PERSATUAN LINGUISTIK MALAYSIA
Bahasa Sendi Tamadun Bangsa

www.plm.org.my

JURNAL LINGUISTIK Vol. 30 (1) Mei 2026 (101-121)

Menuju Leksikografi Pintar: Tinjauan Kaedah Leksikografi dalam Era Digital

*[1]Nurulhuda Mohamad Ali, [2]Norihan Rosli & [3]Fansurina Ramli

[1]nurulhuda@dbp.gov.my, [2]norihan@dbp.gov.my,
& [3]fansurina@dbp.gov.my

Bahagian Perkamusan, Dewan Bahasa dan Pustaka, 50460, Kuala Lumpur

[1]Penulis Koresponden

Tarikh terima : 30 Mac 2026
Received
Terima untuk diterbitkan : 8 Mei 2026
Accepted
Tarikh terbit dalam talian : 31 Mei 2026
Published online

Abstrak

Leksikografi ialah bidang teras dalam dunia linguistik dan teknologi bahasa, yang menumpukan kepada pendokumentasian, analisis, dan penyusunan kata berserta makna untuk menjadi sumber leksikal sesuatu bahasa. Dalam beberapa dekad kebelakangan ini, bidang ini mengalami transformasi signifikan yang didorong oleh perkembangan dan kemajuan dalam perkomputeran linguistik dan kecerdasan buatan (AI). Kertas ini menelusuri evolusi kaedah leksikografi dengan memfokuskan peralihan daripada pendekatan tradisional berasaskan intuisi dan kaedah manual kepada pendekatan korpus dan automasi berbantukan teknologi AI. Melalui tinjauan terhadap kajian terkini, makalah ini meneliti tiga aspek utama: (1) pendekatan tradisional dan perkomputeran dalam leksikografi, (2) peralatan, korpus dan sumber data yang menyokong kerja leksikografi, serta (3) cabaran dan hala tuju pembangunan leksikografi dalam era AI. Analisis makalah ini mendapati amalan leksikografi kini beralih ke arah model hibrid yang menggabungkan kepakaran manusia dengan keupayaan automasi untuk meningkatkan ketepatan dan kebolehskalaan. Kajian kes *Kamus Dewan Perdana* (KDP) turut diketengahkan untuk menggambarkan peralihan daripada penyusunan manual kepada penyusunan digital berbantukan AI. Secara keseluruhannya, tinjauan makalah ini menyumbangkan pemahaman menyeluruh tentang evolusi kaedah leksikografi dan potensi penerapan AI untuk memperkukuh pembangunan kamus bahasa Melayu.

Kata kunci: Leksikografi, Kecerdasan Buatan (AI), sumber linguistik, *Kamus Dewan Perdana*, korpus digital

Towards Smart Lexicography: A Review of Lexicographic Methodologies in the Digital Age

Abstract

Lexicography is a fundamental field within linguistics and language technology that focuses on the documentation, analysis, and systematic organisation of words and their meanings to form lexical resources for a language. In recent decades, the discipline has undergone a significant transformation driven by advances in computational linguistics and artificial intelligence (AI). This paper examines the evolution of lexicographic methodologies, focusing on the transition from traditional intuition-based and manual approaches to corpus-based methods and AI-assisted automation. Drawing on recent studies, this paper investigates three main aspects: (1) traditional and computational approaches in lexicography, (2) tools, corpora, and data resources that support lexicographic work, and (3) the challenges and future directions of lexicographic development in the AI era. The analysis shows that contemporary lexicographic practice is increasingly moving toward a hybrid model that combines human expertise with automated technologies to improve accuracy and scalability. A case study of Kamus Dewan Perdana (KDP) is presented to illustrate the shift from manual dictionary

compilation to digitally assisted lexicography supported by AI. Overall, this review provides an understanding of the evolution of lexicographic methodologies and highlights the potential of AI in strengthening the development of Malay dictionaries.

***Keywords**: Lexicography, Artificial Intelligence (AI), linguistic resources, Kamus Dewan Perdana, digital corpus*

1. Pengenalan: Evolusi Kaedah Leksikografi

Leksikografi merupakan salah satu bidang asas dalam linguistik dan teknologi bahasa, yang melibatkan pendokumentasian, analisis serta penyusunan kata bersama maknanya ke dalam kamus atau sumber leksikal (Pedersen, 2012). Bidang ini memainkan peranan penting dalam proses memahami dan memelihara bahasa, khususnya melalui pembangunan pangkalan data leksikal yang teratur (Ogilvie, 2011). Secara tradisional, kerja leksikografi dilakukan secara manual, di mana pakar bahasa mengumpulkan kata daripada teks, mentakrifkan kata tersebut berdasarkan konteks dan penggunaan, lalu merekodkannya sebagai entri kamus (Tarp, 2008). Proses ini memerlukan tenaga kerja intensif serta masa yang panjang, dengan penelitian rapi terhadap aspek makna, sebutan, golongan kata, contoh penggunaan, dan etimologi yang memainkan peranan penting untuk menentukan asal-usul dan hubungan setiap kata dalam sesuatu bahasa (Taufiq & Aniswal, 2022). Generasi awal kamus bercetak bergantung kepada slip petikan atau kad bahan tulisan tangan serta sistem pemfailan secara manual, namun kini sudah wujud sistem berbantukan kecerdasan buatan (AI) seperti *LLM assistant* di platform *Wordnik*, yang menggunakan input atau arahan *ChatGPT* untuk menghasilkan kata masukan, melabelkan ranah penggunaan, dan mencadangkan takrif awal dalam masa yang singkat. Sistem ini membuktikan bahawa kecerdasan buatan generatif (*generative AI*), walaupun masih pada peringkat prototaip, sudah mampu mereplikakan keseluruhan kaedah kerja tradisional dalam leksikografi atau penyusunan kamus (Wordnik, 2023).

Seiring dengan peredaran masa, amalan leksikografi telah berevolusi dengan integrasi kaedah perkomputeran yang berjaya mengautomasikan aspek penting dalam penyusunan kamus. Kemunculan korpus-korpus (korpora) berskala besar, teknik NLP, dan pembelajaran mesin (ML) telah membolehkan para pekamus menggunakan pelbagai peralatan digital dalam pengekstrakan istilah, penyelesaian kesamaran makna, pengenalpastian kolokasi, serta penjanaan takrif atau glos secara automatik (Atkins & Rundell, 2008; Sinclair, 1991; de Schryver, 2023; Kamarulzaman, Rashidin, Jantan & Md Zain., 2025). Kajian terkini menunjukkan bahawa teknologi kecerdasan buatan generatif seperti *ChatGPT* boleh membantu untuk mendraf entri kamus, menyusun semula takrif yang sesuai bagi pelbagai kumpulan pengguna, serta mencadangkan rujukan silang yang bersesuaian. Sebagai contoh, penilaian pakar dalam kajian Lew (2023) mendapati takrif dan ayat contoh penggunaan yang dijana teknologi *AI* mampu menandingi hasil kepakaran manusia. Hal ini menandakan peranan "arahan yang terancang" (*controlled prompting*) dan pascapengeditan (post-editing) yang semakin penting dalam kerja leksikografi. Inovasi sebegini menjadikan penyusunan kamus lebih cekap dan berskala besar, sekali gus memperlihatkan bagaimana integrasi AI membuka ruang baharu untuk menjadikan kaedah kerja leksikografi lebih efisien, fleksibel, dan berdaya saing untuk memenuhi keperluan pengguna kamus pada zaman moden ini.

Pada masa kini, kaedah leksikografi merangkumi pelbagai pendekatan, daripada yang tradisional sehingga kepada teknik berasaskan AI, yang secara umum boleh dikategorikan kepada (1) leksikografi berasaskan intuisi, (2) leksikografi berasaskan korpus, dan (3) leksikografi perkomputeran (Lew, 2023). Pendekatan asas dan juga tradisional ialah leksikografi berasaskan intuisi. Pendekatan ini bergantung kepada pengetahuan pakar, intuisi, serta kepakaran tatabahasa dalam sesuatu bahasa untuk memilih, mentakrif dan menentukan contoh penggunaan bagi sesuatu kata. Pendekatan berasaskan intuisi ini juga menjadi tunjang utama dalam penyusunan kamus, terutamanya dalam konteks awal atau bagi bahasa yang mempunyai keterhadan dalam sumber linguistik. Namun, pendekatan ini terbatas kerana sering dipengaruhi oleh pertimbangan yang subjektif, iaitu bergantung kepada pandangan atau intuisi individu, serta tidak berasaskan bukti empirik daripada data penggunaan bahasa sebenar yang terakam. Di samping itu, pendekatan ini turut berdepan dengan kesukaran dari segi kebolehskalaan,

kerana apabila jumlah data meningkat, kapasiti manusia menjadi had utama (de Schryver & Prinsloo, 2001). AI kini memperkaya proses penyusunan entri kamus melalui pendekatan *expert-in-the-loop*, iaitu apabila pekamus mengemukakan pertanyaan kepada model bahasa besar (LLM) untuk mendapatkan cadangan seperti frasa, definisi atau maklumat semantik. Cadangan ini kemudiannya dinilai, disemak dan dipilih oleh pekamus berdasarkan kriteria leksikografi yang ditetapkan, tanpa perlu menulis semula atau membina setiap entri sepenuhnya dari awal.

Perkembangan pendekatan leksikografi berasaskan korpus pula menandakan perubahan metodologi yang besar, di mana perubahan tersebut membolehkan para pekamus menganalisis penggunaan bahasa yang autentik dalam koleksi teks yang berstruktur dan berskala besar (Atkins & Rundell, 2008). Pendekatan ini dianggap sebagai pendekatan empirik kerana bersandarkan data terakam yang merupakan bukti penggunaan bahasa sebenar, dan bukan semata-mata bergantung kepada intuisi seorang pekamus (Liu, Mamat & Husain, 2025). Pendekatan ini menyokong penentuan keputusan berdasarkan data yang terakam, misalnya dalam proses menyelesaikan kekaburan makna, mengekstrak kolokasi dan fraseologi, memilih ayat contoh penggunaan, serta menjalankan analisis frekuensi. Hal ini menambah konsistensi dan kesahihan empirikal kepada setiap entri kamus, sekali gus meningkatkan nilai linguistik dan pedagogi kamus itu sendiri (Atkins & Rundell, 2008; Sinclair, 1991). Platform korpus berbentuk dgital seperti *Sketch Engine* kini mampu secara automatik mengenal pasti kolokat penting serta mengelompokkan sinonim rapat, sekali gus menggabungkan teknik konkordans klasik[1] dengan metrik keserupaan semantik berasaskan neural[2], iaitu kaedah pengukuran tahap kesamaan makna antara kata atau frasa yang menggunakan representasi vektor yang dihasilkan oleh rangkaian neural (Herman, 2021).

Bertitik tolak daripada pendekatan leksikografi berasaskan korpus, pendekatan leksikografi berasaskan perkomputeran muncul dalam beberapa dekad kebelakangan ini sebagai tindak balas terhadap inovasi teknologi, serta keperluan terhadap kaedah yang lebih cekap dan boleh dikembangkan untuk mengendalikan data berskala besar. Dalam era AI, pendekatan leksikografi berasaskan perkomputeran merangkumi penjanaan glos (keterangan) menggunakan LLM (de Schryver, 2023), pengekstrakan istilah berasaskan transformer (*transformer-based*[3]) (Lang, Wachowiak, Heinisch, & Gromann, 2021), serta pembangunan glosari digital melalui antara muka NLP yang interaktif (Lew, 2023). Sebagai contoh, sistem *LexicoMatic* memanfaatkan data daripada *Princeton WordNet* dengan memindahkan atau memetakan makna (sensa) setiap kata ke dalam pelbagai bahasa melalui korpus-korpus Wikipedia yang diterjemah secara automatik menggunakan *mBART*. Pendekatan ini membolehkan pembinaan draf WordNet bagi bahasa yang keterhadan sumber linguistik tanpa memerlukan proses terjemahan secara manual (Martelli, Procopio, Barba, & Navigli, 2023). Begitu juga, kajian Malinga, Lupanda, Wa Nkongolo, dan van Deventer (2024) menunjukkan bahawa LLM yang digabungkan dengan teknik AI dapat membantu memulakan dan mengembangkan pembinaan leksikon sentimen bagi pelbagai bahasa (Perancis, Tshiluba, Afrikaans, Sepedi, dan Zulu), termasuk bahasa yang keterhadan sumber linguistik. Pendekatan ini bukan sahaja mengurangkan beban kerja pakar, malah memberikan justifikasi yang jelas bagi setiap skor polariti[4]. Hal ini menunjukkan bahawa pengaplikasian ML dan NLP dapat mempercepat serta mengautomasikan pelbagai aspek penyusunan

[1] **Konkordans klasik** - Teknik analisis korpus tradisional yang memaparkan senarai kata berserta konteks sekelilingnya bagi mengenal pasti pola penggunaan.

[2] **Metrik keserupaan semantik berasaskan neural** (*neural semantic similarity metrics*) - Kaedah pengukuran tahap kesamaan makna antara kata atau frasa menggunakan vektor semantik yang dijana oleh model rangkaian neural.

[3] **Transformer-based** - pendekatan dalam NLP, menggunakan seni bina *Transformer* (Vaswani et al., 2017), berasaskan mekanisme perhatian untuk memahami hubungan antara perkataan dalam ayat.

[4] **Skor polariti** - nilai berangka yang diberikan kepada unit bahasa untuk menunjukkan darjah kecenderungan sentimen sama ada positif, negatif atau neutral.

kamus, di samping meningkatkan kecekapan dan konsistensi, khususnya bagi bahasa yang mempunyai sumber linguistik yang terhad.

Meskipun terdapat pelbagai kemajuan, sebahagian besar penyelidikan dan pembinaan leksikografi masih tertumpu kepada bahasa yang kaya dengan sumber linguistik seperti bahasa Inggeris dan Jerman, yang mendapat manfaat daripada data linguistik berskala besar, infrastruktur perkomputeran yang kukuh, serta peralatan leksikografi yang canggih dan berkredibiliti (Li, Zeng, Qin, Lu, Li, Kudo & Neubig, 2024). Sebaliknya, bahasa dengan sumber linguistik yang terhad seperti bahasa Melayu masih berdepan dengan cabaran yang berterusan, termasuk keterhadan korpus berkualiti tinggi, kekurangan kamus digital, serta akses yang terhad kepada alat dan teknologi berasaskan AI (Li, Zeng, Qin, Lu, Li, Kudo & Neubig, 2024). Oleh itu, usaha leksikografi bahasa Melayu masih banyak bergantung kepada kaedah manual, dengan tahap integrasi yang terhad terhadap teknik perkomputeran moden serta teknologi ML dan AI. Kajian terkini turut menegaskan bahawa tanpa pembangunan platform penilaian terbuka untuk pelbagai bahasa serta sistem berasaskan LLM yang mudah digunakan dan berkos rendah, komuniti bahasa yang keterhadan sumber tidak akan dapat memanfaatkan teknologi tersebut. Keadaan ini sekali gus berpotensi memperluas jurang antara bahasa bersumber tinggi dan bahasa keterhadan sumber dalam bidang leksikografi digital (Pava, Uz Zaman, Meinhardt, Friedman, Truong, Zhang, Cryst & Marivate, 2025; Bella, Helm, Koch & Giunchiglia, 2023).

Sehubungan itu, kertas ini meninjau kajian-kajian berkaitan dengan pendekatan tradisional dan perkomputeran dalam leksikografi serta pengekstrakan istilah, dengan penekanan terhadap bahasa yang mempunyai sumber linguistik terhad. Walaupun pendekatan tradisional seperti leksikografi berasaskan intuisi dan korpus masih relevan serta menjadi asas penting, kajian ini memberi tumpuan khusus kepada perkembangan terkini dalam AI dan ML serta pengaplikasiannya dalam kerja leksikografi. Kaedah moden ini berupaya mengautomasikan pelbagai proses dan mengatasi keterbatasan pendekatan manual atau separa manual, khususnya dalam konteks bahasa Melayu yang masih keterhadan sumber linguistik dan korpus beranotasi.

Melalui penelitian terhadap evolusi daripada amalan tradisional ke arah inovasi berasaskan AI, kertas ini bertujuan mengetengahkan kaedah yang dapat disesuaikan dan dilaksanakan bagi memperkukuh usaha pembangunan leksikografi bahasa Melayu. Dalam hal ini, tiga soalan penyelidikan menjadi panduan kajian kertas ini:

1. Apakah kaedah tradisional dan perkomputeran utama yang digunakan dalam leksikografi dan pengekstrakan istilah?
2. Apakah peralatan, korpus, dan data yang digunakan untuk menyokong kerja leksikografi?
3. Apakah cabaran utama serta hala tuju bagi memajukan pembangunan leksikografi berasaskan AI?

Bahagian seterusnya mengemukakan penelitian terhadap kajian, peralatan, dan metodologi yang relevan merentas spektrum yang sedang berkembang ini.

2. Sorotan Kajian: Landskap Dan Inovasi Kaedah Leksikografi

Bidang leksikografi telah mengalami evolusi yang bermula daripada inskripsi atau catatan awal pada kepingan tanah liat dan buluh, kepada skrol yang diperbuat daripada papirus, manuskrip yang ditulis tangan, kamus bercetak, dan kini platform digital (Tarp & Gouws, 2023). Setiap peringkat perkembangan ini mempengaruhi kaedah dan peralatan yang digunakan dalam penyusunan kamus. Peralihan ke arah pendigitalan membolehkan leksikografi moden menggabungkan pendekatan tradisional dengan kaedah perkomputeran yang canggih seperti analisis korpus, NLP, dan AI (Atkins & Rundell, 2008; Lew, 2023). Sehubungan dengan itu, sorotan kajian ini meneliti pelbagai kajian terkini yang mencerminkan perubahan metodologi leksikografi, khususnya bagaimana gabungan intuisi pakar, data korpus, dan teknik ML digunakan dalam kerja leksikografi, sekali gus memperlihatkan amalan yang boleh disesuaikan untuk pembangunan leksikografi bahasa Melayu.

2.1 Inovasi dan Kaedah Leksikografi: Intuisi → Model Digital Berstruktur

Kajian Tarp dan Gouws (2023) menekankan bahawa pertimbangan ilmu seorang pakar dalam penyusunan kamus tidak dapat digantikan sepenuhnya, namun perlu didokumentasikan dalam bentuk digital yang lebih sistematik dan berstruktur. Kajian ini memfokuskan bahawa kaedah leksikografi kini perlu menggabungkan amalan tradisional, daripada glos tulisan lama, pengekstrakan kosa kata berasaskan abjad, sehinggalah penyusunan kamus secara manual, dengan kaedah perkomputeran, di mana *glossography* dan *dictionography* secara digital memanfaatkan teknologi berasaskan AI untuk menghasilkan glos berasaskan konteks sesuatu kata. Kajian ini juga memperlihatkan bahawa platform dan aplikasi berasaskan AI seperti *DeepL Write* (platform penyuntingan dan penambahbaikan teks), *Ginger* dan *Grammarly* (alat semakan tatabahasa dan gaya bahasa), *LanguageTool* (pengesan kesalahan linguistik pelbagai bahasa), serta *ProWritingAid* (perisian analisis gaya dan struktur penulisan) telah mengintegrasikan data leksikografi ke dalam model AI bagi menyokong kefasihan, ketepatan, dan kejelasan tulisan penggunanya. Sistem sebegini bukan sahaja bergantung kepada data korpus berskala besar, tetapi turut memanfaatkan set data khusus dan bahan rujukan yang distrukturkan dalam pangkalan data moden, yang berbeza secara signifikan daripada repositori leksikon tradisional. Namun, pelbagai kekangan masih dihadapi dalam usaha mengautomasikan tugasan yang lebih rumit seperti penggubalan takrif serta pelabelan sosiolinguistik berdasarkan konteks, selain menyelaraskan kerjasama pelbagai bidang kepakaran untuk melaksanakan integrasi AI secara menyeluruh. Tarp dan Gouws (2023) juga menegaskan bahawa pekamus perlu menyesuaikan dan menstrukturkan semula pangkalan data leksikografi tradisional serta menjalinkan kerjasama strategik bersama pereka bentuk dan pakar teknologi bagi memastikan kawalan editorial terpelihara, ketelusan maklumat terjamin, dan kepercayaan pengguna terus diperkukuh. Akhir sekali, mereka mencadangkan supaya leksikografi didefinisikan semula kepada dua subbidang, iaitu *dictionography* (penghasilan kamus tradisional dan karya rujukan) dan *glossography* (penyediaan dan penyisipan glos), untuk membantu penyelidikan pada masa hadapan tentang kaedah mengekalkan hubungan dua hala yang jelas antara glos yang dijana oleh AI dengan sumber empirikal, demi pembangunan leksikografi digital yang mampan dan berkualiti tinggi. Walaupun kajian ini menyarankan penggabungan amalan tradisional dengan kaedah perkomputeran berasaskan AI, namun pengautomasian sepenuhnya bagi tugasan seperti penggubalan takrif yang khusus dan mendalam serta pelabelan sosiolinguistik masih memerlukan penyeliaan pakar yang berterusan.

Kajian Lugli (2019) meneroka bagaimana unsur intuisi diintegrasikan dalam pendekatan *smart-lexicography* bagi bahasa yang tergolong sebagai bahasa keterhadan sumber, dengan memberikan tumpuan kepada *Buddhist Sanskrit* dan Tibet klasik. Dalam kajian ini, intuisi pakar dijadikan sebagai elemen utama dalam proses membuat keputusan leksikografi, terutamanya ketika menjalankan anotasi teks dan pemilihan ayat contoh penggunaan. Walau bagaimanapun, intuisi tersebut tidak digunakan secara rawak atau berdasarkan pertimbangan subjektif semata-mata, tetapi berpaksikan kepada pembacaan teks yang teliti serta penggunaan alat anotasi digital yang berstruktur dan sistematik. Alat ini membolehkan pertimbangan dan keputusan pakar direkodkan secara konsisten, boleh diulang, dan disertakan dengan tahap keyakinan atau ketidakpastian yang jelas. Projek seperti *Buddhist Translators Workbench* dan *Lexicography in Motion* menunjukkan peralihan daripada pendekatan tradisional yang memfokuskan penyusunan kamus sebagai produk utama kepada pendekatan berasaskan korpus yang menyokong penambahbaikan secara berterusan. Kajian ini menekankan pembinaan korpus beranotasi yang boleh diguna semula bagi pelbagai tujuan, termasuk penghasilan prototaip kamus, pembangunan alat NLP, serta pemvisualan data. Namun, kajian ini masih belum berjaya menangani ketidakserasian antara pendekatan leksikografi tradisional yang berasaskan intuisi dengan keperluan untuk menghasilkan data yang sistematik serta boleh disemak semula. Dalam kajian ini, penyusunan dan pengintegrasian keputusan berasaskan intuisi ke dalam rangka kerja anotasi digital telah menghasilkan satu model yang praktikal dan fleksibel bagi menyokong pembangunan leksikografi yang mampan dalam konteks keterhadan sumber. Pendekatan ini turut menunjukkan bahawa unsur intuisi dapat distrukturkan secara sistematik, seperti yang dibuktikan melalui analisis *Buddhist Sanskrit* dan Tibet klasik. Dalam persekitaran anotasi secara digital yang dibangunkan melalui *Buddhist Translators' Workbench*, tahap keyakinan dan rasional di sebalik setiap keputusan pakar direkodkan secara

berperingkat. Proses ini sekali gus menukar pertimbangan tersirat kepada data berstruktur yang memacu pendekatan leksikografi berasaskan korpus.

Model Hanks, seperti yang dihuraikan dalam kajian Teubert (2013), memperkenalkan pendekatan leksikografi berasaskan pemprosesan perkomputeran, di mana makna (sensa) ditentukan melalui pola sintagmatik yang kerap muncul dalam korpus yang dianalisis, kemudian diperhalusi secara sistematik melalui penerapan intuisi linguistik pakar. Model ini direalisasikan melalui rangka kerja *Corpus Pattern Analysis (*CPA*)* yang mengenal pasti pola tipikal penggunaan kata (norma) dan penyimpangan daripada pola tersebut (eksploitasi). Dalam model ini, intuisi pakar tidak digunakan secara tidak formal, sebaliknya dilaksanakan secara sistematik melalui proses anotasi manual, pelabelan peranan semantik, dan interpretasi pola untuk membezakan norma penggunaan yang signifikan dalam korpus berskala besar seperti *British National Corpus* (BNC). Walaupun belum dapat dilaksanakan secara automatik sepenuhnya, CPA berjaya mengoperasikan pertimbangan semantik manusia secara konsisten. Antara cabaran utama yang dikenal pasti ialah penskalaan proses pengecaman pola serta pengautomasian interpretasi semantik, dua aspek yang amat mencabar khususnya bagi bahasa yang keterhadan sumber. Kajian ini mengesyorkan pembangunan kaedah *semantic parsing*, peningkatan automasi dalam pengecaman pola, serta pembinaan antara muka hibrid yang lebih berkeupayaan untuk menangkap dan merekodkan pertimbangan pakar dengan lebih tepat. Langkah ini penting bagi menyesuaikan kerangka leksikografi berasaskan AI seperti CPA dalam konteks bahasa keterhadan sumber termasuk bahasa Melayu. Kajian ini turut menunjukkan bahawa pola sintagmatik yang kerap muncul dalam BNC hanya memperoleh makna yang signifikan selepas peranan semantik ditandakan secara manual dan norma penggunaan dibezakan daripada bentuk eksploitasi. Secara keseluruhannya, kajian ini menggambarkan peralihan daripada penggunaan intuisi secara tersirat kepada kepakaran yang didokumenkan secara digital, sekali gus menyediakan asas yang telus bagi pelaksanaan automasi bagi bahasa keterhadan sumber.

2.2 Inovasi dan Kaedah Leksikografi: Berasaskan Korpus → Separa Automasi

Setelah intuisi berjaya didokumentasikan sebagai data, korpus berskala besar dan teknologi pencarian pola boleh dimanfaatkan untuk menggubal bahan bagi penyusunan kamus. Kajian oleh Baisa, Blahuš, Cukr, Herman, Jakubíček, Kovář, Medveď, Měchura, Rychlý dan Suchomel (2019) menerangkan sebuah projek leksikografi automatik sepenuhnya berasaskan korpus, di mana korpus daripada Internet dalam bahasa Tagalog yang mengandungi 230 juta token telah dibina dengan menggunakan *Sketch Engine*, *Stanford POS tagger* dan *FastText embedding* untuk menjana kira-kira 45,000 entri. Daripada jumlah tersebut, 15,000 entri disemak dan diedit semula secara manual melalui platform sumber terbuka *Lexonomy*, manakala 30,000 entri dihasilkan secara automatik sepenuhnya. Setiap entri merangkumi pelbagai elemen penting seperti bentuk terbitan, sebutan, makna (sensa) bersama glos, ilustrasi imej (*API Wikimedia Commons/Pixabay/Google*), kolokasi, sinonim/antonim, ayat contoh penggunaan berasaskan korpus (*GDEX- Good Dictionary Examples*), serta terjemahan awal (*Google Translate* dan *Bing*) yang kemudiannya disahkan dan diperkemas oleh pekamus. Bagi memastikan kualiti yang terbaik, seramai tujuh orang editor terjemahan Tagalog-Inggeris bersama penterjemah bahasa Korea telah melaksanakan kaedah kerja editorial yang sistematik dan berstruktur, dengan sokongan latihan serta garis panduan yang lengkap. Projek yang dijalankan selama sembilan bulan ini membuktikan bahawa penghasilan entri kamus secara automatik pada skala besar, yang digabungkan dengan pelibatan tenaga manusia dalam proses pengeditan, merupakan pendekatan yang praktikal dan kos efektif. Walau bagaimanapun, cabaran masih wujud, khususnya dalam pengurusan data dan pengguna yang kompleks, serta keperluan pengeditan intensif bagi entri yang jarang muncul. Pendekatan ini akan diperluas kepada bahasa Urdu dan Lao pada fasa seterusnya, dengan penekanan terhadap pembinaan korpus berskala lebih besar serta penambahbaikan kaedah yang bertujuan untuk mengurangkan kebergantungan kepada tenaga manusia dalam proses pengeditan data. Kajian ini hampir sepenuhnya memfokuskan pendekatan automatik berasaskan korpus dan tidak membincangkan secara menyeluruh amalan leksikografi tradisional, seperti pengelasan sensa secara manual, peranan jawatankuasa editorial dalaman, atau kaedah perbandingan sejarah, mahupun membuat perbandingan antara amalan tersebut dengan kaedah kerja berasaskan pemprosesan komputer.

Pada tahun 2024, Michael Rundell telah menelusuri peralihan bidang leksikografi daripada kaedah manual tradisional seperti pengelasan makna melalui pembacaan baris konkordans, penulisan takrif secara manual, dan penyusunan format untuk penerbitan bercetak, kepada pendekatan berasaskan pemprosesan komputer yang mengautomasikan hampir setiap peringkat penyusunan kamus. Dalam pendekatan baharu ini, korpus berskala besar yang dikumpulkan daripada Internet diproses dan dianalisis secara berperingkat, bermula dengan pemecahan teks kepada unit perkataan, penukaran kepada bentuk dasar, serta penentuan jenis kata seperti kata nama, kata kerja atau kata adjektif, untuk menghasilkan senarai perkataan utama mengikut kekerapan penggunaannya. Analisis kolokasi berasaskan *Word Sketch* serta teknik pengelompokan menyokong pembezaan sensa secara separa automatik, dan algoritma GDEX memudahkan pemilihan ayat contoh. *Sketch Engine* digunakan bagi pengurusan korpus dan analisis *Word Sketch*, manakala GDEX berperanan dalam pengekstrakan ayat contoh penggunaan. Selain itu, platform penyusunan kamus seperti *Lexonomy* dan *TshwaneLex* turut diaplikasikan ke atas pelbagai jenis korpus, daripada data yang mengandungi berbilion perkataan sehinggalah kepada subkorpus yang khusus seperti *British National Corpus* (BNC). Walaupun berjaya mencapai kemajuan, cabaran masih wujud, terutamanya dalam usaha mengautomasikan proses penggubalan takrif serta pelabelan sosiolinguistik yang memerlukan pakar dalam proses pertimbangan editorial.

Kajian Domínguez Vázquez dan Gouws (2023) pula menelusuri perkembangan berterusan dalam bidang leksikografi, daripada amalan tradisional yang melibatkan pemilihan manual elemen berasaskan konteks seperti label, ayat contoh dan kolokasi, yang berfungsi menjelaskan makna sesuatu perkataan dalam struktur mikro kamus, kepada penggunaan pelbagai kaedah berasaskan komputer yang kini mendominasi kerja terminologi. Kaedah tersebut termasuk pencarian konkordans berasaskan korpus (*Sketch Engine*), teknik pemodelan bahasa berasaskan statistik dan rangkaian neural seperti *Word2Vec* dan *FastText*, serta penjana berasaskan peraturan seperti *Xera*, *Combinatoria* dan *CombiContext* yang secara automatik menghasilkan pola valensi serta ayat penggunaan sintetik. Kajian ini turut menyenaraikan pelbagai teknologi dan sumber penting dalam bidang leksikografi moden, termasuk korpus rujukan (*BNC*, *CORGA*, *CREA*), pangkalan data leksikal (*WordNet*, *FrameNet*, *EcoLexicon*), rangkaian semantik (*BabelNet*), sumber beranotasi (*PDEV*, *PropBank*, *NomBank*, *AnCora*), serta repositori terbuka seperti *Wikipedia* dan *Wiktionary*. Selain itu, peralatan analisis korpus dan bantuan terjemahan seperti *OWIDplus*, *SKELL*, *Google Translate* dan *DeepL* turut dimanfaatkan. Namun begitu, masih terdapat cabaran seperti kekurangan korpus yang beranotasi semantik untuk menghubungkan struktur sintaksis dengan makna, kaedah automasi yang sukar membezakan antara hujah yang khusus dengan yang tidak khusus, dan konteks pelbagai dimensi masih sukar untuk diekstrak secara menyeluruh. Justeru, penyelidikan pada masa hadapan perlu menumpukan kepada pembangunan tatacara penilaian yang lebih menyeluruh, ontologi yang lebih terperinci, serta rangka kerja AI yang lebih fleksibel untuk membolehkan pembangunan sumber leksikografi yang interaktif dan berorientasikan pengguna.

2.3 Inovasi dan Kaedah Leksikografi: Penjajaran Makna dan Pengembangan Pelbagai Bahasa

Apabila pasangan terjemahan antara dua bahasa sukar diperoleh, perwakilan makna berbilang bahasa (*multilingual embeddings*) boleh dimanfaatkan untuk menjana kamus dwibahasa dan rangkaian semantik secara automatik. Kajian Marchisio, Saad-Eldin, Duh, Priebe dan Koehn (2022) membandingkan dua pendekatan, (1) penyusunan leksikon dwibahasa secara manual, di mana pekamus membina pasangan terjemahan kata daripada teks selari atau glosari sedia ada, dan (2) pendekatan pengiraan, yang menjana terjemahan secara automatik daripada ruang perwakilan makna ekabahasa. Kaedah pemetaan seperti *Procrustes* digunakan pada peringkat awal untuk memadankan ruang makna antara dua bahasa dengan berpandukan senarai kecil padanan perkataan sebagai rujukan awal, sebelum terjemahan paling sesuai dikesan secara automatik. Pendekatan seterusnya melibatkan algoritma berasaskan graf seperti FAQ dan SGM, yang mewakili hubungan antara perkataan dan mencari padanan yang sepadan antara dua bahasa, termasuk dalam keadaan perbezaan struktur atau corak bahasa yang ketara. Pendekatan GOAT pula diperkenalkan untuk mengoptimumkan proses pemadanan supaya menjadi lebih tepat, khususnya dalam konteks pasangan bahasa yang kekurangan data atau mempunyai struktur yang berbeza. Kajian ini turut menghuraikan proses mewakilkan makna perkataan dalam

bentuk nombor melalui latihan model menggunakan korpus teks berskala besar, pembinaan senarai padanan awal (leksikon benih) daripada perkataan yang serupa dalam kedua-dua bahasa atau daripada kamus kecil, serta penggunaan teknik pembelajaran kendiri untuk meluaskan jumlah padanan terjemahan. Secara keseluruhan, pendekatan ini menunjukkan peralihan daripada kaedah penyusunan kamus secara manual yang memakan masa kepada kaedah pembinaan leksikon yang lebih pantas dan automatik berasaskan teknologi perwakilan makna perkataan.

Martelli, Procopio, Barba dan Navigli (2023) telah memperkenalkan *LexicoMatic*, iaitu sebuah sistem pembinaan dan pengembangan kamus semantik pelbagai bahasa secara automatik. Sistem ini menggabungkan pendekatan tradisional seperti terjemahan dan peluasan makna, iaitu menterjemah *Princeton WordNet* ke dalam bahasa baharu atau membina *wordnet* secara bebas dan menyelaraskannya semula dengan teknik moden berasaskan AI yang dapat memilih makna perkataan yang tepat berdasarkan definisi dan menyelaraskan perkataan merentas bahasa secara automatik. Bagi menyokong kaedah ini, kajian ini menggunakan pelbagai sumber data termasuk *Princeton WordNet* sebagai kamus semantik asas, korpus mengandungi sejuta ayat Wikipedia bahasa Inggeris yang diterjemah menggunakan *mBART50*, set data padanan perkataan, korpus *SemCor* yang dilabel dengan makna perkataan mengikut konteks, serta beberapa sistem pemilihan makna *ESCHER*, penyelarasan diskriminatif *Procopio*, dan rangka kerja transformer. Kajian ini turut mengenal pasti beberapa cabaran utama seperti kekurangan sumber pelbagai bahasa, kos anotasi secara manual yang tinggi, serta perbezaan terjemahan yang menyukarkan penyelarasan automasi. Kajian ini bercadang untuk meluaskan penilaian kepada lebih banyak bahasa, menyesuaikan rangka kerja kepada inventori makna selain *Princeton WordNet*, dan memperbaiki strategi penyelarasan serta penggabungan data bagi meningkatkan ketepatan, keluasan liputan, dan kebolehskalaan dalam pembangunan leksikografi berbantukan AI.

2.4 Inovasi dan Kaedah Leksikografi: Hibrid → Model Bahasa Besar (LLM)

Peralihan zaman serta kemajuan teknologi dalam era AI telah menjadikan pembangunan leksikografi tidak lagi terhad kepada automasi asas, sebaliknya turut merangkumi aspek seperti penulisan takrif, penilaian makna (sentimen) dan pengekstrakan istilah merentas bahasa. Perkembangan ini secara tidak langsung meningkatkan keperluan terhadap pelaksanaan kawalan kualiti yang lebih menyeluruh dan teliti. Kajian McKean dan Fitzgerald (2023) menunjukkan bahawa kaedah tradisional leksikografi masih bergantung kepada proses editorial secara manual dan analisis berasaskan data korpus bagi mengenal pasti entri, frasa dan bentuk kata daripada himpunan teks berskala besar, serta merangka takrif dan contoh penggunaan secara manual oleh pakar. Pada masa yang sama, pendekatan automasi berasaskan AI digunakan untuk mencadangkan peluasan makna entri, menjana senarai frasa, membina draf takrif berdasarkan petikan ayat, menyesuaikan takrif mengikut kumpulan pengguna tertentu, menghasilkan contoh ayat penggunaan, serta menambah label atau sebutan. Kajian ini menggunakan data entri daripada *Wordnik API*, korpus terpilih bagi perkataan baharu dan istilah teknikal, senarai perbendaharaan kata kanak-kanak bagi pendrafan semula takrif, serta platform *Jupyter Notebook* sebagai persekitaran kerja utama. Kajian ini turut mengenal pasti beberapa cabaran utama seperti kesalahan fakta dan halusinasi yang kerap dalam hasil AI, isu etika berkaitan kesan terhadap alam sekitar dan kecenderungan bagi pengulangan, ketidakselarasan hasil antara setiap model, serta implikasi profesional, harta intelek dan peraturan. Antara cadangan penambahbaikan yang dikemukakan termasuk pembangunan tatacara penilaian yang standard untuk menilai keupayaan AI dalam leksikografi, penyesuaian semula model terbuka untuk bidang kamus, pelaksanaan semakan yang menyeluruh oleh pakar bagi setiap peringkat, dan pembentukan penanda aras yang telus untuk memastikan integrasi AI dijalankan secara selamat dan beretika dalam penghasilan kamus.

Kajian Malinga, Lupanda, Wa Nkongolo dan van Deventer (2024) menggabungkan pendekatan leksikografi manual dan automatik untuk membina serta menganalisis leksikon sentimen pelbagai bahasa bagi bahasa Afrika yang keterhadan sumber. Pada peringkat awal, kaedah tradisional digunakan dengan membersihkan leksikon Perancis–Ciluba yang sedia ada secara manual dalam format *Microsoft Excel* melalui proses pengemaskinian ejaan, penyingkiran duplikasi, dan pengesahan terjemahan bersama penutur asli, sebelum menambah 250 entri bahasa Inggeris bagi meluaskan liputan data. Dari sudut automatik, proses diteruskan dengan menggunakan *API Google Translate* untuk pemetaan

terjemahan secara pukal, memanfaatkan *ChatGPT* untuk memberikan skor sentimen mengikut bahasa, serta melatih model ML bagi ramalan sentimen pada peringkat ayat dan aspek tertentu. Pembangunan ini disokong oleh leksikon asal yang mengandungi kira-kira 3,000 entri, leksikon pelbagai bahasa yang diperluas, korpus ujian beranotasi sebanyak 1,000 ayat, serta pelbagai alat analisis bahasa yang membantu dalam pemprosesan data. Kajian turut mengenal pasti beberapa cabaran utama seperti ketidakseimbangan data dan bahasa, kerumitan bentuk kata, dan kesukaran mengekalkan makna sentimen dalam proses terjemahan. Cadangan pada masa hadapan termasuk memperbesar dan mempelbagaikan set data, memperbaiki kaedah penilaian sentimen, serta membangunkan sistem AI yang lebih telus, sensitif terhadap konteks, dan sesuai untuk pembangunan leksikografi bagi bahasa yang keterhadan sumber.

Kajian Xu, Feng, Li dan Dong (2025) meninjau perkembangan dalam pengekstrakan istilah automatik (*Automatic Term Extraction* - ATE) dan menunjukkan bagaimana bidang ini telah berkembang daripada sistem berasaskan peraturan linguistik dan kaedah statistik tradisional kepada model hibrid, algoritma berasaskan graf, pengelasan ML klasik, penanda berasaskan pembelajaran mendalam (DL), dan kini kepada pendekatan berasaskan model bahasa serta ATE pelbagai bahasa yang memanfaatkan transformer. Untuk menyokong pelbagai kaedah tersebut, kajian ini menghimpunkan pelbagai sumber dan alat penting termasuk korpus penanda aras seperti *GENIA*, *FAO*, *SemEval*, *ACTER*, *KP20k* serta koleksi pelbagai bahasa seperti *TTC* dan *Europarl*, selain perisian pengekstrakan seperti *TermoUD, TermoPL* dan *D-Terminer*, serta alat prapemprosesan teks yang digunakan secara meluas. Walaupun kemajuan telah dicapai, kajian menekankan bahawa cabaran utama masih kekal, termasuk banyaknya perkataan berbilang makna, kesukaran menyesuaikan sistem merentas bidang yang berbeza, kebergantungan tinggi terhadap data beranotasi berskala besar dan keperluan perkomputeran yang tinggi, serta kekurangan piawaian penilaian yang menyeluruh dan seragam. Hala tuju kajian ini disarankan untuk memberikan tumpuan kepada penggunaan model bahasa berskala besar, reka bentuk yang lebih berkesan, seni bina yang lebih ringan dan mudah disesuaikan, serta rangka penilaian yang lebih menyeluruh bagi menilai kebolehskalaan, keanjalan dan ketahanan dalam situasi sebenar.

Kajian Pakray, Gelbukh dan Bandyopadhyay (2025) pula meninjau perkembangan NLP bagi bahasa keterhadan sumber dan menghuraikan perkembangan kaedah yang menjadi asas kepada kerja leksikografi dan pengekstrakan istilah. Pada peringkat tradisional, penyusunan kamus masih bergantung kepada penerokaan korpus berasaskan peraturan serta model statistik klasik untuk mengesan pola bentuk kata, menanda kelas kata, dan mengenal pasti calon entri atau istilah. Pada peringkat automatik, pendekatan berasaskan korpus kini menjadi dominan dengan penggunaan model DL dan seni bina pemprosesan lanjutan bagi mengautomatikkan pengekstrakan entri atau istilah, penyelarasan merentas bahasa, dan penilaian makna, sekali gus mempercepat pembinaan kamus bagi bahasa keterhadan sumber. Pendekatan ini juga masih membolehkan penggabungan peraturan linguistik yang dibina oleh pakar, terutamanya dalam situasi di mana sumber data sangat terhad.

Kajian oleh Robert Lew (2024) meneliti bagaimana leksikografi tradisional dan berasaskan perkomputeran berkembang di bawah pengaruh AI. Kaedah tradisional yang berpusat pada kamus bercetak melibatkan pengumpulan perkataan secara manual daripada konteks dan penyusunan mengikut abjad untuk memudahkan pencarian, biasanya digunakan bagi tujuan pemahaman teks, penulisan, dan penterjemahan. Kemunculan leksikografi berasaskan perkomputeran memperkenalkan korpus digital serta pelbagai alat seperti kamus kolokasi, tesaurus, dan analisis berasaskan korpus yang meningkatkan kecekapan capaian dan kegunaan maklumat. Namun, bagi bahasa yang keterhadan sumber seperti bahasa Melayu, kemajuan ini masih terhad disebabkan kekurangan korpus digital, ketiadaan pangkalan data leksikal yang terstruktur, dan perwakilan yang rendah dalam data latihan AI. Walaupun *GPT* dan *DeepL* menunjukkan prestasi yang lebih baik daripada kamus dalam pelbagai tugas, tenaga pekamus masih mempunyai kelebihan dalam bidang yang memerlukan pemahaman budaya dan linguistik yang mendalam, seperti kerja lapangan bersama penutur asli atau pemprosesan bahan bukan digital. Antara proses dalam bidang ini termasuk pengekstrakan frasa berasaskan korpus, penandaan makna, dan penterjemahan mesin pelbagai bahasa, namun kebanyakan teknologi bagi proses-proses ini belum tersedia sepenuhnya atau masih kurang maju untuk bahasa Melayu. Kajian ini menekankan bahawa masa depan leksikografi bergantung kepada model hibrid, iaitu (1) AI membantu meningkatkan

kecekapan, manakala (2) pakar bahasa memastikan ketepatan linguistik, pengurusan data yang beretika, dan pemeliharaan kepelbagaian bahasa. Peranan rentas disiplin ini akan menjadi aspek penting dalam pembangunan AI yang adil dan inklusif bagi bahasa Melayu.

Secara keseluruhan, kajian-kajian ini menunjukkan perubahan ketara dalam pendekatan leksikografi, di mana proses manual tradisional kini semakin dilengkapi, malah ada kalanya digantikan, oleh teknik berasaskan AI yang membolehkan kerja dijalankan dengan skala lebih besar, lebih cekap, dan lebih mudah disesuaikan. Daripada penggunaan intuisi dan penyusunan berasaskan korpus kepada automasi menggunakan model bahasa berskala besar dan alat pelbagai bahasa, bidang ini sedang bergerak pantas ke arah model hibrid yang menggabungkan kepakaran manusia dengan keupayaan mesin. Bagi bahasa keterhadan sumber seperti bahasa Melayu, perkembangan ini membuka laluan yang berpotensi, kerana kaedah tradisional yang berasaskan intuisi dan kerja manual masih penting sebagai asas, manakala alat dan rangka kerja perkomputeran yang muncul mampu mengatasi kekangan lama berkaitan kekurangan data, kebolehskalaan, dan ketepatan makna. Pemanfaatan inovasi ini secara sensitif terhadap konteks dan berlandaskan etika akan menjadi kunci untuk memacu pembangunan leksikografi yang inklusif, mampan, dan berkualiti tinggi.

3. Data dan Korpus Penyelidikan Leksikografi

Bahagian ini membincangkan tinjauan yang menyeluruh terhadap set data dan korpus yang digunakan dalam tiga belas kajian yang telah dianalisis dalam sorotan kajian (Bab 2.0). Kajian-kajian tersebut menunjukkan kepelbagaian pendekatan dalam pembangunan leksikografi dan peristilahan, bermula daripada pembinaan glosari dwibahasa secara tradisional sehinggalah kepada penghasilan kamus berasaskan AI yang lebih canggih. Apabila bidang leksikografi beralih daripada kaedah manual kepada pendekatan automatik dan hibrid, data yang terstruktur, dianotasi, dan berskala besar menjadi semakin penting, bukan sahaja untuk memproses sejumlah besar maklumat leksikal, tetapi juga untuk menjamin ketepatan bahasa, kesahan bukti, dan keserasian merentas bahasa.

Beberapa kajian menggunakan korpus ekabahasa berskala besar dan korpus Internet untuk menyokong kerja leksikografi automatik. Sebagai contoh, kajian Baisa, Blahuš, Cukr, Herman, Jakubíček, Kovář, Medveď, Měchura, Rychlý dan Suchomel (2019) membina korpus Internet yang mengandungi 230 juta perkataan bagi bahasa Tagalog, yang menjadi asas kepada pembangunan entri kamus separa dan sepenuhnya automatik melalui penandaan kelas kata, pemotongan kata dasar, perwakilan makna perkataan dan algoritma pengelompokan. Kajian oleh Rundell (2024) pula menggunakan pelbagai korpus Internet termasuk BNC untuk menghasilkan senarai entri berdasarkan kekerapan dan mengenal pasti pola kolokasi melalui *Sketch Engine* dan *GDEX*. Korpus berskala besar ini bukan sahaja membantu dalam pembezaan makna perkataan dan pengekstrakan contoh, tetapi juga membolehkan pengesanan perkataan baharu dan ungkapan berbilang perkataan secara masa nyata.

Beberapa kajian lain pula menggunakan pangkalan data linguistik beranotasi dan sumber semantik pelbagai bahasa. Domínguez Vázquez dan Gouws (2023) melaporkan penggunaan infrastruktur linguistik kukuh seperti *WordNet* (bagi hubungan makna), *FrameNet* (bagi makna berbingkai), *BabelNet* (bagi pemetaan konsep pelbagai bahasa), *EcoLexicon* (bagi istilah alam sekitar) serta korpus berlabel peranan seperti *PropBank, NomBank*, dan *AnCora*. Sumber ini membolehkan pelaksanaan tugas leksikografi yang kompleks seperti penghasilan takrif berasaskan konteks, pengekstrakan pola valensi, dan pemodelan hubungan peranan semantik. Kajian Martelli, Procopio, Barba dan Navigli (2023) pula memperkenalkan *LexicoMatic*, yang menggunakan *Princeton WordNet* dan korpus terjemahan *mBART50* yang mengandungi lebih sejuta ayat Wikipedia bagi membolehkan pembinaan kamus semantik pelbagai bahasa secara automatik. Proses tersebut turut menggunakan set data berpadanan secara manual serta alat anotasi merentas bahasa berasaskan transformer untuk meningkatkan ketepatan penentuan makna perkataan.

Dalam bidang penyelarasan merentas bahasa dan pembinaan leksikon dwibahasa, kajian oleh Marchisio, Saad-Eldin, Duh, Priebe dan Koehn (2022), dan Malinga, Lupanda, Wa Nkongolo dan van Deventer (2024) menunjukkan bagaimana perwakilan makna pelbagai bahasa dan model neural boleh

menghasilkan leksikon dwibahasa berskala besar. Marchisio, Saad-Eldin, Duh, Priebe dan Koehn (2022) bermula dengan leksikon dwibahasa asas dan korpus ekabahasa berskala besar untuk menyelaraskan ruang makna menggunakan beberapa pendekatan yang berkesan walaupun bagi pasangan bahasa yang berbeza. Malinga, Lupanda, Wa Nkongolo dan van Deventer (2024) pula memulakan kajian dengan leksikon Perancis–Ciluba yang dibersihkan secara manual, sebelum memperluasnya menggunakan *Google Translate, ChatGPT*, dan pelbagai pengelas ML untuk memberikan nilai sentimen dan mengesahkan terjemahan dalam 1,000 ayat ujian beranotasi. Pendekatan ini membuktikan bahawa sistem hibrid mampu menggabungkan ketepatan kerja manusia dengan keupayaan mesin bagi bahasa yang kekurangan sumber.

Korpus beranotasi oleh pakar dan set data berasaskan intuisi sangat penting dalam projek yang melibatkan bahasa klasik atau bahasa keterhadan sumber. Lugli (2019) memberikan tumpuan kepada bahasa *Buddhist Sanskrit* dan Tibet Klasik dengan membina persekitaran anotasi berstruktur yang merakam keputusan pakar, tahap kepastian, dan anotasi. Data ini kemudiannya digunakan semula dalam aplikasi NLP dan alat terjemahan. Kajian oleh McKean dan Fitzgerald (2023) pula menggabungkan set data petikan yang dikendalikan secara manual dengan cadangan automatik melalui GPT-3.5 untuk mengembangkan entri dan membina draf takrif, sekali gus menunjukkan gabungan antara kerja editorial tradisional dengan automasi berasaskan AI.

Bagi pengekstrakan istilah automatik (ATE), Xu, Feng, Li dan Dong (2025) dan Rogers (2012) menggunakan korpus terkenal seperti *GENIA* (bioperubatan), *FAO* (pertanian), *SemEval*, *ACTER*, *KP20k* (kata kunci akademik), *TTC* (pemindahan teknologi), dan *Europarl* (teks parlimen EU). Korpus pelbagai bahasa dan khusus ini menjadi asas untuk mengukur keberkesanan kaedah ATE yang merangkumi analisis statistik, model DL, dan pengekstrakan pelbagai bahasa berasaskan transformer. Beberapa perisian sumber terbuka seperti *TermoUD, TermoPL* dan *D-Terminer* turut digunakan untuk menilai keberkesanan pendekatan klasik, hibrid dan neural.

Bagi bahasa keterhadan sumber, kajian oleh Lew (2024) serta Pakray, Gelbukh dan Bandyopadhyay (2025) menekankan cabaran yang berterusan untuk mengakses korpus digital berkualiti tinggi. Lew menjelaskan bahawa bagi bahasa keterhadan sumber, kebanyakan korpus sedia ada yang biasa digunakan dalam leksikografi adalah dalam bahasa Inggeris. Malah korpus beranotasi berskala besar dan rangkaian pemprosesan bahasa secara *real-time* juga masih belum tersedia atau masih pada tahap awal pembangunan bagi kebanyakan bahasa keterhadan sumber. Pakray, Gelbukh dan Bandyopadhyay (2025) pula mendokumentasikan keterbatasan kaedah klasik serta potensi model berasaskan transformer untuk mengatasi kekurangan data linguistik, namun turut mengingatkan bahawa teknologi ini masih memerlukan asas sumber linguistik yang kukuh seperti korpus dan pangkalan data leksikal yang standard. Korpus yang standard selalunya tidak lengkap atau tidak wujud bagi bahasa yang kurang dikaji.

Perbandingan tentang korpus dalam ketiga belas kajian dipersembahkan dalam Jadual 3.1, yang menunjukkan kepelbagaian dan kecekapan data dalam penyelidikan leksikografi moden, serta mendedahkan jurang besar dalam infrastruktur digital antara bahasa bersumber tinggi dengan bahasa keterhadan sumber. Oleh itu, menangani kekurangan ini akan menjadi langkah penting ke arah pembangunan leksikografi yang adil, inklusif dan bersedia untuk integrasi AI.

Jadual 3.1

Data Kajian Leksikografi

Kajian	Bahasa	Data/Korpus	Medium/Model	Objektif Leksikografi
Tarp & Gouws (2023)	Pelbagai	Korpus hak milik, korpus selari, set data penilaian/set data pengesahan	Grammarly, DeepL Write, ProWritingAid, LanguageTool	Mendefinisikan semula leksikografi kepada *dictionography/glossography*, penjanaan glos berasaskan AI
McKean & Fitzgerald (2023)	Inggeris	Data lema Wordnik, korpus petikan, senarai	GPT-3.5-turbo (melalui API), Jupyter, OpenAI prompts	Mengautomatikkan penulisan takrif, penjanaan contoh, dan peluasan lema

		perbendaharaan kata kanak-kanak		
Baisa, Blahuš, Cukr, Herman, Jakubíček, Kovář, Medveď, Měchura, Rychlý & Suchomel (2019)	Tagalog	Korpus web Tagalog 230 juta token	Sketch Engine, Lexonomy, FastText, GDEX, HDBSCAN, Google/Bing Translate	Penyediaan draf kamus automatik melalui aliran kerja pasca penyuntingan
Marchisio, Saad-Eldin, Duh, Priebe & Koehn (2022)	Pelbagai	Pembenaman monolingual, leksikon dwibahasa asas, korpus dwibahasa	Procrustes, FAQ, SGM, GOAT	Pembinaan leksikon dwibahasa menggunakan perwakilan makna pelbagai bahasa
Malinga, Lupanda, Wa Nkongolo & van Deventer (2024)	Perancis, Ciluba, Inggeris	Leksikon Perancis–Ciluba (~3000 entri), lema Inggeris tambahan, korpus ujian 1,000 ayat	Google Translate API, ChatGPT, BERT, Vader, Integrated Gradients	Pembinaan leksikon sentimen pelbagai bahasa untuk bahasa Afrika kurang sumber
Rogers (2012)	Inggeris	Korpus perubatan, korpus khusus bidang	Perisian pengekstrakan istilah (tidak dinyatakan), analisis semantik	Pengekstrakan istilah dan pengurusan istilah khusus bidang
Lew (2024)	Inggeris, Melayu	Korpus kolokasi, data terjemahan, sumber berasaskan korpus	GPT, DeepL, alat penandaan korpus	Menyokong model leksikografi hibrid manusia–AI terutamanya untuk bahasa kurang sumber
Rundell (2024)	Inggeris	Korpus Internet berskala besar (berbilion token), BNC	Sketch Engine, GDEX, Lexonomy, TshwaneLex, Word Sketch	Leksikografi pasca penyuntingan; penyahkaburan makna dan pemilihan contoh dibantu AI
Domínguez Vázquez & Gouws (2023)	Sepanyol, Inggeris	BNC, CORGA, CREA, FrameNet, WordNet, BabelNet, Wikipedia, Wiktionary	Sketch Engine, Word2Vec, fastText, CombiContext, Xera, Combinatoria	Penjanaan data kontekstual dan pengekstrakan pola valensi
Martelli, Procopio, Barba dan Navigli (2023)	Pelbagai	Princeton WordNet, korpus sintetik (>1 juta ayat Wikipedia diterjemah), SemCor, 300 pasangan sejajar	ESCHER, mBART50, Procopio aligner, Hugging Face Transformers	Pembinaan kamus semantik pelbagai bahasa melalui penyelarasan dan pemilihan makna
Xu, Feng, Li & Dong (2025)	Inggeris, Pelbagai	GENIA, FAO, SemEval, ACTER, TTC, Europarl, KP20k	TermoUD, TermoPL, D-Terminer, TF-IDF, PMI, Transformers	Tinjauan perkembangan ATE daripada kaedah berasaskan peraturan kepada pengekstrakan berasaskan model bahasa besar
Pakray et al. (2025)	Pelbagai (low-resource languages)	Korpus selari, korpus teks mentah	CRF, HMM, BiLSTM, XLM-R, mBERT, CNN, model perhatian	Automasi penandaan kelas kata, penyelarasan istilah, dan pembinaan kamus bagi bahasa yang kurang sumber
Lugli (2019)	Buddhist Sanskrit, Tibet klasik	Korpus beranotasi daripada Buddhist Translators	Alat anotasi dengan tahap kepastian berperingkat,	

		Workbench, kamus prototaip	Lexicography in Motion	Perolehan intuisi berstruktur bagi bahasa sejarah yang kurang sumber

4. Kaedah Penilaian Leksikografi

Bahagian ini membincangkan kaedah penilaian utama yang telah dilaksanakan kajian-kajian yang telah diteliti. Apabila kaedah menjadi semakin automatik dan dipacu oleh data, pemahaman tentang pelbagai bentuk penilaian yang biasa digunakan dalam bidang leksikografi menjadi semakin penting. Jadual 4.1 berikut memperincikan penilaian utama yang lazim digunakan dalam leksikografi:

Jadual 4.1

Kaedah Penilaian Utama

Jenis	Keterangan
Kualitatif	Penilaian yang dilakukan secara manual oleh pakar dengan membaca, menyunting, dan menilai kualiti kandungan kamus. Contohnya, menilai sama ada sesuatu takrif "betul", ayat contoh jelas, atau terjemahan memberi makna yang tepat. Penilaian ini tidak bergantung kepada nombor atau formula, tetapi berdasarkan pertimbangan linguistik.
Kuantitatif	Penilaian yang menggunakan nombor, formula, dan skor. Contohnya, bilangan ketepatan, bilangan kadar capaian, atau sejauh mana ketepatan dalam mengelakkan kesilapan. Penilaian ini memberikan tumpuan kepada pengukuran prestasi secara statistik.
Konseptual/Deskriptif	Penilaian yang berbentuk perbincangan atau ulasan, bukan ujian. Fokus diberikan kepada fungsi, ciri-ciri yang berguna, aspek yang mungkin belum lengkap, serta cabaran yang dikenal pasti, tanpa menjalankan eksperimen rasmi atau mengumpul data berangka. Pendekatan ini menyerupai artikel ulasan yang menghuraikan idea dan pemerhatian.

Dalam leksikografi tradisional, penilaian lazimnya dijalankan secara kualitatif dengan bergantung kepada semakan pakar untuk menilai ketepatan, kejelasan, dan keberkesanan pedagogi sesuatu entri kamus. Penilaian ini merangkumi pengesahan takrif, label tatabahasa, dan contoh penggunaan, yang biasanya dilakukan melalui jawatankuasa editorial atau perbincangan bersama pakar. Namun, dengan perkembangan leksikografi berasaskan perkomputeran dan pengekstrakan istilah, kaedah penilaian kini menjadi lebih pelbagai. Sistem leksikografi moden sering menggabungkan penilaian kualitatif dengan metrik kuantitatif seperti *Inter-Annotator Agreement* (IAA), ketepatan makna, liputan data, serta penunjuk prestasi yang diadaptasi daripada bidang carian maklumat dan ML seperti ketepatan, kadar capaian, dan skor F1. Sebahagian kajian turut menggunakan penilaian konseptual atau deskriptif untuk memberikan pandangan teori tentang reka bentuk alat, tahap kegunaan, dan potensi aplikasi.

Kajian-kajian yang dihuraikan dalam Bab 2 menggambarkan kepelbagaian pendekatan ini. Beberapa kajian menggunakan semakan manusia secara berstruktur dan penyuntingan pakar selepas proses automatik untuk menilai kandungan yang dihasilkan. Sebagai contoh, Lugli (2019) dan Teubert (2013) menggunakan anotasi linguistik terperinci, penandaan peranan semantik, dan penjejakan tahap keyakinan untuk menstrukturkan keputusan intuitif. Dalam kajian Baisa, Blahuš, Cukr, Herman, Jakubíček, Kovář, Medveď, Měchura, Rychlý dan Suchomel (2019), editor melaksanakan rangka proses pengeditan yang teliti, termasuk IAA, latihan editorial, dan semakan berkumpulan sistematik bagi 15,000 entri.

Kaedah kuantitatif pula lebih menonjol dalam kajian yang menggunakan ML atau pendekatan berasaskan perwakilan makna. Marchisio, Saad-Eldin, Duh, Priebe dan Koehn (2022) menilai prestasi perwakilan makna pelbagai bahasa melalui ketepatan padanan paling sesuai. Martelli, Procopio, Barba

dan Navigli (2023) menilai hasil sistem berdasarkan ketepatan, kadar capaian, dan ketepatan dalam penentuan makna perkataan. Xu, Feng, Li dan Dong (2025) menilai model pengekstrakan istilah dengan korpus beranotasi seperti *GENIA, ACTER* dan *KP20k* menggunakan metrik standard seperti ketepatan, kadar capaian dan skor F1. Malinga, Lupanda, Wa Nkongolo dan van Deventer (2024) menggunakan pengelas ML klasik dan model berasaskan *BERT* untuk penandaan sentimen pelbagai bahasa, dengan prestasi dilaporkan melalui skor penilaian standard dan alat tafsiran keputusan.

Beberapa kajian lain menumpukan kepada analisis kesilapan secara kualitatif dalam sistem yang melibatkan semakan kepakaran manusia. McKean dan Fitzgerald (2023) menilai hasil GPT-3.5 secara manual dengan memberikan tumpuan kepada ketepatan fakta, kesalahan maklumat, dan kesinambungan editorial. Rundell (2024) turut menekankan pengeditan tanpa bergantung kepada ukuran berangka. Domínguez Vázquez dan Gouws (2023), Lew (2024), dan Pakray et al. (2025) pula menjalankan penilaian berbentuk konseptual atau deskriptif, dengan memberikan pandangan tentang keupayaan teknologi, kekayaan makna, serta cabaran praktikal integrasi AI tanpa menjalankan ujian empirikal baharu. Jadual 4.2 di bawah merumuskan pendekatan penilaian yang digunakan dalam ketiga belas kajian tersebut, mengikut kategori kaedah penilaian dan berfungsi sebagai rujukan perbandingan yang ringkas.

Jadual 4.2

Ringkasan Kaedah Penilaian Kajian Leksikografi

Kajian	Kaedah Penilaian	Jenis
Baisa, Blahuš, Cukr, Herman, Jakubíček, Kovář, Medveď, Měchura, Rychlý & Suchomel (2019)	*Inter-annotator Agreement* (IAA), pengeditan semula (secara manual) secara berstruktur, perbandingan kelompok, latihan editor	Kualitatif + Kuantitatif
Rundell (2024)	Penyemakan editorial, pengesahan manual terhadap hasil model bahasa	Kualitatif
Domínguez Vázquez & Gouws (2023)	Penilaian konseptual berpaksikan alat (tiada ujian empirikal dilaporkan)	Konseptual/Deskriptif
Martelli, Procopio, Barba dan Navigli (2023)	Ketepatan, kadar capaian, penentuan makna perkataan, ketepatan penjajaran	Kuantitatif
Marchisio, Saad-Eldin, Duh, Priebe & Koehn (2022)	Ketepatan padanan terdekat, ketepatan pemetaan berasaskan graf, penilaian perwakilan makna	Kuantitatif
McKean & Fitzgerald (2023)	Penilaian berasaskan semakan manusia, pengesanan kesalahan (halusinasi, bias), pengesahan petikan secara manual	Kualitatif
Malinga, Lupanda, Wa Nkongolo & van Deventer (2024)	Metrik pengelas ML (F1, ketepatan, kadar capaian), prestasi BERT, analisis keputusan	Kuantitatif
Lugli (2019)	Pemantauan rasional penanda, tahap kepastian, kebolehulangan, penggunaan semula dalam alat NLP	Kualitatif
Hanks via Teubert (2013)	Analisis Corak Korpus (CPA), penandaan manual norma dan eksploitasi, penandaan peranan semantik	Kualitatif
Xu, Feng, Li & Dong (2025)	Perbandingan dengan korpus piawai (GENIA, FAO, ACTER), ketepatan, kadar capaian, skor F1	Kuantitatif
Rogers (2012)	Pengesahan pengekstrakan berasaskan korpus, perbandingan domain, contoh berbeza	Kualitatif + Kuantitatif
Lew (2024)	Perbincangan konseptual tentang kebolehgunaan, ketersediaan alat, liputan linguistik	Konseptual/Deskriptif

Pakray et al. (2025)	Pengumpulan tinjauan berasaskan laporan prestasi model (tiada ujian langsung)	Konseptual/Deskriptif

Secara keseluruhan, kepelbagaian pendekatan ini menunjukkan bahawa penilaian leksikografi yang kukuh memerlukan gabungan pelbagai kaedah. Semakan pakar secara kualitatif memastikan ketepatan bahasa dan kesesuaian konteks, manakala penilaian kuantitatif menambah unsur objektif. Penilaian konseptual pula memperkayakan bidang ini dengan menyediakan teori serta mengenal pasti kekangan praktikal. Seiring perkembangan teknologi leksikografi yang semakin automatik, gabungan ketiga-tiga pendekatan ini menjadi penting untuk menjamin kebolehpercayaan, kemudahan penggunaan, dan integriti linguistik.

4.1. Perbincangan: Sintesis Kaedah, Peralatan Dan Hala Tuju Leksikografi

Kertas ini mengemukakan satu tinjauan menyeluruh yang menelusuri perkembangan bidang leksikografi dan pembangunan peristilahan, daripada kaedah tradisional yang dilakukan secara manual kepada kaedah moden yang automatik dan berbantukan AI. Berdasarkan tiga belas kajian yang telah dianalisis, perbincangan berikut menghuraikan persoalan utama yang menjadi panduan kepada penyelidikan ini.

4.1.1 Perbandingan Kaedah Tradisional dan Berasaskan AI

Kaedah leksikografi secara tradisional banyak bergantung kepada intuisi pakar, di mana pekamus mentafsir dan mentakrifkan perkataan berdasarkan pengetahuan linguistik, bacaan petikan, dan kefasihan bahasa. Pendekatan berasaskan intuisi ini masih meluas digunakan bagi bahasa keterhadan sumber seperti bahasa Melayu yang bergantung kepada kepakaran manusia.

Kemunculan teknologi digital dan korpus telah membawa kepada kebangkitan leksikografi berasaskan korpus sebagai pendekatan utama. Melalui pendekatan ini, pekamus menggunakan koleksi teks yang berstruktur sebagai bukti empirik untuk membantu membuat keputusan berkaitan dengan makna perkataan, kolokasi, contoh ayat, dan kekerapan penggunaan. Platform seperti *Sketch Engine, Word Sketch*, dan *GDEX* memainkan peranan penting untuk menjayakan pendekatan ini. Dengan teknologi digital dan berbantukan AI, dunia leksikografi mengalami kemajuan yang pesat, termasuk:

- Analisis corak korpus (Teubert, 2013),
- Pembinaan leksikon dwibahasa berasaskan perwakilan makna (Marchisio, Saad-Eldin, Duh, Priebe & Koehn, 2022),
- Penggunaan LLM untuk penulisan takrif dan penandaan sentimen (McKean & Fitzgerald, 2023; Malinga, Lupanda, Wa Nkongolo & van Deventer,2024),
- Pengekstrakan istilah berasaskan transformer (Xu, Feng, Li & Dong, 2025),
- Pengaplikasian rangkaian neural dan pengembangan kamus pelbagai bahasa (Martelli, Procopio, Barba & Navigli, 2023).

Pendekatan ini membolehkan penyusunan entri kamus berlaku dengan lebih pantas, pengecaman corak secara automatik, dan penjajaran berskala merentas bahasa, sekali gus meningkatkan kecekapan penyusunan kamus berbanding kaedah tradisional.

4.2 Infrastruktur Data dan Peralatan Digital dalam Leksikografi

Pelbagai kaedah dan data telah digunakan untuk menyokong pelaksanaan kerja leksikografi, antaranya:

- Korpus:
    - Korpus berasaskan Internet (*230M-token Tagalog corpus)*

    - *British National Corpus* (*BNC*)
    - Korpus Wikipedia yang diterjemah melalui *mBART50*
    - Korpus bidang khusus (*GENIA, FAO, Europarl*)
- Sumber leksikal dan pangkalan data semantik:
    - *WordNet, FrameNet, BabelNet, EcoLexicon*
    - *PropBank, SemCor, AnCora*
- Peralatan berbantukan AI:
    - *Sketch Engine, Lexonomy, GDEX, TermoPL*
    - *Transformers* (*BERT, mBERT, XLM-R*)
    - *FastText, Word2Vec*
    - Algoritma penjajaran makna (*GOAT, Procrustes, SGM*)
    - Model berasaskan AI (*GPT-5.0, GPT-4.0*, *GPT-3.5, DeepL Write*)

Sumber ini menyokong pelbagai fungsi termasuk pengekstrakan contoh ayat, penentuan makna, pengelompokan sinonim, penjajaran pelbagai bahasa, dan pengekstrakan istilah. Hasil sorotan kajian menekankan bahawa korpus yang kaya, terstruktur dan dianotasi merupakan asas penting dalam pembinaan sebuah leksikon.

4.3 Cabaran dan Hala Tuju Leksikografi Berbantukan AI

Walaupun pelbagai kemajuan telah dicapai, beberapa cabaran utama masih wujud, terutamanya bagi bahasa Melayu:

- Kekurangan sumber data: Kekurangan korpus yang teranotasi untuk menyokong ML,
- Had Automasi: Penulisan takrif dan pelabelan sosiolinguistik masih memerlukan penilaian pakar dan sukar untuk diautomasikan sepenuhnya,
- Halusinasi: Model LLM seperti *ChatGPT* kadangkala menghasilkan maklumat yang kelihatan betul tetapi sebenarnya tidak tepat,
- Akses kepada teknologi terkini: Teknologi seperti *LexicoMatic* dan *ESCHER* masih belum boleh diakses secara terbuka dan belum disesuaikan untuk bahasa bukan Eropah,
- Rangka Penilaian: Tiada piawaian penilaian yang seragam, berskala, dan pelbagai bahasa untuk kandungan leksikografi yang dijana oleh AI.

Hala tuju leksikografi memberikan tumpuan kepada model hibrid yang menggabungkan kecekapan automasi dengan semakan pakar, termasuk:

- Memperluas set data dan leksikon pelbagai bahasa sebagai sumber terbuka,
- Menyesuaikan model bahasa berskala besar dengan korpus bidang khusus,
- Membangunkan kaedah penilaian yang sesuai untuk bahasa keterhadan sumber,
- Mewujudkan antara muka AI yang menghubungkan hasil dengan bukti sumber secara telus,
- Menggalakkan kerjasama dalam kalangan pakar bahasa, teknologi dan dasar.

Kesimpulannya, AI menawarkan pendekatan yang menyokong leksikografi moden, terutamanya dalam konteks pelbagai bahasa dan bahasa keterhadan sumber. Namun, peranan kepakaran manusia masih tidak dapat digantikan. Bidang ini perlu terus mengimbangi inovasi dengan ketelitian linguistik, kualiti data dan kepekaan budaya untuk membina sumber leksikal yang tepat dan inklusif.

4.4 Kajian Kes: Evolusi Penyusunan *Kamus Dewan Perdana* (KDP), Manual → Berbantukan AI

Penyusunan *Kamus Dewan Perdana* (KDP) oleh Dewan Bahasa dan Pustaka (DBP) menggambarkan peralihan daripada pendekatan tradisional kepada kaedah berasaskan perkomputeran. Pada peringkat awal, entri dan subentri dikumpulkan secara manual dan bergantung kepada teks bercetak, kad petikan, serta intuisi pakar. Kaedah ini kaya dengan nilai linguistik tetapi memakan masa yang lama, kos dan tenaga kerja yang besar.

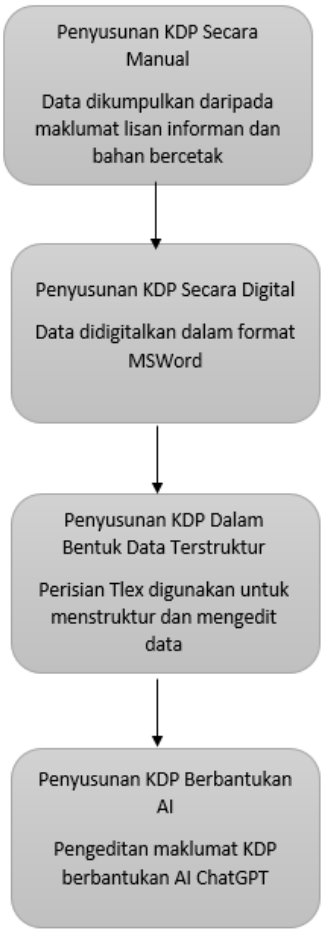
Pada awal tahun 2000-an, kemunculan peralatan digital dan capaian data secara dalam talian mula mengubah kaedah kerja penyusunan kamus. Kaedah penyusunan kamus mula beralih kepada proses pendigitalan apabila entri KDP didigitalkan dalam format *Microsoft Word* (*MsWord*), namun format *MsWord* tidak terstruktur dan mengambil masa yang masih lama untuk disemak semula. Keperluan bagi pengurusan data yang lebih sistematik telah mendorong pasukan leksikografi DBP mengaplikasikan *TshwaneLex* (*TLex*) sebagai perisian utama dalam penyusunan kamus, khususnya KDP. *TLex* membolehkan peralihan kaedah penyusunan kamus daripada bentuk fail teks yang statik kepada pangkalan data dinamik, di mana entri, subentri, takrif, kelas kata, label penggunaan dan contoh penggunaan dapat dianotasikan secara teratur dan mudah disemak semula tanpa perlu mengambil masa yang lama.

Dengan kemajuan AI yang pesat, edisi kedua KDP yang kini dalam proses editorial sudah mula menggunakan *ChatGPT* sebagai alat bantuan penulisan dan penyelidikan. Hal ini menandakan peralihan penting dalam proses penyusunan kamus kepada pendekatan leksikografi secara hibrid. Namun begitu, kepakaran manusia masih kekal sebagai panduan utama walaupun AI digunakan untuk memperhalusi draf bagi takrif, contoh penggunaan, dan penambahbaikan makna. Pendekatan ini berjaya meningkatkan produktiviti, khususnya dalam penyusunan entri yang bersifat teknikal serta memudahkan penjanaan keterangan awal dalam perbincangan editorial.

Walaupun berbantukan AI, proses penilaian KDP masih berasaskan semakan kualitatif yang memerlukan kepakaran manusia. Semua entri dan subentri KDP diperiksa secara teliti oleh pakar dan pekamus untuk memastikan ketepatan bahasa, kesesuaian konteks, serta pematuhan kepada piawaian tatabahasa DBP. Selain itu, teknik CPA turut diaplikasikan bagi menganalisis struktur sintagmatik dan corak penggunaan korpus bahasa Melayu, sekali gus memastikan keputusan leksikal berpaksikan sumber bukti empirikal, iaitu data sebenar penggunaan bahasa Melayu.

Kejayaan DBP dalam menerbitkan KDP pada tahun 2020 telah berjaya membuktikan bahawa kaedah tradisional bersama-sama teknologi digital dan berbantukan AI boleh digabungkan secara berkesan dalam pembinaan sebuah kamus bahasa Melayu yang bersifat komprehensif. Melalui pengumpulan data secara manual, penyusunan kamus yang terstruktur dengan menggunakan perisian *TLex*, dan pengeditan draf berbantukan AI seperti *ChatGPT*, DBP telah memperlihatkan satu pendekatan yang praktikal dan moden dalam penyusunan sebuah kamus. Evolusi berperingkat ini selari dengan trend global dan menekankan kepentingan kepakaran manusia yang tetap menjadi keutamaan walaupun bidang leksikografi kini semakin bersifat automatik.

Rajah 4.1
Evolusi Kaedah Penyusunan Kamus Dewan Perdana



5. Kesimpulan

Kertas ini telah meneliti perkembangan kaedah, peralatan, dan cabaran yang membentuk bidang leksikografi dan pembangunan peristilahan moden, khususnya dalam konteks kemajuan AI dan perkomputeran linguistik. Berdasarkan analisis terhadap tiga belas kajian yang mewakili pelbagai pendekatan, perbandingan menyeluruh antara kaedah tradisional dan moden, termasuk struktur data yang menyokongnya dan strategi penilaian yang digunakan untuk memastikan kualiti serta kebolehgunaan dapat ditinjau. Tiga soalan penyelidikan menjadi rangka kepada tinjauan ini, menyumbang kepada pemahaman yang lebih jelas tentang arah perkembangan semasa bidang leksikografi.

Pertama, dalam proses membincangkan kaedah tradisional dan perkomputeran, hasil analisis menunjukkan evolusi peralihan bersejarah penyusunan kamus berasaskan kepakaran dan intuisi manusia kepada pendekatan berasaskan korpus dan automasi berbantukan AI. Walaupun kaedah tradisional masih relevan dalam proses memberikan ketelitian linguistik dan kepekaan budaya, pendekatan ini terhad bagi keupayaan memproses data yang berskala besar dalam tempoh yang lebih singkat. Sebaliknya, pendekatan berasaskan perkomputeran dan berbantukan AI seperti analisis pola korpus, pengekstrakan istilah, penjajaran pelbagai bahasa, dan penggunaan model LLM membolehkan penjanaan takrif secara automatik, penentuan makna perkataan, serta peluasan kandungan merentas bahasa. Inovasi ini membuktikan bahawa AI tidak menggantikan peranan seorang pekamus, tetapi memperkasakan keupayaan untuk memproses saiz data yang lebih besar, memperluas liputan bahasa, dan menghasilkan sumber leksikal yang berkualiti tinggi serta lebih tepat.

Kedua, tinjuan ini juga meneroka pelbagai peralatan, korpus, dan data yang menyokong kerja leksikografi bagi kesemua kajian yang telah diteliti. Sumber tersebut termasuk korpus ekabahasa berskala besar seperti data bahasa daripada Internet dan *British National Corpus* (BNC), data pelbagai bahasa seperti *Europarl* dan *mBART50-parallel Wikipedia*; serta pangkalan data leksikal seperti *WordNet*, *FrameNet*, *BabelNet*, dan *EcoLexicon*. *Sketch Engine*, GDEX, *FastText*, *Lexonomy*, serta pelbagai model berasaskan transformer seperti *BERT* dan *mBERT* memainkan peranan penting dalam analisis korpus, pengekstrakan istilah, penjanaan glos, dan pembinaan leksikon dwibahasa. Ketersediaan dan kualiti sumber ini sangat penting untuk membolehkan pembangunan leksikografi berskala besar dan tepat, namun kebanyakan peralatan tersebut masih belum dibangunkan sepenuhnya atau tidak dapat diakses bagi bahasa yang kekurangan sumber.

Ketiga, analisis kertas ini turut menonjolkan cabaran utama dan hala tuju pembangunan leksikografi berbantukan AI. Antara cabaran yang dikenal pasti ialah kekurangan sumber data, kesukaran mengautomasikan tugasan linguistik yang kompleks, kecenderungan bias dan kesilapan fakta dalam model LLM, serta kekurangan piawaian penilaian yang inklusif. Walaupun begitu, masa depan leksikografi bergantung kepada pembangunan pendekatan hibrid yang menggabungkan kekuatan teknologi dengan penilaian pakar linguistik. Usaha pada masa hadapan perlu memberikan tumpuan kepada pembangunan korpus pelbagai bahasa sebagai sumber terbuka, penyesuaian model dengan data yang khusus budaya, dan reka bentuk sistem yang mengekalkan pelibatan manusia dalam proses penilaian. Penekanan perlu diberikan kepada kerjasama antara disiplin, pengurusan data yang beretika, serta reka bentuk AI yang inklusif bagi memenuhi keperluan komuniti bahasa Melayu.

Seiring dengan dapatan sorotan kajian, evolusi penyusunan KDP memberikan contoh yang nyata tentang perubahan metodologi ini. Pada peringkat awal, entri dan subentri KDP dikumpulkan secara manual melalui proses editorial tradisional yang melibatkan bahan bercetak dan perakaman lisan. Namun, dengan kemunculan Internet dan alat pemprosesan digital pada tahun 2000-an, pendekatan tersebut berkembang kepada pendigitalan entri dalam format *MsWord*, yang memudahkan penyuntingan dan penyimpanan tetapi masih tidak mempunyai struktur data leksikal yang sistematik. Pengenalan perisian *TLex* menandakan satu titik perubahan penting dengan membolehkan entri disusun dalam persekitaran terstruktur dan dapat dianotasi dengan tag *XML*. Perkembangan terkini menyaksikan penggunaan alat berbantukan AI seperti *ChatGPT* dalam penyediaan edisi kedua KDP, khususnya untuk cadangan ayat contoh dan pengeditan draf awal. Walaupun kemajuan teknologi ini telah membawa banyak manfaat, proses penilaian dalam KDP kekal berasaskan semakan pakar secara kualitatif, disokong oleh CPA bagi memastikan keputusan leksikal berpaksikan penggunaan data bahasa sebenar.

Keseluruhannya, kertas ini mengesahkan bahawa masa hadapan leksikografi bergantung kepada gabungan yang seimbang antara kepakaran manusia, kaedah tradisional dan inovasi teknologi. Walaupun AI menawarkan keupayaan untuk mempercepat kerja-kerja leksikografi, peranan pekamus tetap penting dan kekal relevan untuk menjamin ketepatan bahasa, integriti budaya, dan nilai pendidikan. Bagi bahasa Melayu, hala tuju pada masa hadapan perlu memberikan tumpuan kepada penggunaan teknologi moden dan pembangunan ekosistem data yang khusus mengikut keperluan bahasa bagi menyokong pembangunan sumber leksikal yang mampan serta bermakna.

7. Rujukan

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P., & Suchomel, V. (2019). Automating dictionary production: A Tagalog–English–Korean dictionary from scratch. *Proceedings of the eLex 2019 Conference*, 805–818.

Bella, G., Helm, P., Koch, G., & Giunchiglia, F. (2023). Towards Bridging the Digital Language Divide. *arXiv*. https://doi.org/10.48550/arXiv.2307.13405

de Schryver, G.-M., & Prinsloo, D. J. (2001). Corpus-based activities versus intuition-based compilations by lexicographers: The Sepedi lemma-sign list as a case in point. *Nordic Journal of African Studies*, *10*(3), 374–398.

de Schryver, G.-M. (2023). Artificial intelligence in lexicography: Charting the uncharted. *International Journal of Lexicography*, *36*(3), 303–327. https://doi.org/10.1093/ijl/ecad014

Domínguez Vázquez, M. J., & Gouws, R. H. (2023). The definition, presentation and automatic generation of contextual data in lexicography. *International Journal of Lexicography*, 36(3), 233–259. https://doi.org/10.1093/ijl/ecac020

Herman, O. (2021). Precomputed word embeddings for 15+ languages. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing* (RASLAN 2021) (pp. 41–46). Tribun EU.

Kamarulzaman, U., Rashidin, R., Jantan, Z. & Md Zain., N. A. (2025). Pola Bahasa Korpus Ekolinguistik: Isu COVID-19. *Jurnal Linguistik*, *29*(1): 145–156.

Kamus Dewan Perdana. (2020). *Dewan Bahasa dan Pustaka*.

Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. (2021). Transforming term extraction: Transformer based approaches to multilingual term extraction across domains. In Findings of the Association for Computational Linguistics: *ACL-IJCNLP 2021* (pp. 3607–3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.316

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the 22nd EURALEX International Congress: Lexicography and Semantics* (pp. 303–312). Znanstvena založba Filozofske fakultete, Univerza v Ljubljani.

Lew, R. (2024). Dictionaries and lexicography in the AI era. In T. Fontenelle (Ed.), *Lexicography and AI* (pp. 23–42). De Gruyter. https://doi.org/10.1515/9783110798081-002

Li, Z., Zeng, J., Qin, Z., Lu, Y., Li, S., Kudo, T., & Neubig, G. (2024). Language Ranker: A metric for quantifying LLM performance across high- and low-resource languages. arXiv preprint arXiv:2404.11553.

Liu, X., Mamat, R., & Husain, S. (2025). A corpus-based critical discourse analysis of approximators in Chinese and Malaysian diplomatic discourse. *Jurnal Linguistik*, 29(2), 120–136.

Lugli, L. (2019). Smart lexicography for low-resource languages: Lessons learned from Buddhist Sanskrit and Classical Tibetan. In Proceedings of eLex 2019.

Malinga, M., Lupanda, I., Wa Nkongolo, M., & van Deventer, P. (2024). A multilingual sentiment lexicon for low-resource language translation using large language models and explainable AI. arXiv Preprint, arXiv:2411.04316. https://doi.org/10.48550/arXiv.2411.04316

Marchisio, K., Saad-Eldin, A., Duh, K., Priebe, C., & Koehn, P. (2022). Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. arXiv preprint, arXiv:2210.14378.

Martelli, F., Procopio, L., Barba, E., & Navigli, R. (2023, November). LexicoMatic: Automatic creation of multilingual lexical-semantic dictionaries. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 820–833). Association for Computational Linguistics.

McKean, E., & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In Proceedings of ASIALEX 2023 (pp. 2–10).

Merx, R., Vylomova, E., & Kurniawan, K. (2024). Generating bilingual example sentences with large language models as lexicography assistants. arXiv preprint arXiv:2401.03182.

Ogilvie, S. (2011). Linguistics, lexicography, and the revitalization of endangered languages. *International Journal of Lexicography*, *24*(4), 389–404. https://doi.org/10.1093/ijl/ecr019

Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*(2), 183–197. doi:10.1017/nlp.2024.33

Pava, J., Uz Zaman, H. B., Meinhardt, C., Friedman, T., Truong, S. T., Zhang, D., Cryst, E., & Marivate, V. (2025, April 22). Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low Resource Language Contexts [White paper]. Stanford HAI.

Pedersen, B. S. (2012). Lexicography in language technology. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson* (pp. 31–46). John Benjamins.

Rogers, M. (2012). Corpus linguistics and lexicography: Context, selection and interpretation. *The Journal of Specialised Translation*, (17), 244–249.

Rundell, M. (2024). Automating the creation of dictionaries: Are we nearly there? *International Journal of Lexicography*, *37*(1), 1–22. https://doi.org/10.1093/ijl/ecad028

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Tarp, S. (2008). The Third Leg of Two-legged Lexicography. *Hermes – Journal of Language and Communication Studies*, *40*, 117–138.

Tarp, S., & Gouws, R. H. (2023). A necessary redefinition of lexicography in the digital age: Glossography, dictionography and implications for the future. Lexikos, 33(Afrilex series 33), 425–447. https://doi.org/10.5788/33-1-1826

Taufiq, M. N., & Aniswal, A. G. (2022). Penentuan kognat kata antara bahasa: Penilaian semula dari sudut etimologi. *Jurnal Linguistik*, *26*(2), 10–25.

Teubert, W. (2013). Lexical norms and their elaboration in the work of Patrick Hanks: A review article. *International Journal of Lexicography*, *26*(4), 437–452. https://doi.org/10.1093/ijl/ect033

Wordnik. (2023). Wordnik's LLM Assistant prototype. https://blog.wordnik.com/wordniks-llm-assistant-experiment

Xu, K., Feng, Y., Li, Q., Dong, Z. (2025). Survey on terminology extraction from texts. *Journal of Big Data, 12*, 29. https://doi.org/10.1186/s40537-025-01077-x

Biodata Penulis

Nurulhuda Mohamad Ali merupakan Perancang Bahasa Gred S10 di Bahagian Perkamusan, Dewan Bahasa dan Pustaka sejak tahun 2012. Beliau memiliki Ijazah Sarjana Muda Pengurusan Teknologi Multimedia dari Universiti Multimedia serta Diploma Penterjemahan Profesional (Kepujian Tinggi) dari Persatuan Penterjemah Malaysia. Beliau secara langsung terlibat dalam penyusunan dan pendigitalan *Kamus Dewan Perdana Edisi Pertama*. Beliau juga turut terlibat dalam penyusunan *Kamus Dewan Perdana Edisi Kedua*. Bidang kepakaran dan penulisan beliau meliputi leksikografi korpus, kecerdasan buatan (AI) dan pemprosesan bahasa tabii (NLP) bagi bahasa Melayu. Sejak tahun 2023, beliau turut berkhidmat sebagai Penasihat Penterjemahan bagi NN Translation Services Sdn. Bhd. Kini, beliau sedang melanjutkan pengajian Sarjana Sains dalam bidang NLP di Universiti Multimedia.

Norihan binti Rosli merupakan seorang Perancang Bahasa Gred S10 di Dewan Bahasa dan Pustaka. Beliau merupakan graduan Ijazah Sarjana Muda Bahasa Gunaan: Bahasa Melayu Komunikasi Profesional dari Universiti Teknologi MARA (UiTM). Beliau telah berkhidmat di Bahagian Perkamusan, Dewan Bahasa dan Pustaka sejak tahun 2015. Sepanjang tempoh perkhidmatannya, beliau telah terlibat dalam penyusunan *Kamus Dewan Perdana* dan merupakan salah seorang editor bagi *Kamus Etimologi Bahasa Melayu Dewan*. Kini, beliau terlibat dalam penyusunan *Kamus Dewan Edisi Kedua* dan *Kamus Sekolah Dewan*. Bidang kepakaran beliau meliputi leksikografi ekabahasa dan etimologi bagi bahasa Melayu.

Fansurina binti Ramli merupakan seorang Perancang Bahasa di Dewan Bahasa dan Pustaka. Beliau merupakan graduan Ijazah Sarjana dalam Kejuruteraan Awam dan Infrastruktur dari IUT Le Havre dan Université Joseph Fourier, Perancis. Beliau memulakan kerjayanya di Dewan Bahasa dan Pustaka pada tahun 2021 dan telah berkhidmat di Bahagian Majalah selama tiga tahun. Pada tahun 2024, beliau mula bertugas di Bahagian Perkamusan di bawah Unit Dwibahasa. Kini, beliau terlibat dalam penyusunan *Kamus Perancis–Melayu Dewan Edisi Kedua* dan *Kamus Melayu–Jerman Dewan*. Bidang pengkhususan beliau meliputi leksikografi dwibahasa yang melibatkan penyusunan kamus bahasa asing kepada bahasa Melayu.