Sanskara Ilmu Sosial dan Humaniora
Vol. 2, No. 03, Agustus 2025, hal. 177-185
ISSN: 3031-7789, DOI: 10.58812/sish.v2.i03

Pemetaan Bibliometrik Penelitian Global tentang *Corpus Linguistics*

Loso Judijanto
IPOSS Jakarta, losojudijantobumn@gmail.com

Info Artikel

Article history:

Received Agu, 2025
Revised Agu, 2025
Accepted Agu, 2025

Kata Kunci:

Analisis Bibliometrik; *Deep Learning*; Linguistik Korpus; Pemrosesan Bahasa Alami; Tren Penelitian

Keywords:

Bibliometric Analysis; *Corpus Linguistics*; *Deep Learning*; *Natural Language Processing*; *Research Trends*

ABSTRAK

Penelitian ini bertujuan memetakan lanskap penelitian global di bidang *corpus linguistics* dengan menggunakan pendekatan bibliometrik berbasis data dari basis data Scopus. Analisis dilakukan terhadap publikasi periode 2000–2025 dengan fokus pada tren publikasi, kolaborasi internasional, kata kunci dominan, dan keterkaitan tematik. Data dianalisis menggunakan perangkat lunak VOSviewer dan Bibliometrix untuk menghasilkan visualisasi jejaring *co-authorship*, *co-occurrence*, dan *co-citation*. Hasil penelitian menunjukkan bahwa *corpus linguistics* berada pada pusat interaksi antara linguistik tradisional dan teknologi komputasional, dengan keterkaitan kuat terhadap *natural language processing systems*, *semantics*, *deep learning*, dan *large language models*. Selain itu, topik seperti *low resource languages*, *contrastive learning*, dan *speech recognition* muncul sebagai bidang yang tengah berkembang dan berpotensi menjadi fokus utama riset di masa depan. Temuan ini memberikan implikasi praktis bagi peneliti, industri teknologi bahasa, dan pembuat kebijakan, serta kontribusi teoritis dalam memperluas pemahaman *corpus linguistics* sebagai bidang interdisipliner yang dinamis.

ABSTRACT

This study aims to map the global research landscape in corpus linguistics using a bibliometric approach based on data from the Scopus database. The analysis was conducted on publications for the period 2000-2025 with a focus on publication trends, international collaboration, dominant keywords, and thematic linkages. Data were analyzed using VOSviewer and Bibliometrix software to produce visualizations of co-authorship, co-occurrence, and co-citation networks. The results show that corpus linguistics is at the center of the interaction between traditional linguistics and computational technology, with strong links to natural language processing systems, semantics, deep learning, and large language models. In addition, topics such as low resource languages, contrastive learning, and speech recognition emerge as emerging areas that have the potential to be a major focus of future research. The findings provide practical implications for researchers, the language technology industry, and policy makers, as well as theoretical contributions in expanding the understanding of corpus linguistics as a dynamic interdisciplinary field.

Corresponding Author:

Name: Loso Judijanto
Institution: IPOSS Jakarta
Email: losojudijantobumn@gmail.com

1. PENDAHULUAN

Perkembangan teknologi digital telah mengubah lanskap penelitian linguistik secara signifikan dalam beberapa dekade terakhir. Salah satu bidang yang mendapatkan manfaat terbesar dari kemajuan ini adalah *corpus linguistics*, yakni pendekatan linguistik yang berfokus pada analisis data bahasa dalam bentuk korpora atau kumpulan teks besar yang dapat diproses secara elektronik. Sejak awal kemunculannya pada pertengahan abad ke-20, *corpus linguistics* berkembang pesat berkat meningkatnya kapasitas komputasi dan meluasnya ketersediaan data digital. Berbagai sumber data seperti berita daring, media sosial, dokumen resmi, hingga karya sastra kini dapat diakses dan dianalisis secara otomatis, memungkinkan peneliti untuk mengidentifikasi pola bahasa, frekuensi kata, kolokasi, dan fenomena linguistik lain dengan skala yang jauh lebih besar dibandingkan metode tradisional (Gries, 2009; McEnery, 2019).

Seiring waktu, *corpus linguistics* tidak hanya menjadi ranah kajian murni (*pure linguistics*), tetapi juga berkembang menjadi alat penting dalam *applied linguistics*, pengajaran bahasa, terjemahan, leksikografi, *natural language processing* (NLP), dan bahkan analisis budaya melalui bahasa. Misalnya, dalam pengajaran bahasa Inggris sebagai bahasa kedua (EFL/ESL), penggunaan korpora telah membantu guru dan peneliti memahami penggunaan kosakata dan struktur kalimat yang umum ditemui dalam bahasa autentik (Kennedy, 2014; Leech, 2014). Di ranah leksikografi, kamus modern seperti *Collins COBUILD* atau *Oxford English Corpus* sangat bergantung pada analisis korpus untuk memastikan keakuratan entri kata dan contoh penggunaannya. Penerapan ini membuktikan bahwa *corpus linguistics* memiliki relevansi yang luas lintas bidang dan disiplin, serta dapat menjadi dasar kebijakan bahasa yang berbasis bukti.

Dalam konteks akademik global, perkembangan *corpus linguistics* menunjukkan pola yang kompleks dan dinamis. Pertama, bidang ini bersifat sangat interdisipliner, melibatkan kolaborasi antara ahli bahasa, ilmuwan komputer, pendidik, dan bahkan pakar ilmu sosial. Kedua, penyebaran penelitian ini bersifat internasional, dengan pusat-pusat riset di Eropa, Amerika Utara, dan Asia yang saling berinteraksi dalam menghasilkan pengetahuan baru. Namun, meskipun terdapat banyak publikasi di jurnal bereputasi tinggi, distribusi tema, tren topik, pola kolaborasi, serta peta penyebaran geografis penelitian *corpus linguistics* belum sepenuhnya terdokumentasi dalam kajian yang bersifat komprehensif dan kuantitatif.

Salah satu pendekatan yang dapat memberikan gambaran menyeluruh tentang perkembangan ini adalah metode bibliometrik. Analisis bibliometrik mampu memetakan tren publikasi, mengidentifikasi penulis dan institusi yang paling produktif, memetakan jaringan kolaborasi, serta mengungkap topik-topik dominan dalam suatu bidang ilmu (McEnery & Hardie, 2011; Reppen & Simpson-Vlach, 2019). Melalui teknik seperti *co-citation analysis*, *keyword co-occurrence*, dan analisis tren temporal, para peneliti dapat melihat bagaimana pengetahuan di bidang *corpus linguistics* berkembang, bergeser, dan terhubung dengan disiplin ilmu lainnya. Meskipun metode ini telah banyak digunakan di berbagai bidang, seperti kedokteran, ilmu perpustakaan, atau ilmu komputer, penerapannya untuk memetakan bidang *corpus linguistics* secara global masih relatif terbatas.

Keterbatasan kajian terdahulu ini menimbulkan kebutuhan mendesak akan pemetaan menyeluruh yang mampu memberikan pandangan strategis bagi pengembangan riset ke depan. Bagi akademisi, pemetaan ini dapat menjadi panduan dalam memilih arah penelitian yang relevan. Bagi pembuat kebijakan dan pengelola pendidikan tinggi, informasi ini dapat membantu dalam perencanaan program studi, pembiayaan riset, dan pembangunan jejaring internasional. Selain itu, pemetaan global juga dapat membantu mengidentifikasi celah penelitian, misalnya dalam studi *corpus linguistics* untuk bahasa-bahasa yang kurang terwakili, atau dalam integrasi teknologi AI dan *machine learning* dalam analisis bahasa.

Permasalahan yang diangkat dalam penelitian ini berangkat dari kenyataan bahwa meskipun *corpus linguistics* telah berkembang menjadi bidang yang mapan dan berpengaruh, hingga kini belum tersedia analisis bibliometrik yang secara sistematis memetakan perkembangan

globalnya. Tidak banyak kajian yang mengungkap siapa aktor utama dalam bidang ini, negara atau institusi mana yang menjadi pusat inovasi, topik-topik apa yang mendominasi diskursus ilmiah, serta bagaimana kolaborasi lintas negara terjalin. Akibatnya, pemahaman kita mengenai peta pengetahuan *corpus linguistics* masih parsial dan cenderung terfragmentasi. Berdasarkan kesenjangan tersebut, tujuan penelitian ini adalah melakukan pemetaan bibliometrik menyeluruh terhadap publikasi ilmiah di bidang *corpus linguistics* dengan cakupan global.

2. METODOLOGI

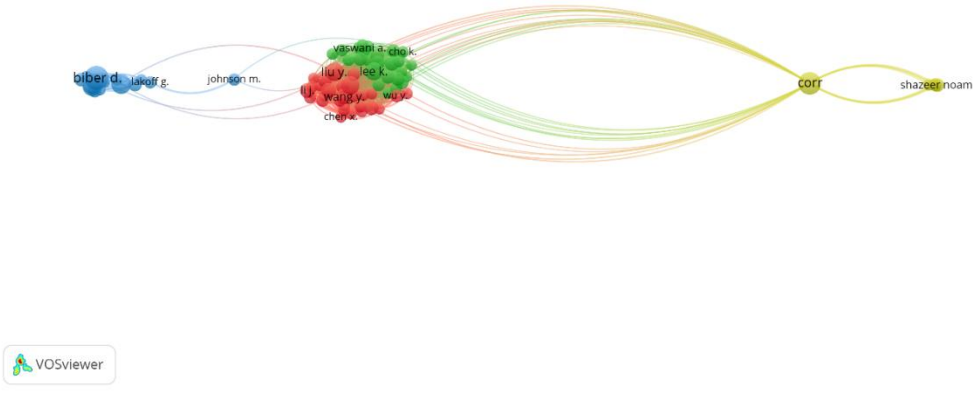
Penelitian ini menggunakan pendekatan bibliometrik dengan desain deskriptif kuantitatif untuk memetakan tren dan pola penelitian global dalam bidang *corpus linguistics*. Pendekatan ini dipilih karena mampu memberikan gambaran menyeluruh mengenai distribusi publikasi, pola sitasi, jaringan kolaborasi, dan topik dominan yang muncul dalam literatur ilmiah (Donthu et al., 2021). Data penelitian diperoleh dari basis data Scopus sebagai salah satu pangkalan data bibliografi terbesar dan bereputasi internasional, yang mencakup publikasi dari berbagai disiplin ilmu dan wilayah geografis. Penggunaan Scopus didasarkan pada kelengkapan metadata publikasi yang memungkinkan analisis bibliometrik secara komprehensif, termasuk informasi penulis, afiliasi, kata kunci, sitasi, dan referensi.

Pengambilan data dilakukan dengan merumuskan kata kunci pencarian yang relevan, yakni "corpus linguistics" dan istilah turunannya, yang kemudian disesuaikan dengan operator logis (*Boolean operators*) untuk memastikan cakupan luas namun tetap relevan. Pencarian dibatasi pada publikasi berjenis artikel jurnal dan prosiding konferensi yang telah melalui proses *peer review*, guna menjamin kualitas akademik sumber data. Rentang waktu penelitian ditetapkan antara 2000 hingga 2025, agar mencakup perkembangan kontemporer sekaligus tren historis dalam *corpus linguistics*. Seluruh metadata yang diperoleh diunduh dalam format CSV dan BibTeX untuk dianalisis lebih lanjut. Analisis bibliometrik dilakukan menggunakan perangkat lunak VOSviewer. VOSviewer digunakan untuk menghasilkan visualisasi peta pengetahuan seperti *co-authorship network*, *keyword co-occurrence*, dan *co-citation analysis*.

3. HASIL DAN PEMBAHASAN

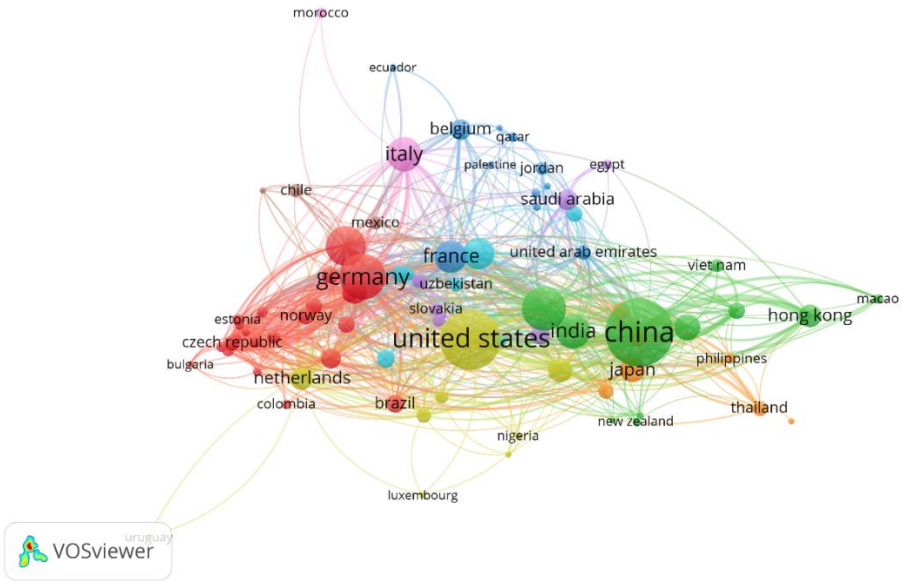
3.1 Hasil

a. Co-Authorship Analysis



Gambar 1. Visualisasi Penulis
Sumber: Data Diolah

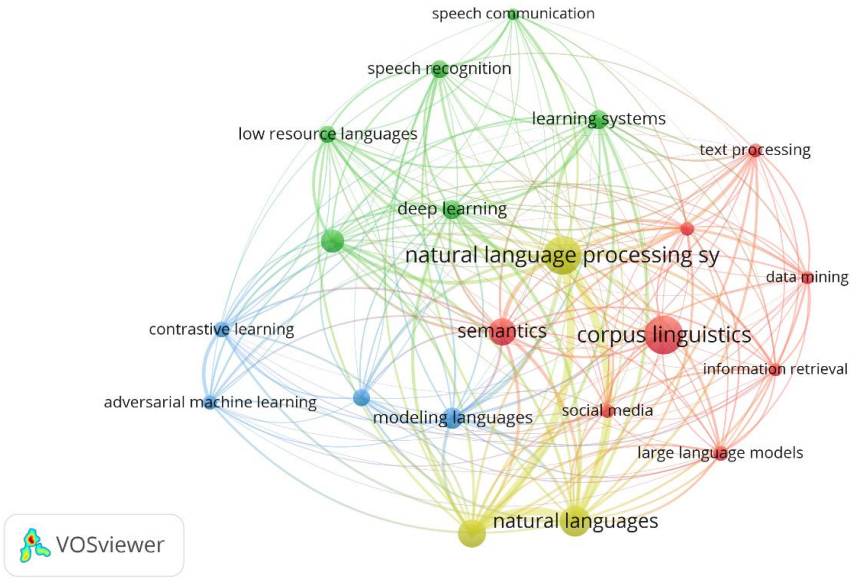
Menampilkan peta jejaring kolaborasi penulis dalam bidang *corpus linguistics*, dengan pembagian klaster warna yang merepresentasikan kelompok kolaborasi yang saling terhubung. Klaster biru di sisi kiri menunjukkan tokoh seperti Biber D. dan Lakoff G., yang cenderung berkolaborasi dalam lingkup penelitian linguistik deskriptif dan analisis korpus tradisional. Klaster hijau dan merah di bagian tengah merepresentasikan kelompok penulis yang lebih padat interaksi kolaborasinya, seperti Vaswani A., Cho K., Liu Y., dan Wang X., yang kemungkinan besar terkait pada tema pengembangan model bahasa dan teknologi NLP berbasis korpus. Di sisi kanan, node berwarna kuning seperti Corr dan Shazeer Noam mengindikasikan tokoh yang menjadi penghubung lintas klaster, berperan sebagai titik integrasi antara penelitian linguistik berbasis korpus dan inovasi teknologi *deep learning*.



Gambar 2. Visualisasi Negara
Sumber: Data Diolah

Gambar 2 menunjukkan peta kolaborasi internasional dalam penelitian *corpus linguistics*, di mana ukuran lingkaran mewakili jumlah publikasi suatu negara dan ketebalan garis mengindikasikan intensitas kolaborasi. Terlihat bahwa United States, China, Germany, United Kingdom, dan France menjadi pusat utama dalam jejaring global, dengan keterhubungan yang kuat ke berbagai negara lain. Klaster hijau mengelompokkan negara-negara Asia seperti China, India, Japan, dan Hong Kong yang saling terhubung erat, sedangkan klaster merah didominasi oleh negara-negara Eropa seperti Germany, Netherlands, dan Norway. Klaster biru mencakup negara-negara seperti France, Belgium, dan Saudi Arabia, sedangkan klaster ungu memperlihatkan koneksi Italia dengan negara-negara Mediterania dan Timur Tengah.

b. Keyword Co-Occurrence Analysis



Gambar 3. Visualisasi Jaringan
Sumber: Data Diolah

Gambar 3 menggambarkan peta keterkaitan tematik dalam penelitian corpus linguistics dan bidang-bidang terkait, di mana setiap node merepresentasikan kata kunci penelitian dan ukuran node mencerminkan frekuensi kemunculannya. Terlihat bahwa natural language processing systems berada di posisi sentral dengan ukuran node besar, mengindikasikan bahwa bidang ini menjadi fokus utama dan memiliki keterkaitan luas dengan berbagai topik lain. Posisi sentral ini menunjukkan bahwa corpus linguistics saat ini sangat erat hubungannya dengan pengembangan dan penerapan sistem pemrosesan bahasa alami, terutama dalam konteks komputasi modern.

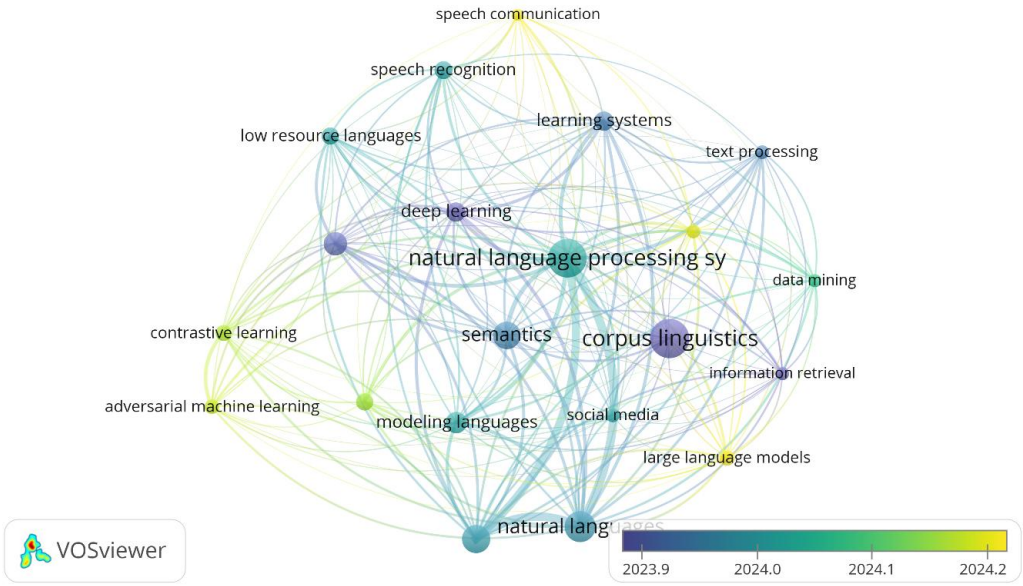
Klaster merah di sisi kanan menampilkan keterhubungan erat antara *corpus linguistics* dengan topik seperti *text processing*, *data mining*, *information retrieval*, *social media*, *semantics*, dan *large language models*. Ini menunjukkan bahwa penelitian berbasis korpus kini semakin mengarah pada pemrosesan teks skala besar, pengambilan informasi, dan pemanfaatan model bahasa besar seperti GPT, BERT, dan sejenisnya. Hubungan yang kuat dengan *semantics* juga mencerminkan bahwa analisis makna tetap menjadi komponen penting dalam studi berbasis korpus, khususnya untuk memahami konteks dan penggunaan bahasa.

Klaster hijau di bagian atas memperlihatkan fokus pada *speech communication*, *speech recognition*, *learning systems*, *deep learning*, dan *low resource languages*. Hubungan ini mengindikasikan bahwa penelitian *corpus linguistics* juga berkembang ke arah pemrosesan bahasa lisan (speech) dan pembelajaran mesin mendalam (*deep learning*), serta mulai memberikan perhatian pada bahasa-bahasa dengan sumber daya terbatas. Tren ini penting karena menunjukkan bahwa teknologi NLP tidak hanya difokuskan pada bahasa global, tetapi juga mulai merambah bahasa minoritas yang sebelumnya kurang terlayani secara digital.

Klaster biru di sisi kiri bawah mencakup topik seperti *contrastive learning*, *adversarial machine learning*, dan *modeling languages*. Kehadiran topik-topik ini menandakan adanya integrasi metode pembelajaran mesin tingkat lanjut dalam penelitian bahasa berbasis korpus, termasuk pembelajaran kontras yang digunakan

untuk meningkatkan kinerja model, serta teknik *adversarial* untuk menguji ketahanan dan keamanan model NLP. Fokus ini menegaskan pergeseran *corpus linguistics* dari sekadar analisis linguistik ke arah pengujian dan optimasi model kecerdasan buatan.

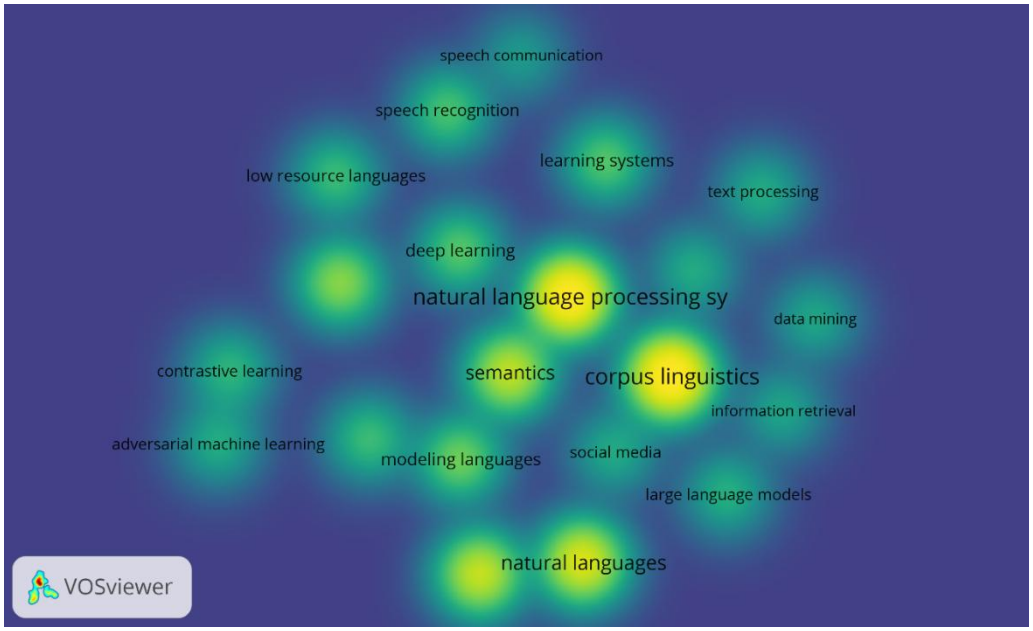
Klaster kuning yang terhubung langsung ke pusat memperlihatkan istilah seperti *natural languages*, yang menjadi penghubung antara berbagai klaster. Node ini menunjukkan bahwa konsep bahasa alami tetap menjadi dasar bagi seluruh penelitian, meskipun fokus kajiannya bervariasi dari linguistik murni hingga pemrosesan bahasa berbasis AI. Secara keseluruhan, peta ini menegaskan bahwa *corpus linguistics* berada di persimpangan antara linguistik tradisional dan teknologi kecerdasan buatan modern, dengan jembatan penghubung yang kuat menuju *deep learning*, pemrosesan bahasa lisan, dan model bahasa besar, sekaligus membuka peluang riset interdisipliner yang semakin luas.



Gambar 4. Visualisasi Overlay
Sumber: Data Diolah

Gambar 4 menunjukkan perkembangan temporal kata kunci penelitian di bidang *corpus linguistics* dan area terkait, dengan gradasi warna dari biru (lebih lama) ke kuning (lebih baru). Node seperti corpus linguistics, natural language processing systems, dan semantics muncul dalam warna biru keunguan, menandakan bahwa topik-topik ini merupakan fondasi penelitian yang sudah lama menjadi fokus. Sementara itu, kata kunci seperti speech communication, learning systems, dan text processing terlihat dalam warna kuning, yang mengindikasikan tren penelitian yang lebih baru dan sedang berkembang pesat pada periode akhir 2023 hingga awal 2024.

Kata kunci berwarna hijau, seperti contrastive learning, low resource languages, dan data mining, merepresentasikan area penelitian yang berada pada fase transisi, di mana minat mulai meningkat dalam beberapa tahun terakhir. Misalnya, contrastive learning menjadi penting seiring meningkatnya penerapan metode pembelajaran mesin terkini dalam pengolahan bahasa, sedangkan penelitian pada low resource languages menunjukkan arah pengembangan NLP yang lebih inklusif dan berorientasi pada keberagaman bahasa. Posisi node-node ini yang terhubung ke pusat seperti natural language processing systems mengindikasikan bahwa inovasi dalam area ini memiliki keterkaitan erat dengan inti penelitian linguistik berbasis korpus.



Gambar 5. Visualisasi Densitas
Sumber: Data Diolah

Gambar 5 menunjukkan tingkat kepadatan kemunculan kata kunci dalam publikasi terkait *corpus linguistics* dan bidang-bidang yang beririsan. Warna kuning merepresentasikan area dengan kepadatan tertinggi, yang berarti kata kunci tersebut sering muncul dan menjadi fokus utama penelitian. Terlihat bahwa natural language processing systems, corpus linguistics, semantics, dan natural languages memiliki intensitas paling tinggi, menandakan posisi sentralnya dalam diskursus akademik. Sementara itu, warna hijau menunjukkan kepadatan menengah, terlihat pada kata kunci seperti deep learning, learning systems, information retrieval, dan data mining, yang juga berperan penting namun dengan frekuensi sedikit lebih rendah dibanding kata kunci inti.

Kata kunci dengan kepadatan rendah, ditandai dengan gradasi hijau kebiruan, seperti contrastive learning, adversarial machine learning, low resource languages, dan speech communication, mengindikasikan bahwa meskipun topik-topik ini sedang berkembang, intensitas publikasinya masih relatif lebih sedikit dibandingkan tema utama. Namun, kehadirannya di peta menunjukkan potensi perluasan arah penelitian di masa depan, terutama dalam mengintegrasikan metode pembelajaran mesin canggih dan fokus pada bahasa dengan sumber daya terbatas.

c. Citation Analysis

Tabel 1. Artikel yang Paling Banyak Dikutip

Sitasi	Penulis dan Tahun	Judul
308	(Javaid et al., 2023)	*ChatGPT for healthcare services: An emerging stage for an innovative perspective*
220	(Wagner et al., 2023)	*Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap*
215	(Chen et al., 2024)	*Benchmarking Large Language Models in Retrieval-Augmented Generation*
213	(Navigli et al., 2023)	*Biases in Large Language Models: Origins, Inventory, and Discussion*
190	(Ranathunga et al., 2023)	*Neural Machine Translation for Low-resource Languages: A Survey*

Sitasi	Penulis dan Tahun	Judul
146	(Deng et al., 2023)	*Large Language Models Are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models*
101	(Zhang et al., 2023)	*Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models*
96	(Chai, 2023)	*Comparison of text preprocessing methods*
85	(Latif & Zhai, 2024)	*Fine-tuning ChatGPT for automatic scoring*
67	(Kresevic et al., 2024)	*Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework*

Sumber: Scopus, 2025

3.2 Implikasi Praktis

Hasil pemetaan bibliometrik ini memberikan wawasan strategis bagi berbagai pemangku kepentingan di bidang *corpus linguistics* dan teknologi bahasa. Pertama, bagi peneliti dan akademisi, peta kata kunci inti seperti *natural language processing systems*, *corpus linguistics*, dan *semantics* menunjukkan area yang menjadi pusat penelitian, sehingga dapat dijadikan rujukan untuk mengembangkan riset yang relevan dan memiliki peluang sitasi tinggi. Kedua, bagi praktisi industri teknologi bahasa, temuan mengenai keterkaitan *deep learning*, *information retrieval*, dan *large language models* menandakan potensi aplikasi langsung dalam pengembangan produk berbasis AI, seperti chatbot, sistem terjemahan otomatis, dan analisis sentimen. Ketiga, bagi pembuat kebijakan dan penyandang dana riset, adanya kata kunci seperti *low resource languages* menggarisbawahi pentingnya pendanaan dan dukungan untuk riset yang mengangkat bahasa-bahasa minoritas, sehingga dapat memperluas akses teknologi secara inklusif dan mengurangi kesenjangan digital.

3.3 Kontribusi Teoritis

Penelitian ini memberikan kontribusi signifikan terhadap pemahaman teoretis mengenai posisi dan perkembangan *corpus linguistics* dalam ekosistem ilmu pengetahuan global. Pemetaan bibliometrik menunjukkan bahwa bidang ini tidak lagi berdiri sendiri sebagai cabang linguistik murni, melainkan telah bertransformasi menjadi disiplin interdisipliner yang terintegrasi erat dengan ilmu komputer, kecerdasan buatan, dan ilmu data. Dengan mengidentifikasi klaster kata kunci seperti *speech recognition*, *adversarial machine learning*, dan *contrastive learning*, studi ini memperluas kerangka teoretis *corpus linguistics* ke arah yang lebih komputasional. Selain itu, penelitian ini menegaskan bahwa konsep inti seperti *semantics* dan *natural languages* tetap menjadi landasan teoritis, namun kini diperkaya oleh model-model pembelajaran mesin mutakhir yang memungkinkan analisis bahasa dalam skala besar dan lintas bahasa.

3.4 Limitasi

Walaupun hasil studi ini memberikan gambaran komprehensif, terdapat beberapa keterbatasan yang perlu dicatat. Pertama, sumber data penelitian ini hanya menggunakan basis data Scopus, sehingga publikasi yang terindeks di pangkalan data lain seperti Web of Science, Dimensions, atau Google Scholar mungkin belum sepenuhnya terwakili. Kedua, penggunaan kata kunci pencarian yang spesifik seperti "corpus linguistics" dan variasinya dapat menyebabkan *bias coverage*, di mana publikasi relevan dengan istilah berbeda tidak masuk dalam analisis. Ketiga, analisis bibliometrik yang dilakukan bersifat kuantitatif dan visual, sehingga tidak menggali secara mendalam konten dan metodologi dari setiap publikasi. Oleh karena itu, penelitian lanjutan yang mengombinasikan pendekatan bibliometrik dengan tinjauan literatur sistematis (*systematic literature review*) akan memberikan pemahaman yang lebih kaya baik dari sisi tren maupun substansi kajian.

4. KESIMPULAN

Berdasarkan hasil pemetaan bibliometrik, dapat disimpulkan bahwa penelitian global di bidang *corpus linguistics* menunjukkan dinamika yang semakin interdisipliner, dengan keterkaitan erat terhadap bidang *natural language processing*, *semantics*, dan teknologi berbasis pembelajaran mesin seperti *deep learning* dan *large language models*. Topik inti seperti *information retrieval*, *text processing*, dan *data mining* menegaskan peran *corpus linguistics* sebagai jembatan antara analisis bahasa tradisional dan inovasi komputasional mutakhir. Sementara itu, munculnya fokus pada *low resource languages*, *contrastive learning*, dan *speech recognition* mencerminkan arah pengembangan ke ranah inklusivitas bahasa dan teknologi bahasa lisan. Temuan ini tidak hanya memetakan lanskap riset saat ini, tetapi juga membuka peluang strategis bagi pengembangan teori, aplikasi industri, dan kebijakan riset di masa depan, dengan menekankan pentingnya kolaborasi lintas disiplin dan lintas negara untuk menjawab tantangan serta memanfaatkan peluang dalam ekosistem linguistik global.

DAFTAR PUSTAKA

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553.

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762.

Deng, Y., Xia, C. S., Peng, H., Yang, C., & Zhang, L. (2023). Large language models are zero-shot fuzzers: Fuzzing deep-learning libraries via large language models. *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 423–435.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241.

Javaid, M., Haleema, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards & Evaluations*, *3*(1).

Kennedy, G. (2014). *An introduction to corpus linguistics*. Routledge.

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L. S., & Shung, D. L. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digital Medicine*, *7*(1), 102.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210.

Leech, G. (2014). The state of the art in corpus linguistics. *English Corpus Linguistics*, 8–29.

McEnery, T. (2019). *Corpus linguistics*. Edinburgh University Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, *15*(2), 1–21.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1–37.

Reppen, R., & Simpson-Vlach, R. (2019). Corpus linguistics. In *An introduction to applied linguistics* (pp. 91–108). Routledge.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X.-Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. *Proceedings of the Fourth ACM International Conference on AI in Finance*, 349–356.