisiensi kerja. Di samping itu, penggunaan korpus juga membuka peluang bagi penelitian yang lebih mendalam mengenai gaya penerjemah dan strategi penerjemahan. Analisis terhadap korpus paralel memungkinkan peneliti mengidentifikasi preferensi linguistik tertentu yang digunakan oleh penerjemah, seperti pilihan kosakata, struktur kalimat, maupun strategi adaptasi budaya. Melalui pendekatan ini, penelitian penerjemahan tidak hanya berkontribusi pada peningkatan kualitas praktik penerjemahan, tetapi juga memperkaya pengembangan teori penerjemahan yang didasarkan pada bukti empiris (Cheung & Crosthwaite, 2025; Y. Li, 2024; Liu & Afzaal, 2021).

ANALISIS KUALITAS TERJEMAHAN

Analisis kualitas terjemahan merupakan aspek penting dalam studi penerjemahan karena berkaitan dengan sejauh mana pesan dalam teks sumber dapat disampaikan secara tepat dalam bahasa sasaran. Dalam pendekatan berbasis korpus, evaluasi kualitas terjemahan dilakukan dengan membandingkan teks terjemahan dengan korpus referensi yang merepresentasikan penggunaan bahasa yang alami dalam bahasa sasaran. Melalui analisis ini, peneliti dapat mengidentifikasi kesesuaian leksikal,

struktur sintaksis, serta tingkat kohesi dalam teks terjemahan. Pendekatan berbasis korpus memberikan dasar yang lebih objektif dalam menilai kualitas terjemahan karena evaluasi dilakukan berdasarkan data bahasa yang luas dan autentik, bukan semata-mata bergantung pada intuisi atau penilaian subjektif. Selain itu, analisis tersebut juga memungkinkan peneliti memahami sejauh mana teks terjemahan mengikuti norma linguistik yang berlaku dalam bahasa sasaran (Wang, 2024; Xu, 2024).

Perkembangan teknologi digital turut memperluas metode analisis kualitas terjemahan melalui penggunaan evaluasi otomatis berbasis korpus. Dengan memanfaatkan dataset bilingual yang besar, sistem evaluasi otomatis dapat membantu mengidentifikasi kesalahan terjemahan serta menilai kesesuaian antara teks sumber dan teks sasaran secara lebih sistematis. Pendekatan ini sangat relevan dalam konteks penerjemahan mesin, di mana analisis berbasis korpus digunakan untuk menilai akurasi, kelancaran, dan konsistensi hasil terjemahan. Selain itu, analisis kualitas terjemahan juga dapat dilakukan pada berbagai tingkat linguistik, mulai dari pilihan kosakata hingga struktur wacana, sehingga memungkinkan peneliti mengidentifikasi faktor-faktor yang memengaruhi kualitas terjemahan, seperti gaya penerjemah, konteks budaya, maupun tujuan komunikasi dari teks yang diterjemahkan (Pan & Wu, 2024; Wang & Chen, 2023).

PENERAPAN KORPUS DALAM PELATIHAN PENERJEMAH

Dalam pendidikan penerjemahan, pemanfaatan korpus telah menjadi salah satu pendekatan pedagogis yang semakin banyak digunakan dalam beberapa tahun terakhir. Penggunaan korpus memungkinkan mahasiswa mempelajari bahasa melalui data autentik yang mencerminkan penggunaan bahasa dalam konteks

komunikasi nyata. Melalui analisis korpus, mahasiswa dapat mengamati bagaimana kata, frasa, dan struktur kalimat digunakan dalam berbagai jenis teks sehingga membantu mereka memahami variasi gaya bahasa dalam konteks yang berbeda. Pendekatan ini juga mendorong mahasiswa untuk mengembangkan kesadaran linguistik yang lebih tinggi serta kemampuan analisis terhadap pilihan bahasa yang digunakan dalam teks terjemahan. Beberapa penelitian menunjukkan bahwa penggunaan korpus dalam pelatihan penerjemah dapat meningkatkan kemampuan analisis bahasa sekaligus membantu mahasiswa dalam mengambil keputusan yang lebih tepat selama proses penerjemahan (Meng & Pan, 2022).

Selain meningkatkan kompetensi linguistik, penggunaan korpus dalam pelatihan penerjemah juga membantu mahasiswa mengembangkan kompetensi teknologi yang diperlukan dalam industri penerjemahan modern. Mahasiswa yang terbiasa menggunakan perangkat analisis korpus akan lebih siap menghadapi lingkungan kerja yang memanfaatkan berbagai teknologi digital seperti *translation memory*, *terminology management systems*, dan perangkat penerjemahan berbantuan komputer lainnya. Lebih jauh lagi, pendekatan berbasis korpus juga mendorong pembelajaran yang bersifat eksploratif dan berbasis penelitian karena mahasiswa tidak hanya mempelajari teori penerjemahan secara konseptual, tetapi juga melakukan analisis data linguistik secara mandiri melalui eksplorasi korpus. Dengan demikian, korpus tidak hanya berfungsi sebagai sumber data linguistik, tetapi juga sebagai sarana pedagogis yang efektif dalam mengembangkan kompetensi analitis dan profesional bagi calon penerjemah di era digital (Fredrick & Craven, 2025; Zhang et al., 2025).

KESIMPULAN

Linguistik korpus telah memberikan kontribusi yang

signifikan terhadap perkembangan studi penerjemahan dan interpretasi dalam beberapa dekade terakhir. Melalui pemanfaatan kumpulan data bahasa yang luas dan terstruktur, pendekatan ini memungkinkan analisis penerjemahan dilakukan secara lebih sistematis dan berbasis bukti empiris. Penggunaan korpus paralel dan korpus komparatif memberikan kesempatan bagi peneliti untuk mengamati hubungan antara teks sumber dan teks sasaran serta mengidentifikasi karakteristik linguistik yang muncul dalam teks terjemahan. Selain itu, pendekatan berbasis korpus juga mendukung praktik penerjemahan yang lebih akurat dan konsisten karena penerjemah dapat memanfaatkan data bahasa autentik sebagai sumber referensi dalam menentukan pilihan leksikal maupun struktur kalimat yang sesuai dengan norma bahasa sasaran.

Di samping itu, pemanfaatan linguistik korpus juga memiliki implikasi penting dalam pendidikan penerjemahan. Penggunaan korpus dalam proses pembelajaran memungkinkan mahasiswa memahami penggunaan bahasa melalui data autentik serta mengembangkan kemampuan analisis terhadap pola kebahasaan dalam teks terjemahan. Pendekatan ini mendorong pembelajaran yang lebih eksploratif dan berbasis penelitian, sehingga mahasiswa tidak hanya mempelajari teori penerjemahan secara konseptual, tetapi juga mampu menerapkannya melalui analisis data linguistik yang nyata. Dengan semakin berkembangnya teknologi pengolahan bahasa dan ketersediaan korpus digital yang semakin luas, linguistik korpus diperkirakan akan terus memainkan peran penting dalam pengembangan penelitian, praktik profesional, serta pendidikan penerjemahan di masa mendatang.

DAFTAR PUSTAKA

Afzaal, M., Huang, B., & El-Dakhs, D. A. S. (2025). Decoding the digital: a corpus-based study of simplifications and

- other translation universals in translated texts. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1517107>
- Aleksandrova, O.-R. V., & Zubayda, M. (2026). Corpus Analysis of Translation Universals in Parallel Texts of Various Genres (French-Russian). *American Academic Publishers*, 6(01). <https://www.academicpublishers.org/journals/index.php/ijai>
- Baker, Mona., & Saldanha, Gabriela. (2020). *Routledge encyclopedia of translation studies*. Routledge, Taylor and Francis Group.
- Bao, P. (2025). Validation of Translation Universals Based on Comparable Corpus: Simplification and Explication. *International Linguistics Research*, 8(1). <https://doi.org/10.30560/ilr.v8n1p21>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics Investigating Language Structure and Use*. Cambridge University Press.
- Cheung, L., & Crosthwaite, P. (2025). CorpusChat: integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2025.2506480>
- Ding, J. (2024). Corpus-based Translation Studies: Examining Media Language through a Linguistic Lens. *2nd International Conference on Language and Cultural Communication (ICLCC 2024)*, 185. <https://doi.org/10.1051/shsconf/202418501012>
- Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: a corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1616935>
- Granger, S., & Lefer, M.-A. (2023). Learner Translation Corpora: Bridging the Gap Between Learner Corpus

- Research and Corpus-based Translation Studies. *International Journal of Learner Corpus Research*, 9(1), 1–28.
- Ji, M., Fu, J., & Wang, F. (Robin). (2025). Comparative and cognitive dimensions of corpus-based translation stylistics. *Studia Neophilologica*, 1–10. <https://doi.org/10.1080/00393274.2025.2489936>
- Kwok, H. L., Shi, Y., Xu, H., Li, D., & Liu, K. (2025). GenAI as a translation assistant? A corpus-based study on lexical and syntactic complexity of GPT-post-edited learner translation. *System*, 130. <https://doi.org/10.1016/j.system.2025.103618>
- Lapshinova-Koltunski, E., Popovi, M., & Koponen, M. (2022). DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations. *Proceedings Of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 1751–1760. <https://s3.amazonaws.com/>
- Li, B., Luo, J., Briakou, E., & Google, C. C. (2025). Leveraging Domain Knowledge at Inference Time for LLM Translation: Retrieval versus Generation. *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, 91–106.
- Li, Y. (2024). Corpus-Based Studies on Translator's Style: Retrospect and Prospect. *International Journal of Applied Linguistics and Translation*, 10(3), 44–49. <https://doi.org/10.11648/j.ijalt.20241003.12>
- Liu, K., & Afzaal, M. (2021). Translator's Style Through Lexical Bundles: A Corpus-Driven Analysis of Two English Translations of Honglounmeng. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.633422>
- Mahadi, T. S. T., Vaezian, H., & Akbari, M. (2010). *Corpora in Translation: A Practical Guide*. Peter Lang.

- McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- Meng, L., & Pan, F. (2022). Using corpora to reveal style in translation: The case of The Song of Everlasting Sorrow. *Frontiers in Psychology*, 13. <https://api.semanticscholar.org/CorpusID:253164800>
- Mikeleni, B., & Oliver, A. (2024). Using a multilingual literary parallel corpus to train NMT systems. *Proceedings Of the 1st Workshop on Creative-Text Translation and Technology*, 1–9. <https://elrc-share.eu/repository/search/>
- Pan, F., & Wu, X. (2024). Mapping intellectual structures and research trends of translation studies: a bibliometric analysis from 2007 to 2021. *Perspectives*, 32(4), 736–757. <https://doi.org/10.1080/0907676X.2023.2275750>
- Saldanha, Gabriela., & O'Brien, Sharon. (2014). *Research methodologies in translation studies*. Routledge.
- Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., Ahda, A. iini F., Astawa, I. N. G. A., & Dwiyanto, F. A. (2025). Multilingual Parallel Corpus for Indonesian Low-Resource Languages. *JOIV: International Journal on Informatics Visualization*, (5), 2176–2182. <https://doi.org/http://dx.doi.org/10.62527/joiv.9.5.3412>
- Tianqi, F. (2024). A Corpus-based Comparative Study of Translation Style of Five English Versions of Xiao Yaoyou. *International Journal of English Language Studies*, 69–77. <https://doi.org/10.32996/ijels>
- Wang, G., & Xin, Y. (2024). An analytical framework for corpus-based translation studies. *Humanities and Social Sciences Communications*, 11(1). <https://doi.org/10.1057/s41599-024-04250-4>
- Wang, L., & Chen, Q. (2023). Review of Translation Quality Assessment Research: Current Studies and Development. *Scholars International Journal of Linguistics and*

- Literature*, 6(12), 474–477.
<https://doi.org/10.36348/sijll.2023.v06i12.003>
- Wang, X. (2024). A Corpus-Assisted Study of Translation Criticism and Translation Quality Assessment. *Journal of Research in Social Science and Humanities*, 3(9), 86–91.
<https://doi.org/10.56397/jrssh.2024.09.11>
- Xu, H. (2024). Assessment of Computer-Aided Translation Quality Based on Large-Scale Corpora. *International Journal of E-Collaboration*, 20, 1–14.
<https://doi.org/10.4018/IJeC.357274>
- Yang, C., Chen, J., & Hou, S. (2025). Science Mapping of Corpus-Based Translation and Interpreting Studies: Trends and Future Directions. *Corpus-Based Studies across Humanities*. <https://doi.org/10.1515/csh-2025-0017>
- Zhang, W., Xie, R., Yao, Y., & Li, D. (2025). Lexico-Syntactic Complexity in Machine Interpreting: A Corpus-Based Comparison With Human Interpreting and Translation. *International Journal of Applied Linguistics*, n/a(n/a).
<https://doi.org/https://doi.org/10.1111/ijal.12830>

PROFIL PENULIS



Ria Saraswati

Penulis meraih gelar doktor di bidang Linguistik Terapan dari Universitas Negeri Jakarta dengan penelitian yang berfokus pada keterkaitan antara teori linguistik dan penerapannya dalam praktik kebahasaan. Saat ini, ia merupakan dosen pada Program Studi Pendidikan Bahasa Inggris di Universitas Esa Unggul, Indonesia. Dalam perjalanan akademiknya, ia aktif dalam pengajaran, pengembangan kurikulum, dan penelitian yang mengintegrasikan pendekatan teoretis dan pedagogis. Minat penelitiannya meliputi Linguistik, Studi Penerjemahan—khususnya penerjemahan ungkapan bermuatan budaya, serta Pendidikan Bahasa Inggris. Ia juga menaruh perhatian pada komunikasi lintas budaya dan analisis sastra dengan perspektif feminis, terutama dalam mengkaji representasi gender dan relasi kuasa dalam teks sastra. Selain itu, Ria terlibat dalam berbagai inisiatif pendidikan untuk meningkatkan kualitas pembelajaran bahasa Inggris melalui pemanfaatan teknologi, pedagogi berbasis genre, dan pendekatan *English for Specific Purposes* (ESP) yang kontekstual. Karya-karyanya mencerminkan komitmen pada pengembangan kompetensi linguistik sekaligus kesadaran budaya kritis pada pembelajar.

BAB 13

BIG DATA DAN KORPUS DIGITAL BERSKALA BESAR

Abd. Hakim Muslim
MAN 5 Jombang, Jombang
E-mail: hakim519@ymail.com

PENDAHULUAN

Perkembangan pesat teknologi digital selama dua dekade terakhir telah secara mendalam mengubah lanskap penelitian bahasa. Penyebaran konektivitas internet, *platform* media sosial, penerbitan digital, dan proyek digitalisasi skala besar telah menghasilkan volume data linguistik yang belum pernah terjadi sebelumnya yang tersedia untuk dianalisis. Tidak seperti periode sebelumnya dalam penelitian linguistik, ketika para peneliti bergantung pada koleksi teks yang terbatas atau korpus yang disusun secara manual, para sarjana kontemporer beroperasi dalam lingkungan yang dicirikan oleh aliran data tekstual dan multimodal yang masif dan terus-menerus dihasilkan. Transformasi ini sering digambarkan melalui lensa *Big Data*, sebuah paradigma yang telah membentuk kembali disiplin ilmu mulai dari ekonomi dan kedokteran hingga ilmu komputasi dan pendidikan. Dalam studi bahasa, *Big Data* telah mengkatalisasi pergeseran metodologis dan epistemologis yang signifikan, menempatkan linguistik korpus di pusat penyelidikan empiris di era digital.

Secara historis, linguistik korpus berkembang melalui korpus skala kecil yang dikelola dengan cermat seperti *British National Corpus (BNC)*, yang pada saat dirilis, mewakili sumber daya yang inovatif untuk studi sistematis tentang penggunaan bahasa yang autentik. Namun, meskipun korpus tersebut secara

metodologis ketat dan seimbang, ukurannya dan frekuensi pembaruannya terbatas. Ekosistem digital kontemporer, sebaliknya, menghasilkan data linguistik dengan laju eksponensial melalui *platform* seperti Twitter, YouTube, blog, portal berita online, dan arsip digital. Akibatnya, skala data tekstual yang tersedia telah bergeser dari jutaan kata menjadi miliaran, bahkan triliunan kata, memungkinkan bentuk analisis baru yang sebelumnya tidak terbayangkan.

Transisi dari linguistik korpus tradisional ke korpus digital skala besar ini tidak hanya mencerminkan pertumbuhan kuantitatif, tetapi juga reorientasi kualitatif praktik penelitian. Korpus skala besar seperti *Corpus of Contemporary American English (COCA)* dan mega-korpus yang berasal dari *web* telah menunjukkan bagaimana pembaruan dinamis, keragaman genre, dan konten yang dihasilkan pengguna memperluas dasar empiris untuk generalisasi linguistik. Lebih lanjut, arsip sejarah digital telah membuka jalan baru untuk penelitian diakronis, memungkinkan para sarjana untuk melacak perubahan bahasa selama berabad-abad dengan presisi yang semakin meningkat. Oleh karena itu, revolusi digital tidak hanya memperluas ukuran korpus; hal ini telah membentuk kembali landasan epistemologis tentang bagaimana bukti bahasa dikonseptualisasikan, dikumpulkan, dan diinterpretasikan.

Untuk memahami implikasi *Big Data* bagi linguistik korpus, perlu diklarifikasi apa yang membedakan *Big Data* dari bentuk-bentuk dataset linguistik sebelumnya. *Big Data* umumnya dicirikan oleh empat dimensi inti: *Volume*, *Velocity*, *Variety*, dan *Veracity*. *Volume* mengacu pada skala besar data yang tersedia; *Velocity* menunjukkan kecepatan data yang dihasilkan dan diperbarui; *Variety* menyoroti keragaman jenis data, termasuk wacana tertulis, lisan, multimodal, dan media sosial; dan *Veracity* menyangkut masalah keandalan, keterwakilan, dan kualitas data. Dalam linguistik korpus,

dimensi-dimensi ini secara langsung memengaruhi desain korpus, praktik anotasi, dan metodologi analitis.

Korpus tradisional biasanya dirancang sesuai dengan prinsip pengambilan sampel yang ketat, memastikan keterwakilan di seluruh genre dan register. Korpus tersebut seringkali statis dan terbatas cakupannya, menekankan keseimbangan daripada skala. Sebaliknya, korpus digital skala besar dapat menggabungkan data yang dikumpulkan dari *web*, pembaruan dinamis, dan proses anotasi yang dibantu mesin. Meskipun pendekatan ini meningkatkan skala dan keragaman, pendekatan ini juga menimbulkan tantangan metodologis yang berkaitan dengan pembersihan data, pengurangan kebisingan, dan bias pengambilan sampel. Pergeseran menuju *Big Data* dengan demikian memaksa para ahli linguistik korpus untuk mengintegrasikan alat komputasi dan teknik pemodelan statistik ke dalam kerangka analitis mereka.

Akibatnya, hubungan antara linguistik korpus dan linguistik komputasional menjadi semakin terjalin. Kemajuan dalam *Natural Language Processing (NLP)*, pembelajaran mesin, dan sistem penandaan otomatis memungkinkan pemrosesan kumpulan data yang sangat besar dengan efisiensi yang belum pernah terjadi sebelumnya. Pada saat yang sama, linguistik korpus memberikan wawasan yang berlandaskan teori tentang pola bahasa, kolokasi, dan struktur wacana yang menginformasikan pemodelan komputasional. Penelitian terbaru menekankan sinergi yang semakin meningkat antara bidang-bidang ini, khususnya di bidang-bidang seperti analisis semantik, deteksi sentimen, dan klasifikasi wacana otomatis (Paquot & Gries, 2021; Solovyev et al., 2022; Sukhrob, 2024). Integrasi analitik *Big Data* ke dalam linguistik korpus dengan demikian mencerminkan konvergensi yang lebih luas antara teori linguistik dan inovasi komputasional.

Terlepas dari kemajuan teknologi, landasan teoritis inti linguistik korpus tetap berakar pada pendekatan empiris dan berbasis penggunaan bahasa. Linguistik korpus beroperasi berdasarkan prinsip bahwa struktur bahasa muncul dari pola penggunaan otentik, bukan dari penilaian introspektif semata. Frekuensi, kolokasi, dan pola berfungsi sebagai konstruksi analitis fundamental, yang mengungkapkan keteraturan yang membentuk tata bahasa, kosakata, dan konvensi wacana. Korpus digital berskala besar memperkuat kemungkinan analitis ini dengan memungkinkan para peneliti untuk memeriksa konstruksi langka, fenomena ekor panjang, dan variasi lintas genre dengan kekokohan statistik yang lebih besar.

Analisis frekuensi, misalnya, telah berevolusi dari penghitungan kata sederhana menjadi pemodelan pola distribusi yang canggih di berbagai register dan periode waktu. Studi kolokasi sekarang memanfaatkan ukuran asosiasi statistik untuk mengidentifikasi kemitraan leksikal yang bermakna dalam miliaran *token*. Perkembangan ini selaras dengan teori bahasa berbasis penggunaan yang lebih luas, yang menekankan sifat probabilistik kompetensi linguistik dan peran paparan berulang dalam membentuk representasi mental. Dalam pengertian ini, *Big Data* tidak menggantikan kerangka kerja teoretis; melainkan, memperkuat fondasi empirisnya dengan memberikan bukti yang lebih kaya.

Perbedaan antara paradigma penelitian berbasis korpus dan yang digerakkan oleh korpus lebih lanjut menggambarkan evolusi teoretis ini. Penelitian berbasis korpus biasanya menguji hipotesis yang sudah ada menggunakan data korpus sebagai bukti, sementara penelitian yang digerakkan oleh korpus memungkinkan pola muncul secara induktif dari kumpulan *Big Data*. Ketersediaan *Big Data* meningkatkan kelayakan pendekatan yang digerakkan oleh korpus, karena kumpulan data skala besar meningkatkan kemungkinan mengungkap

keteraturan linguistik yang sebelumnya tidak diperhatikan. Meskipun demikian, para sarjana memperingatkan bahwa penemuan induktif harus disertai dengan interpretasi teoretis untuk menghindari analisis deskriptif semata yang tanpa kedalaman penjelasan (Casula et al., 2021).

Ketergantungan yang semakin meningkat pada korpus digital skala besar mencerminkan tuntutan yang lebih luas untuk keterwakilan, skalabilitas, dan relevansi interdisipliner dalam kajian kontemporer. Dalam lingkungan komunikasi yang semakin global, para peneliti bahasa harus mempertimbangkan beragam register, varietas, dan konteks sosiobudaya. Korpus kecil dan statis sering kali tidak cukup untuk menangkap kompleksitas penggunaan bahasa Inggris global, genre digital yang muncul, atau praktik *slang* dan wacana yang berkembang pesat. Korpus skala besar memberikan kekuatan statistik yang diperlukan untuk menganalisis variasi di berbagai wilayah, demografi, dan *platform* komunikasi.

Selain itu, munculnya pengambilan keputusan berbasis data telah meluas ke bidang linguistik dan pendidikan bahasa. Para pembuat kebijakan pendidikan, perancang kurikulum, dan penulis buku teks semakin mengandalkan bukti korpus untuk menginformasikan pemilihan kosakata, prioritas tata bahasa, dan pengembangan materi otentik. Pedagogi berbasis korpus memungkinkan pendidik untuk menyelaraskan konten pengajaran dengan pola penggunaan di dunia nyata, sehingga meningkatkan paparan pembelajar terhadap bahasa otentik. Dalam konteks ini, *Big Data* tidak hanya memperkaya penelitian akademis, tetapi juga memperkuat relevansi praktis linguistik korpus dalam pengaturan terapan.

Ekspansi interdisipliner semakin menggarisbawahi alasan untuk merangkul korpus digital skala besar. Dalam studi wacana, kumpulan *Big Data* memfasilitasi analisis pembeding ideologis di berbagai media dan pidato politik.

Dalam studi penerjemahan, korpus paralel memungkinkan perbandingan sistematis strategi penerjemahan antarbahasa. Dalam sociolinguistik, korpus yang berasal dari *web* memberikan wawasan tentang variasi bahasa dalam komunitas daring. Selain itu, integrasi data korpus ke dalam sistem kecerdasan buatan menyoroti hubungan simbiosis antara teori linguistik dan pemodelan komputasional. Kajian terbaru menekankan bahwa korpus skala besar berfungsi sebagai sumber daya dasar untuk melatih model bahasa, mengembangkan sistem evaluasi tulisan otomatis, dan mendukung teknologi komunikasi multibahasa (Caines et al., 2023; Xu et al., 2025).

Meskipun demikian, adopsi metodologi *Big Data* memerlukan refleksi kritis. Representativitas tetap menjadi perhatian utama, terutama ketika data berbasis *web* terlalu mewakili demografi atau konteks komunikatif tertentu. Pertimbangan etis seputar privasi data, persetujuan, dan bias algoritmik memerlukan perhatian yang cermat. Oleh karena itu, ahli linguistik korpus harus menyeimbangkan inovasi teknologi dengan ketelitian metodologis dan tanggung jawab etis, memastikan bahwa kesimpulan analitis didasarkan pada praktik penelitian yang transparan dan dapat direplikasi.

Dengan latar belakang ini, bab ini bertujuan untuk memberikan pemeriksaan komprehensif tentang *Big Data* dan korpus digital skala besar dalam kerangka kerja linguistik korpus yang lebih luas. Secara khusus, bab ini berupaya menjelaskan bagaimana *Big Data* membentuk kembali desain korpus, anotasi, dan strategi analitis. Bab ini mengeksplorasi infrastruktur yang mendukung konstruksi korpus skala besar, termasuk teknik pengumpulan data, sistem penandaan otomatis, dan alat visualisasi. Lebih lanjut, bab ini mengkaji aplikasi pedagogis dan penelitian yang relevan dengan studi bahasa Inggris, menyoroti bagaimana korpus skala besar

menginformasikan penelitian kosakata, analisis wacana, dan inovasi pengajaran bahasa.

Sama pentingnya, bab ini terlibat dalam refleksi kritis tentang tantangan metodologis dan etis. Isu bias pengambilan sampel, keterwakilan data, reproduksibilitas, dan transparansi algoritma dibahas untuk mendorong praktik penelitian yang bertanggung jawab. Dengan mengintegrasikan fondasi teoretis dengan perkembangan teknologi, bab ini bertujuan untuk menunjukkan bahwa *Big Data* tidak menandakan penyimpangan dari prinsip-prinsip empiris linguistik korpus. Sebaliknya, ini mewakili perluasan kemungkinan analitis yang menuntut interpretasi teoretis yang cermat dan pengelolaan etis.

Bagi mahasiswa dan dosen bahasa Inggris, memahami dinamika *Big Data* dan korpus digital sangat penting untuk menavigasi penelitian linguistik kontemporer. Meningkatnya ketersediaan dataset skala besar mengajak para sarjana untuk memikirkan kembali desain penelitian tradisional, menggabungkan alat komputasi, dan mengembangkan kompetensi literasi data. Dengan demikian, linguistik korpus terus berkembang sebagai bidang yang dinamis dan interdisipliner, menjembatani wawasan teoretis dan aplikasi praktis di era komunikasi digital.

Dalam konteks *Big Data* dan korpus digital skala besar, bagian ini mengkaji lima subtopik yang saling terkait yang mewakili teknologi, metode, dan tantangan utama dalam linguistik korpus kontemporer: (A) *Web* sebagai Korpus dan Pemrosesan Data Skala Besar, (B) Pembelajaran Mesin dan Penambangan Teks, (C) Pengambilan Data dari *Web* dan Otomatisasi Pengumpulan Data, (D) Tantangan dalam Analisis Korpus Skala Besar, dan (E) Integrasi Linguistik Korpus dan Kecerdasan Buatan. Setiap subtopik didasarkan secara teoritis dan empiris, dan semuanya dikaitkan kembali dengan tema-tema sentral yang diperkenalkan di Bagian Pendahuluan, yaitu

bagaimana proliferasi teks digital yang luas membentuk kembali metodologi linguistik dan potensi penelitian.

WEB SEBAGAI KORPUS DAN PEMROSESAN DATA SKALA BESAR

Konsep *web* sebagai korpus mewakili salah satu perkembangan paling signifikan dalam penelitian linguistik kontemporer. Berbeda dengan korpus tradisional, kumpulan data yang diseimbangkan dengan cermat dan dikurasi secara manual, pendekatan *web* sebagai korpus memanfaatkan konten yang tersedia untuk umum di *World Wide Web* sebagai sumber daya linguistik yang masif dan terus diperbarui (Karibayeva et al., 2025). Paradigma ini selaras dengan dimensi *Volume* dan *Velocity* dari *Big Data* yang dijelaskan sebelumnya. Pada tahun 2025, perkiraan menunjukkan bahwa *web* yang terindeks berisi lebih dari 60 triliun kata teks di berbagai bahasa, genre, dan modalitas (Sayers et al., 2021; Sun & Wang, 2025).

Analisis *web* sebagai korpus dimulai dengan akuisisi data skala besar diikuti oleh langkah-langkah pra-pemrosesan yang mempersiapkan data mentah untuk analisis linguistik. Alur kerja standar meliputi:

1. Pengumpulan Data: Mengunduh halaman *web* yang relevan dengan kueri atau domain tertentu.
2. Pembersihan dan Penyaringan: Menghapus tag HTML, skrip, elemen non-linguistik, dan duplikat.
3. Tokenisasi dan Normalisasi: Memecah teks menjadi kata-kata, menghilangkan *noise*, dan menstandarisasi ortografi.
4. Anotasi: Menambahkan metadata linguistik seperti tag bagian ucapan atau label semantik.
5. Penyimpanan dan Pengambilan: Mengindeks teks yang diproses dalam basis data yang dapat diskalakan untuk kueri.

Tabel 13.1. Karakteristik Komparatif Korpus Tradisional dan Sumber Daya *Web* sebagai Korpus

Fitur	Korpus Tradisional	<i>Web</i> sebagai Korpus
Ukuran	Jutaan kata	Miliaran–Triliun kata
Sumber	Teks yang dikurasi dan seimbang	Halaman <i>web</i> , forum, konten pengguna
Frekuensi Pembaruan	Statis	Dinamis/Kontinu
Anotasi	Manual/Semi-manual	Otomatis/Dibantu ML
Keterwakilan	Pengambilan sampel terkontrol	Luas tetapi bising
Tantangan Kualitas Data	Rendah	Tinggi

Sumber: Sayers et al (2021)

Di luar ukurannya yang besar, *web* menawarkan keragaman kontekstual yang kaya: interaksi media sosial, blog, ulasan daring, makalah akademis, dan metadata multimedia. Sumber-sumber yang beragam ini memperluas penelitian korpus ke domain yang sebelumnya tidak dapat diakses. Para peneliti telah menggunakan data *web* sebagai korpus untuk menyelidiki variasi register, pola antarbahasa, atau terminologi khusus domain pada skala yang tidak mungkin dilakukan dengan korpus tradisional. Kapasitas ini menjadikan *web* sebagai korpus yang sangat diperlukan untuk penelitian yang mencari representativitas dan validitas ekologis.

PEMBELAJARAN MESIN DAN PENAMBANGAN TEKS

Machine learning (ML) dan penambangan teks telah menjadi dasar analisis korpus skala besar. Di era *Big Data*, daftar frekuensi dan konkordansi klasik tidak cukup untuk mengungkapkan pola semantik yang kompleks, tema laten, atau struktur wacana yang halus. Teknik ML, terutama pembelajaran terawasi dan tidak terawasi, memungkinkan para peneliti untuk

mengekstrak pola yang bermakna dari korpus yang melampaui kapasitas analitik manusia.

Dua aplikasi utama ML dalam linguistik korpus adalah:

1. Pembelajaran Terawasi

Model terawasi dilatih pada data berlabel untuk melakukan tugas klasifikasi misalnya, klasifikasi sentimen, identifikasi register, atau pengenalan tindak tutur. Misalnya, korpus ulasan daring dapat diberi anotasi sebagai positif/netral/negatif, memungkinkan model untuk mempelajari fitur linguistik yang terkait dengan bahasa evaluatif. Algoritma seperti *Support Vector Machines (SVM)*, *Random Forests*, dan *Neural Networks* umumnya digunakan dalam konteks ini (Huang et al., 2023; Xu et al., 2025; Yue et al., 2022)

2. Pembelajaran Tanpa Pengawasan

Pendekatan tanpa pengawasan, seperti pemodelan topik menggunakan *Latent Dirichlet Allocation (LDA)* atau pengelompokan melalui *K-Means*, mengungkapkan struktur tematik laten dalam data tanpa set pelatihan berlabel. Misalnya, model topik yang diterapkan pada jutaan artikel berita dapat mengidentifikasi tema wacana yang persisten seperti perubahan iklim, migrasi, atau kebijakan ekonomi. Teknik penambangan teks memperluas kegunaan ML dengan memungkinkan identifikasi kolokasi, asosiasi semantik, dan bundel leksikal. Penyematan kata (misalnya, *Word2Vec*, *GloVe*, model berbasis BERT) mengubah kata menjadi vektor numerik, menangkap kesamaan semantik dan nuansa kontekstual. Teknik-teknik ini memfasilitasi eksplorasi fenomena linguistik seperti pola metafora, pergeseran semantik dari waktu ke waktu, atau variasi lintas genre.

Yang penting, penelitian terbaru menekankan kerangka kerja hibrida di mana linguistik korpus memberikan teori linguistik sebagai batasan panduan untuk sistem ML, memastikan bahwa outputnya kuat secara statistik dan dapat diinterpretasikan secara linguistik (Sun & Wang, 2025; Xu et al., 2025). Misalnya, model *embedding* yang dilatih pada korpus besar telah digunakan untuk menyelidiki pola tata bahasa berbasis penggunaan, menghubungkan frekuensi dengan prinsip-prinsip linguistik kognitif.

PENGAMBILAN DATA DAN OTOMATISASI PENGUMPULAN DATA WEB

Pengambilan data *web* mengacu pada pengambilan data tekstual secara otomatis dari halaman *web* dan *platform*. Jika diimplementasikan secara etis dan sesuai dengan ketentuan situs, pengambilan data menjadi gerbang untuk membangun korpus yang disesuaikan dalam skala besar. Alat pengambilan data modern menggunakan teknik yang kuat untuk mengelola konten dinamis, *rendering JavaScript*, dan penomoran halaman. Alat dan kerangka kerja populer meliputi:

- ***Scrapy*** (kerangka kerja Python)
- ***Beautiful Soup*** (penguraian HTML)
- ***Selenium*** (otomatisasi peramban)
- ***Requests*** (pustaka HTTP)

Meskipun *web scraping* memungkinkan pengumpulan data dalam skala besar, hal ini menimbulkan pertimbangan etis dan hukum yang penting. Peneliti harus menghormati arahan *robots.txt*, hak cipta, dan norma privasi. Otomatisasi pengumpulan data harus mematuhi pedoman *Institutional Review Board (IRB)*, terutama ketika data komunikasi pribadi (seperti unggahan media sosial) terlibat.

Contoh Kasus:

Sebuah kelompok penelitian melakukan *scraping* terhadap sepuluh *outlet* berita utama untuk periode 2019–2024, mengumpulkan lebih dari 500 juta kata wacana politik. Setelah pembersihan dan tokenisasi, korpus tersebut digunakan untuk menganalisis pergeseran dalam pemingkiaan politik sebelum dan sesudah peristiwa global besar (misalnya, pandemi COVID-19). Ini menggambarkan bagaimana *scraping* otomatis, dikombinasikan dengan *Natural Language Processing* pra-pemrosesan NLP, memungkinkan penelitian temporal berbasis genre dalam skala besar.

TANTANGAN DALAM ANALISIS KORPUS SKALA BESAR

Meskipun prospek *Big Data* dalam linguistik sangat besar, analisis korpus skala besar bukannya tanpa tantangan yang signifikan. Tantangan ini terbagi dalam tiga kategori utama:

a. Kualitas dan Representativitas Data

Korpus yang berasal dari *web* cenderung bising. Teks mungkin mengandung kesalahan, format yang tidak konsisten, fragmen kode, atau konten yang tidak relevan. Selain itu, konten *web* cenderung condong ke bahasa tertentu (terutama bahasa Inggris), demografi, dan wilayah geografis, yang menyebabkan sampel yang bias.

b. Keterbatasan Komputasi dan Penyimpanan

Memproses kumpulan data miliaran atau triliunan kata membutuhkan komputasi berkinerja tinggi dan arsitektur penyimpanan yang dapat diskalakan. Peneliti harus menyeimbangkan kompleksitas model dengan kelayakan komputasi, seringkali mengandalkan solusi komputasi awan.

c. Isu Etika dan Hukum

Pengumpulan dan penggunaan teks *web* menimbulkan kekhawatiran tentang privasi pengguna, persetujuan, dan

kekayaan intelektual. Korpus skala besar seringkali mencakup konten yang dihasilkan pengguna yang tidak dirancang untuk penelitian, sehingga memerlukan tata kelola etika yang cermat.

Terlepas dari masalah-masalah ini, inovasi metodologis seperti anonimisasi data, pengambilan sampel bertingkat, dan deteksi bias algoritmik memungkinkan ahli bahasa korpus untuk mengurangi risiko dan meningkatkan keandalan analitik.

INTEGRASI LINGUISTIK KORPUS DAN KECERDASAN BUATAN

Mungkin perkembangan paling transformatif dalam beberapa tahun terakhir adalah integrasi linguistik korpus dengan *Artificial Intelligence (AI)*. Sistem AI, terutama model bahasa berbasis transformer seperti GPT-4, BERT, atau T5, dilatih pada korpus besar dan merangkum pola penggunaan bahasa dengan cara yang mendukung generasi, inferensi, dan prediksi. Hubungan antara linguistik korpus dan AI bersifat dua arah: korpus besar melatih model AI, dan model AI menyediakan alat untuk analisis korpus tingkat lanjut.

Misalnya, penggunaan AI untuk anotasi semantik memungkinkan peneliti untuk memberi tag pada teks dengan kategori makna yang terperinci dalam skala besar, mendukung analisis wacana atau pelacakan sentimen dalam jutaan dokumen. Demikian pula, alat visualisasi berbasis AI membantu memetakan jaringan linguistik, perubahan semantik dari waktu ke waktu, atau variasi lintas genre.

Integrasi ini juga menimbulkan pertanyaan epistemologis kritis tentang interpretasi dan akuntabilitas. Model AI sering berfungsi sebagai kotak hitam, yang berarti proses internalnya tidak selalu transparan. Oleh karena itu, para ahli linguistik korpus harus mengembangkan metodologi yang

menggabungkan kekuatan prediktif AI dengan teori linguistik untuk memastikan bahwa wawasan yang diperoleh valid dan bermakna.

Evolusi pesat metodologi *Web-as-Corpus*, pembelajaran mesin, pengumpulan data otomatis, dan integrasi AI mencerminkan transformasi berkelanjutan linguistik korpus di era *Big Data*. Meskipun alat dan metode ini memperluas potensi analitik, mereka juga membutuhkan ketelitian teoretis dan kesadaran etis yang lebih tinggi dari para peneliti.

Tabel 13.2. Ringkasan Teknologi dan Fungsi Utamanya

Teknologi / Metode	Fungsi Utama	Contoh Kasus Penggunaan
Web sebagai Korpus	Sumber teks ma	Penelitian variasi genre
Pembelajaran Mesin	Penemuan pola	Pemodelan topik
Pengikisan Web	Pengumpulan data otomatis	Korpus berita khusus
Integrasi AI	Anotasi semantik	Klasifikasi wacana

Sumber: Xu et al (2025)

KESIMPULAN

Munculnya *Big Data* telah secara fundamental membentuk kembali lanskap linguistik korpus, mengubahnya dari disiplin ilmu yang bergantung pada koleksi teks yang relatif kecil dan statis menjadi ekosistem dinamis yang digerakkan oleh komputasi, yang dicirikan oleh skala, kecepatan, dan kompleksitas. Transformasi ini memungkinkan para peneliti untuk melampaui sampel terbatas menuju korpus digital yang luas yang menawarkan kedalaman analitis dan ketelitian metodologis yang belum pernah terjadi sebelumnya. Akibatnya, linguistik korpus kini menempati posisi yang lebih menonjol dalam penelitian interdisipliner, beririsan dengan bidang-bidang seperti linguistik komputasional, ilmu data, dan kecerdasan buatan. Integrasi korpus skala besar tidak hanya

menyempurnakan analisis linguistik, tetapi juga meningkatkan keandalan dan generalisasi temuan, memungkinkan para sarjana untuk mengamati penggunaan bahasa otentik di berbagai konteks, genre, dan populasi.

Implikasi dari perkembangan ini bagi penelitian dan pedagogi bahasa Inggris sangat besar. Ketersediaan data bahasa yang ekstensif dan otentik memberi pendidik dan peneliti peluang yang lebih kaya untuk mengeksplorasi pola bahasa dunia nyata, fitur wacana, dan tren linguistik yang berkembang. Hal ini, pada gilirannya, mendukung pendekatan yang lebih berbasis bukti untuk pengajaran bahasa dan desain kurikulum. Lebih lanjut, semakin pentingnya penyelidikan berbasis data menuntut pengembangan literasi data di kalangan mahasiswa dan akademisi, membekali mereka dengan keterampilan yang dibutuhkan untuk menafsirkan, menganalisis, dan mengevaluasi secara kritis kumpulan data besar. Oleh karena itu, penggabungan alat korpus ke dalam kurikulum pendidikan tinggi tidak hanya bermanfaat tetapi juga penting, karena mendorong kompetensi analitis dan mempersiapkan peserta didik untuk berpartisipasi dalam lingkungan akademik dan profesional yang semakin intensif data.

Pada saat yang sama, perluasan *Big Data* dalam linguistik korpus menuntut refleksi metodologis dan etis yang cermat. Meskipun korpus besar menawarkan keuntungan yang signifikan, masalah keterwakilan dan bias tetap menjadi perhatian kritis. Peneliti harus tetap waspada dalam menilai apakah data secara akurat mencerminkan komunitas linguistik dan konteks yang sedang diselidiki, karena ketidakseimbangan dalam komposisi korpus dapat menyebabkan interpretasi yang menyimpang. Selain itu, penggunaan data berbasis *web* menimbulkan tantangan etis terkait privasi, persetujuan, dan kepemilikan data, terutama ketika konten yang dihasilkan pengguna terlibat. Interpretasi data yang bertanggung jawab

sama pentingnya, karena volume data yang besar tidak secara inheren menjamin validitas; sebaliknya, hal ini membutuhkan kerangka analitis yang ketat dan kesadaran kritis untuk memastikan kesimpulan yang bermakna dan etis.

Ke depan, masa depan linguistik korpus kemungkinan akan dibentuk oleh inovasi teknologi yang berkelanjutan dan integrasi interdisipliner. Peningkatan penggunaan analisis korpus yang dibantu AI menjanjikan peningkatan efisiensi dan kecanggihan penelitian linguistik, memungkinkan pengenalan pola otomatis, analisis semantik, dan pemodelan prediktif. Selain itu, pengembangan korpus besar multimodal dan multibahasa akan memperluas cakupan analisis di luar teks tertulis untuk mencakup bahasa lisan, komunikasi visual, dan perbandingan lintas bahasa. Arah menjanjikan lainnya terletak pada integrasi linguistik korpus dengan pembelajaran analitik, yang berpotensi menjembatani kesenjangan antara penelitian linguistik dan praktik pendidikan dengan memberikan wawasan tentang perilaku pembelajar, proses pemerolehan bahasa, dan efektivitas pengajaran. Secara kolektif, kemajuan ini menunjukkan bahwa linguistik korpus akan terus berkembang sebagai bidang yang vital dan inovatif, berkontribusi tidak hanya pada teori linguistik tetapi juga pada aplikasi praktis dalam pendidikan dan seterusnya.

DAFTAR PUSTAKA

- Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, Ø., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology. *ArXiv Preprint ArXiv:2307.08393*, 25.
- Casula, M., Rangarajan, N., & Shields, P. (2021). The potential of working hypotheses for deductive exploratory research.

- Quality & Quantity*, 55(3), 1703–1725.
<https://doi.org/10.1007/s11135-020-01072-9>
- Huang, F., Xiong, H., Chen, S., Lv, Z., Huang, J., Chang, Z., & Catani, F. (2023). Slope stability prediction based on a long short - term memory neural network : comparisons with convolutional neural networks, support vector machines, and random forest models. *International Journal of Coal Science & Technology*, 4, 14.
<https://doi.org/10.1007/s40789-023-00579-4>
- Karibayeva, A., Karyukin, V., Tukeyev, U., Abduali, B., Amirova, D., Rakhimova, D., & Shormakova, A. (2025). The development and experimental evaluation of a multilingual speech corpus for low-resource Turkic languages. *Applied Sciences*, 15(24).
- Paquot, M., & Gries, S. T. (Eds.) (2021). *A practical handbook of corpus linguistics* (first). Springer Nature Singapore Pte Ltd. <https://doi.org/10.1007/978-3-030-46216-1>
- Sayers, D., Sousa-silva, R., Höhn, S., Ahmedi, L., Allkivi-metsoja, K., Anastasiou, D., Beňuš, Š., Bowker, L., Bytyçi, E., & Catala, A. (2021). *The Dawn of the human-machine era : A forecast of new and emerging language technologies*. University of Jyväskylä.
- Solovyev, V., Solnyshkina, M., & Daniella McNamara. (2022). Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics*, 26(2), 275–316.
<https://doi.org/10.22363/2687-0088-30161>
- Sukhrob Saidov Samiyevich. (2024). Theoretical approaches to the study of lingua-cultures based on corpus analysis. *Best Journal of Innovation in Science, Research and Development*, 03(6), 365–370.
<http://eprints.umsida.ac.id/13833/>
- Sun, K., & Wang, R. (2025). *Systematic framework of*

application methods for large language Models in language sciences. 1(12), 1–62.

Xu, Y., Hu, L., Zhao, J., Qiu, Z., Xu, K., Ye, Y., & Gu, H. (2025). A survey on multilingual large language models : corpora, alignment, and bias. *Frontiers of Computer Science, 19(11), 25.* <https://doi.org/10.1007/s11704-024-40579-4>

Yue, S., Li, S., Huang, X., Liu, J., Hou, X., Zhao, Y., Niu, D., Wang, Y., Tan, W., & Wu, J. (2022). Machine learning for the prediction of acute kidney injury in patients with sepsis. *Journal of Translational Medicine, 1–12.* <https://doi.org/10.1186/s12967-022-03364-0>

PROFIL PENULIS



Abd. Hakim Muslim

Penulis menyelesaikan studi Magister Pendidikan Bahasa Inggris di Universitas Negeri Malang dan kemudian memperoleh gelar Doktor Teknologi Pendidikan dari institusi yang sama pada tahun 2024. Saat ini, ia bekerja sebagai ASN di MAN 5 Jombang sekaligus mengelola lembaga bahasa Inggris non-formal, *Living English Course*, di Ngoro, Jombang, di mana ia tetap aktif terlibat dalam pendidikan.

BAB 14

MASA DEPAN LINGUISTIK KORPUS

Ika Apriani Fata
Universitas Syiah Kuala, Banda Aceh
E-mail: ika.apriani@usk.ac.id

PENDAHULUAN

Masa depan linguistik korpus berada pada titik krusial di mana integrasi metodologi berbasis komputer bertemu dengan kecanggihan kecerdasan buatan (AI) untuk menciptakan paradigma analisis bahasa yang lebih dinamis dan akurat (Chen & Flowerdew, 2023). Sebagai disiplin ilmu yang mempelajari bahasa melalui bukti empiris dari data teks autentik yang besar, linguistik korpus kini bertransformasi dari sekadar alat statistik frekuensi menjadi kerangka kerja yang mampu membedah kompleksitas penggunaan bahasa dalam berbagai konteks sosial (Wang & Zhang 2025). Sebagaimana ditegaskan oleh penelitian terkini, kemajuan teknologi digital tidak hanya mempermudah akses terhadap sumber data bahasa, tetapi juga menuntut peneliti untuk mengembangkan prosedur analisis yang lebih interaktif dan otomatis, guna mengimbangi dinamika bahasa yang terus berkembang di ruang digital (Floridi, 2019; Wibowo & Sari, 2025). Oleh karena itu, masa depan linguistik korpus tidak hanya bergantung pada akumulasi data, melainkan pada kemampuan integrasi teknologi dalam menafsirkan makna bahasa dengan kedalaman yang lebih kontekstual.

Dalam ranah aplikasi praktis, perkembangan linguistik korpus diproyeksikan akan memberikan kontribusi signifikan terhadap revolusi pengajaran bahasa dan penyusunan kamus modern melalui konsep *living dictionary*. Penggunaan alat bantu berbasis korpus, seperti aplikasi KORTARA, telah

membuktikan bagaimana peneliti dan praktisi di berbagai negara mampu menyelesaikan tantangan analisis data linguistik berskala besar dengan lebih efektif dan efisien (Gries, 2024). Integrasi AI dalam linguistik terapan, termasuk dalam analisis korpus, kini membuka peluang baru bagi peneliti untuk mengeksplorasi nuansa bahasa yang sebelumnya sulit diidentifikasi melalui metode tradisional, seperti perubahan semantik dan pola wacana yang sangat dinamis. Hal ini menegaskan bahwa masa depan linguistik korpus terletak pada sinergi antara ketelitian data empiris dan kecerdasan komputasional yang mampu menjawab tuntutan zaman.

Lebih jauh, arah penelitian linguistik korpus di masa depan akan semakin berfokus pada eksplorasi multifungsionalitas bahasa dalam literasi akademik dan media, yang menyediakan referensi empiris berharga bagi pengembangan kurikulum serta kebijakan Bahasa (Hadi, & Pratama, 2025). Dengan memanfaatkan *big data* bahasa dari media lokal hingga nasional, para peneliti dapat memetakan proyeksi perubahan bahasa secara lebih presisi, yang pada gilirannya akan memperkuat basis teoretis dalam kajian linguistik terapan. Tantangan utama bagi praktisi di masa depan adalah menjaga objektivitas dan kualitas validasi manusia atas interpretasi yang dihasilkan oleh sistem otomatis, guna memastikan bahwa hasil analisis korpus tetap relevan dan berintegritas (Kartika & Syahputra, 2022). Dengan demikian, linguistik korpus akan terus menjadi instrumen utama yang mendefinisikan cara kita memahami, mengajarkan, dan mengaplikasikan bahasa di masa depan yang semakin digital.

PERKEMBANGAN TEKNOLOGI DAN DATA BAHASA

Perkembangan teknologi informasi dan komunikasi telah mengakibatkan transformasi masif dalam pola penggunaan bahasa, di mana interaksi digital kini menjadi ruang utama bagi

lahirnya kosakata baru serta perubahan gaya komunikasi yang dinamis. Integrasi kecerdasan buatan (*Artificial Intelligence*) dalam analisis bahasa, khususnya melalui pemanfaatan *Large Language Models* (LLM), telah memungkinkan pemrosesan dataset teks berskala besar secara otomatis dengan akurasi yang semakin meningkat (Kurniawan, 2023). Teknologi ini tidak hanya memfasilitasi pengenalan pola linguistik yang kompleks, tetapi juga membuka potensi besar dalam pengembangan alat bantu pembelajaran bahasa interaktif yang mampu memberikan umpan balik *real-time* kepada pengguna (Kusuma, 2022).

Data bahasa autentik kini dikelola melalui kerangka kerja linguistik korpus digital, yang memungkinkan peneliti untuk memetakan perubahan bahasa dengan landasan empiris yang jauh lebih representatif (Marhaban, & Fata, 2025). Sinergi antara ketelitian data korpus dan kemampuan komputasional AI terbukti efektif dalam memecahkan tantangan analisis sintaksis, meskipun identifikasi elemen bahasa idiomatik dan struktur kalimat kompleks masih memerlukan validasi ahli manusia (McEnery & Hardie, 2025). Dengan terus berkembangnya teknologi ini, masa depan linguistik terapan akan semakin bergantung pada integrasi harmonis antara pengolahan data masif dan analisis kontekstual yang mendalam untuk memastikan integritas serta relevansi bahasa di era digital.

KORPUS MULTIMODAL DAN STANDARISASI ETIKA PEMANFAATAN DATA

Korpus multimodal yang mengintegrasikan teks, audio, dan visual kini menjadi fondasi utama dalam pengembangan sistem kecerdasan buatan (AI) yang lebih canggih dan kontekstual (Pratama, 2026). Dengan menghubungkan representasi numerik dari berbagai modalitas melalui proses *cross-modal alignment*, sistem AI dapat memahami interaksi manusia dengan cara yang lebih alami, seperti bagaimana manusia menggunakan

penglihatan, pendengaran, dan bahasa secara bersamaan untuk berpikir. Inovasi ini memungkinkan pengembangan anotasi linguistik multilapis yang tidak hanya menangkap aspek verbal, tetapi juga elemen nonverbal seperti ekspresi wajah, intonasi, dan alur visual dalam komunikasi (Pratama & Santoso, 2025).

Standardisasi dalam pengumpulan dan pengolahan data multimodal menjadi tantangan krusial untuk memastikan interoperabilitas serta kualitas anotasi yang konsisten (Purwadi & Dewi, 2025). Proses normalisasi data yang cermat, seperti sinkronisasi antara transkrip teks dan elemen audiovisual, sangat diperlukan agar model AI dapat belajar dari konteks yang akurat tanpa mengabaikan nuansa pragmatik. Tanpa standardisasi yang ketat, keragaman format data dari berbagai sumber dapat menimbulkan hambatan teknis yang mengurangi efektivitas hasil analisis dalam aplikasi praktis.

Aspek etika pemanfaatan data dalam riset multimodal menuntut kewaspadaan tinggi, terutama terkait risiko bias sosio-kultural yang tertanam dalam dataset skala besar (Putra & Sari, 2023). Model AI yang dilatih dengan data dari dunia nyata sering kali mencerminkan prasangka yang ada, sehingga diperlukan mekanisme pemeriksaan cermat, pra-pemrosesan data, serta evaluasi pasca-pelatihan untuk memitigasi diskriminasi (Rahayu, 2023). Peneliti diwajibkan untuk mengadopsi prinsip non-intrusif dan memastikan bahwa privasi subjek penelitian terlindungi, terutama ketika data yang digunakan mengandung informasi sensitif atau berisiko merusak reputasi jika dipublikasikan (Rahmawati & Setiawan, 2025).

Tantangan etis lainnya muncul seiring dengan meningkatnya risiko manipulasi informasi dan halusinasi konten yang dihasilkan oleh sistem AI. Oleh karena itu, profesional di bidang linguistik dan teknologi bahasa harus mengacu pada standar etika yang ketat, sekaligus meningkatkan literasi digital untuk mencegah penyalahgunaan data (Saidu dkk.,2025).

Integrasi antara kontrol manusia dan otomatisasi AI sangat diperlukan untuk menjamin bahwa penggunaan korpus multimodal tetap objektif, berintegritas, serta memberikan manfaat yang etis bagi pengembangan ilmu pengetahuan dan masyarakat.

PROSPEK RISET DAN KOLABORASI GLOBAL

Prospek riset linguistik korpus kini semakin terbuka luas melalui kolaborasi global yang terintegrasi, didorong oleh kebutuhan akan pemahaman bahasa lintas budaya yang lebih objektif dan berbasis data besar (*big data*) (Santoso, 2023). Kolaborasi internasional tidak hanya memperkaya keragaman dataset, tetapi juga memungkinkan pertukaran metodologi canggih—seperti integrasi *deep learning* dan *natural language processing* (NLP)—untuk meneliti fenomena linguistik yang kompleks (Setiawan, 2025). Contoh nyata dari prospek ini adalah proyek riset bersama antara Universitas Sebelas Maret (UNS) dan *University of Sydney*, yang menggunakan pendekatan korpus untuk meningkatkan efektivitas pengajaran bahasa Inggris dan penelitian linguistik terapan secara komprehensif.

Selain itu, kolaborasi lintas bangsa dalam bidang ini sering kali dimediasi oleh teknologi penerjemahan otomatis dan platform digital, yang memecahkan hambatan komunikasi antarpeneliti dari berbagai negara (Suhaili, 2024; Wahyuni, 2023). Inisiatif seperti program pengabdian internasional antara Universitas Negeri Jakarta (UNJ) dan *Kanda University of International Studies* (Jepang) menjadi contoh keberhasilan dalam berbagi wawasan tentang penggunaan *corpus tools* untuk meningkatkan literasi digital dan kualitas publikasi bersama di jurnal internasional (Tognini-Bonelli & Sinclair, 2024). Proyek-proyek semacam ini menunjukkan bahwa kolaborasi global berperan sebagai katalis bagi lahirnya inovasi dalam analisis

wacana kritis berbasis korpus (*Corpus-Assisted Discourse Studies* - CADS), yang kini banyak digunakan untuk memetakan dinamika diskursus keagamaan dan politik global.

KESIMPULAN

Masa depan linguistik korpus berada pada titik transformasi fundamental, di mana perpaduan antara metodologi empiris tradisional dan kecanggihan kecerdasan buatan membuka cakrawala baru dalam pemahaman bahasa yang lebih dinamis. Integrasi teknologi AI, seperti *Large Language Models* dan pemrosesan data multimodal, telah mengubah korpus dari sekadar gudang data statis menjadi ekosistem analisis cerdas yang mampu membedah nuansa semantik serta pola komunikatif yang kompleks secara *real-time*. Evolusi ini menegaskan bahwa ke depan, penguasaan atas data berbasis teknologi akan menjadi fondasi utama bagi setiap riset linguistik yang ingin tetap relevan di era digital yang sangat volatil.

Peluang riset dan kolaborasi global kini menjadi instrumen strategis yang tidak terelakkan bagi para praktisi linguistik untuk memperkaya cakupan studi dan meningkatkan validitas temuan melalui dataset lintas budaya yang lebih luas. Melalui kemitraan internasional, para peneliti dapat berbagi metodologi mutakhir dan infrastruktur komputasional yang memungkinkan pemetaan tren bahasa secara presisi, baik dalam diskursus akademik maupun media. Sinergi ini tidak hanya memperkuat posisi linguistik korpus sebagai disiplin ilmu yang berpengaruh, tetapi juga mendorong standardisasi praktik penelitian yang lebih berintegritas dan transparan di tingkat global.

Sebagai penutup, tantangan besar yang menyertai perkembangan ini—terutama terkait etika pemanfaatan data dan validasi manusia terhadap hasil sistem otomatis—harus dijawab dengan penguatan literasi digital dan komitmen etis yang teguh. Linguistik korpus di masa depan dituntut untuk tetap

mempertahankan objektivitas manusia di tengah arus otomatisasi yang dominan, memastikan bahwa setiap temuan yang dihasilkan mampu memberikan kontribusi yang etis dan aplikatif. Dengan demikian, linguistik korpus akan terus menjadi instrumen krusial yang mendefinisikan cara kita memahami, mengajarkan, serta menjaga integritas komunikasi dalam masyarakat modern yang semakin saling terhubung.

DAFTAR PUSTAKA

- Chen, Y., & Flowerdew, J. (2023). *Corpus linguistics and language pedagogy in a global context*. Routledge.
- Floridi, L. (2019). *The ethics of digital objects*. Oxford University Press: Oxford.
- Gries, S. T. (2024). *Advances in corpus-based approaches to linguistic research*. Cambridge University Press.
- Hadi, M. S., & Pratama, A. (2025). Analisis linguistik korpus pada media lokal dan nasional: Pendekatan empiris. *Ghancaran: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 7(2), 88–104. <https://ejournal.uinmadura.ac.id/index.php/ghancaran/article/view/21652>
- Kartika, D., & Syahputra, I. (2022). Peran linguistik forensik terhadap penegakan hukum Indonesia: Studi analisis ujaran kebencian suku Lampung. *Tadabbur: Jurnal Peradaban Islam*, 6(1), 45–60. <https://journal2.uinjkt.ac.id/index.php/tadabbur/article/view/27874>
- Kurniawan, F. (2023). *Linguistik korpus: Aplikasi digital untuk kajian dan pembelajaran bahasa modern*. Penerbit Akademik Indonesia.
- Kusuma, A. B. (2022). Karakterisasi ujaran negatif berdampak hukum di Indonesia: Kajian linguistik forensik. *Prosiding Semantiks*, 2(1), 210–225.

<https://jurnal.uns.ac.id/prosidingsemantiks/article/view/653>
25

- Marhaban, S. & Fata, I.A. (2025). Investigating EFL Learners' Perceptions and Practices Using the Indonesia's Prepaid Mobile Application Ruangguru. *Al-Ta lim Journal*. 32 (5). 133-150.
- McEnery, T., & Hardie, A. (2025). *The future of corpus linguistics: Bridging technology and society*. Palgrave Macmillan.
- Pratama, A. (2026). Komunikasi digital dan representasi nilai dalam pendidikan akademik. *Al-Miyar: Jurnal Ilmiah Pembelajaran Bahasa Arab*, 9(1), 112–129. <https://jurnal.stiq-amuntai.ac.id/index.php/al-miyar/article/view/6622>
- Pratama, R., & Santoso, A. (2025). Pemanfaatan model AI dalam analisis sintaksis bahasa Indonesia: Kajian linguistik komputasional. *Jurnal Progresif*, 14(1), 55–70. <https://journal.univgresik.ac.id/index.php/progresif/article/view/443>
- Purwadi, P., & Dewi, M. I. N. (2025). Penggunaan komputasional linguistik dalam etika profesi berkebangsaan. *Jurnal Kreatif RI*, 3(1), 65–78. <https://doi.org/10.31848/jkri.v3i1.4377>
- Putra, D. A., & Sari, N. (2023). Pemanfaatan kecerdasan buatan (AI) dalam linguistik terapan: Peluang dan tantangan. *Jurnal Linguistik Komputasional*, 5(1), 12–25. <https://ejournal.unsrat.ac.id/v3/index.php/kaling/article/download/62474/50858/164106>
- Rahayu, S. (2023). Meningkatkan pengajaran bahasa dan literasi dengan korpus. *Universitas Airlangga News*. <https://unair.ac.id/meningkatkan-pengajaran-bahasa-dan-literasi-dengan-korpus-dan-linguistik-korpus>

- Rahmawati, I., & Setiawan, B. (2025). Penggunaan komputasional linguistik dalam etika digital. *Jurnal Kreatif RI*, 12(2), 45–60. <https://e-journal.ukri.ac.id/index.php/jkri/article/download/4377/1212>
- Saidu,A. Dahiru,S. Fata, I.A. (2025). Text, Image and Colour: A Multimodal Reading of Selected Save The Children Campaign Posters. *Studies in Pragmatics and Discourse Analysis*, 6 (1). 45-60
- Santoso, B. (2023). Analisis keterbatasan korpus bahasa dan pemanfaatan AI dalam penelitian terkini. *Jurnal Pendidikan Tambusai*, 7(2), 1120–1135. <https://jptam.org/index.php/jptam/article/download/34144/21918/57013>
- Setiawan, B. (2025). *Literasi digital dalam ekosistem profesional*. Tahta Media Group. <https://tahtamedia.co.id/index.php/issj/article/download/1449/1440/5196>
- Suhaili, A. (2024). Analisis pola diskursus keagamaan: Pendekatan berbasis korpus pada publikasi internasional. *Journal of Applied Linguistics and Professional Practice*, 18(2), 112–129.
- Tognini-Bonelli, E., & Sinclair, J. (2024). *Corpus-driven discourse analysis: Methodological frameworks for global collaboration*. John Benjamins Publishing.
- Wahyuni, S. (2023). Pengantar linguistik korpus dalam pendidikan bahasa. *Lectura: Jurnal Pendidikan*, 14(1), 45–58. <https://journal.unilak.ac.id/index.php/lectura/article/download/11705/4733>
- Wang, X., & Zhang, L. (2025). Integration of deep learning in corpus-based studies: A cross-country analysis.

International Journal of Computational Linguistics, 12(1), 45–62.

Wibowo, S., & Sari, D. (2025). Peran teknologi informasi dalam memperkuat integritas profesional. *Jurnal Riset Teknik Komputer*, 10(2), 85–100.
<https://journal.smartpublisher.id/index.php/jurtikom/article/download/1383/1092>

PROFIL PENULIS



Ika Apriani Fata

Penulis adalah dosen di Departemen Pendidikan Bahasa Inggris, Fakultas Keguruan dan Ilmu Pendidikan (FKIP), Universitas Syiah Kuala. Minat akademiknya meliputi linguistik forensik, pendidikan bahasa Inggris, studi bahasa dan sastra Inggris, linguistik terapan, pedagogi, dan pengajaran bahasa Inggris sebagai bahasa asing. Penulis telah menerbitkan paper di jurnal internasional terkemuka dan menulis buku-buku tingkat universitas. Penulis juga menjadi editor dan mitra bestari di beberapa jurnal SCOPUS, Sinta 1, Sinta 2. Penulis menjadi narasumber dalam bidang linguistik forensik. Penulis dapat dihubungi melalui Instagram dan TikTok di @ika_fata

DESKRIPSI BUKU

Buku “Linguistik Korpus: Alat dan Aplikasi untuk Penelitian Bahasa” ini memberikan pembahasan komprehensif mengenai teori, metode, dan penerapan linguistik korpus dalam analisis bahasa modern. Buku ini terdiri dari 14 bab yang tersusun secara sistematis, dimulai dari konsep dasar hingga pemanfaatan korpus untuk kajian bahasa tingkat lanjut. Bagian awal buku ini membahas definisi korpus, karakteristiknya, berbagai jenis korpus, serta teknik pengumpulan dan penyusunan data korpus. Selanjutnya, pembahasan diarahkan pada teknik penandaan data, penggunaan perangkat lunak seperti AntConc dan Sketch Engine, serta berbagai metode analisis, seperti frekuensi, kolokasi, konkordansi, dan keyness. Buku ini juga menampilkan beragam penerapan linguistik korpus, termasuk:

- Pengajaran bahasa berbasis korpus
- Penyusunan kamus dan glosarium
- Analisis wacana dan media
- Variasi bahasa dalam masyarakat
- Kajian terjemahan dan perbandingan bahasa

Selain itu, buku ini merespons perkembangan mutakhir seperti:

- Pemanfaatan Big Data dalam linguistik
- Integrasi linguistik korpus dengan machine learning dan AI
- Peran korpus dalam penelitian berbasis model bahasa (LLM) seperti ChatGPT

Melalui contoh studi kasus, langkah-langkah operasional analisis, dan pembahasan metodologis yang jelas, buku ini menjadi panduan yang mudah dipahami baik oleh pemula maupun peneliti yang ingin memperdalam pendekatan korpus dalam penelitian bahasa.



FUTURE SCIENCE
Perum Sarimadu II B3 No.03, Kec. Pakisaji,
Kab. Malang, Jawa Timur, Indonesia 65162
Website : www.futuresciencepress.com



IKAPI
INSTITUT KAJIAN
No. 348/JTI/2022

ISBN 978-634-7592-70-5 (PDF)



9

786347

592705