Pembinaan sesuatu data korpus, khususnya data korpus bersaiz kecil bagi meneliti aspek bahasa yang tertentu bergantung kepada tujuan kajian. Hal ini kerana tujuan kajian akan menentukan teks yang dipilih untuk dijadikan data korpus (Kennedy, 1998:71). Umumnya, untuk membangunkan sesuatu data korpus beberapa aspek perlu diambil kira, iaitu dari segi:

i. pemilihan teks,

ii. jumlah teks yang memadai untuk sesuatu kajian, dan

iii. saiz data korpus yang perlu dibangunkan

(Biber, 1993: 243-257 dan McEnery, Xiao dan Tono. 2006: 125).

Di samping tiga aspek di atas, bahagian ini turut membincangkan bentuk teks yang digunakan dalam kajian ini.

3.2.1 Pemilihan Teks

Kriteria pemilihan sesuatu teks untuk dijadikan data korpus terbahagi kepada dua, iaitu kriteria luaran dan dalaman. Kriteria luaran melibatkan situasi, iaitu genre atau laras sesuatu teks, manakala kriteria dalaman melibatkan taburan ciri-ciri linguistik. Antara kedua-dua kriteria ini, kriteria eksternal lebih diutamakan (Sinclair, 1995; Atkins, Clear dan Ostler, 1992: 5-6; dan Biber, 1993:256 dalam McEnery, Xiao dan Tono, 2006: 14). Di samping aspek genre, pemilihan teks juga bergantung kepada aspek linguistik yang ingin diteliti, iaitu tujuan sesuatu kajian dijalankan. Aspek penelitian yang berbeza menyebabkan data korpus yang dibina juga berbeza-beza. Ini dapat dilihat dalam kajian oleh

Johansson dan Oksefjell (1996), Suad Awab (1999) dan Yuanwen (2002), yang membentuk data korpus selari dengan keperluan kajian mereka. Oleh itu, kajian ini yang meneliti kata sifat bahasa Melayu memilih teks atau genre yang tinggi frekuensi kata sifat di dalamnya.

Sebenarnya kajian yang meneliti golongan kata dan kaitannya dengan teks atau genre masih belum dilakukan terhadap bahasa Melayu. Bagaimanapun, bagi bahasa Inggeris, bahasa Russia dan beberapa bahasa yang lain, kajian perkaitan antara kekerapan golongan kata dengan sesuatu genre telah dilakukan. Oleh itu, penelitian terhadap bahasa Russia dan bahasa Inggeris ini dijadikan asas dalam menentukan pemilihan genre untuk meneliti kata sifat bahasa Melayu. Dalam bahasa Russia misalnya, berdasarkan kajian oleh Hoffmann (1995, dalam Rayson, Wilson dan Leech, 2005: 299), kata nama, kata sifat dan kata sendi mempunyai kekerapan yang tinggi dalam genre sains, manakala kata kerja, kata adverba dan kata ganti nama banyak terdapat dalam genre yang bersifat imaginasi. Hal ini menyamai kajian yang dilakukan oleh Nakumura (1991) yang meneliti frekuensi tag golongan kata bahasa Inggeris berdasarkan lima belas genre yang disimpan di dalam korpus LOB. Dengan menggunakan *Hayashi's Quantification Method Type III* hasil yang didapati oleh beliau ialah genre yang bermaklumat mempunyai kekerapan kata nama, kata sifat dan kata sendi nama yang tinggi berbanding golongan kata yang lain. Sementara itu, kata kerja, adverba dan kata ganti nama mempunyai frekuensi yang tinggi dalam genre imaginasi (dalam Rayson, Wilson dan Leech, 2005: 299).

Begitu juga kajian yang dilakukan oleh Biber (1998) turut mendapati bahawa kata sifat lebih kerap ditemui dalam genre yang bermaklumat atau ilmiah,

manakala kata kerja, ganti nama dan adverba dalam genre imaginasi. Hal ini dengan jelas menunjukkan bahawa dalam bahasa Inggeris dan bahasa Rusia, kata sifat mempunyai frekuensi yang tinggi dalam genre ilmiah berbanding genre-genre yang bersifat imaginasi (Rayson, Wilson dan Leech, 2005: 303).

Berasaskan kajian-kajian tersebut, maka kajian ini turut memilih salah satu genre ilmiah dalam meneliti kata sifat, iaitu karya biografi. Biografi dikategorikan sebagai bacaan yang ilmiah kerana biografi tergolong dalam sumber bacaan bukan fiksyen (Nik Anuar Nik Mahmud, 2005: 1). Tambahan pula, penulisan biografi amat menitikberatkan fakta, iaitu karya yang dihasilkan berasaskan kepada sesuatu kajian dan bukan rekaan atau imaginasi penulis semata-mata. Bahkan terdapat beberapa biografi yang dihasilkan untuk memenuhi keperluan ijazah pertama hinggalah ke peringkat ijazah kedoktoran (Monir Yaacob, 2005:1). Dalam kajian ini, karya biografi ini terdiri daripada siri biografi "wanita menulis untuk wanita" yang dihasilkan oleh Nik Safiah Karim dan Rokiah Talib (lihat 1.3 dan 3.2.2) telah dipilih.

3.2.2 Jumlah Teks

Terdapat lima buah buku dihasilkan dalam siri "wanita menulis untuk wanita" dan kesemua buku tersebut telah dipilih sebagai data kajian. Buku-buku tersebut ialah *Siti Hasmah : Citra Wanita Dua Zaman, Ibu Enjah Ibu Mithali, Tan Sri Fatimah : Potret Seorang Pemimpin, Wan Mas Wan Ibrahim – Ibu Mithali Ke-2* dan *Tan Sri Zaleha Ismail : Aspirasi dan Perjuangan.* Buku ini dihasilkan

sebagai penghargaan serta penelitian terhadap sumbangan wanita dalam pembangunan negara. Empat daripada lima buah buku biografi ini dihasilkan oleh Nik Safiah Karim dan Rokiah Talib, manakala sebuah lagi buku, iaitu *Siti Hasmah : Citra Wanita Dua Zaman* dihasilkan oleh kumpulan penulis daripada Persatuan Siswazah Wanita Malaysia (PSWM) dan Nik Safiah Karim serta Rokiah Talib turut terlibat dalam penghasilan buku ini.

Lima buah buku yang dipilih ini memadai untuk kajian ini kerana buku-buku ini dianggap mempunyai jumlah kata sifat yang tinggi. Hal ini kerana untuk menggambarkan seseorang tokoh, khususnya tokoh wanita, maka banyak kata sifat digunakan. Tambahan pula, kesemua teks ini telah mendapat kebenaran daripada kedua-dua penulis biografi tersebut. Kebenaran atau hak cipta[10] ini amat penting bagi menerbitkan sesuatu bahan dalam bentuk data korpus kerana untuk mendapatkan kebenaran ini amat sukar diperoleh walaupun sesuatu teks itu hanya diperlukan untuk dijadikan sebagai sumber kajian yang tidak komersial sifatnya (McEnery, Xiao dan Tono, 2006: 72 dan Kennedy, 1998: 76-77). Di samping itu, dua daripada lima teks ini adalah dalam bentuk *softcopy* yang memudahkan pembinaan data korpus.

3.2.3 Saiz Data Korpus

[10] Kajian ini tidak berjaya mendapatkan teks biografi dalam bentuk *softcopy* daripada Pusat Dokumentasi Melayu, Dewan Bahasa dan Pustaka dan juga bahan daripada data korpus bahasa Melayu di Bahagian Penyelidikan disebabkan isu hak cipta (berdasarkan beberapa perbincangan dengan pegawai-pegawai yang berkenaan pada Julai – Oktober 2006).

Dari segi saiz, data korpus yang dibina daripada lima buah buku biografi ini mengandungi sejumlah 157,328 patah perkataan. Jadual yang berikut menyenaraikan bilangan token yang terdapat dalam setiap teks.

Jadual 3.1
Bilangan Perkataan bagi Setiap Teks

Bil.	Teks	Jumlah Perkataan
1	Siti Hasmah : Citra Wanita Dua Zaman	36,153
2	Ibu Enjah Ibu Mithali	27,755
3	Tan Sri Fatimah : Potret Seorang Pemimpin	37,651
4	Wan Mas Wan Ibrahim – Ibu Mithali Ke-2	19,617
5	Tan Sri Zaleha Ismail : Aspirasi dan Perjuangan	36,152
	JUMLAH	**157,328**

Walaupun jumlah ini dianggap kecil dalam kajian korpus, tetapi jumlah ini adalah memadai dalam penelitian kata sifat bahasa Melayu kerana jumlah ini mampu menggambarkan kata sifat yang terdapat dalam genre ini. Tambahan pula, kajian ini merupakan kajian rintis dalam penelitian golongan kata bahasa Melayu, maka jumlah ini dirasakan wajar. Menurut Biber, saiz data korpus yang digunakan dalam kajian tatabahasa adalah kecil berbanding kajian leksikal. Hal ini disebabkan kajian tatabahasa perlu meneliti distribusi/taburan sesuatu perkataan. Bahkan menurut Leech (1991, dalam McEnery, Xiao dan Tono, 2006: 72), saiz tidak begitu penting. Data korpus yang terdiri daripada 1000 perkataan juga memadai sekiranya data korpus berkenaan mengandungi contoh yang memadai bagi aspek linguistik yang diteliti. Bahkan, dalam kajian golongan kata bahasa Melayu, Knowles dan Zuraidah Mohd. Don (2006: 11) turut menggunakan sampel

sebanyak 120,000 patah perkataan bahasa Melayu yang diambil daripada empat buah teks.

Di samping itu, saiz data korpus juga dipengaruhi oleh bentuk data yang digunakan untuk membangunkan data korpus. Disebabkan kajian ini menggunakan pengimbas optik untuk membina data korpus, maka data korpus yang dibangunkan tidak boleh bersaiz besar. Hal ini adalah kerana data yang diimbas menggunakan mesin pengimbas optik mempunyai banyak kesalahan berbanding data yang terdapat dalam bentuk *mechine-readable.* Ini telah menyebabkan masa yang digunakan untuk menghasilkan sesuatu data adalah lama kerana setiap bahan perlu diedit terlebih dahulu bagi memastikan ketepatannya.

Saiz yang kecil juga disebabkan data yang diteliti perlu dianotasi secara manual. Ini kerana data korpus bahasa Melayu masih tidak mempunyai pelabelan golongan kata secara automatik. Justeru aspek ini turut mempengaruhi saiz data korpus yang dihasilkan.

3.2.4 Bentuk Teks

Teks yang digunakan dalam kajian ini merupakan teks tulisan dalam bentuk buku. Bagi membolehkan teks ini disimpan di dalam komputer, maka teks ini perlu dipindahkan ke dalam bentuk yang boleh dibaca oleh komputer, iaitu dalam bentuk *machine-readable.* Disebabkan dua buah teks sahaja yang terdapat dalam bentuk *softcopy*, maka teks yang selebihnya perlu diimbas menggunakan

mesin pengimbas optik. Penggunaan mesin pengimbas optik ini memerlukan teks yang telah diimbas diedit semula bagi memastikan ketepatannya. Walaupun penggunaan mesin pengimbas optik ini lebih cepat berbanding penaipan semula teks, tetapi mesin pengimbas ini sering melakukan kesilapan semasa membaca sesuatu huruf. Antara bentuk kesilapan yang tipikal ialah:

o	berubah menjadi	*a*	contoh	:	*mendorang,arang*
d	berubah menjadi	*cl*	contoh	:	*clatang*, *menclampingi*
l	berubah menjadi	*1*	contoh	:	*1ama, me1akukan*
e	berubah menjadi	*c*	contoh	:	*kccekapan,eita-cita*
m	berubah menjadi	*in*	contoh	:	*mainpu, inasyarakat*

Bagi mengesan kesilapan ejaan, maka penyemak ejaan, iaitu Dewan Eja Pro telah digunakan. Penyemak ejaan ini dapat menyemak kesilapan ejaan dan memberikan beberapa cadangan bagi bentuk ejaan yang betul. Bagaimanapun, keseluruhan ejaan perlu diteliti satu demi satu secara manual kerana walaupun ejaan tersebut dianggap betul oleh penyemak ejaan tetapi dari segi konteksnya terdapat beberapa perkataan yang masih salah. Contohnya antara perkataan *satu* dan *sate;calon* dan *talon;* penyemak ejaan menganggap kedua-dua bentuk ini adalah betul. Walaupun, dari segi konteksnya bentuk yang betul bagi teks tersebut ialah perkataan *satu* dan *calon,* tetapi *sate* dan *talon* tidak ditandai sebagai salah kerana perkataan ini juga merupakan perkataan bahasa Melayu. Justeru bagi memastikan ejaan yang digunakan bertepatan dengan konteksnya, maka setiap ejaan perlu diteliti.

3.3 Perisian *WordSmith*

Kajian ini menggunakan program *WordSmith* dan daripada program ini perisian *WordList* digunakan. *WordList* ini boleh membantu penghasilan *batch word list.* Daripada perisian ini, output dihasilkan dalam tiga format yang berbeza, iaitu:

i. analisis statistik
ii. senarai tatatingkat frekuensi (frequency ranked word list)
iii. senarai kata mengikut urutan abjad.

(Bowler dan Pearson, 2002:109)

Daripada data korpus ini, kajian ini telah menggunakan senarai kata (*wordlister*), iaitu perisian ini telah menyenaraikan keseluruhan token yang terdapat dalam teks yang dikaji. Perkataan disenaraikan dalam *wordlister* ini berdasarkan kepada urutan abjad atau urutan kekerapan perkataan (Sinclair, 1991: 31) (lihat Lampiran 1 dan 2). Walau bagaimanapun, dalam lampiran ini hanya dipaparkan sepuluh halaman pertama senarai perkataan ini sahaja. Daripada *wordlister* ini (Lampiran 1 dan 2) didapati sebanyak 9342 *types* yang disenaraikan. Disebabkan kajian ini menggunakan data korpus yang terdiri daripada data mentah (data yang masih belum mempunyai pelabelan kelas kata), maka berdasarkan taksonomi kata sifat oleh Dixon (1982)[11] sebanyak 367 kata sifat diperolehi daripada senarai kata tersebut, iaitu bersamaan dengan 3.92 peratus, tetapi yang diteliti hanyalah 282 kata (Lampiran3). Daripada jumlah ini, kata sifat yang paling tinggi kekerapannya ialah perkataan *besar*, iaitu hadir sebanyak 269 kali dan ini diikuti oleh perkataan *baik* (208 kali), dan *kecil* (180

[11] Berdasarkan kriteria semantik, sintaktik dan morfologi, Dixon (1982:16) menggolongkan kata sifat kepada 7 subgolongan, iaitu ukuran (*besar, kecil*), keadaan/sifatan (*panas, berat*), warna (*merah,, putih*), perasaan (*gembira, pandai*), usia/waktu (*baharu,tua*), nilai (*baik, miskin*) dan kecepatan (*perlahan, cepat*).

kali). Selebihnya kurang daripada 180 kali. Daripada jadual ini, sebanyak empat perkataan yang mempunyai kekerapan antara 100 kali hingga 179 kali, manakala 360 lagi mempunyai kekerapan antara sekali hingga 99 kali. Ini menunjukkan bahawa kemampuan sesuatu kata sifat untuk hadir berulang kali dalam data korpus adalah rendah.

Daripada jumlah ini, didapati bahawa jumlah kata sifat yang paling banyak terdapat di dalam teks ini ialah kata sifat yang kekerapannya kurang daripada 10 kali, iaitu sebanyak 256 kata sifat (69.75 peratus). Hal ini telah mempengaruhi pemilihan kata sifat yang diteliti. Oleh itu, kata sifat yang dijadikan tumpuan analisis ialah kata sifat yang mempunyai kekerapan dua kali dan ke atas, iaitu sebanyak 282 kata sifat (76.83 peratus) (lihat Lampiran 4). Pemilihan kekerapan dua kali ke atas bertujuan agar kajian ini meliputi hampir keseluruhan kata sifat yang diteliti kerana bagi kata sifat yang terdiri daripada kata terbitan, kemampuan untuk kata sifat ini hadir dalam kekerapan yang tinggi amat sedikit. Kata sifat berawalan *ter-* misalnya, sejumlah enam kata terbitan ini hadir dengan kekerapan dua kali. Justeru, bagi membolehkan kajian ini lebih menyeluruh, maka kekerapan dua kali dan ke atas dianggap wajar.

3.4 Konkordans

Berdasarkan maklumat yang diperoleh daripada *wordlister* ini, maka konkordans dijana untuk meneliti taburan kata sifat yang hadir di dalam teks. Konkordans yang dihasilkan mempunyai rentang (span) $\pm$ 4 dan ini bererti kira-

kira empat atau lima perkataan hadir di kiri dan kanan kata kunci. Konkordans ini kemudiannya diisih sama ada isih tengah, isih kanan atau isih kiri. Isih tengah bererti kata kunci disusun mengikut urutan abjad (*centre*), manakala bagi isih kiri, perkataan pertama sebelum kata kunci disusun mengikut urutan (*left 1*). Begitu juga dengan isih kanan, iaitu perkataan pertama selepas kata kunci diisih mengikut urutan abjad (*right 1*). Yang berikut merupakan contoh baris-baris konkordans bagi perkataan *besar*.

18	sangsi." Kenyataan terakhir ini begitu	**besar** ertinya kepada Siti Hasmah adik-be
19	pada 1 Ogos 1956 diikuti dengan berinai	**besar** sehingga ke hari bersanding pada 5
20	di atas pelamin. Pada malam berinai	**besar**, pengantin lelaki hadir sama tetap
21	tal **Besar** Kuala Lumpur. Bagi berinai	**besar** pula, pengantin perempuan berpakai
22	i ini tidak berancang untuk berkeluarga	**besar**. Dua orang anak pun sudah cukup, k
23	gu dengan Rafidah dan jika ini berlaku,	**besar** kemungkinan beliau akan menjadi pa
24	hawa tauke-tauke yang mempunyai bot-bot	**besar** itu sebenarnya tidak punya lesen u
26	angkat, walhal keluarganya sudah cukup	**besar**. Sebenarnya hal demikian biasa bag
27	u diluahkan oleh mummy ialah," Kau dah	**besar** panjang, tinggi pelajaran, tinggi
122	sanya, hingga mereka mengadakan kenduri	**besar** menyembelih lembu dan menjemput YB
124	an kad kepada perwakilan. Beliau ketawa	**besar** mengenangkan modal beliau cuma RM2
125	engan Ketua Polis Negara, Zaleha ketawa	**besar** mengimbas kembali episod tersebut.
126	h juga berasa geram apabila kuasa-kuasa	**besar** mengenakan tindakan terhadap negar
127	asa bimbang dengan kata dua kuasa-kuasa	**besar** terhadap isu-isu yang melanda nega
236	aleha. Kedua-dua anak muda ini terkejut	**besar** dan tidak percaya bahawa orang yan
237	san persatuan. Tentunya beliau terkejut	**besar** kerana beliau terkenal dengan cara
261	ka itu, kami tidak mempunyai modal yang	**besar** untuk membuka klinik," cerita Siti
262	bahawa latar belakang keluarganya yang	**besar** banyak mempengaruhi beliau, yang s
263	suatu keluarga atau kumpulan rakan yang	**besar**. Semua penuntut mengambil berat te
264	Hasmah melalui dua titik perubahan yang	**besar**. Pertama ialah apabila Mahathir me
265	han UNIFEM kerana sumbangan beliau yang	**besar** terhadap kesihatan wanita dan kana

Contoh ini menunjukkan bahawa konkordans ini diisih kiri (*L1*). Berdasarkan baris konkordans di atas, kata sifat *besar* akan ditentukan kehadiran dan fungsinya, iaitu sama ada hadir dalam gatra penerang bagi kata inti untuk berfungsi sebagai penerang nama dan penerang kerja, hadir dalam gatra predikat untuk berfungsi sebagai predikat, atau hadir dalam gatra penerang predikat untuk berfungsi sebagai penerang kepada predikat (lihat Bab 4).

Di sebabkan kajian ini menggunakan baris konkordans, maka analisis dilakukan pada tahap atau peringkat sintaksis. Bagaimanapun, dalam menganalisis kata sifat yang berfungsi sebagai penerang, maka tahap analisis ialah pada tahap

frasa (lihat 4.1), manakala bagi fungsi predikat dan penerang predikat, analisis adalah pada tahap klausa atau ayat.

3.5 Kesimpulan

Dalam kajian ini, kata sifat bahasa Melayu diteliti berdasarkan data korpus yang dibangunkan. Data korpus sejumlah 150,000 patah perkataan adalah memadai kerana kajian ini merupakan kajian terhadap golongan kata, iaitu kata sifat. Daripada data korpus tersebut dijana baris-baris konkordans dan daripada baris konkordans ini diteliti aspek kehadiran kata sifat dalam binaan sintaksis. Berdasarkan kepada kehadiran kata sifat ini, barulah ditentukan fungsi-fungsi kata sifat tersebut, sama ada berfungsi sebagai penerang nama atau sebagai predikat atau penerang predikat.