# BAB 1

# PENDAHULUAN

## 1.1 Pengenalan

Linguistik merupakan kajian bahasa secara saintifik dan 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 tidak 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 inilah baru 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 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 (antaranya seperti kajian yang dilakukan oleh Sinclair (1991, 1997), Biber, Conrad dan Reppen (1996), Garside, Leech dan McEnery (1997), Williams (2002), Hunston (2002), Mason dan Hunston (2004) dan McEnery, Xiao dan

[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.*

Tono (2006), tetapi juga bahasa-bahasa lain seperti bahasa Itali (Gavioli, 1997); Jerman (Dodd, 1997 dan Jones, 1997); Rusia (Azarova, dan Sinopalnikova, 2004, Lashevskaja dan Shemanaeva, 2008); Perancis (Inkster, 1997); Jepun (Chujo, Utiyama dan Miura, 2006 dan Cao dan Nishina, 2007) dan tidak ketinggalan bahasa Melayu (Norliza Jamaluddin (2000), Imran Ho Abdullah (2008), Zaharani Ahmad (2008), Knowles dan Zuraidah Mohd. Don (2003, 2006, 2008). Dalam kajian ini, data korpus berkomputer turut dijadikan sebagai asas kajian bagi meneliti perlakuan kata sifat bahasa Melayu. Kini, data korpus digunakan sebagai data kajian bahasa dalam hampir semua cabang linguistik seperti semantik, sintaksis, sosiolinguistik, leksikografi dan morfologi (McEnery dan Wilson, 2001:1) Data korpus digunakan untuk menghuraikan aspek-aspek bahasa secara terperinci, menguji dan akhirnya membuat kesimpulan tentang rumus atau teori bahasa (Tognini-Bonelli, 2001: 65 ).

Disebabkan dalam kajian ini penelitian kata sifat berasaskan kepada data korpus, maka aspek data korpus ini dibincangkan dengan agak mendalam, di samping skop kajian, permasalahan kajian dan tujuan kajian ini dilaksanakan.

## 1.2 Data Korpus

Bagi menghuraikan data korpus dengan lebih lanjut, maka dalam bahagian ini perbincangan dibahagikan kepada empat aspek, iaitu definisi data korpus, pendekatan “corpus based” dalam kajian linguistik, jenis data korpus dan peranti

analisis data korpus. Kesemua aspek ini dibincangkan dalam bahagian ini kerana aspek ini saling berkait dan mempengaruhi kajian yang dijalankan.

### 1.2.1 Definisi Data Korpus

Istilah korpus berasal daripada bahasa Latin yang bermaksud *body* dan jamak bagi *corpus* ialah *corpora* (Knowles dan Zuraidah Mohd. Don, 2006: 9). Istilah ini dipilih kerana "corpus" diaplikasikan kepada "*...any body of texts, however small and homogeneous, which the linguist assembles for the purpose of analysis...*" (Butler, 2004:150). Bagaimanapun, dalam kajian ini istilah yang digunakan ialah korpus atau data korpus sama ada dalam bentuk tunggal atau jamak. Dalam bidang linguistik, istilah korpus linguistik merupakan istilah moden yang muncul pada tahun 1980-an dan telah dijadikan metodologi secara meluas dalam kajian linguistik. Pada umumnya istilah ini digunakan bagi merujuk data bahasa yang terbentuk secara tabii yang dijadikan asas dalam penyelidikan linguistik (Leech, 1997:1). Data bahasa ini boleh terdiri daripada data tulisan atau data lisan.

Selepas penghasilan Korpus Brown pada tahun 1960, iaitu data korpus pertama yang dianggap sebagai data korpus moden disebabkan data korpus ini dihasilkan dalam bentuk elektronik dan diproses oleh komputer bagi tujuan kajian linguistik dan kejuruteraan bahasa, maka pelbagai definisi telah dibentuk berserta dengan penambahan beberapa ciri bagi menggambarkan data korpus dengan tepat. Yang berikut merupakan tujuh definisi data korpus yang diberikan oleh Renouf;

Francis; the Expert Advisory Group on Language Engineering Standards (dalam Butler, 2004:150); Kennedy; Tognini-Bonelli; Bowler dan Pearson; dan McEnery, Xiao dan Tono.

> *A collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research.*
>
> (Renouf, 1988:1)

> *A corpus is a collection of texts assumed to be representative of a given language, dialect or other subset of a language to be used for linguistics analysis.*
>
> (Francis, 1992 : 17)

> *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.*
>
> (Butler, 2004: 150)

> *Collection of texts in an electronic database. ...necessarily consist of structured collection of text specifically compiled for linguistic analysis, that they are large or that they attempt to be representative of a language as a whole.*
>
> (Kennedy, 1998: 3)

> *Collection of texts assumed to be representative of a given language put togather so that it can be used for linguistics analysis.*
>
> (Tognini-Bonelli, 2001: 2)

> *A large collection of authentic text, that have been gathered in electronic form according to a specific set of criteria*.
>
> (Bowler dan Pearson, 2002: 9)

> *A collection of sampled text, written and spoken, in machine-readable form which may be annotated with various forms of linguistics information.*
>
> (McEnery, Xiao dan Tono, 2006: 4)

Berdasarkan definisi di atas, didapati kesemuanya menyatakan bahawa data korpus merupakan himpunan data bahasa atau teks. Data bahasa ini boleh sahaja dalam bentuk lisan atau tulisan atau kedua-duanya sekali yang tersimpan di dalam komputer dan berfungsi sebagai sampel bahasa bagi penyelidikan linguistik. Dalam definisi yang diberikan di atas, terdapat tiga aspek penting dalam menjelaskan data korpus, iaitu *data yang autentik, data dalam bentuk elektronik* dan *data yang bersaiz besar*.

Autentik atau ketulenan di sini bererti data yang dikumpulkan merupakan data sebenar yang dihasilkan dalam komunikasi sesama manusia. Ini menunjukkan bahawa data yang diselidiki bukan direka oleh penghasil data korpus melainkan data tersebut ialah data yang terbentuk secara tabii. Data seperti ini penting dalam penyelidikan linguistik kerana data ini dapat menunjukkan perilaku sebenar sesuatu bahasa semasa bahasa tersebut digunakan. Data korpus yang dijadikan sampel bahasa ini membolehkan pengkaji mendapat gambaran sebenar tentang sesuatu bahasa bagi membentuk model atau rumus bahasa.

Data yang tersimpan sebagai korpus adalah dalam bentuk elektronik. Selain istilah elektronik, data korpus sering dikaitkan dengan frasa *machine-*

*readability*. Hal ini bermaksud pengendalian data korpus sebahagian besarnya dilakukan oleh mesin, iaitu komputer. Penggunaan komputer dalam bidang korpus linguistik telah meningkatkan keupayaan pemprosesan data. Komputer digunakan untuk penstoran, pemerolehan (menerusi imbasan optik atau *on line* daripada laman web tertentu), pemprosesan dan penganalisisan data. Hal ini secara tidak langsung menjadikan kerja-kerja *pencarian, pemilihan, pengisihan* dan *pemformatan* data dapat dilakukan dengan mudah dan cepat. Data dalam bentuk elektronik ini boleh dimanipulasikan dengan menggunakan perisian-perisian tertentu yang dikenali sebagai peranti analisis korpus (seperti *wordlister* dan *concordanser*).

Ciri elektronik ini memberi kesan dari segi metodologi kerana ciri ini memberi dimensi baharu dalam penelitian linguistik. Data tidak sahaja dihuraikan secara kualitatif, malahan aspek kuantitatif juga digunakan dan ini menjadikan kajian linguistik benar-benar bersifat empirikal. Di samping itu, pemprosesan yang dilakukan secara automatik ini adalah konsisten dan tepat serta mampu mengelakkan *human bias* dalam penganalisisan data. Disebabkan sebahagian besar penyelidikan data korpus kini menggunakan komputer, maka istilah *data korpus berkomputer* telah digunakan bagi kajian yang berasaskan data korpus.

Dari segi saiz data pula penggunaan komputer telah membolehkan data disimpan dalam skala yang lebih besar. Contohnya, Korpus Brown yang merupakan data korpus pertama yang dihasilkan dalam bentuk *machine readable* telahmerekodkan sejumlah satu juta patah perkataan, iaitu data yang dikumpulkan

daripada 500 buah teks dan setiap teks mengandungi 2000 patah perkataan (Garside, Leech dan McEnery, 1997: 1).

Berbanding dengan data korpus yang dihasilkan sebelum ini, iaitu pada tahun 1950-an dan tahun-tahun sebelumnya, kebanyakan ujaran atau bahan tulisan dicatatkan pada helaian-helaian kertas dan disimpan di dalam kotak kasut (McEnery, Xiao dan Tono, 2006: 3). Jumlah data yang dijadikan kajian adalah sedikit dan tidak dapat menggambarkan bahasa secara keseluruhan. Walau bagaimanapun, sejak penghasilan Korpus Brown ini, jumlah data korpus sentiasa ditingkatkan, contohnya data yang terdapat di dalam Bank of English mempunyai 524 juta perkataan (McEnery, Xiao dan Tono, 2006: 9), British National Corpus (BNC) 100 juta patah perkataan (Garside, 1997: 1) dan CONCORD sebanyak 5 juta. Saiz yang besar ini menjadikan data korpus representatif terhadap sesuatu teks.

### 1.2.2 Pendekatan "Corpus Based"

Pendekatan "corpus based" atau pendekatan berasaskan data korpus bererti pendekatan yang memanfaatkan data korpus untuk menguji, memberikan contoh-contoh yang autentik dan menghuraikan secara terperinci sesuatu aspek bahasa yang sebelum ini telah dibentuk oleh teori-teori tertentu. Oleh itu, menerusi pendekatan ini, teori-teori yang dibentuk sebelum kewujudan data korpus diteliti semula dengan menggunakan sejumlah data yang besar kuantitinya (Tognini-Bonelli, 2001: 65). Jika sebelum ini penelitian bahasa lebih merupakan refleksi

ahli linguistik terhadap pengalaman bahasanya sendiri dan berasaskan data manual yang kecil kuantitinya, tetapi dalam pendekatan berasaskan data korpus, huraian sesuatu bahasa merupakan hasil daripada data sebenar yang tinggi jumlahnya dan digunakan oleh penutur natif bahasa itu sendiri.

Hal ini secara tidak langsung menunjukkan bahawa pendekatan ini secara tipikalnya menjadikan teori yang sedia ada sebagai titik permulaan sesuatu penelitian bahasa dan teori tersebut diperbetulkan dan disemak serta diperincikan berdasarkan kepada bukti yang sedia ada, iaitu data korpus berkomputer (McEnery, Xiao dan Tono, 2006: 10). Ciri penting dalam analisis data berasaskan komputer ini ialah penganalisisan data dilakukan secara empirikal, penggunaan sebahagian besar komputer dalam menganalisis data dan teknik analisis dilakukan secara kuantitatif dan / atau kualitatif (Biber, Conrad dan Reppen, 1996: 116). Bentuk analisis seperti ini mampu menyediakan analisis yang andal dan secara khusus pendekatan ini membolehkan pola asosiasi yang lebih kompleks dan sistematik diteliti. Penghasilan data seperti ini membolehkan kajian berasaskan data korpus berkomputer meneliti pola asosisasi, sama ada asosiasi leksikal atau asosiasi tatabahasa.

Secara relatifnya penyelidikan berasaskan data korpus yang menggunakan berkomputer masih baharu. Bagaimanapun, kajian yang menggunakan data korpus sebagai asas penyelidikan bahasa adalah seusia dengan persoalan mengenai bahasa itu sendiri (Tognini-Bonelli, 2001: 50 dan Aijmer dan Altenberg, 1996: 1). 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).

Menurut McEnery, dan Wilson (2001: 3-4), data korpus sebenarnya telah digunakan sejak tahun 1800 lagi dalam bidang linguistik, misalnya dalam kajian perolehan bahasa oleh Preyer (1889) dan Stern (1924), kajian terhadap konvensi ejaan oleh Kading (1897), kajian pedagogi bahasa oleh Fries dan Traver (1940) serta Bongers (1947) dan kajian linguistik bandingan oleh Eaton pada 1940. Data korpus juga telah lama digunakan sebagai asas dalam huraian tatabahasa, khususnya tatabahasa bahasa Inggeris.

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) (dalam Kennedy, 1998:17-18). Kajian yang lebih bersistematik berasaskan data korpus telah dilakukan oleh Charles C. Fries (1973). 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 (Kennedy, 1998:17-18).

Berdasarkan kajian-kajian tersebut, dapat dikatakan bahawa kajian linguistik pada peringkat awal menyerupai data korpuskerana 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. Boas (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 (dalam McEnery, Xiao dan Tono, 2006: 3).

Walaupun beberapa kajian di atas 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 kepada kajian bahasa Indo-Eropah sahaja, tetapi telah tersebar kepada 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 kepada 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 barulah 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 (Radford, 1994: 5).

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 (McEnery dan Wilson, 2001: 5-6). 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 atypical 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 (Tognini-Bonelli, 2001: 2), 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 (1992:8), 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 kepada teori yang menggambarkan realiti bahasa dan model bahasa yang munasabah bergantung sepenuhnya kepada 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. Bahagian bab yang berikut ini akan membincangkan pendekatan berasaskan data korpus dan kritikan terhadap pendekatan berasaskan intuisi semata-mata.

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 (Hunston, 2002:3).

Tambahan pula, data korpus yang dibina memenuhi kriteria tertentu bagi menjadikannya sebagai bentuk yang representatif bagi sesuatu bahasa (Sinclair, 1992: 6). 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.

Berbanding dengan intuisi, 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 ditentu sahkan disebabkan

sukar untuk mengesahkan introspektif seseorang kerana intuisi tidak boleh dilihat atau diteliti (McEnery dan Wilson, 2001: 11).

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.

Berbanding data korpus, intuisi pada umumnya mampu memberi maklumat yang segera terhadap dua perkara sahaja, iaitu makna sesuatu kata secara terpisah dan bentuk ayat yang sempurna juga secara terpisah (Sinclair, 1997:3). 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 dan Wilson (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 (Tognini-Bonelli, 2001: 49) 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-pengkaji yang berbeza bagi menghasilkan model bahasa yang lebih berwibawa (Sampson, 2005: 19). Aspek yang lebih penting lagi, 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 (Tognini-Bonelli, 2001: 49-51). 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/frasa/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 "...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 di atas, Chomsky (1962, dlm. Leech, 1996:8) juga dengan tegas menolak data korpus kerana pada pendapatnya wujud "*skewedness"* terhadap data yang diteliti. Menurut beliau,

> *"Any natural corpus will be skewed. Some sentances 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".*

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. Salah satu cara yang digunakan adalah dengan cara membina data korpus dalam skala yang besar. Saiz yang besar ini, yang mencapai jutaan patah perkataan membolehkan bentuk lazim dan bentuk yang tidak lazim dipaparkan. Hal ini terlihat dalam data korpus pertama yang dihasilkan dalam bentuk digital, iaitu Korpus Brown (1960) yang terdiri daripada satu juta perkataan, manakala korpus-korpus lain yang dihasilkan pada dalam tempoh 30 tahun selepas itu memperlihatkan peningkatan saiz yang amat besar, misalnya Korpus BBC terdiri daripada 120 juta perkataan, Korpus Birmingham (20 juta), Korpus Dutch (60 juta), Korpus Perancis (190 juta) dan Korpus Jerman (27 juta) (Mohd. Zulkifli Bahari, 1993: 32).

Di samping itu, *skewedness* juga dapat dielakkan dengan cara 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. Misalnya, data korpus *COBUILD* yang dibangunkan bagi tujuan penghasilan kamus telah menetapkan beberapa kriteria untuk menjadikannya representatif, antaranya bahan yang dipilih perlulah menggambarkan bahasa dalam pelbagai bidang atau laras; terdiri daripada 75 peratus bahasa tulisan dan 25 peratus bahasa lisan; keutamaan kepada data prosa berbanding puisi dan terdiri daripada bahasa baku yang digunakan oleh penutur natif yang berusia 16 tahun ke atas (Sinclair, 1988: 2-3). Bagi kajian ini

pula, tiga kriteria digunakan dalam pembinaan data korpus bagi mengelakkan *skewedness*, iaitu kriteria pemilihan teks, jumlah teks, dan saiz data korpus (lihat 3.2).

Di samping itu, bagi mengelakkan data korpus ini sama seperti arkib yang merupakan koleksi secara rawak bagi sesuatu data, maka data korpus yang disusun perlu 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 perbahanan yang memuatkan segala teks di dalamnya, tetapi perlu memenuhi kriteria-kriteria tertentu bagi membolehkan data tersebut representatif terhadap bahasa yang diwakilinya. Sinclair (1996, dalam McEnery, Xiao dan Tono, 2006:4) 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".*

Justeru, reka bentuk dan kerelevanan data korpus sebagai asas kajian dalam linguistik telah menyebabkan linguistik pada masa ini, khususnya selepas tahun-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 appropriate. 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.*

### 1.2.3 Jenis Data Korpus

Dalam bahagian ini, data korpus dibahagikan kepada empat kelompok berdasarkan ciri-cirinya, iaitu data korpus umum dan data korpus khusus; data korpus tulisan dan data korpus lisan; data korpus ekabahasa dan data korpus neka bahasa; dan data korpus terbuka dan data korpus tertutup (Bowler dan Pearson, 2002: 11).

#### 1.2.3.1 Data Korpus Umum dan Data Korpus Khusus

Data korpus umum ialah data korpus yang mewakili bahasa secara keseluruhannya. Oleh itu, data korpus ini terdiri daripada data lisan dan data tulisan yang merangkumi pelbagai jenis teks, iaitu laporan akhbar, radio, televisyen, bahan-bahan ilmiah, fiksyen, dan efemeral. Jenis data korpus seperti ini seboleh mungkin menggambarkan keseluruhan aktiviti bahasa yang terdapat dalam sesuatu komuniti bahasa.

Data korpus khusus pula ialah data korpus yang berfokus kepada aspek-aspek tertentu bahasa, misalnya tertumpu kepada bahasa yang digunakan dalam

bidang perundangan atau bidang ekonomi atau bidang politik. Ini bererti bahasa yang termuat dalam data korpus tersebut tertumpu kepada sesuatu laras sahaja.

### 1.2.3.2 Data Korpus Tulisan dan Data Korpus Lisan

Data korpus tulisan ialah data korpus yang memuatkan data daripada bahasa tulisan sahaja, begitu juga data korpus lisan yang terdiri daripada data lisan sahaja. Contoh data korpus lisan ialah London-Lund Corpus of Spoken English (Aijmer dan Altenberg, 1996:1). Bagaimanapun, kebanyakan data korpus terdiri daripada gabungan data lisan dan tulisan seperti British National Corpus dan Bank of English.

### 1.2.3.3 Data Korpus Ekabahasa dan Data Korpus Neka Bahasa

Data korpus ekabahasa merupakan data yang terdiri daripada satu bahasa sahaja seperti data korpus di Dewan Bahasa dan Pustaka, manakala data korpus neka bahasa terdiri daripada teks dua bahasa atau lebih. Data korpus neka bahasa ini dapat dibahagikan kepada dua jenis, iaitu data korpus selari (*paralel corpus*)dan data korpus perbandingan (*comparable corpus*). Data korpus selari ialah data korpus yang terdiri daripada teks dalam sesuatu bahasa dan diterjemahkan ke dalam bahasa-bahasa yang lain. Contohnya teks bahasa Melayu seperti *Seroja Masih di Kolam* telah diterjemahkan dalam bahasa Inggeris dan Jepun. Bagi data korpus perbandingan pula, data korpus ini terdiri daripada teks

dalam kategori yang sama seperti laporan teknikal, manual komputer dan sebagainya, tetapi ditulis dengan menggunakan bahasa yang berbeza.

#### 1.2.3.4 Data Korpus Terbuka dan Data Korpus Tertutup

Data korpus terbuka terdiri daripada data korpus yang sentiasa ditambah datanya dari semasa ke semasa. Oleh itu, saiz data korpus ini sentiasa meningkat dan tidak terhad. Data korpus ini yang turut dikenali sebagai *monitor corpus* kerap digunakan dalam bidang leksikografi. Data korpus seperti ini penting dalam leksikografi kerana kemampuannya merekodkan bahasa dalam konteks semasa bagi membolehkan makna perkataan diperbaharui dari semasa ke semasa. Contohnya data korpus Bank of English telah dijadikan asas dalam penyusunan Collins COBUILDS (McEnery dan Wilson, 2001: 22-24). Data korpus tertutup pula ialah data korpus yang mempunyai saiz tertentu dan saiz ini tidak boleh ditambah. Data korpus seperti ini biasanya digunakan dalam penyelidikan linguistik.

### 1.2.4 Peranti Analisis Data Korpus

Data korpus yang disimpan di dalam komputer telah membolehkan peranti analisis korpus digunakan untuk menganalisis data. Pada umumnya, peranti analisis korpus ini terdiri daripada dua ciri utama, iaitu:

i. penjanaan senarai perkataan dan

ii. penjanaan konkordans.

Antara perisian yang kerap digunakan untuk penghasilan *wordlister* dan *concordancer* ialah *WordSmith Tools*.

Bagi fungsi penyenaraian perkataan, *WordSmith Tools* membolehkan *batch word list* dihasilkan. Berdasarkan senarai perkataan ini, analisis statistik dapat dilakukan terhadap data korpus, iaitu senarai perkataan ini boleh mengira seluruh token yang terdapat dalam sesuatu teks, malahan jumlah kekerapan sesuatu *type* yang hadir di dalam teks tersebut juga boleh dihitung menerusi senarai perkataan ini. Senarai perkataan ini boleh disusun mengikut urutan abjad atau urutan frekuensi. Senarai perkataan ini membolehkan kajian dilakukan terhadap sesuatu teks atau kajian perbandingan yang melibatkan beberapa teks. Nisbah antara type dan token juga boleh dilakukan.

Konkordans pula digunakan untuk meneliti kehadiran perkataan tertentu dalam konteks terdekatnya. Format yang sering digunakan ialah *key words in context (KWIC)*. Berdasarkan format ini, kata yang diteliti hadir di tengah-tangah baris konkordans dan ditunjukkan konteksnya, iaitu dengan perkataan yang hadir di bahagian kanan dan kirinya. Sama seperti senarai perkataan, konteks yang hadir di kiri atau di kanan kata kunci juga disusun mengikut urutan abjad bagi memudahkan penelitian terhadap pola sesuatu kata. Peranti analisis data korpus ini akan dibincangkan dengan lebih lanjut dalam bab 3.

## 1.3 Skop Kajian

Kajian ini terhad kepada aspek golongan kata sifat bahasa Melayu yang terdapat dalam lima buah karya biografi (lihat 3.2.2). Bagi menjelaskan kata sifat dari segi fungsinya, maka kajian ini menggunakan karya biografi sebagai data kajian bagi menjadikan kajian ini lebih empirikal. Karya biografi dipilih kerana biografi tergolong dalam genre ilmiah. Karya ilmiah dijadikan asas penelitian kata sifat bahasa Melayu kerana berdasarkan beberapa kajian terhadap bahasa Inggeris dan bahasa Rusia, kata sifat banyak terdapat dalam karya ilmiah (Nakamura, 1991 dan Rayson, Wilson dan Leech, 2005). Dalam kajian ini, lima buah karya biografi yang dihasilkan oleh Nik Safiah Karim dan Rokiah Talib dijadikan data korpus dan data korpus ini mengandungi sejumlah 157,000 patah perkataan.

Bagi meneliti kata sifat, kriteria sintaktik, morfosintaktik dan morfologi digunakan. Walaupun aspek semantik atau makna merupakan salah satu ciri penentu sesuatu golongan kata, tetapi ciri ini tidak ditekankan dalam kajian ini kerana penentuan kata sifat lebih bergantung pada aspek fungsian (lihat kajian yang dilakukan oleh Dik, 1997; Hengeveld, 1992b; Bhat, 1994 Beck, 2002; dan Croft, 2003 pada bahagian 2.2.2.4). Daripada data korpus ini, kata sifat diteliti dari segi fungsinya kerana dalam kajian ini fungsi dijadikan asas utama dalam menentukan golongan kata sifat. Bagi meneliti fungsi kata sifat, maka kehadiran kata sifat dalam baris-baris konkordans dilihat dari sudut posisinya dalam untaian sintaktik, iaitu berdasarkan kepada distribusi sintaktik. Distribusi sintaktik ini diperjelaskan menerusi gatra dan unsur pengisi. Oleh itu, dalam kajian ini kata

sifat ditentukan berdasarkan kriteria sintaktik, di samping kriteria morfosintaktik dan morfologi (lihat 4.3.3 dan 4.4).

Berdasarkan gatra dan unsur pengisi, tiga fungsi kata sifat ditemui dalam data korpus yang dibangunkan, iaitu kata sifat yang berfungsi sebagai penerang nama dan penerang kerja, kata sifat yang berfungsi sebagai predikat dan kata sifat yang berfungsi sebagai penerang predikat. Walaupun dalam bahasa Melayu, kata sifat juga berfungsi sebagai pelengkap, iaitu kehadirannya selepas kata kerja tak transitif berpelengkap, tetapi dalam kajian ini aspek ini tidak diteliti kerana didapati jumlahnya terlalu rendah[2] berbanding tiga fungsi yang lain. Daripada ketiga-tiga fungsi ini, kesimpulan kajian ini pada akhirnya adalah berkaitan dengan perlakuan kata sifat yang berkait dengan teks ini sahaja dan kesimpulan ini bagaimanapun tidak menggambarkan keseluruhan kata sifat bahasa Melayu. Oleh itu, karya biografi ini merupakan model atau perwakilan bahasa Melayu dan tidak menjelaskan kata sifat secara keseluruhannya. Hal ini demikian kerana aspek kata sifat yang diteliti merupakan milik data yang diteliti dan data ini tidak mampu menggambarkan keseluruhan kata sifat bahasa Melayu (lihat Knowles dan Zuraidah Mohd. Don, 2006:12-14).

[2] Dalam kajian ini, daripada sejumlah 100 kata sifat pertama yang tertinggi kekerapannya dalam data korpus (Lampiran 3), hanya terdapat 15 sahaja kata sifat yang berfungsi sebagai pelengkap kata kerja tak transitif (lihat Lampiran 12). Dalam fungsi ini, kekerapan kehadirannya sebagai pelengkap kata kerja tak transitif juga agak rendah, iaitu kekerapan paling tinggi ialah kata sifat *berat* (11 kali perulangan), diikuti oleh kata sifat *kuat* dan *baik* (5 kali perulangan), *tinggi* dan *kerap* (2 perulangan) dan bagi kata sifat yang lainnya setiap kata sifat ini hanya memaparkan satu contoh sahaja. Oleh itu, kajian ini tidak meneliti kata sifat dalam fungsi ini kerana sukar untuk melihat pola kehadirannya dalam ayat.

## 1.4 Permasalahan Kajian

Walaupun dalam bahasa Melayu, khususnya buku-buku tatabahasa awal telah memuatkan huraian berkaitan kata sifat, tetapi masih terdapat ketidaksepakatan dalam menentukan kata sifat sehingga kini. Antaranya ialah aspek istilah, pendefinisian, perimbuhan, penggolongan dan kombinasi perkataan yang hadir bersama-sama kata sifat. Dari segi istilah, terdapat tiga istilah yang digunakan bagi merujuk kata sifat bahasa Melayu, iaitu istilah *keadaan nama* atau *rupa nama,kata sifat* dan *adjektif.* Istilah yang berbeza ini timbul akibat daripada tanggapan yang berbeza terhadap kata sifat. Istilah *keadaan nama* dan *rupa nama* digunakan oleh Mohd. Shah Yusof (1922) dan Md. Said Sulaiman (1937) kerana golongan kata ini dianggap sebagai kata yang menjelaskan keadaan atau sifat sesuatu kata nama sahaja dan tidak golongan kata yang lain.

Bagaimanapun, Za'ba (1940) menggunakan istilah *sifat* berbanding *keadaan nama* atau *rupa nama.* Bagi Za'ba (1940), *sifat* bukan sahaja golongan kata yang menerangkan kata nama, tetapi turut menerangkan perkataan selain kata nama. Oleh itu, Za'ba menggolongkan kata sifat kepada dua golongan, iaitu sifat nama (penerang nama) dan sifat kata (penerang selain kata nama). Istilah *sifat* turut digunakan oleh ahli linguistik moden seperti Abdullah Hassan (1986), Arbak Othman (1981) dan Asmah Hj. Omar (1993b), tetapi istilah ini hanya merujuk kepada fungsi kata sifat sebagai penerang kata nama sahaja (Abdullah Hassan, 1986: 35). Berbeza dengan Mees (1969), Asraf (1978) dan Nik Safiah et al., (1995), istilah *adjektif* digunakan dalam menjelaskan kata sifat kerana istilah ini

selaras dengan istilah linguistik moden *adjective* bagi golongan kata yang menerangkan kata nama.

Dari segi definisi pula, bahasa Melayu yang penggolongan katanya mengikut acuan bahasa Inggeris[3] turut mendefinisikan kata sifat sebagai golongan kata *yang menerangkan kata nama* (Mohd. Shah Yusuf, 1922; Abdullah Talib, 1928; Syed Muhammad Othman Yahya, 1936, Md. Said Sulaiman, 1937 dan Za'ba 1940). Pendefinisian ini memperlihatkan bahawa definisi yang diterapkan kepada kata sifat adalah berdasarkan fungsinya dan ia berbeza daripada definisi yang diberikan kepada kata nama dan kata kerja kerana kedua-dua golongan ini didefinisikan berdasarkan makna leksikal kata berkenaan. Definisi kata sifat ini menimbulkan persepsi yang berbeza-beza terhadap kata sifat dan akhirnya menimbulkan permasalahan. Pertama ialah definisi seperti ini menjadikan kata nama sebagai asas dalam menentukan golongan kata sifat. Oleh itu, untuk mendefinisikan kata sifat, definisi kata nama perlulah jelas. Umumnya, kata nama ditakrifkan sebagai kata yang menamakan sesuatu benda, perkara, hal dan sebagainya. Bagaimanapun, definisi kata nama juga masih kabur kerana definisi ini tidak dapat menjelaskan dengan tepat golongan kata nama (lihat 2.2.1). Oleh itu, pendefinisian kata sifat yang bergantung pada definisi kata nama menjadi tidak tepat disebabkan oleh kekaburan dalam definisi kata nama.

Kedua, definisi seperti ini juga terlalu luas cakupannya. Hal ini disebabkan dalam menjalankan fungsi sebagai penerang kata nama, kata sifat bukanlah merupakan golongan kata tunggal yang menjadi penerang kepada kata nama.

[3] Definisi tipikal kata sifat menurut Nesfield ialah *a word used to qualify a noun* (Rastall, 1995: 25)

Contohnya dalam bahasa Melayu terdapat frasa *air kopi, air mendidih* dan *air panas*. Perkataan *kopi, mendidih* dan *panas* kesemuanya berfungsi sebagai penerang kata nama, tetapi dari segi penggolongan, ketiga-tiga perkataan tersebut berbeza golongannya, iaitu *kopi* merupakan golongan kata nama, *mendidih* merupakan kata kerja, manakala *panas* merupakan kata sifat. Oleh itu, dalam menjalankan fungsi penerang, kata sifat bukanlah merupakan satu-satunya golongan kata yang menerangkan kata nama, tetapi kata nama dan kata kerja juga menjalankan fungsi yang sama. Di samping itu, definisi kata sifat sebagai penerang juga tidak dapat menjelaskan kategori kata sifat dengan tepat. Contohnya frasa yang berikut:

| | | |
|---|---|---|
| *kerajaan Siam* | - | *upacara kerajaan* |
| *kepercayaan hatinya* | - | *seorang dayang kepercayaan* |

Mees (1969: 77)

Berdasarkan contoh ini, perkataan yang bergaris dalam lajur di sebelah kiri dengan jelas merupakan kata nama kerana perkataan tersebut merupakan kata yang diterangkan. Bagaimanapun, dalam lajur kanan, perkataan yang sama sukar ditentukan golongan katanya dan jika didasarkan kepada definisi, kesemua perkataan ini menepati definisi kata sifat. Oleh itu, pendefinisian kata sifat sebagai penerang nama tidak dapat menjelaskan golongan kata sifat dengan tepat kerana definisi ini tidak dapat memberikan maklumat tentang posisi dan unsur pengisi yang boleh mengisi sesuatu posisi. Dalam penelitian golongan kata, posisi merupakan aspek penting dalam menjelaskan sesuatu golongan kerana posisi merupakan fungsi sesuatu kata. Sesuatu posisi itu boleh diisi oleh golongan kata yang berbeza.

Dari aspek morfologi pula, pengimbuhan kata sifat turut menimbulkan permasalahan. Dalam bahasa Melayu, kata sifat dikatakan mempunyai dua imbuhan yang tipikal, iaitu imbuhan *ter-* dan *se-* [4] (lihat Asmah Hj. Omar, 1962: 448, 1993a: 169; Yock Fang, L., 1967: 139; Yock Fang, L., dan Abdullah Hassan, 1994: 44-45; dan Nik Safiah Karim, 1995: 223-224). Kedua-dua imbuhan ini, merupakan imbuhan utama kata sifat kerana dari segi maknanya, kedua-dua imbuhan ini membawa maksud perbandingan atau pemeringkatan kata/superlatif. Aspek ini sesuai dengan makna yang didukung oleh kata sifat kerana kata sifat pada umumnya merupakan golongan kata yang mengandungi darjah pemeringkatan (darjah kepalingan).

Contoh:

| Kata Dasar | - | Superlatif | - | Perbandingan |
|---|---|---|---|---|
| *besar* | - | *terbesar* | - | *sebesar* |
| *cetek* | - | *tercetek* | - | *secetek* |
| *muda* | - | *termuda* | - | *semuda* |

Walaupun menurut Asraf (1978:560) imbuhan merupakan penanda bagi sesuatu golongan kata, tetapi dalam penentuan kata sifat, imbuhan sukar untuk dijadikan sebagai kriteria utama disebabkan daripada penelitian terhadap data korpus berkomputer, didapati dalam menjalankan fungsi tipikalnya sebagai penerang, kata sifat lazimnya terdiri daripada kata dasar. Di samping itu, sehingga kini dalam bahasa Melayu masih belum ada kajian yang meneliti korelasi antara imbuhan kata sifat dan fungsi tipikalnya dalam ayat. Yang sering dihuraikan hanyalah contoh ayat yang terdiri daripada kata sifat berimbuhan dan dijelaskan

[4]Di samping imbuhan *ter-* dan *se-,* imbuhan *ke-, ber-, pe-,me-, me-...-kan, ber-...an, ke-...-an,-isme, -wi, -iah, -el-, -er-,* dan *–em-* juga dianggap sebagai imbuhan kata sifat. Bagaimanapun imbuhan ini tidak seproduktif imbuhan *ter-* dan *se-* (lihat Asmah Hj. Omar, 1993: 169; Yock Fang, L., dan Abdullah Hassan, 1994: 44-45; dan Nik Safiah Karim, 1995: 223-224).

maknanya sama ada sebagai kata yang mendukung makna darjah kesangatan atau sebagai kata yang mendukung maksud perbandingan (lihat Asmah Hj. Omar, 1962: 448, 1993a: 169; Yock Fang, L., 1967: 139; Yock Fang L., dan Abdullah Hassan, 1994: 44-45; dan Nik Safiah Karim, 1995: 223-224 ).

Dari segi penggolongan kata sifat pula, sejak awal lagi terdapat perbezaan pendapat sama ada kata sifat tergolong sebagai golongan utama atau sebagai subgolongan sama ada subgolongan kata nama atau kata kerja. Za'ba (1940, 1958), Mees (1969), Arbak Othman (1981) dan Nik Safiah Karim (1995) mengkategorikannya sebagai golongan utama, manakala Marsden (1812) meletakkannya di bawah subgolongan kata nama, sementara Abdullah Hassan (1986) dan Asmah Haji Omar (1993b) mengkategorikannya dalam subgolongan kata karyaan atau kata kerja. Bagi ahli bahasa yang meletakkan kata sifat sebagai golongan utama beranggapan bahawa kata sifat ini mempunyai ciri atau kriteria yang tersendiri yang berbeza daripada golongan kata yang lain, sementara bagi pendapat yang meletakkan kata sifat sebagai subgolongan kata nama pula beranggapan bahawa kata sifat ini merupakan kata yang berfungsi sebagaipenerang kata nama. Justeru itu, disebabkan kata sifat ini lazim hadir bersama-sama kata nama, maka kata sifat ini sesuai dikategorikan di bawah kata nama. Pengkategorian kata sifat di bawah kata karyaan atau kata kerja pula disebabkan oleh perilaku kata sifat ini yang menyamai perilaku kata kerja, iaitu fungsinya sebagai predikat ayat (Asmah Hj. Omar, 1993a:146). Misalnya:

(i) kedudukan kata sifat yang sama seperti kata kerja tak transitif apabila hadir di dalam ayat.

Contoh:

| | |
|---|---|
| Pemuda itu tidur. (KKttr) | Rumah itu cantik. (KS) |
| Adik menangis. (KKttr) | Mahasiswa rajin. (KS) |

(ii) kedua-dua kata sifat dan kata kerja tak transitif boleh didahului oleh kata bantu.

Contoh:

| | |
|---|---|
| Adik *masih* makan. (KK) | Dia *masih* cantik. (KS) |
| Anak itu *sudah* tidur. (KK) | Anak itu *sudah* pandai. (KS) |
| Kita *boleh* pergi (KK) | Kita *boleh* kaya. (KS) |
| Kamu *harus* datang. (KK) | Kamu *harus* rajin. (KS) |

Abdullah Hassan (1973:371), turut menyatakan bahawa kata sifat tergolong di bawah kata kerja tak transitif statif kerana kata sifat merupakan kata yang menerangkan sesuatu keadaan dan penggolongan seperti ini menyamai penggolongan kata sifat dalam bahasa Inggeris dan menurut Robins (1964:266),

> *In several languages, ...Malay and Japanese are examples, the translation equivalents of many adjective in European Languages are best regarded formally as a subclass of intransitive verbs.*

Dari segi kombinasi kata sifat pula, lazimnya kata intensiti merupakan kategori kata yang sering hadir bersama-sama kata sifat. Kata intensiti ini dianggap sebagai penanda tipikal bagi kata sifat kerana hampir semua kata sifat boleh digandingkan bersama-sama kata intensiti (Asmah Hj. Omar (1962: 448-449), Yock Fang, L. (1967: 308, 1983: 132-140, 1994: 5-6, 41-42 dan Asraf, 1978:560). Contoh: *sangat jahat, paling cantik, sungguh manis, amat tinggi* dan

*kecil* *sekali*. Aspek ini menyebabkan kata intensiti dijadikan sebagai penguji keahlian bagi golongan kata sifat (Asraf Abdul Wahab, 1978:560). Bagaimanapun, apabila meletakkan kata intensiti sebagai penanda tipikal kata sifat, secara tidak langsung telah membataskan golongan kata lain yang boleh digandingkan bersama-sama kata sifat. Oleh itu, kekompleksan kata sifat tidak dapat dijelaskan dengan menyeluruhkerana dalam bahasa Melayu, kata intensiti bukan merupakan satu-satunya bentuk kata yang hadir bersama-sama kata sifat. Di samping itu, walaupun kata intensiti dijadikan sebagai penanda golongan dan penguji tipikal bagi kata sifat, tetapi aspek ini masih belum dapat menjelaskan korelasi antara fungsi kata sifat dan kehadirannya bersama-sama kata intensiti atau bentuk kata yang lain.

Berdasarkan permasalahan di atas, maka penelitian yang mendalam terhadap kata sifat perlu dilakukan. Bagi menjelaskan persoalan-persoalan ini, maka fungsi yang dijalankan oleh kata sifat perlu diteliti. Fungsi ini perlu diteliti kerana istilah, definisi, imbuhan, penggolongan dan kekompleksan kata sifat adalah berdasarkan fungsi yang dijalankan oleh golongan kata ini. Bagi meneliti fungsi ini, sejumlah data korpus diperlukan kerana daripada data ini akan diperoleh perilaku kata sifat apabila hadir di dalam frasa, klausa atau ayat. Dengan meneliti fungsi kata sifat ini boleh diperoleh maklumat berkaitan fungsi tipikal kata sifat, korelasi antara fungsi dan aspek morfologinya, korelasi antara fungsi dan kekompleksan kata sifat dan perbezaan antara kata sifat dengan bentuk kata yang lain apabila hadir dalam fungsi yang sama.

## 1.4 Tujuan Kajian

Pada umumnya kajian ini bertujuan untuk meneliti kata sifat berdasarkan data korpus berkomputer. Daripada data korpus yang dijana ini barulah dijelaskan perlakuan kata sifat bahasa Melayu. Secara khusus, kajian ini meneliti tiga fungsi kata sifat dalam untaian sintaktik, iaitu :

i. menelilti fungsi kata sifat sebagai unsur penerang bagi sesuatu kata inti. Dalam fungsi ini akan diteliti kata sifat dan unsur penerang lain yang boleh hadir dalam gatra penerang. Dalam menjalankan fungsi ini aspek yang turut diteliti ialah golongan kata inti yang diterangkan oleh kata sifat, sama ada kata nama sahaja atau golongan-golongan kata yang lain. Bagi menggambarkan kata sifat dalam fungsi tipikalnya ini, kata sifat juga dibezakan daripada kata nama dan kata kerja, iaitu golongan kata yang juga menjalankan fungsi penerang. Oleh itu, di sini konstrastif antara kata sifat dengan kata nama dan kata kerja turut diteliti.

ii. menjelaskan fungsi predikat, iaitu fungsi tipikal bagi kata kerja. Dengan meneliti fungsi ini, aspek yang ingin diteliti ialah kata sifat sebagai unsur yang boleh mengisi gatra predikat.

iii. menjelaskan fungsi penerang predikat. Dalam fungsi ini, akan diteliti perilaku kata sifat, iaitu ciri yang menentukan kata sifat dalam fungsi penerang predikat.

Berdasarkan ketiga-tiga fungsi ini akan ditunjukkan fungsi yang paling lazim bagi kata sifat. Dari sini barulah diketahui fungsi tipikal kata sifat bahasa Melayu.

## 1.5 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, jumlahnya yang besar, terdapat dalam bentuk elektronik dan disusun secara sistematik (Hunston, 2002: 14). 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 ini menggunakan data korpus sebagai asas dalam penelitian aspek tatabahasa, khususnya penggolongan kata kerana data korpus berkomputer pada ketika ini merupakan kecenderungan baharu bagi ahli linguistik dalam meneliti aspek bahasa. Data korpus kini bukan sahaja digunakan dalam penelitian terhadap bahasa Inggeris, malahan bahasa Russia, China Jepun, Drybal dan Catalan. Bahkan pada tahun-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 Mohd. Don (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 (2005). Justeru, penggunaan data korpus sebagai asas dalam kajian penggolongan kata sifat adalah relevan. Hal ini juga penting seperti yang diperkatakan oleh Sinclair (1997), "*...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).