Kajian Linguistik Bahasa Melayu Bersumberkan Korpus. Kekuatan dan Kelemahannya

Hasmidar Hassan
Universiti Brunei Darussalam, Brunei
hasmidar.hassan@ubd.edu.bn

Published: 12 Jun 2023

To cite this article (APA): Hassan, H. (2023). Kajian Linguistik Bahasa Melayu Bersumberkan Korpus. Kekuatan dan Kelemahannya. *PENDETA*, *14*(1), 23–34. https://doi.org/10.37134/pendeta.vol14.1.fa.2.2023

To link to this article: https://doi.org/10.37134/pendeta.vol14.1.fa.2.2023

Abstrak

Korpus merupakan sumber yang amat bernilai sebagai sumber data dalam kajian linguistik bahasa Melayu kini. Sejak akhir 90-an, korpus mula dimanfaatkan oleh ramai pengkaji linguistik bahasa Melayu dalam kajian peringkat sarjana dan doktor falsafah. Terkini, Kamus Dewan Perdana (2020) juga memanfaatkan korpus dan laman sesawang sebagai sumber data. Data raya daripada pelbagai sumber ini membolehkan pelbagai data diperoleh dan dimanfaatkan untuk tujuan kajian linguistik bahasa Melayu dalam tempoh yang singkat. Namun, penggunaan korpus sebagai sumber data menuntut pendekatan, strategi dan matlamat yang betul bagi membolehkan dapatan yang diperoleh empiris dan boleh dipertanggungjawabkan. Data korpus yang dimanfaatkan bukanlah satu jaminan bahawa hasil atau dapatan yang diperoleh merupakan dapatan yang sahih yang tidak boleh dipertikaikan dengan alasan 'itulah bentuk bahasa yang digunakan oleh penutur jati bahasa Melayu' dan 'data korpus merupakan data autentik yang dikutip berdasarkan konteks penggunaan yang sebenar'. Penggunaan korpus sebagai sumber data dalam kajian tatabahasa pula menjadi semakin kompleks khususnya yang melibatkan pembentukan rumus dalam bidang morfologi atau sintaksis. Oleh itu, makalah ini diketengahkan bagi membincangkan dengan perinci tentang keupayaan dan kekangan korpus sebagai sumber data kajian linguistik bahasa Melayu agar keghairahan pengkaji menggunakan korpus dalam kajian linguistik bahasa Melayu dapat diwajarkan.

Kata kunci: korpus, linguistik, bahasa Melayu, kekuatan, kelemahan

Abstract

Corpus is a very valuable resource in the linguistic study of the Malay language. Since the late 90s, the corpus began to be used by many Malay linguistics researchers in their master's and doctoral studies. Most recently, the Kamus Dewan Perdana (2020) also utilizes corpus and websites as data sources. This immense amount of data from various sources allows a variety of data to be obtained in a short period of time and be utilized for the purpose of researching the Malay language. However, the use of corpus as data requires the right approach, strategy, and goals to enable empirical and accountable findings. The corpus data used is not a guarantee that the results or findings obtained are authentic findings that cannot be disputed on the grounds that 'that is the form of the language used by native Malay speakers' and 'corpus data is authentic data collected based on the actual context of use'. The use of corpus as a data source in the study of grammar is becoming increasingly complex, especially those involving the formation of rules in the field of morphology or syntax. Therefore, this paper aims to discuss in detail the capabilities and the limitations of the corpus as a source of data for the study of Malay linguistics so that the enthusiasm of researchers to use the corpus in the study of Malay linguistics can be justified.

Keywords: *corpus, linguistics, Malay language, strength, weaknesses*

PENGENALAN

Kajian linguistik bahasa Melayu yang memanfaatkan korpus sebagai sumber data telah memberikan impak yang sangat besar dan merevolusikan penyelidikan dalam pelbagai subbidang linguistik bahasa Melayu. Penggunaan korpus sebagai sumber data merupakan satu pendekatan yang amat bernilai kerana sifatnya yang autentik dan keupayaan korpus menyediakan data yang banyak dalam masa yang singkat. Menyedari sumbangan dan kelebihan korpus terhadap penyelidikan linguistik bahasa Melayu, semakin ramai penyelidik di negara Malaysia yang memanfaatkan korpus sama ada korpus yang tersedia di pangkalan data Dewan Bahasa dan Pustaka (http://sbmb.dbp.gov.my/korpusdbp/), dan *Malay Concordance Project* (https://mcp.anu.edu.au/) ataupun korpus yang dijana sendiri dengan menggunakan perisian seperti *Antconc*. Di sebalik peningkatan penyelidikan berasaskan korpus di Malaysia, perlu juga dibincangkan sejauh manakah keupayaan dan kesahan korpus yang ada terhadap penyelidikan subbidang linguistik di Malaysia.

Latar Belakang Perkembangan Korpus[1]

Penggunaan koleksi atau himpunan teks dalam kajian linguistik bukanlah satu idea baru pada hari ini. Kerja–kerja menyusun dan menyenaraikan semua perkataan dalam teks yang tertentu mengikut konteks telah dimulakan sejak 1920-an lagi. Hari ini kerja–kerja sebegini dikenali sebagai '*concordancing'*. Senarai kekerapan kata daripada teks atau daripada himpunan teks dihasilkan dan digunakan untuk kajian pemerolehan bahasa, sintaksis, semantik dan linguistik perbandingan. Walaupun istilah linguistik korpus tidak digunakan pada masa itu, namun kajian yang dilakukan pada masa itu menyamai kajian berasaskan korpus pada hari ini, kecuali pada masa itu pengkaji bahasa tidak menggunakan komputer.

Menurut Rundell (1996), kini korpus telah bergerak dari margin pinggiran untuk mengisi posisi utama sebagai sumber kajian bahasa. Hal ini menunjukkan perubahan dalam dua aspek iaitu, ideologi (perubahan besar ke arah kajian bahasa secara empiris dan statistik) dan juga dalam teknologi, yang kini mampu mencapai kuasa memproses yang begitu hebat dan mengagumkan dan kuasa penyimpanan data yang banyak pada kos yang rendah .

Hari ini, linguistik korpus berkait rapat dengan penggunaan komputer, dan hakikat sebenarnya ialah istilah 'Linguistik Korpus' bagi kebanyakan sarjana bermaksud 'penggunaan koleksi teks boleh baca berasaskan komputer untuk tujuan kajian bahasa'. Dengan kata lain, linguistik korpus ini hanya membawa cara baru dalam memperlakukan data bahasa, dan caranya diperkuatkan dengan teknologi pengkomputeran (Asmah Hj Omar 2005).

Korpus pada zaman moden ini, yang pertama berbentuk elektronik boleh baca ialah '*Brown Corpus of Standard American English'* yang diusahakan oleh Francis dan Kucera (1979). Kerja–kerja untuk mengumpulkan teks ini mengambil masa hampir dua dekad untuk dilengkapkan. Korpus ini mengandungi satu juta perkataan berdasarkan teks bahasa Inggeris Amerika yang dicetak pada tahun 1961. Untuk menjadikan korpus ini lebih mudah dirujuk, ia dibahagikan kepada beberapa bahagian seperti laporan akhbar, kemahiran dan hobi, agama, saintifik, fiksyen dan sebagainya. Kini, korpus ini dianggap sangat kecil dan sudah ketinggalan zaman. Namun manfaatnya cukup besar kerana susun atur korpus ini telah disalin semula oleh penyusun korpus lain. LOB**,** *Lancaster-Oslo-Bergen*, korpus (bahasa Inggeris British) dan *Kolhapur Corpus* (bahasa Inggeris Indian) merupakan dua contoh korpus yang dibuat sepadan dengan korpus Brown. Kedua–duanya mengandungi 1 juta kata bahasa bertulis (500 teks, 2,000 kata setiap teks) yang dibahagikan kepada 15 kategori seperti korpus Brown.

Menjelang pertengahan 1970-an, korpus lain, misalnya *Birmingham Collection of English Texts* (BCET) telah menghimpunkan 7.3 juta dan bertambah menjadi 20 juta kata pada tahun 1985. Perkembangan korpus yang lebih besar kelihatan berlaku pada awal tahun 1990-an, pembinaan *British*

[1] Korpus di sini bermaksud 'korpus komputer' , iaitu 'himpunan teks digital yang dikumpulkan berdasarkan kriteria tertentu' (Rusli Abdul Ghani, 2004: 1).

National Corpus merupakan peristiwa yang cukup signifikan. Korpus ini telah berjaya menyediakan himpunan pangkalan data korpus yang jauh lebih besar, merangkumi teks tulisan dan lisan sebanyak 100 juta kata yang telah dilengkapkan pada tahun 1994 (Sinclair 1991).

Di Malaysia, tugas menghimpunkan teks untuk tujuan kajian bahasa diusahakan oleh pihak Dewan Bahasa dan Pustaka. Himpunan teks yang melibatkan pembangunan pangkalan data bermula pada tahun 1983 di bawah Projek Analisis Teks Secara Berkomputer (Zaiton Ab. Rahman, 1987). Projek ini menyasarkan data teks sebesar dua juta kata melalui teknik pensampelan ala korpus Brown. Bagaimanapun, apabila saiznya belum pun mencecah setengah juta, kriteria pensampelan diabaikan dan teks lengkap mula dikumpulkan untuk mengambil kira keperluan perkamusan dan kajian bahasa yang memerlukan konteks yang lebih luas dan wacana yang utuh (Rusli Abdul Ghani, 2004).

Menurut Rusli (2004), objektif pembinaan pangkalan data korpus yang digariskan dalam Sasaran Kerja Utama DBP 2001-2005 ialah pengumpulan sebanyak 30 juta kata, dan menjadikan jumlah kumulatifnya sebanyak 120 juta kata pada tahun 2005 (selain itu, di bawah program pembinaan sistem korpus, sebuah sistem korpus yang baru akan dibina sebagai ganti sistem sedia ada yang dibina melalui kerjasama dengan Universiti Sains Malaysia pada tahun 1994). Data korpus ini terdiri daripada teks bertulis yang merangkumi teks Melayu lama (daripada hikayat dan kitab) dan teks moden yang diambil terutamanya daripada sumber buku, akhbar, dan majalah. Korpus lisan masih dalam perancangan kerana penandaan yang diperlukan untuk korpus lisan jauh lebih rumit daripada korpus tulisan dan tidak mampu ditangani pada masa ini.

Tambah beliau, pembinaan pangkalan data korpus DBP seperti pembinaan pangkalan data korpus lain juga, bertujuan menyediakan suatu prasarana penelitian yang objektif dan autentik sifatnya kepada para penyelidik bahasa Melayu supaya dapatan yang diperoleh daripada kajian berdasarkan korpus ini dapat mencerminkan perilaku tipikal kata dan frasa bahasa Melayu dalam persekitaran penggunaannya yang sebenar dan dapat pula dijadikan asas untuk penyusunan kamus, tatabahasa dan buku–buku bahasa yang lainnya.

Kini kiraan mutakhir data teks yang terkumpul dalam pangkalan data DBP sudah pun melebihi 137 juta kata dengan enam subkorpus iaitu buku, majalah, akhbar, efemeral, teks tradisional dan kertas kerja. Daripada enam subkorpus ini, didapati bahawa bahan akhbar merupakan komponen yang paling besar yang terdapat dalam data korpus. Bahan korpus akhbar ini termasuklah Berita Minggu, Metro Ahad, Pedoman rakyat, Harian metro, Berita Harian, Harakah dan Utusan Malaysia. Oleh itu langkah awal yang perlu dilakukan ialah meneliti dan menghuraikan data yang besar ini supaya apa–apa kajian yang dilakukan dan sebarang dapatan bukan saja sahih dalam batas cakupan data yang dikaji malah boleh mewakili penggunaan sebenar bahasa Melayu.

Definisi Korpus

Secara umumnya, korpus boleh ditakrifkan sebagai "himpunan makalah (tulisan dan sebagainya) mengenai sesuatu perkara tertentu atau kumpulan bahan untuk kajian (seperti kumpulan contoh penggunaan kata dan lain–lain)"

(Kamus Dewan Edisi Keempat, 2010)

Bagaimanapun, dalam konteks linguistik moden, korpus mempunyai pengertian tambahan sebagai bahan yang "terbacakan dan terolahkan komputer."

Perkataan korpus sebenarnya diterbitkan daripada perkataan Latin yang bermaksud 'badan', yang boleh digunakan untuk merujuk sebarang bentuk teks bertulis atau lisan. Bagaimanapun, dalam konteks linguistik moden ini, istilah ini digunakan untuk merujuk koleksi ataupun himpunan teks yang banyak, yang mewakili satu sampel kepelbagaian atau penggunaan yang tertentu sesuatu bahasa yang disampaikan dalam bentuk mesin boleh baca.

Mengikut *The Oxford Companion to the English Language*, (1992), korpus didefinisi sebagai koleksi teks, lisan dan/atau bertulis yang disusun atur berdasarkan satu set kriteria yang jelas dan tentu. Perkataan korpus ini berasal daripada perkataan Latin yang bermaksud 'badan'. Dalam bentuk jamaknya, biasanya disebut korpora. Dalam kajian linguistik dan leksikografi, sesuatu bentuk teks, ujaran, atau spesimen lain dianggap sebagai lebih kurang mewakili sesuatu bahasa, dan biasanya disimpan dalam bentuk pangkalan data elektronik. Kini, korpus komputer boleh menyimpan berjuta–juta perkataan, yang cirinya boleh dianalisis dengan cara penandaan atau pencirian (tambahan pengenalpastian dan penggolongan perkataan dan pembentukan lain) dan penggunaan program konkordans. Korpus linguistik mengkaji data sebarang korpus.

Crystal (1991) pula mendefinisi korpus sebagai koleksi data linguistik, sama ada teks bertulis atau transkripsi ucapan yang dirakamkan yang boleh digunakan sebagai titik permulaan pemerian linguistik atau sebagai cara mengesahkan hipotesis tentang bahasa. Hal yang hampir serupa turut dinyatakan oleh Cook (2003) yang mendefinisi korpus sebagai "*the systematic analysis and description of extensive databanks of language which has actually occurred in use*",

Sinclair (1991) dengan ringkasnya mendefinisi korpus sebagai koleksi teks bahasa yang terhasil secara semula jadi, dipilih untuk menandakan kepelbagaian bahasa.

Berdasarkan definisi yang diberikan di atas, maka korpus dapatlah ditakrifkan sebagai koleksi atau himpunan teks yang dikumpulkan mengikut prinsip–prinsip yang tertentu untuk tujuan yang tertentu. Korpus ini amat bernilai kerana komponennya membolehkan suatu kenyataan yang menyeluruh dibuat terhadap sesuatu bahasa. Kenyataan pengkategorian juga boleh dilakukan kerana korpus ini disusun dengan teratur mengikut kategori yang tertentu.

Rasional Penggunaan Data Korpus dalam Kajian Linguistik Bahasa Melayu

Secara amnya, ramai pengkaji dan sarjana berpandangan bahawa penggunaan data korpus dalam kajian linguistik ini adalah perlu kerana dapatan dan kesimpulan yang dibuat akan bersifat objektif dan empiris. Kajian tanpa sumber yang sahih dan nyata, yang dapat mewakili penggunaan bahasa penutur asli bahasa Melayu, tentunya akan mengimplikasikan dapatan yang tidak sahih dan tidak empiris juga. Hal ini sejajar dengan pendapat para sarjana sebelum ini, misalnya Leech (1992) (dlm. McEnery dan Wilson 1996) yang menjelaskan bahawa kepentingan korpus dalam kajian bahasa sejajar dengan kepentingan data yang empiris. Data yang empiris membolehkan pengkaji bahasa membuat kenyataan yang objektif, dan bukannya kenyataan yang bersifat subjektif atau yang berdasarkan persepsi kognitif dalaman seseorang individu terhadap bahasa. Jelas Rundell (1996: 4) pula, satu fakta penting yang perlu diingat tentang korpus ini ialah, nilai sebarang korpus sebagai sumber data linguistik berkait rapat dengan kandungannya. Dengan kata lain, generalisasi yang dibuat tentang bahasa berasaskan bukti korpus adalah sebaik bukti korpus yang digunakan pengkaji (ertinya input yang baik akan menghasilkan output yang baik juga).

Selain itu, kajian bahasa yang berasaskan korpus membolehkan pengkaji mengenal pasti dan menjelaskan kepelbagaian produktiviti bahasa, ataupun *'permissble variety'* istilah yang diberikan oleh Sinclair. Contohnya, kajian bagi kata nama X yang pertama dalam struktur a (n) X bagi Y membolehkan pengkaji menyingkap peluang–peluang yang produktif dalam bahasa (1996).

Aston (dlm**.** Wichmann 1997) pula menjelaskan kepentingan korpus dalam kajian bahasa dari segi pengetahuan skema. Menurut beliau kajian bahasa yang berasaskan korpus membolehkan pengkaji menyingkap pelbagai aturan tentu dalam wacana dan ruang lingkup yang boleh dikaji dengan gabungan yang tetap dan separa tetap. Selain itu, korpus juga membolehkan kajian perbezaan antara gabungan sintagmatik pada tahap makna (maklumat/struktur retorik) dan bentuk (kolokasi/koligasi/semantik/pragmatik/prosodik) dan gabungan paradigmatik antara situasi dan makna (genre dan laras), antara makna dan bentuk (lakuan tuturan yang lazim dan prosedur rujukan) dan antara situasi dan bentuk (formula rutin, terminologi teknikal) dilakukan .

Sinclair (dlm. Wichmann et al. 1997: 35-36) menjelaskan bahawa korpus juga menyajikan sumber yang besar terhadap kajian makna atau semantik. Tambah beliau, kajian makna berasaskan korpus membolehkan perbezaan bentuk dan makna dihuraikan dengan lebih terperinci. Makna mempunyai kesan yang sangat besar terhadap struktur. Jika satu perkataan mempunyai dua makna, maka bolehlah dijangkakan bahawa kata itu juga mempunyai sekurang–kurangnya dua struktur. Hal ini hanya dapat dijelaskan dengan mengkaji contoh penggunaan bahasa.

Korpus juga membolehkan satu pendekatan yang bersifat objektif dilakukan dalam kajian semantik, dengan mengambil kira ketaktentuan dan ketaksejajaran. Mindt (1991) telah menunjukkan bagaimana korpus boleh digunakan untuk menyediakan kriteria yang objektif dalam menentukan makna pada istilah linguistik. Beliau telah mengetengahkan persoalan tentang 'bagaimana makna istilah kerap kali dihuraikan dengan merujuk intuisi ahli bahasa itu sendiri iaitu, pendekatan yang paling rasional sedangkan perbezaan semantik yang tergabung dalam teks dengan konteks yang khusus – sintaksis, morfologi dan prosodi, dan dengan mengambil kira persekitaran entiti linguistik, satu penanda objektif yang empiris bagi perbezaan semantik yang tertentu boleh diperhatikan dan dijelaskan.

Korpus juga memainkan peranan yang penting dalam kajian bidang pragmatik. Menurut Myers (dlm. McEnery dan Wilson, 2001), korpus merupakan sumber kajian yang penting dalam bidang pragmatik kerana bidang ini bergantung pada aspek konteks dan jika sampel yang digunakan hanya sebahagian kecil daripada korpus, maka ia tidak akan menggambarkan konteks sosial dan konteks teks yang sebenar. Oleh yang demikian, dengan adanya kumpulan data korpus yang semakin banyak kini, maka diharapkan kajian dalam bidang pragmatik yang bersifat analisis kuantitatif berdasarkan korpus akan menarik minat pengkaji bahasa untuk menerokainya.

Selain bidang semantik, pragmatik dan bidang–bidang yang dijelaskan sebelum ini, korpus juga memberikan sumbangan yang besar terhadap kajian tatabahasa. Kajian ketatabahasaan misalnya kajian sintaksis, dan kajian leksikal, merupakan kajian yang paling kerap menggunakan korpus. Korpus telah menjadi sumber yang paling berguna bagi kajian sintaksis kerana korpus berpotensi memberikan representatif kepelbagaian bahasa yang tinggi kuantitinya dan peranan korpus itu sendiri yang bersifat data empiris bagi ujian hipotesis yang diperoleh daripada teori tatabahasa. Antara kajian tatabahasa yang melibatkan analisis data kuantitatif bersumberkan korpus dilakukan oleh Schmied (1993), yang mengkaji klausa relatif dan kajian yang lebih sistematik tentang kekerapan ketatabahasaan oleh Oostdijk dan de Haan (1994a) yang bertujuan menganalisis kekerapan pelbagai jenis klausa dalam bahasa Inggeris (McEnery & Wilson 1991).

Dalam kajian ini khususnya, korpus merupakan sumber kajian yang amat sesuai memandangkan korpus ini dapat menyediakan sumber data yang banyak dan dapat diperoleh dalam masa yang singkat.

Berdasarkan huraian di atas, ternyata bahawa korpus merupakan sumber data yang cukup penting untuk dimanfaatkan oleh pelbagai bidang dalam kajian linguistik. Hal ini disebabkan oleh beberapa faktor yang didukung oleh korpus itu sendiri, dan faktor–faktor tersebut dapat disimpulkan seperti berikut :

i. Percontohan dan pembilangan. Korpus dijadikan sampel kerana dapat mewakili satu populasi dan sebarang dapatan yang diperoleh daripada korpus boleh digeneralisasikan pada populasi yang lebih besar. Oleh yang demikian pembilangan dalam linguistik korpus lebih bermakna daripada bentuk-bentuk lain dalam pembilangan linguistik kerana ia dapat menjelaskan kepelbagaian bahasa dan tidak terhad pada aspek yang sedang dianalisis. Korpora atau korpus menawarkan sampel perwakilan bagi bahasa atau ragam bahasa tertentu dan ahli bahasa boleh menganalisis korpus untuk memerhati corak, aliran dan variasi dalam penggunaan bahasa, yang membolehkan penemuan yang lebih tepat dan komprehensif.

ii. Mudah untuk dicapai dan digunakan. Kebanyakan korpus sudah tersedia oleh pihak yang bertanggungjawab terhadap pengembangan bahasa, sama ada badan kerajaan mahupun pihak universiti dan disediakan sama ada pada harga yang rendah atau percuma. Apabila korpus telah diperoleh, biasanya data korpus ini mudah digunakan dengan menggunakan program konkordans.

iii. Data yang diperluas. Banyak korpus yang telah diperluas dengan maklumat linguistik tambahan seperti ulasan bahagian pertuturan, pecahan–pecahan ayat dan transkripsi prosodi. Oleh itu, data yang diperbaik daripada korpus yang diberikan catatan tambahan lebih mudah dan lebih spesifik daripada data yang tidak diberikan catatan tambahan.

iv. Data yang sebenar/autentik. Bagi kebanyakan bahagian, data korpus ialah data raya yang sebenar dan realistik, tidak dipantau dan menggambarkan penggunaan bahasa dalam konteks sosial yang sebenar. Oleh sebab itulah korpus menyediakan sumber data yang digunakan oleh penutur bahasa Melayu secara semula jadi dan merangkumi pelbagai laras, dialek dan gaya bahasa untuk dikaji.

v. Analisis Kuantitatif: Korpus atau korpora membolehkan penyelidik meneliti fenomena linguistik secara besar-besaran berdasarkan pendekatan kuantitatif. Mereka boleh menjalankan analisis statistik, mengukur frekuensi dan mengenal pasti corak yang mungkin tidak kelihatan dalam set data yang lebih kecil atau anekdot.

vi. Kajian Jangka Panjang: Korpora boleh disusun dalam tempoh yang panjang, memboleh kajian jangka panjang dapat dilakukan. Keadaan ini membolehkan ahli bahasa menyelidiki perubahan bahasa, evolusi, dan corak diakronik, menjejaki perkembangan linguistik dari semasa ke semasa.

vii. Pengujian Hipotesis: Penyelidikan berasaskan korpus membolehkan pengkaji bahasa menguji hipotesis dan mengesahkan teori secara empiris. Mereka boleh membandingkan pemerhatian mereka dengan data daripada korpus untuk menentukan kesahihan dakwaan mereka, yang menyebabkan kesimpulan yang lebih mantap dan boleh dipercayai dapat dibuat.

Kelemahan Korpus dalam Kajian Linguistik Bahasa Melayu

Dalam keghairahan kita memanfaatkan korpus dalam kajian dan penyelidikan linguistik bahasa Melayu, kita harus menyedari beberapa kelemahan korpus dan berusaha mengatasinya dengan sokongan pendekatan dan prosedur lain bagi memantapkan kajian dan dapatan kajian.

Secara amnya, kelemahan korpus sebagai sumber kajian linguistik berdasarkan hakikat bahawa wujudnya hal seperti berikut :

i. Perwakilan yang Bias : Walaupun terdapat usaha untuk mewujudkan korpus yang dapat menjadi perwakilan bagi satu set data yang besar, kemungkinan korpus ini masih berat sebelah. Demografi, kumpulan sosial atau wilayah geografi tertentu mungkin kurang diwakili, yang menyebabkan penemuan yang tidak lengkap atau bias .

ii. Kekurangan Konteks: Korpus biasanya membentangkan data bahasa secara berasingan, tanpa konteks penuh situasi komunikatif kecuali pengkaji menjana sendiri korpus tersebut daripada satu teks dan memanfaatkan korpus tersebut sebagai sumber data. Ketiadaan maklumat kontekstual ini mungkin mengehadkan pemahaman penggunaan bahasa dan tafsiran fenomena linguistik tertentu yang menyeluruh.

iii. Kesukaran untuk Mengakses Fenomena yang luar biasa atau terpencil: Fenomena linguistik yang luar biasa atau jarang berlaku mungkin sukar untuk dijana daripada korpus kerana kekerapan kejadiannya yang rendah. Ahli bahasa mungkin perlu menggunakan kaedah alternatif, seperti temu bual atau teknik pengumpulan data khusus untuk mengkaji fenomena tersebut. Dengan kata lain, walaupun korpus merupakan data raya, namun masih ada bentuk bahasa yang mewakili penutur tertentu yang tidak dicakupi.

iv. Had Pengumpulan Data: Penyusunan korpus boleh memakan masa dan intensif sumber. Proses pengumpulan, mengisih dan menganotasi sejumlah besar data memerlukan usaha

dan kepakaran yang ketara. Selain itu, untuk mendapatkan persetujuan bagi penggunaan data dan memastikan privasi data boleh menimbulkan cabaran etika.

v. Kepelbagaian Linguistik: Penggunaan bahasa sangat berubah-ubah, dipengaruhi oleh pelbagai faktor seperti konteks, perbezaan individu dan faktor sosial. Korpus mungkin tidak memberikan rangkaian penuh variasi linguistik, mengehadkan kebolehgeneralisasian penemuan pada konteks atau populasi yang berbeza.

Kelemahan korpus dalam kajian linguistik yang disenaraikan secara am ini mungkin kelihatan lebih jelas jika dibincangkan mengikut subbidang linguistik seperti berikut :

Isu dalam Kajian Morfologi dan Sintaksis.

Penggunaan korpus dalam kajian morfologi merupakan satu pendekatan yang baik kerana data raya yang terjana dan terhimpun dalam pangkalan data dapat dimanfaatkan untuk kajian tentang penggunaan imbuhan, variasi imbuhan, pembentukan kata, binaan ayat dan ragam ayat yang digunakan oleh pengguna bahasa Melayu. Selain itu, korpus juga dapat memperlihatkan variasi penyebaran dan kekerapan struktur ayat merentas pelbagai konteks.

Namun, suka diingatkan bahawa bukan semua data ini terlahir daripada pengguna yang sempurna penguasaan bahasanya dan bukan semua bentuk kata yang digunakan oleh penutur bahasa Melayu yang dijana dan terhimpun dalam pangkalan data korpus itu dapat mewakili bentuk yang baku dan bermakna seperti lazimnya. Dengan kata lain, bahasa yang termuat dalam korpus itu mungkin juga terhasil daripada penutur bukan Melayu atau bukan penutur jati bahasa Melayu. Oleh itu, sekiranya pengkaji menemukan dapatan baharu, perlu juga diingatkan bahawa apa-apa yang ditemui itu bersifat khusus konteks dan tidaklah mewakili keseluruhan set data yang lengkap. Penggunaan korpus tidak dinafikan merupakan sumber data yang sangat membantu pengkaji membuat penyelidikan dan analisis yang lebih lengkap dan menyeluruh tetapi dapatan itu masih terhad pada parameter skop kajian dan bukanlah dapatan yang dapat digeneralisasikan. Sekiranya imbuhan, bentuk kata atau ayat yang ditemui itu jelas salah atau berbeza dari segi rumus atau tatabahasa dan maknanya yang sedia ada, walaupun penggunaannya agak kerap dalam korpus yang dijana (yang berkemungkinan dijana daripada teks oleh penutur yang sama atau bukan penutur jati), apakah bentuk itu harus diterima dengan alasan bahawa korpus itu data yang terhasil daripada penggunaan bahasa dalam konteks yang sebenar? Sekiranya pandangan ini diterima, pastinya kebolehterimaan dapatan itu bersifat sementara juga kerana dalam tempoh beberapa tahun akan datang, penggunaan bahasa oleh penutur mungkin akan berubah juga. Kita sedia maklum bahawa bahasa ini bersifat dinamik dan penggunaan bahasa sangat dipengaruhi oleh faktor persekitaran. Perubahan ini berlaku secara diakronik dan juga secara sinkronik. Oleh itu, apa juga dapatan pengkaji dalam bidang morfologi dan sintaksis yang bersumberkan korpus dapat dianggap sebagai variasi (pada struktur permukaan) dan bentuk serta struktur yang sempurna atau yang menjadi dasar sama ada morfem atau binaan frasa dan ayat diyakini masih kekal sama. Perlu diingat juga bahawa ketika kajian dan penelitian tatabahasa dibuat oleh Za'ba, Asmah Hj Omar, Nik Safiah et al., dan pengkaji lain yang hampir sezaman mereka, data yang menjadi sumber kajian tidak rencam. Dengan kata lain, persekitaran dan aktiviti manusia yang wujud ketika itu tidak mewujudkan laras yang rencam seperti zaman ini. Maka, yang kian banyak berubah berkemungkinan dari aspek leksikal dan istilah yang muncul akibat bidang yang kian berkembang dan pengaruh struktur ayat dan kata bahasa asing juga.

Dalam kebanyakan kajian tatabahasa yang memanfaatkan korpus, didapati bahawa kebanyakan pengkaji sebenarnya tidak memberikan dapatan yang bertentangan dengan rumus yang sedia ada, sebaliknya dengan cara yang cukup berhemah, para pengkaji mengemukakan dapatannya sebagai maklumat tambahan pada rumus atau tatabahasa yang sedia ada berdasarkan pendekatan dan teori yang terkini.

Isu dalam Kajian Semantik dan Pragmatik

Penggunaan korpus dalam kajian bidang semantik dan pragmatik ternyata memberikan manfaat yang luar biasa kepada pengkaji kerana sifat korpus itu yang dapat memperlihatkan kekerapan dan penggunaan bahasa mengikut konteks dan situasi yang tertentu. Namun, perlu diingat juga bahawa data yang diperoleh itu sangat bersifat khusus penutur dan bentuk bahasa yang digunakan dalam konteks komunikasi atau dalam bentuk perbualan merupakan bentuk bahasa yang pragmatik. Dengan kata lain, ujaran yang disampaikan itu bukanlah bentuk yang sempurna, ujarannya mengalami banyak pengguguran dan makna ujarannya juga adakala bersifat taksa.

Menurut Myers (dlm. McEnery dan Wilson 2001), korpus merupakan sumber kajian yang penting dalam bidang pragmatik kerana bidang ini bergantung pada aspek konteks. Namun, jika sampel yang digunakan hanya sebahagian kecil daripada korpus, maka ia tidak akan menggambarkan konteks sosial dan konteks teks yang sebenar.

Kecenderungan kajian pragmatik yang memfokuskan kesantunan, yang memanfaatkan korpus yang dijana daripada novel atau cerpen merupakan kaedah yang tidak betul kerana bahasa yang santun yang termuat dalam korpus ialah rekaan dan datangnya daripada penulis novel itu sendiri. Ternyata banyak kajian linguistik bahasa Melayu yang mengkhusus kajian kesantunan terlepas pandang isu ini. Yang perlu diselidiki ialah ungkapan yang santun atau tidak santun yang disampaikan oleh penutur sebenar dan bukannya ungkapan yang termuat dalam karya rekaan (novel atau cerpen). Hal yang perlu diselidiki ialah apakah faktor yang mencetus penggunaan bahasa yang tidak santun dan apakah strategi kesantunan yang digunakan oleh penutur sebenar dalam konteks menjaga atau mengancam air muka pendengar. Oleh itu, korpus yang dimanfaatkan mestilah korpus yang dapat memberikan maklumat konteks yang penuh daripada satu wacana atau perbualan yang lengkap oleh penutur yang sebenar dan bukannya data yang terpisah-pisah (sekiranya korpus daripada pelbagai genre dimanfaatkan).

Oleh itu, dalam keghairahan menggunakan korpus sebagai sumber kajian, para pengkaji perlu sedar bahawa ketika mereka memanfaatkan korpus dalam kajian pragmatik, makna yang diperoleh sangat dipengaruhi oleh konteks dan dipengaruhi oleh hajat penutur juga (*speaker's intention*). Dengan kata lain, dapatan daripada kajian pragmatik yang memanfaatkan korpus sebenarnya merujuk maksud penutur dan bersifat sangat khusus penutur dan konteks dan tidak boleh digeneralisasikan.

Isu dalam Penyusunan Kamus Dewan Perdana Bahasa Melayu

Penggunaan korpus dalam penyusunan kamus bukanlah suatu pendekatan baharu. Seawal tahun 80-an, Dewan Bahasa dan Pustaka sudah mula membina pangkalan data korpus untuk dimanfaatkan dalam penyusunan kamus. Terkini, Kamus Dewan Perdana juga memanfaatkan korpus dalam merakam dan menyenaraikan kosa kata yang digunakan oleh penutur bahasa Melayu (termasuk dialek dan kata pinjaman).

Dalam usaha untuk menyusun kamus, kata atau leksikal ini diperoleh daripada korpus yang bahasanya digunakan dalam konteks rasmi mahupun tidak rasmi (contohnya bahasa basahan, bahasa pasar, dialek, slanga, dan sebagainya). Korpus memudahkan usaha pengkaji menjana data dan mengisih data yang banyak dalam masa yang singkat. Korpus juga membolehkan variasi penggunaan perkataan yang merentas daftar, dialek atau domain yang berbeza dikenal pasti. Hal ini membolehkan kamus memasukkan berbilang makna, sinonim atau bentuk perkataan alternatif, memastikan pengguna boleh memahami dan menggunakan bahasa dalam pelbagai konteks.

Menyedari hakikat korpus itu dijana daripada sumber yang pelbagai, iaitu bentuk bahasa yang baku dan tidak baku, maka kamus haruslah dimengerti sebagai 'buku' yang mengumpulkan dan merakam segala bentuk kata berserta makna kamusnya yang digunakan oleh penutur bahasa Melayu dan bukanlah buku rujukan tentang tatabahasa bahasa Melayu. Malahan, Kamus Dewan Perdana turut

merakam kosa kata pinjaman dan kosa kata dialek dan kosa kata daripada bahasa serumpun, khususnya bahasa Indonesia.

Walaupun penyusunan Kamus Dewan Perdana bertujuan untuk merekodkan kata yang digunakan dalam dunia Melayu dan etimologi setiap kata tersebut, aspek tatabahasa juga turut termuat dalam KDP khususnya bagi menunjukkan golongan kata setiap entri atau leksikalnya, makna dan contoh penggunaannya dalam ayat yang dipetik daripada pelbagai laras, genre dan sumber.

Namun, timbul beberapa persoalan yang terdapat dalam KDP (2020) yang menuntut penjelasan pihak yang terlibat dalam penyusunan KDP ini. Berdasarkan tinjauan sepintas lalu, satu entri yang menimbulkan persoalan ialah entri 'tersepakkan'[2]. KDP menggolongkan kata 'tersepakkan' sebagai kata kerja transitif (kkt) sama seperti 'tersepak' (kkt). Entri ini bagaimanapun tidak ditemui dalam KD Edisi Keempat kecuali kata akarnya sahaja, iaitu 'sepak'. Dalam KDP, makna kamus atau takrifan yang diberikan bagi 'tersepak' dan 'tersepakkan' adalah seperti berikut :

tersepak	kt menyepak (sso atau sst) dengan tidak sengaja: Seorang lelaki lari menyelamatkan diri selepas tersepak sebutir bom. Tiba-tiba kaki Minah tersepak tunggul.
tersepakkan	kt tersepak akan sst. Lelaki itu melunjurkan kakinya membuatkan Zara tersepakkan kaki panjangnya. Dua tiga kali dia tersungkur kerana tersepakkan batu-batu yang ada di situ.

(KDP, 2021:2056).

Takrifan yang diberikan KDP menunjukkan 'tersepak' dan 'tersepakkan' tidak berbeza dari segi maknanya. 'Tersepak' diberi maksud 'menyepak seseorang atau sesuatu **dengan tidak sengaja'** dan 'tersepakkan' diberi maksud '**tersepak** akan sesuatu'. Kedua-dua entri ini mendukung maksud tidak sengaja kerana '**tersepak** akan sesuatu' bagi 'tersepakkan' sama maknanya dengan 'menyepak (sso atau sst) dengan **tidak sengaja'**.

Yang menjadi persoalan di sini ialah mengapakah wujud dua bentuk kata, iaitu 'tersepak' dan 'tersepakkan' dengan maksud yang sama? Dari segi morfologinya, imbuhan teR- juga membentuk kata kerja transitif terbitan tetapi dengan maksud tidak sengaja manakala imbuhan meN- membentuk kata kerja transitif terbitan dengan maksud melakukan sesuatu perbuatan dan dilakukan dengan sengaja oleh pelaku (Nik Safiah Karim et al., 2010 dan Asmah, 2015).

Makna imbuhan yang didukung teR- membolehkan 'tersepak' ditulis semula sebagai '**menyepak** sesuatu dengan tidak sengaja'(walaupun bersifat oksimoron) dan bagi 'tersepakkan' pula, kata ini boleh ditulis semula sebagai '**menyepakkan** sesuatu dengan tidak sengaja'. Bagaimanapun, bagi kata kerja transitif terbitan yang berimbuhan meN-...-kan yang diterbitkan daripada kata kerja yang secara lahiriahnya atau secara semula jadinya ialah kata kerja transitif, makna kata kerja transitif tersebut tidak lagi mendukung makna 'melakukan sesuatu pada objek' sebaliknya mendukung makna 'benefaktif' atau 'manfaat orang lain'. Antara kata kerja yang secara lahiriahnya atau secara semula jadinya ialah kata kerja transitif ialah tulis (apa?), baca (apa?), beli (apa?), makan (apa?), masak (apa?), dan sepak (apa?). Kata tanya 'apa?' dalam kurungan menunjukkan perlunya objek hadir selepas kata kerja tersebut.

Apabila kata kerja yang secara lahiriahnya atau secara semula jadinya ialah kata kerja transitif dan menerima imbuhan teR-...-kan, maka kata kerja transitif ini juga mendukung maksud yang sama, iaitu makna benefaktif atau untuk manfaat orang lain dan dilakukan **dengan tidak sengaja**. Hal ini dikatakan demikian kerana imbuhan teR- bagi kata 'tersepak' boleh ditulis semula sebagai '**menyepak dengan tidak sengaja'**.

[2] Kertas kerja tentang entri ini telah dibentangkan dalam Kolokium Perkamusan Melayu: *Kamus Dewan Perdana* Khazanah Unggul Negara, 27 September 2022 di Dewan Bahasa dan Pustaka.

Kata 'tersepakkan' cuba dijejaki dalam korpus Dewan Bahasa dan Pustaka, namun hanya dua ayat yang ditemui berdasarkan 1000 perkataan yang dijana.

Carian Kata: tersepakkan

Bil. Konkordans: 2 konkordans dijumpai.

1) Sa-masa ia hendak berjalan di-dalam istana itu ia **tersepakkan** Rangga Jiwa dan Sibutatal. | HIKAYAT CHEKEL WENENG PATI.Sastera, 1965

2) Apabila babi-babi lain berjalan dan tersangkut atau **tersepakkan** hulu lembing itu babi yang sakit itu akan meraung kesakitan. | SENGALANG BURONG.Sastera, 1990

Dua ayat yang terjana ini membuktikan bahawa 'tersepakkan' bukanlah bentuk lazim yang digunakan oleh penutur bahasa Melayu pada zaman ini dan penggunaannya amat terhad. Data ini sewajarnya diabaikan dan tidak diketengahkan kerana bentuk ini ternyata tidak memberikan suatu makna yang jelas (berdasarkan penggunaan apitan teR-…-kan yang dibincangkan dalam perenggan sebelum ini). Satu fakta lain yang perlu diperhati dalam penggolongan kata 'tersepakkan' di atas ialah ayat yang dijana merupakan ayat yang dipetik daripada karya klasik, iaitu Hikayat Chekel Weneng Pati dan Sengalang Burong. Pastinya penulisan karya sastera sebegini sebenarnya tidak terikat pada hukum tatabahasa yang ketat dan ciri ini merupakan antara ciri laras bahasa sastera.

Dalam hal ini, mungkin juga wajar dipertimbangkan bahawa apa-apa yang terjana daripada korpus tidak semestinya direkodkan sebagai leksikon bahasa Melayu kerana bentuk-bentuk yang tidak lazim dan bentuk yang tidak mewakili makna yang logik [berdasarkan penggunaan apitan teR-…-kan] seharusnya tidak digalakkan untuk digunakan oleh pengguna bahasa Melayu. Harus juga disedari bahawa bentuk bahasa yang terhasil dalam karya sastera ditunjangan 'lesen puitis' atau 'kebebasan berkarya' (*poetic license*) yang bermaksud penulis ini mempunyai kebebasan untuk menggunakan bahasa yang menyimpang daripada peraturan bahasa konvensional apabila menulis atau berkarya dengan tujuan untuk menghasilkan kesan puitis.

Kelemahan data daripada korpus ini sebenarnya disedari oleh penyusun KDP berdasarkan petikan berikut :

> "Data korpus daripada laman sesawang menimbulkan beberapa isu, antaranya sama ada penyumbang data merupakan penutur asli bahasa Melayu atau tidak, kesahihan sumber data, kerancuan bahasa, serta pertindanan dan perulangan data. Oleh itu, penyusun mengambil sikap berhati-hati dan membuat penilaian serta tapisan sebelum memilih data untuk dianalisis.
>
> (Kamus Dewan Perdana, 2021: xiv)

Selain sebab dan alasan yang diberikan bagi menolak 'tersepakkan', satu lagi alasan yang difikirkan harus dipertimbangkan juga ialah tidak ada keselarasan antara imbuhan apitan 'teR-…-kan' dan maknanya yang diberikan KDP dengan mana-mana buku tatabahasa bahasa Melayu yang utama di Malaysia. Tatabahasa Dewan (2010) misalnya menyenaraikan apitan yang membentuk kata kerja hanyalah meN-... -kan, beR- ... -kan, beR- ... -an, di- ... -kan, meN-... -i, di- ... -i, memper- ... -kan, memper- ... -I, ke- ... -an, diper- ... -kan dan diper- ... -i. Imbuhan apitan 'teR-...-kan' tidak tersenarai sebagai salah satu apitan yang membentuk kata kerja transitif terbitan. Pastinya, jika KDP ini dirujuk oleh guru dan pelajar di sekolah, maklumat ini akan diterima tanpa perbahasan lanjut. Pengguna awam berpendapat bahawa apa-apa yang terkandung dalam kamus yang diterbitkan oleh Dewan Bahasa dan Pustaka adalah betul dan boleh digunakan dalam penulisan ilmiah. Ramai yang masih tidak memahami konsep dan takrifan kamus yang sebenar.

Selain isu 'tersepakkan' terdapat entri lain yang turut menimbulkan persoalan dan semuanya berpuncakan korpus yang dimanfaatkan untuk menyusun KDP ini. Seperti yang dinyatakan di atas, korpus yang dijana terdiri daripada pelbagai sumber dan genre. Bahasa yang digunakan juga pelbagai, iaitu bahasa baku dan juga bahasa basahan, bahasa pasar dan slanga.

Usaha merekodkan kosa kata yang terdiri daripada kata asal bahasa Melayu, dialek, bahasa basahan, slanga, pinjaman dan sebagainya sememangnya menjadi matlamat utama kebanyakan kamus, namun, aspek penggolongan kata, makna kamus atau takrifan dan contoh ayat yang diberikan bagi setiap entri dalam kamus haruslah juga mengarah pada pemahaman yang selari dengan maklumat yang terkandung dalam buku tatabahasa yang menjadi pegangan pengguna dan penutur bahasa Melayu.

KESIMPULAN

Korpus yang rencam daripada pelbagai sumber sememangnya dapat memberikan data yang banyak dan pelbagai. Namun, hakikat bahawa bukan semua pengguna bahasa menggunakan bentuk bahasa yang sempurna harus diambil kira. Selain itu, sebanyak mana juga data yang dijana oleh mana-mana pangkalan data korpus, korpus tersebut masih tidak mewakili penggunaan yang lengkap, kepelbagaian bahasa yang mungkin digunakan, domain yang pelbagai atau laras yang komprehensif. Kemungkinan juga korpus yang dijana tidak seimbang dan bais dari segi genre dan laras bahasanya yang tertentu. Sifat bahasa yang sangat dinamik akibat perkembangan disiplin dan bidang tertentu juga akan menyebabkan bentuk kata baharu terutamanya kata pinjaman dengan makna yang tertentu juga akan berubah dan bertambah. Dalam hal ini, korpus yang dijana mempunyai jangka masa yang tertentu dan mungkin juga menjadi tidak relevan pada masa yang akan datang. Oleh itu, korpus harus sentiasa dikemas kini, dan dijana daripada sumber atau teks baharu bagi membolehkan penyelidikan dilakukan dengan merujuk korpus terkini.

Penggunaan korpus bagi kajian linguistik tidak dinafikan amat bermanfaat dan mempunyai banyak kelebihan. Contohnya, korpus membolehkan penggunaan bahasa yang autentik diperoleh, kekerapan penggunaan kata atau leksikal dapat dicerap berserta kolokasinya dan maklumat konteks serta variasi penggunaan kata atau ayat dapat dikenal pasti. Bagaimanapun, cara dan pendekatan yang digunakan oleh pengkaji untuk memanfaatkan korpus ini harus selari dengan matlamat kajiannya.

Jelas Rundell (1996: 4), satu fakta penting yang perlu diingat tentang korpus ini ialah, nilai sebarang korpus sebagai sumber data linguistik berkait rapat dengan kandungannya. Dengan kata lain, generalisasi yang dibuat tentang bahasa berasaskan bukti korpus adalah sebaik bukti korpus yang digunakan pengkaji (ertinya input yang baik akan menghasilkan output yang baik juga).

Rujukan

Abdullah Hassan et al. (2006). Sintaksis. Siri Pengajaran dan pembelajaran Bahasa Melayu. Kuala Lumpur : PTS Publications &Distributor Sdn. Bhd

Asmah Hj. Omar. (2015). Nahu Melayu Mutakhir (Edisi Kelima). Kuala Lumpur : Dewan Bahasa dan Pustaka.

Aston, G. (1997). Enriching the learning environment: Corpora in ELT. Dlm. A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora,* 51-64. London: Longman.

Cook, G. (2003). *Applied linguistics: Oxford introductions to language study* (H. G. Widdowson, Series Ed.). Oxford: Oxford Univeresity Press.

Crystal, D. 1991. A Dictionary of linguistics and phonetics. Dicapai pada 9 Jun 2023 daripada http://www.natcorp.ox.ac.uk/what/whatis.html (13 Mac 2004)

Halliday, M.A.K, et al.(2004). *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

Hanks, P. (2009). The impact of corpora on dictionaries. Dlm. P. Baker (Ed.), *Contemporary corpus linguistics*, (214 – 236). New York: Continuum International Publishing Group.

Hishamudin Isam, Faizah Ahmad dan Mashetoh Abd. Mutalib. (2014). Wajaran Penggunaan Data Korpus dalam Penulisan Ilmiah: Dimensi Baharu Sukatan Pelajaran Bahasa Melayu Sijil Tinggi Pelajaran Malaysia (STPM). Jurnal Pendidikan Bahasa Melayu – JPBM (Malay Language Education Journal – Mylej). Vol. 4, Bil. 2 (Nov. 2014): 67-77

Kamus Dewan Bahasa Inggeris – Bahasa Melayu. (1992). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kamus Dewan (Edisi Keempat). (2010). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kamus Dewan Perdana. (2020). Kuala Lumpur: Dewan Bahasa dan Pustaka.

McEnery & Wilson. (1996). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T & Wilson, A. (2001). *Corpus linguistics. An introduction*. Ed. ke-2. Edinburgh: Edinburgh University Press.

Mindt 1991. Corpora and Semantics. Dicapai pada 9 Jun 2023 daripada http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/corpus4/4fra1.htm (7 Januari 2004)

Nik Safiah Karim et al. (2010). Tatabahasa Dewan (Edisi Ketiga). Kuala Lumpur : Dewan Bahasa dan Pustaka.

Oostdijk, N & Haan, P. de. 1994. Clause patterns in modern British English: A corpus-based (quantitative) study. *ICAME Journal* **18:** 41–79.

Oxford Advanced Learner's Dictionary. Tentative. Dicapai pada 10 Jun 2023 daripada https://www.oxfordlearnersdictionaries.com/definition/english/tentative?q=tentative)

Rundell, M. (1996). The corpus of the future and the future of the corpus. Dicapai pada 10 Jun 2023 daripada http//www.ruf.rice.edu/~barlow/futcrp.html

Rusli Abdul Ghani. (2004). Pangkalan data korpus DBP: Perancangan, pembinaan dan pemanfaatan. Kertas kerja Seminar Sehari Linguistik (SKALI 05) UKM-DBP. Bangi, 12-13 April 2005

Schmied. 1993. Begriffsglossar und index zu ulrichs von zatzikhoven lanzelet. Dlm. McEnery,T & Wilson, A. c*orpus linguistics. An introduction,* hlm. 104-109. Ed. ke-2. Edinburgh: Edinburgh University Press.

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

Sinclair, J. (1997). Corpus evidence in language description. Dlm. Wichmann, A et al. (pnyt.), *Teaching and language Corpora,* hlm. 27-39. Longman: London & New York.

Siti Aeisha Joharry dan Hajar Abdul Rahim. (2014). Corpus Research In Malaysia: A Bibliographic Analysis. Kajian Malaysia, Vol. 32, Supp. 1, 2014, 17–43 Penerbit Universiti Sains Malaysia

Williams, Geoffrey. (2023). From meaning to words and back: Corpus linguistics and specialised lexicography. 91-106 dicapai daripada https://doi.org/10.4000/asp.1320

Hasmidar bt Hassan

Penulis ialah penolong profesor di Jabatan Bahasa Melayu dan Linguistik, Universiti Brunei Darussalam. Sebelum ini, beliau merupakan pensyarah kanan di Pusat Pengajian Ilmu Kemanusiaan, Universiti Sains Malaysia, Pulau Pinang.