PENELITIAN TATABAHASA: ANTARA DATA KORPUS DENGAN INTUISI

Norliza Jamaluddin
(Malaysia)

Pendahuluan

Linguistik merupakan kajian bahasa secara saintifik manakala objek yang menjadi kajian dalam bidang linguistik ialah bahasa, iaitu penelitian terhadap struktur bahasa. Apabila meneliti struktur bahasa, pengkaji linguistik sebenarnya meneliti cara sesuatu bahasa disusun dalam pemikiran penutur aslinya, iaitu individu yang menghasilkan ujaran-ujaran yang lazim dan bersistematik. Untuk meneliti pemikiran manusia bukan merupakan aspek yang mudah dan hal ini menjadi masalah dasar dalam tatabahasa. Tambahan pula untuk menghuraikan sesuatu tatabahasa, pengkaji sebenarnya memerlukan bukti bahasa yang dihasilkan dan daripada bukti ini barulah model atau rumus bahasa dibentuk.

Sejak awal pengkajian bahasa, terdapat usaha untuk mengumpulkan bukti bahasa sama ada menerusi pembacaan terhadap sesuatu teks (data tulisan) atau menerusi bahasa yang didengari daripada penutur asli (data lisan). Pada peringkat awal ini, pengumpulan data bahasa amat kecil jumlahnya. Bagaimanapun jumlah ini semakin meningkat apabila terhasilnya pita rakaman pada tahun 1950-an. Edward Sapir dan William Labov misalnya telah menggunakan metode rakaman apabila meneliti bahasa Red Indian (Biber dan Finegan, 1996:207). Pada tahun 1960-an pula iklan televisyen telah dijadikan sebagai salah satu sumber dalam penyelidikan linguistik. Kajian seperti ini telah dilakukan oleh Leech dengan menyalin 617 iklan untuk dijadikan data korpus pertamanya (Svartik, 1996:4). Bagaimanapun, kuantiti data yang dijadikan asas kajian masih sedikit bagi membolehkan kajian linguistik ketika itu benar-benar andal. Hal ini menunjukkan bahawa kajian linguistik pada ketika itu masih kekurangan sumber bukti bagi memaparkan struktur-struktur yang

lazim[1] dalam bahasa, walaupun diketahui bahawa bentuk lazim ini penting dalam menghuraikan bahasa.

Penelitian terhadap bentuk-bentuk lazim yang hadir berulang kali di dalam data yang dikumpulkan amat penting bagi membolehkan kesimpulan dibuat terhadap bahasa yang dikaji. Dalam bidang linguistik, data yang dikumpulkan, disimpan dan disusun untuk kajian bahasa diistilahkan sebagai data korpus. Disebabkan kebanyakan data korpus pada masa ini tersimpan dalam bentuk elektronik dengan menggunakan perisian tertentu, maka istilah data korpus berkomputer telah digunakan. Kini kajian berasaskan data korpus berkomputer telah menjadi kajian arus perdana (Leech, 1996:9; Svartik, 1996:3–13). Menurut Sinclair (1991:1),

> *"Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular".*

Penggunaan data korpus berkomputer bukan sahaja terhadap kajian bahasa Inggeris, tetapi juga bahasa-bahasa lain seperti bahasa Rusia, Perancis, Jepun, Catalan, Drybal dan tidak ketinggalan bahasa Melayu. Bagaimanapun data korpus bukan merupakan satu daripada cabang linguistik, tetapi merupakan metodologi dalam kajian linguistik kerana data korpus digunakan sebagai data kajian bahasa dalam hampir semua cabang linguistik seperti semantik, sintaksis, sosiolinguistik, leksikografi dan morfologi. Dalam metodologi, data korpus digunakan untuk menghuraikan aspek-aspek bahasa secara terperinci, menguji dan akhirnya membuat kesimpulan tentang rumus atau teori bahasa.

[1] Istilah yang digunakan oleh Sinclair (1997:28) dalam kajian korpus linguistik bagi bentuk yang lazim ialah *regularities*, manakala McEnery dan Wilson (2001:9) menggunakan istilah *recursive.*

Antara Data dengan Intuisi

Pendekatan berasaskan data bererti pendekatan yang mengutamakan data berbanding teori dalam penyelidikan bahasa. Ini bererti rumus bahasa dihasilkan berdasarkan pada data korpus yang diteliti. Data korpus ini berbeza daripada data korpus sebelumnya kerana pendekatan berasaskan data ini terdiri daripada data yang tulen, data yang tidak dipilih-pilih bagi memenuhi tujuan tertentu, data yang terdiri daripada jumlah yang besar, data yang disusun secara sistematik dan data yang tidak dianotasi terlebih dahulu berdasarkan pada teori-teori yang wujud sebelumnya (Hunston dan Gill, 1999:14).

Secara relatifnya penyelidikan berasaskan data korpus yang menggunakan komputer masih baharu. Bagaimanapun, kajian yang menggunakan data korpus sebagai asas penyelidikan bahasa adalah seusia dengan persoalan mengenai bahasa itu sendiri (Tognini, 2001:50). Kajian awal yang dianggap berasaskan data korpus ialah kajian yang dilakukan oleh Alexander Cruden. Dengan menjadikan kitab Injil sebagai data korpus, beliau telah menerbitkan *Cruden's Concordance* pada tahun 1736 (Kennedy, 1999:13–14).

Dalam bidang linguistik, data korpus sebenarnya telah digunakan sejak tahun 1800 lagi, misalnya dalam kajian perolehan bahasa, konvensi ejaan, pedagogi bahasa, linguistik bandingan, sintaksis dan semantik. Dalam kajian perolehan bahasa misalnya, rekod bahasa yang dicatatkan di dalam diari oleh ibu bapa kanak-kanak berkenaan sekitar tahun 1876 hingga 1926 telah dijadikan asas dalam kajian linguistik. Ini dapat dilihat dalam kajian yang dilakukan oleh Preyer (1889) dan Stern (1924). Sementara itu, dalam kajian terhadap konvensi ejaan, Kading (1897) telah menggunakan data korpus sejumlah 11 juta patah perkataan dalam bahasa Jerman. Data korpus ini digunakan bagi menyemak taburan frekuensi huruf dan urutan huruf dalam bahasa Jerman. Dalam bidang pedagogi bahasa pula, data korpus digunakan untuk meneliti pedagogi bahasa asing, misalnya kajian oleh Fries dan Traver (1940) dan kajian Bongers (1947). Dalam kajian linguistik bandingan, Eaton (1940) telah membandingkan makna bagi perkataan dalam bahasa Belanda,

Perancis, Jerman dan Itali, manakala dalam kajian sintaksis dan semantik, Fries (1952) telah menghuraikan tatabahasa bahasa Inggeris berdasarkan data korpus (McEnery dan Wilson, 2001:3–4).

Pada awal abad ke-20, akhbar dan novel kerap kali dijadikan sebagai sumber untuk menggambarkan ciri-ciri dan binaan tatabahasa pada ketika itu. Ini dapat dilihat dalam kajian yang dijalankan oleh Jespersen (1909-1949), Poutsma (1926–1929) dan Kurisinga (1931–1932). Kajian yang lebih bersistematik berasaskan data korpus telah dilakukan oleh Charles C. Fries (1952). Buku *The Structure of English* yang dihasilkan oleh Fries ini telah menggunakan sejumlah 250,000 patah perkataan daripada rakaman perbualan telefon dan beliau melakukan analisis secara manual.

Berdasarkan kajian-kajian tersebut, dapat dikatakan bahawa kajian linguistik pada peringkat awal menyerupai data korpus kerana huraian bahasa adalah berdasarkan data. Walaupun ahli-ahli bahasa pada ketika itu tidak menggelar diri mereka sebagai ahli linguistik korpus, tetapi asas yang digunakan untuk menghuraikan bahasa ialah data korpus. Baos (1940), misalnya dianggap sebagai ahli linguistik lapangan. Begitu juga ahli-ahli linguistik struktural yang lain seperti Sapir, Newman, Bloomfield dan Pike yang turut menggunakan data korpus untuk menghuraikan bahasa pada ketika itu (McEnery, Xiao dan Tono, 2006:3).

Walaupun beberapa kajian tadi menunjukkan bahawa bidang linguistik telah menggunakan data korpus seawal abad ke-18 lagi, namun kajian yang berasaskan data korpus ini mempunyai sejarah yang “tenggelam timbul”. Dari segi perkembangan data korpus pula, walaupun pada peringkat awal (tahun-tahun 1800) pendekatan yang digunakan untuk menghuraikan bahasa dan pembentukan model bahasa berasaskan data korpus, tetapi pada akhir tahun 1950-an sehingga 1980-an, pendekatan ini semakin kurang popular dan menjadi pendekatan yang marginal akibat daripada desakan bahawa linguistik tidak memerlukan data empirikal (Sampson, 2005:16).

Hal ini dapat dilihat menerusi kritikan-kritikan Chomsky yang menolak penggunaan data korpus dalam pembentukan model bahasa dan huraian bahasa. Sehinggalah baru-baru ini, iaitu dalam tahun 1980-an, data korpus berkembang semula selari dengan perkembangan dalam bidang pengkomputeran. Bahkan, sejak akhir-akhir ini dapat dilihat peningkatan dalam bidang linguistik korpus yang berlaku secara besar-besaran dan perkembangan ini tidak terbatas pada kajian bahasa Indo-Eropah sahaja, tetapi telah tersebar pada bahasa-bahasa lain di dunia.

Sebenarnya perkembangan data korpus, terutamanya pada dekad-dekad terakhir ini banyak dipengaruhi oleh kritikan Chomsky terhadap data yang digunakan dalam kajian linguistik. Kritikannya ini telah menyebabkan pengguna data korpus perlu memastikan bahawa data korpus yang digunakan itu adalah seimbang dan representatif. Oleh itu, bagi menghuraikan data korpus dengan lebih lanjut, penting dinyatakan terlebih dahulu kritikan-kritikan Chomsky dan pertentangan antara Chomsky dan ahli linguistik korpus.

Chomsky pada dasarnya membahagikan bahasa pada dua kategori, iaitu “kecekapan” (*competence*) dan “perlakuan” (*performance*). Kecekapan berbahasa dikatakan sebagai pengetahuan berbahasa penutur-pendengar, manakala perlakuan merupakan bentuk penggunaan bahasa yang sebenar dalam situasi sebenar. Kecekapan ini melibatkan pemerolehan seperangkat rumus yang terhad bilangannya yang menjana seperangkat ayat yang tidak terhad jumlahnya. Kecekapan ini ada di dalam minda seseorang penutur asli, iaitu penutur asli mempunyai kebolehan membentuk, menyebut dan mentafsir bahasa. Penutur asli dikatakan tidak mengetahui rumus bahasanya sendiri, tetapi hanya mempunyai pengetahuan akliah, iaitu pengetahuan luar sedar terhadap rumus bahasanya dalam membentuk struktur bahasa. Disebabkan dalam kajian bahasa penelitian dilakukan terhadap bagaimana bahasa distrukturkan maka menurut Chomsky, aspek yang diteliti ialah kecekapan bahasa dan bukan perlakuan (Radford, 1994:2). Oleh itu, analisis bahasa dilakukan bagi mengetahui dan menghuraikan kecekapan bahasa penutur asli.

Perlakuan bahasa menurut Chomsky, tidak boleh dijadikan alat dalam penelitian linguistik kerana perlakuan merupakan gambaran yang tidak lengkap bagi kecekapan bahasa. Perlakuan bahasa dipengaruhi oleh faktor-faktor luaran semasa sesuatu ujaran dihasilkan atau dilafazkan. Ini bererti dalam kajian linguistik aspek yang diteliti ialah struktur bahasa yang ada dalam minda seseorang dan hal ini telah mewujudkan pergantungan terhadap intuisi. Persoalan antara kecekapan dan perlakuan bahasa sebagai asas penyelidikan linguistik inilah yang menjadi perdebatan antara Chomsky dengan ahli linguistik korpus.

Bagi ahli linguistik korpus, huraian bahasa perlu berdasarkan perlakuan kerana perlakuan dikaitkan dengan penghasilan bahasa secara luar sedar. Menurut golongan yang mementingkan data sebagai asas penyelidikan bahasa, kecekapan tidak boleh dicapai secara terus, tetapi menerusi tiga cara, iaitu berdasarkan input, refleksi dan output. Input merupakan bahan atau data bahasa yang terdapat dalam minda penutur asli dan dari sinilah tatabahasa terbentuk, manakala refleksi pula dapat dilakukan dengan pengkaji meneliti kecekapan bahasanya sendiri, sementara output merupakan perlakuan bahasa penutur asli itu sendiri (Cook, 1969:2) .

Cara yang pertama, iaitu berdasarkan input merupakan aspek yang sukar untuk dilakukan kerana tidak mudah untuk seseorang pengkaji meneliti pemikiran seseorang penutur asli dalam membentuk dan menyusun bahasanya sendiri. Sementara, bagi melakukan refleksi pula, aspek ini dianggap peranti yang lemah dalam meneliti bahasa kerana bahasa pada prinsipnya dihasilkan secara luar sedar, iaitu penutur asli tidak sedar terhadap rumus yang dihasilkan semasa mengujarkan bahasa. Walaupun analisis menunjukkan bahawa seseorang pengkaji mampu menghasilkan bahasa yang betul, tetapi sukar untuk pengkaji itu meneliti bahasa berasaskan perasaannya sendiri dan ini mampu mengundang *human bias*.

Aspek yang terakhir, iaitu berdasarkan output merupakan pilihan terbaik (Cook, 1969:2) kerana penelitian terhadap bahasa dilakukan terhadap perlakuan sebenar penutur asli sesuatu bahasa dan bukan berasaskan penilaian terhadap bahasa yang dihasilkan oleh pengkaji itu sendiri. Oleh itu, bagi mendapatkan data bahasa yang dihasilkan secara luar sedar ini, pengkajian bahasa sebenarnya dilakukan terhadap perlakuan dan daripada perlakuan ini baharulah model bahasa dibina. Perlakuan bahasa ini diperoleh daripada data korpus yang dikumpulkan oleh seseorang pengkaji.

Perbezaan pendapat antara Chomsky dengan ahli linguistik korpus ini seterusnya menyebabkan pertentangan sama ada huraian bahasa perlu berasaskan kepada:

i. pemerhatian terhadap data yang dibentuk secara rekaan, atau

ii. pemerhatian terhadap data yang terbentuk secara tabii.

(McEnery dan Wilson, 2001: 5–6)

Chomsky yang membentuk model bahasa berasaskan minda merupakan golongan rasionalis dan matlamat utama golongan ini adalah untuk mencapai kemunasabahan kognitif. Menurut fahaman rasionalis, aspek bahasa yang dijadikan asas dalam pembentukan teori perlulah sempurna dan betul struktur ayatnya. Oleh itu, untuk meneliti ayat yang sempurna dan betul strukturnya maka data yang dikaji biasanya terdiri daripada ayat-ayat yang diperoleh daripada informan yang kebanyakannya terdiri daripada ahli linguistik itu sendiri. Ahli-ahli linguistik ini mencipta atau mereka ayat mereka sendiri. Tambahan pula sebagai penutur asli, ahli linguistik dikatakan berkeupayaan untuk membuat penilaian terhadap sempurna atau tidak dan berstruktur atau tidak ayat yang dihasilkan.

Berbeza dengan ahli linguistik korpus yang menganggap diri mereka sebagai golongan empirikal, maka huraian bahasa perlulah berasaskan aspek bahasa yang terbentuk secara tabii dan di luar sedar, iaitu penekanan terhadap perlakuan bahasa. Mereka berpendapat bahawa “*Our focus should be on what happens, not what we think*

should happen". Hal ini berbeza daripada golongan rasionalis kerana mereka dikatakan "...*represent micro fragments of data, they are typical and most often would be unlikely to occur as real life language"* (Suad, 1999:32). Bahkan, bagi menjadikan linguistik sebagai salah satu disiplin saintifik sama seperti disiplin-disiplin yang lain, maka aspek empirikal amat diutamakan. Titik permulaan bagi penyelidikan empirikal adalah berdasarkan data yang autentik, iaitu diperoleh daripada data yang wujud secara tabii dan secara tipikalnya menerusi pemerhatian data korpus. Berasaskan pemerhatian terhadap data korpus ini, maka teori bahasa dapat dibentuk. Menurut Leech (1991:8), data korpus merupakan *explicandum* bagi linguistik. Oleh itu, model bahasa dibina berdasarkan pemerhatian terhadap data korpus (McEnery dan Wilson, 2001:6).

Pemerhatian terhadap data yang berbeza-beza ini telah menghasilkan dua pendekatan yang berbeza (McEnery, Xiao dan Tono, 2006:6–8; Hunston, 2002:20; Aarts, 1996:46–47) iaitu:

i. pendekatan berasaskan intuisi (*intuition-based approach*)

ii. pendekatan berasaskan data korpus (*corpus-based approach*)

Chomsky yang telah mengubah objek linguistik daripada huraian bahasa yang abstrak sifatnya pada teori yang menggambarkan realiti bahasa dan model bahasa yang munasabah bergantung sepenuhnya pada pendekatan berasaskan intuisi. Sebaliknya bagi golongan yang mementingkan kajian empirikal, penelitian bahasa perlu berdasarkan sumber bukti dan data korpus merupakan bukti bagi bahasa yang dihuraikan itu.

Dalam pendekatan yang berasaskan data korpus, bahasa yang dihuraikan dapat mewakili keseluruhan atau sebahagian besar penutur natif. Menerusi data yang dikumpulkan, ahli linguistik sebenarnya mengumpulkan dan memaparkan pengalaman berbahasa penulis atau penutur bahasa berkenaan dan bentuk bahasa ini merupakan sudut pandangan masyarakat umum tentang bahasa yang digunakan. Perkara ini dapat dilakukan

kerana data korpus yang dibina biasanya mempunyai saiz yang besar, iaitu terdiri daripada ratusan juta perkataan, terdiri daripada pelbagai bidang dan laras serta diambil daripada pelbagai sumber, contohnya data yang termuat dalam *monitor corpus*. Saiz yang besar ini telah menghasilkan pola bahasa yang bermakna serta dapat memberikan gambaran tentang sifat bahasa secara konsisten dan membolehkan pengkaji mengeksploitasi data yang telah ada sepenuhnya dengan menggunakan perisian-perisian tertentu. Perisian ini membantu pengkaji meneliti cara sesuatu bahasa digunakan.

Tambahan pula, data korpus yang dibina memenuhi kriteria tertentu bagi menjadikannya sebagai bentuk yang representatif bagi sesuatu bahasa. Contohnya, bagi projek korpus COBUILD, telah ditetapkan bahawa salah satu kriteria yang perlu dalam membangunkan data korpus ialah data tersebut terdiri daripada buku-buku fiksyen dan bukan fiksyen yang laris jualannya (Renouf, 1988:3). Hal ini menjadikan sebaran bagi bentuk bahasa yang dihasilkan meluas dan diterima atau dibaca oleh sebahagian besar anggota masyarakat bagi bahasa berkenaan. Justeru itu, penggunaan data korpus melibatkan sudut pandangan masyarakat secara umum berbanding pandangan peribadi seseorang pengkaji.

Dalam kajian berdasarkan intuisi ini telah menyebabkan seseorang pengkaji itu hanya meneliti bahasa yang dihasilkan oleh sekelompok kecil pengguna kerana data diperoleh daripada dirinya sendiri atau daripada beberapa informan yang lain. Tambahan pula, intuisi dikatakan boleh dipengaruhi oleh dialek atau sosiolek seseorang (McEnery, Xiao dan Tono, 2006:6–7). Oleh itu, bentuk bahasa yang dianggap betul dan sempurna oleh seseorang pengkaji atau seseorang penutur asli mungkin tidak diterima oleh penutur asli yang lain. Walaupun intuisi mereka betul, tetapi bahasa yang dibentuk tidak mewakili penutur natif secara keseluruhannya. Bahkan bahasa yang dibentuk berdasarkan intuisi pengkaji itu sendiri sukar untuk ditentukan sahkan disebabkan sukar untuk mengesahkan introspektif seseorang kerana intuisi tidak boleh dilihat atau diteliti.

Di samping itu, data korpus bukan sekadar memaparkan data yang banyak, tetapi dengan menggunakan peranti analisis data korpus seperti senarai perkataan dan konkordans, data yang dianalisis disusun secara sistematik. Hal ini bukan sahaja membolehkan bentuk-bentuk tipikal yang terdapat dalam bahasa ditunjukkan, malahan penganalisisan data dapat dilakukan secara tuntas (Bowler dan Pearson, 2002:13; Hunston, 2002:20; McEnery dan Wilson, 2001:15–16). Menerusi baris-baris konkordans, pola bagi perkataan *interested* dan *interesting* misalnya dapat diteliti, iaitu *interested* mempunyai pola *interested* + *in*, manakala *interesting* + kata nama. (Hunston, 2002:9). Berdasarkan bentuk tipikal yang ditunjukkan oleh baris-baris konkordans, maka seseorang pengkaji boleh melakarkan pola bahasa yang diteliti. Tambahan pula, peranti analisis data korpus bukan sekadar menunjukkan kolokasi perkataan, tetapi juga frekuensi kata dan frasa.

Intuisi pada umumnya mampu memberi maklumat segera terhadap dua perkara sahaja, iaitu makna sesuatu kata secara terpisah dan bentuk ayat yang sempurna juga secara terpisah. Walaupun intuisi mampu memberikan beberapa contoh kata yang hadir bersama-sama perkataan *intersting* dan *interested*, tetapi sukar untuk ditunjukkan polanya. Di samping itu, contoh kata yang ditunjukkan sukar untuk disokong oleh realiti kerana tidak diketahui sama ada bentuk yang dinyatakan itu benar-benar wujud atau hanya dibentuk dalam fikiran pengkaji bahasa itu sahaja. Menurut Lewis (2001:127) "*We all tend to have confidence in our intuitions about language, but unfortunately the empirical evidence sometimes shows that our intuitions are seriously flawed".*

Tambahan pula, tanpa sokongan bukti realiti, model bahasa berasaskan minda sukar untuk ditunjukkan secara kuantitatif. Walaupun Chomsky berpendapat bahawa analisis kuantitatif tidak memberi sebarang makna dalam kajian bahasa, tetapi menurut McEnery (2001:16) aspek kuantitatif merupakan peranti analisis yang mempunyai kesan yang besar dalam kajian linguistik. Bahkan, menerusi data korpus aspek bahasa yang diteliti mampu

ditafsirkan secara kualitatif dan sebahagian lagi berdasarkan kuantitatif bagi menjadikannya lebih andal berbanding kajian-kajian sebelum ini.

Di samping itu, intuisi juga tidak mampu untuk memberikan maklumat kolokasi, prosodi dan frasa. Dalam bahasa Inggeris, misalnya kata adverba yang berkolokasi dengan kata adjektif seperti *acutely aware, keenly felt, paintfully clear* dan *readily available* sukar untuk diketahui menerusi intuisi (Granger, 1998 dalam Hunston, 2002:20). Bagi maklumat kolokasi misalnya, tidak semua kolokasi kata dapat dinyatakan dengan tepat menerusi intuisi. Intuisi biasanya mampu memberikan kolokasi bagi perkataan yang biasa digunakan, tetapi agak sukar untuk menyatakan kolokasi kata yang jarang berlaku. Tambahan pula, menerusi intuisi sukar untuk seseorang menyatakan kolokasi yang tipikal bagi sesuatu kata secara tepat.

Pendekatan data korpus juga tidak menyebabkan pengkaji mencipta ayat-ayat tertentu, tetapi berdasarkan bukti yang terdiri daripada data yang autentik. Di samping itu, data tersebut tidak dipilih oleh peneliti bahasa bagi disesuaikan dengan teori yang digunakan oleh mereka. Di sini, data korpus dianggap sebagai bukti empirikal dalam bidang linguistik. Bahkan, saiz data yang besar mampu menyediakan bukti bahasa yang digunakan oleh penurut natif dan dapat diterima sebagai bentuk bahasa sebenar serta bebas daripada *human bias* (Biber, Conrad, dan Reppen, 1998:3). Berdasarkan bukti yang ditunjukkan, maka sesuatu aspek bahasa boleh dikaji berulang-ulang kali oleh pengkaji yang berbeza bagi menghasilkan model bahasa yang lebih berwibawa. Aspek yang lebih penting lagi ialah data korpus turut memaparkan bukti-bukti bahasa yang sebelum ini tidak diketahui, iaitu aspek bahasa yang tidak wujud dalam penghasilan teori sebelum ini. Berdasarkan data bahasa yang autentik ini barulah model bahasa dibentuk. Hal ini turut memungkinkan pengklasifikasian dan pengkategorian bahasa yang dibuat sebelum ini boleh diselidiki semula (Murison-Bowie, 1996:182 dan Suad Awab, 1999:33).

Berbeza dengan pendekatan berasaskan intuisi, penyelidik telah mencipta sesuatu bentuk bahasa (contohnya ayat) yang dirasakan betul dan tepat menurut sudut pandangan mereka. Hal ini perlu dilakukan oleh penyelidik bagi mematuhi teori yang telah ada dalam sesuatu bahasa. Penyelidik sebenarnya mereka perkataan atau frasa atau ayat dan bentuk bahasa yang direka itu disesuaikan dengan teori yang telah ada. Hal ini demikian kerana mereka menggunakan konsep **teori dahulu** dan kemudian baru disesuaikan dengan bahasa yang dikaji (Azhar, 1993:9–10, 40; Knowles dan Zuraidah Mohd. Don, 2006:5–6). Justeru, mereka tidak memerlukan data korpus dalam menghuraikan bahasa dan bergantung sepenuhnya terhadap intuisi mereka sendiri atau bertanyakan sama ada bentuk bahasa yang dihasilkan itu tepat dan sempurna daripada orang lain. Di sini, penyelidik membentuk model bahasa berdasarkan intuisinya sendiri berbanding penelitian terhadap bahasa yang dihasilkan oleh penutur natif. Kesannya teori yang dihasilkan menghuraikan sesuatu yang tidak wujud kerana berasaskan ayat-ayat yang dibinanya sendiri. Menurut Stubb (1996:29), ahli linguistik seperti ini dikatakan sebagai *"...judge and jury their own theory hardly a basis for objective comment"*. Secara tidak langsung pendekatan ini telah menafikan huraian atau bentuk sebenar sesuatu bahasa. Oleh itu bahasa yang dihuraikan sebenarnya merupakan bahasa yang dihasilkan oleh pengkaji itu sendiri.

Di samping kritikan-kritikan yang telah dinyatakan tadi, Chomsky (1962) juga dengan tegas menolak data korpus kerana pada pendapatnya wujud *skewedness* terhadap data yang diteliti. Menurut Chomsky,

> *"Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list"* (dlm. Leech, 1996:8).

Beliau juga memberikan beberapa contoh yang menyebabkan wujudnya *skewedness* dalam data korpus, misalnya ayat "*I live in New York*" lebih kerap dimuatkan berbanding "*I live in Dayton, Ohio*" kerana New York lebih lazim dan lebih dikenali berbanding Dayton.

Skewedness ini mungkin ada kebenarannya pada peringkat awal korpus kerana korpus yang disebut sebagai korpus kotak kasut hanyalah bersaiz kecil dan sukar untuk diteliti pola yang berulang secara manual. Bahkan data korpus seperti ini sering kali tidak memuatkan bentuk bahasa yang jarang digunakan atau mengeluarkan bentuk-bentuk yang terlalu lazim. Potensi untuk sesuatu data korpus menjadi *skewed* telah cuba dielakkan dalam penghasilan data korpus berkomputer pada masa ini.

Data korpus seboleh mungkin menekankan aspek representatif terhadap data yang disusun. Bagi menghasilkan data korpus yang representatif ini, data korpus yang disusun perlu meliputi cakupan yang luas, iaitu terdiri daripada pelbagai penulis, laras, genre dan bentuk bagi memberikan gambaran yang tepat tentang seluruh populasi bahasa. Bagaimanapun, bagi mengelakkan data korpus ini sama seperti arkib yang merupakan koleksi secara rawak bagi sesuatu data, maka data korpus disusun berdasarkan kriteria-kriteria tertentu. Menurut Leech (1992:116),

> *"…computer corpora are rarely haphazard collections of textual material: they are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) representative of some language or text type".*

Ini menunjukkan bahawa data korpus disusun bukan semata-mata sebagai data yang memuatkan segala teks di dalamnya, tetapi perlu memenuhi kriteria tertentu bagi membolehkan data tersebut representatif terhadap bahasa yang diwakilinya. Sinclair (1996) turut menekankan aspek representatif sesuatu data korpus, iaitu dengan menyatakan bahawa,

> "*a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language"*
>
> (McEnery, Xiao dan Tono, 2006:4)

Justeru, reka bentuk dan kerelevanan data korpus sebagai asas kajian dalam linguistik telah menyebabkan linguistik pada masa ini, khususnya selepas tahun 1980-an, banyak menggunakan data korpus berkomputer untuk meneliti sifat bahasa. Menurut Sinclair (1992:5),

> "...*so much language available on record, particularly written language in electronic form..., our theory and descriptions should be re-examined to make sure they are approapriate. We have exprienced not only a quantitative change in the amount of language data available for study, but a consequent qualitative change in the relation between data and hypothesis".*

Kesimpulan

Dapat disimpulkan bahawa pendekatan data korpus perlu dijadikan asas kajian linguistik pada masa ini disebabkan data yang terdapat di dalam data korpus merupakan data yang tulen, jumlah yang besar, terdapat dalam bentuk elektronik dan disusun secara sistematik. Ini menunjukkan bahawa data yang diteliti merupakan bentuk bahasa sebenar yang dihasilkan oleh penutur natif dan dapat ditangani secara automatik. Hal ini membolehkan penelitian bahasa dilakukan dalam pelbagai aspek dan dilakukan secara mendalam. Perkara ini penting kerana data korpus berkomputer dapat mengelakkan pengkaji daripada menggunakan intuisinya semata-mata dalam membentuk rumus atau model bahasa.

Kajian linguistik yang menggunakan data korpus sebagai asas dalam penelitian aspek bahasa merupakan kecenderungan baharu bagi ahli linguistik. Data korpus kini bukan sahaja digunakan dalam penelitian terhadap bahasa Inggeris, malahan bahasa Rusia, China, Jepun, Drybal dan Catalan. Bahkan pada tahun kebelakangan ini, iaitu pada penghujung tahun 1990-an sehingga kini, dalam bidang linguistik Melayu, data korpus berkomputer telah mula digunakan dalam penyelidikan linguistik. Antaranya kajian yang dilakukan oleh Norliza Jamaluddin (2000) dan Knowles, dan Zuraidah (2003, 2006 dan 2008).

Selain itu, beberapa seminar berkaitan dengan data korpus berkomputer turut diadakan, seperti Seminar Adverba Bahasa Melayu (2004), dan Seminar Kajian Bahasa dan Korpus: Dimensi Linguistik Semasa pada 2005. Justeru itu, penggunaan data korpus sebagai asas dalam kajian tatabahasa pada masa ini adalah relevan. Hal ini juga penting seperti yang diperkatakan oleh Sinclair (1997) berkaitan dengan salah satu aspek tatabahasa, iaitu kajian golongan kata sebagai, "...*that word-classes have to be completely rethought in the light of corpus evidence of the similarity of words in their corpus patterns"* (Butler, 2004:154).

Bibliografi

Aarts, J., 1996. "*Intuition-based and Observation-based Grammars"* dlm. Aijmer, K dan Altenberg (ed.) *English Corpus Lingusitics: Studies in Honour of Jan Svartivik.*, B. London and New York: Longman.

Azhar Simin, 1993. *Sintaksis Wacana "Yang" dalam Bahasa Melayu.* Kuala Lumpur. Dewan Bahasa dan Pustaka.

Biber, D., Conrad, S. dan Reppen, R., 1996. "*Corpus-Based Investigations of Language Use"* dlm. *Annual Review of Applied Linguistics.* Vol. 16. USA: Cambridge University Press.

Biber, D. dan Edward Finegan, 1996. "On the Exploitation of Computerized Corpora in Variation Studies" dlm. Aijmer, K dan Altenberg, B. (ed.) *English Corpus Lingusitics: Studies in Honour of Jan Svartivik.* London and New York: Longman.

Bowler, L dan Pearson, J., 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London and New York: Routledge Taylor and Francis.

Butler, C.S., 2004. "*Corpus Studies and Functional Linguistics Theories"* dlm. *Function of Language*. Vol. II. Issue 2. Amsterdam: John Benjamins Publishing Company.

Cook, W.A., 1969. *Introductionn to Tagmermic Analysis*. New York: Holt. Rinehart and Winston Inc.

Garside, R., Leech, G. dan McEnery, T., 1997. *Corpus Annotation: Linguistics Information From Corpus Text Corpora.* England: Longman.

Hunston, S. and Gill Francis, 1999. *Pattern Grammar: Corpus Driven Approach to the Lexical of English*. Amsterdam: John Banjamins Publishing Company.

Hunston, S., 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kennedy, G., 1998. *An Introduction to Corpus Linguistics.* Edinburgh: Pearson Education Limited.

Knowles, G. dan Zuraidah Mohd Don, 2006. *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Leech. G., 1992. "Corpora and Theories of Linguistics Performance" dlm. Svartvik, J. (ed.) *Direction in Corpus Linguistics.* Berlin: Mouton de Gruyter.

Leech, G., 1996. "*The State of the Art in Corpus Linguistics"* dlm. Aijmer, K dan Altenberg, B. (ed.) *English Corpus Lingusitics: Studies in Honour of Jan Svartivik.* London and New York: Longman.

Lewis, M., 2001. "*Language in the Lexical Approach"* dlm. *Teaching Collocation: Further Developments in the Lexical Approach.* England: Language Teaching Publications.

McEnery, T. dan Andrew Wilson, 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, T., Richard Xiao dan Yukio Tono, 2006. *Corpus-Based Language Studies : An Advanced Resource Book*. London and New York: Routledge Taylor and Francis

Murison-Bowie, S., 1996. *"Linguistic Corpora and Language Teaching"* dlm. *Annual Review of Applied Linguistics.* Vol. 16. USA: Cambridge University Press.

Radford, A., 1994. *Tatabahasa Transformasi* (terj.) Noor Ein Mohd. Noor dan Zaiton Abdul Rahman. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Renouf, A., 1988. "*Corpus Development"* dlm. *Looking Up.* London and Glasgow: Collins ELT.

Sampson, G., 2005. "*Quantifying the Shift Towards Empirical Methods"* dlm. *International Journal Of Corpus Linguistics.* Amsterdam: John Banjamins Publishing Company.

Sinclair, J.M., 1991. *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

Sinclair, J.M., 1992. *"Trust the Text"* dlm. *Advances in Systemic Linguistics: Recent Theory and Practice.* (ed.) Davies, M. dan Ravelli, L. London: Pinter Ltd.

Sinclair, J.M., 1997. *"Corpus Evidance in Language Description"* dlm. *Teaching and Language Corpora.* London dan New York*:* Longman.

Suad Awab, 1999. *Multi-Word Units in a Corpus-Based Study of MoU: Modal Multi-Word Units.* Tesis Ph.D. Lancaster University.

Svartvik,J., 1996. "*Corpora are becoming Mainstream"* dlm. *Using Corpora for Language Research.* London and New York: Longman.

Tognini, E., 2001. *Corpus Linguistics at Work.* Amsterdam; John Benjamins Publishing Company.