# Pangkalan Data Korpus DBP:
# Perancangan, Pembinaan dan Pemanfaatan

Rusli Abdul Ghani (rusli@dbp.gov.my)
Norhafizah Mohamed Husin( hafizah@dbp.gov.my)
Chin Lee Yim (chin@dbp.gov.my)
**DEWAN BAHASA DAN PUSTAKA MALAYSIA**

Abstrak

Kertas ini menghuraikan aspek perancangan dan pembinaan pangkalan data korpus DBP dari mula pembinaannya pada 1980-an hingga ke awal 2004. Tumpuan kertas ini, bagaimanapun, adalah terhadap perancangan, pembinaan dan pemanfaatan pangkalan data teks bahasa Melayu yang dibangunkan di Dewan Bahasa dan Pustaka Malaysia khusus untuk penelitian berasaskan korpus. Dari sudut perancangan dan pembinaan sistem pangkalan data kami melakarkan kriteria sistem yang perlu diambil kira dalam menyiapkan reka bentuk pembinaan pangkalan data. Dari sudut data pula, yang utama ialah aspek kriteria pemilihan dan tipologi teks. Aspek ini kami jabarkan berdasarkan jumlah dan jenis teks, waktu teks dihasilkan, aspek pengarang, wadah teks dan sebagainya kerana ini memberikan kesahihan dan kewajaran kepada penyelidikan yang berasaskan korpus. Dalam bahagian akhir kami membincangkan beberapa aspek penelitian bahasa yang boleh dilaksanakan dengan menggunakan korpus yang besar, seimbang dan yang dapat mencerminkan penggunaan sebenar bahasa Melayu oleh penutur aslinya.

## 1.0 PENDAHULUAN

Semuanya bermula dengan Korpus Universiti Brown (Francis dan Kučera 1964). Korpus[1] yang mencetuskan penelitian linguistik berasaskan korpus ini masih digunakan sehingga sekarang, tentunya dengan beberapa pembaikan termasuk diberikan penandaan pada tahun 1979 dan tersedia dalam enam versi dari yang asal hingga ke versi keenam iaitu *Brown MARC form* yang disediakan oleh Universiti Stanford (untuk perincian lihat http://helmer.aksis.uib.no/icame/brown/bcm.html).

[1] ***Korpus*** di sini bersinonim dengan 'korpus komputer' dan membawa maksud "himpunan teks digital yang dikumpulkan berdasarkan kriteria tertentu". Dalam kertas ini, demi kepraktisan, kami menggunakan istilah 'korpus' sebagai bermaksud "korpus komputer" melainkan dinyatakan sebaliknya.

Korpus Brown ini asalnya terdiri daripada sejuta kata bahasa Inggeris AS. Korpus ini terbina daripada 500 sampel teks, setiap satu sebesar 2000 kata, dipetik daripada pelbagai genre. Sejuta kata boleh dianggap sangat besar mengingat akan kekangan perkakasan dan upaya pemprosesan yang ada pada era itu.

Bagaimanapun, menjelang pertengahan 1970-an saiz korpus lain seperti *Birmingham Collection of English Texts* (BCET) membesar daripada 7.3 juta kata kepada 20 juta menjelang tahun 1985 diikuti dengan pangkalan korpus lain yang jauh lebih besar, seperti British National Corpus (http://www.natcorp.ox.ac.uk/), dengan teks tulisan dan lisan sebesar 100 juta kata.

Di Dewan Bahasa dan Pustaka pula, usaha awal pemanfaatan himpunan teks dalam penelitian bahasa melibatkan pembangunan pangkalan data pada 1983 di bawah Projek Analisis Teks Secara Komputer (Zaiton Ab. Rahman 1987). Projek ini mensasarkan data teks sebesar dua juta kata melalui teknik pensampelan *à la* korpus Brown. Namun, tatkala saiznya belum pun mencecah setengah juta, kriteria pensampelan diabaikan dan teks lengkap mula dikumpulkan untuk mengambil kira keperluan perkamusan dan kajian bahasa yang memerlukan konteks yang lebih luas dan wacana yang utuh.

## 2.0 PEMBINAAN PANGKALAN DATA KORPUS

Huraian bahagian ini hanya menyentuh secara umum tiga aspek pembinaan pangkalan data korpus sahaja kerana tumpuan kertas ini adalah pada tipologi data teks itu sendiri. Yang pertama ialah objektifnya, kedua reka bentuk awal dan ketiga reka bentuk pangkalan data korpus DBP yang seimbang dan representatif.

### 2.1 Objektif dan Tujuan

Objektif pembinaan pangkalan data korpus yang digariskan dalam Sasaran Kerja Utama DBP 2001–2005 adalah pengumpulan sebanyak 30 juta kata, lalu menjadikan jumlah kumulatifnya sebanyak 120 juta kata pada tahun 2005 (Selain itu, di bawah program

pembinaan sistem korpus, sebuah sistem korpus yang baru akan dibina sebagai ganti sistem sedia ada yang dibina melalui kerjasama dengan Universiti Sains Malaysia pada tahun 1994).

Data korpus ini terdiri daripada teks tulisan yang merangkumi teks Melayu lama (daripada hikayat dan kitab) dan teks moden yang diambil terutamanya daripada sumber buku, akhbar, dan majalah. Korpus lisan masih dalam perancangan kerana penandaan yang diperlukan untuk korpus lisan jauh lebih rumit daripada korpus tulisan dan tidak tertangankan buat masa itu.

Tujuan utama pembinaan pangkalan data korpus ini adalah untuk menyediakan suatu prasarana penelitian yang objektif dan autentik sifatnya kepada para penyelidik bahasa Melayu supaya dapatan yang diperoleh daripada kajian berdasarkan korpus ini dapat mencerminkan peri laku tipikal kata dan frasa bahasa Melayu dalam persekitaran penggunaannya yang sebenar dan dapat pula dijadikan asas untuk penyusunan kamus, tatabahasa dan buku-buku bahasa yang lainnya.

Pangkalan data ini juga akan disediakan dengan kemudahan capaian melalui Internet dan World Wide Web supaya lebih ramai penyelidik di dalam dan di luar negara dapat memanfaatkannya.

Kini kiraan mutakhir data teks yang terkumpul dalam pangkalan data DBP sudah pun melebihi 100 juta kata. Kata tinggal kata dan angka yang besar ini tidak memberi erti apa-apa andainya tidak diteliti dan dikaji. Langkah awal yang perlu dilakukan adalah meneliti dan menghuraikan data yang besar ini supaya apa-apa kajian yang dilakukan dan sebarang dapatan bukan sahaja sahih dalam batas cakupan data yang dikaji tetapi boleh ditentuluarkan untuk mewakili penggunaan sebenar bahasa Melayu.

## 2.2 Reka Bentuk Awal: Pangkalan Data Teks

Pangkalan data korpus DBP pada awalnya direka bentuk sebagai arkib teks (juga dikenali sebagai pangkalan data teks) dan di DBP sendiri arkib ini sering kali disalahertikan sebagai ‘korpus’ atau ‘pangkalan data korpus’.

Konsep ‘arkib’, ‘koleksi teks’ ‘korpus’, ‘korpus komputer’, ‘sub-korpus’ dan ‘kutipan’ dibezakan dalam kertas ini dan dalam kerja-kerja penelitian di Bahagian Penyelidikan Bahasa, DBP berdasarkan takrifan yang terdapat dalam “*Preliminary Recommendations on Corpus Typology*” EAG–TCWG–CTYP/P (Sinclair 1996) dan dilakarkan di bawah seperti yang berikut:

- ***Korpus*** ialah kumpulan cebisan bahasa (atau teks lengkap) yang dipilih dan disusun mengikut kriteria linguistik[2] yang eksplisit untuk digunakan sebagai sampel sesuatu bahasa;
- ***Korpus komputer*** ialah korpus yang diberi penanda, kod dan diformatkan secara piawai[3] serta dapat dicapai dan diproses dengan komputer (dalam linguistik korpus, ‘korpus komputer’ disingkatkan kepada ‘korpus’ sahaja kerana sudah tersirat dalam wacananya);
- ***Sub-korpus*** merupakan bahagian daripada korpus yang lebih besar dan mempunyai semua ciri korpus atau boleh juga merupakan “… *a dynamic selection from a corpus during on-line analysis*.” (Atkins et al. 1992);
- ***Koleksi*** dan ***arkib*** merujuk kepada set atau kumpulan teks yang tidak perlu dipilih atau disusun mengikut kriteria linguistik dan lantaran itu berbeza daripada korpus (dalam korpus linguistik ‘arkib’ merujuk kepada himpunan teks elektronik dan dikenali juga sebagai pangkalan data teks);
- ***Kutipan*** (*citation*) ialah contoh invidu sesuatu kata dalam konteks penggunaannya dan kumpulan kutipan ini *tidak boleh* dianggap sebagai korpus melainkan sekadar himpunan kutipan sahaja.

[2] Kriteria linguistik ini merangkumi aspek pelaku, waktu, persekitaran teks atau cebisan bahasa yang dihasilkan dan fungsi komunikatif masing-masing (Kučera dan Francis 1967; Sinclair 1988; Atkins et al. 1992).

[3] Dalam kes data DBP, kami menggunakan SGML untuk penandaan minimal. Untuk pangkalan data korpus mutakhir kami menerima pakai garis panduan *Text Encoding Initiative* (http://www.tei-c.org/)

Berbeza dengan arkib teks yang lain (sebagai contoh *Oxford Text Archive* [4] atau *Gutenberg Project* [5] ), pangkalan data teks DBP ini dilengkapi sistem untuk memproses teks yang dipilih. Teks boleh diproses untuk memperagakan baris konkordans dan boleh dianalisis untuk mempamerkan maklumat statistik seperti kekerapan kata dan jumlah kata.

Ada dua sebab mengapa reka bentuk ini terpilih dengan sendirinya. Pertama, atas tujuan kepraktisan. Teks digital perlu dikumpul dengan banyak dalam waktu yang sesingkat mungkin supaya himpunan teks tersebut boleh segera dimanfaatkan untuk kerja perkamusan.

Lantaran itu, pengumpulannya pada peringkat awal pembinaan adalah lebih bersifat oportunistik. Mana-mana teks terbitan DBP (buku, majalah, kertas kerja) yang sudah tersedia dalam bentuk digital akan dimasukkan dalam pangkalan data dan mana-mana teks digital yang ada pada penerbit lain dibekalkan secara gratis atau dibeli (seperti data akhbar) secara pukal. Data selebihnya ditaip semula atau diimbas dan dibaca pruf supaya keandalan teks itu melebihi 95%. Dengan demikian, semua teks digital bahasa Melayu layak diarkibkan tanpa perlu ada kriteria pemilihan khusus.

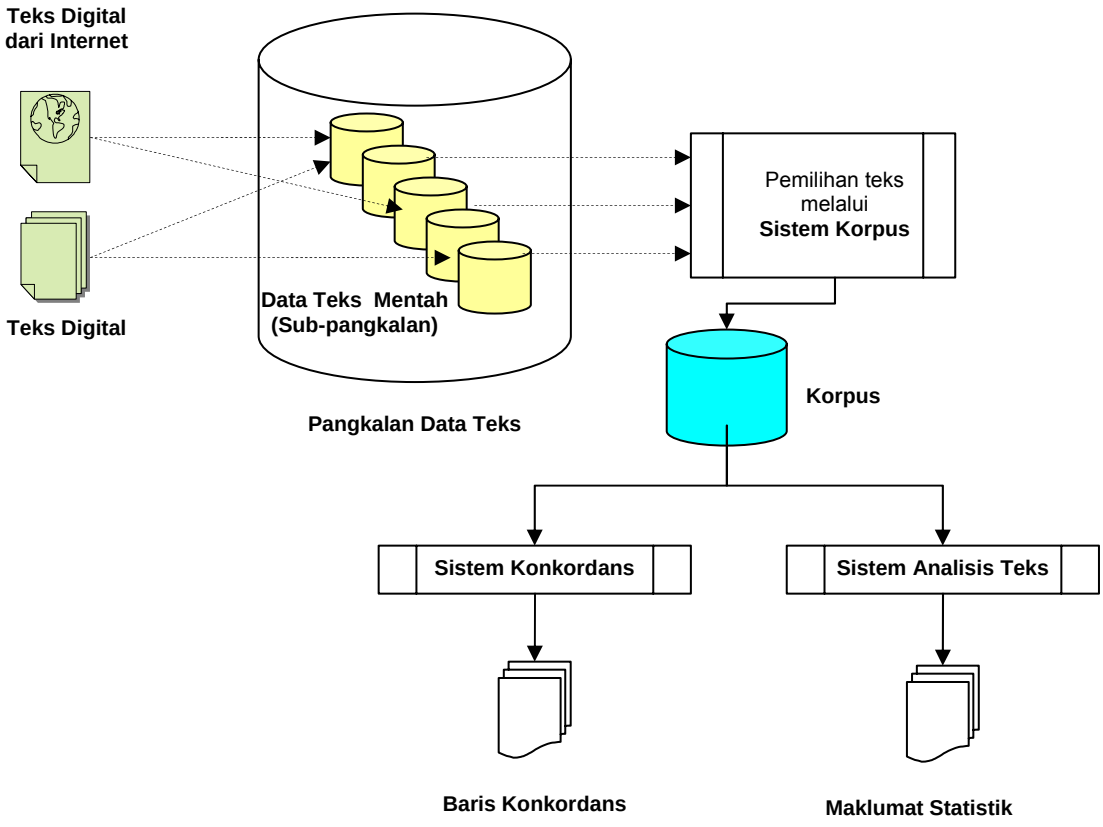
Teks ini disimpan dalam pangkalan data yang berasingan (disebut sub-korpus tetapi sebetulnya sub-arkib atau sub-pangkalan). Pangkalan kecil ini diberikan nama berdasarkan jenis terbitan (buku, majalah, akhbar, efemera), jenis teks (teks lama atau tradisional, terjemahan) atau genre (drama, puisi).

Pemecahan pangkalan ini didorong oleh batas perkakasan dan keperluan untuk mengasingkan teks berdasarkan wadah terbitan. Data yang besar dalam sesuatu pangkalan tidak mampu diproses oleh sistem dan perlu dipecahkan kepada sub-pangkalan. Sebagai contoh, data buku perlu disimpan dalam sub-pangkalan db1, db2, db3 … dan seterusnya kerana jika disatukan maka pemprosesan dan keseluruhan sistem akan ‘tergantung’. Atas sebab kekurangan ini dan kekurangan lain maka satu sistem baru sedang dibina di bawah projek pembinaan Sistem Bahasa Melayu Bersepadu.

[4] http://ota.ahds.ac.uk/
[5] http://promo.net/pg/

Sebab yang kedua ialah konsep dan kriteria korpus bahasa Melayu yang seimbang dan representatif belum dapat dijelaskan pada waktu itu: 'Seimbang' yang bagaimana dan 'representatif' bagi apa? Lantaran itu, sebagai dasar kami memberikan pengguna pangkalan data teks itu kebebasan untuk mentakrifkan sendiri kriteria berpandukan skop kajian masing-masing.

Dengan demikian apa yang dinamakan Pangkalan Data Korpus DBP [6] itu sebenarnya belum lagi sepenuhnya 'korpus' tetapi masih merupakan sebuah arkib atau pangkalan teks. Namun demikian, daripada pangkalan ini, teks-teks dapat dipilih berdasarkan kriteria linguistik tertentu untuk dijadikan korpus oleh peneliti dan diproses untuk kegunaan peneliti itu sendiri.




**RAJAH 1 PANGKALAN DATA DBP**

[6] Merupakan nama output bagi projek dalam Sasaran Kerja Utama DBP 2001-2005 yang hanya akan terealisasi dengan terbinanya Sistem Bahasa Melayu Bersepadu (2005?).

Pada hemat kami, reka bentuk pangkalan DBP yang sedia ada ini memberikan keluwesan kepada para penyelidik untuk mentakrifkan kriteria korpus penelitian masing-masing tanpa dikekang dan dipaksa menerima kriteria DBP. Carta alir pangkalan data ini boleh digambarkan seperti dalam Rajah 1.

### 2.3 Reka Bentuk Semasa: Pangkalan Data Korpus Bahasa Melayu DBP

Daripada pangkalan data yang sedia ada ini nanti teks-teks akan disarikan berdasarkan kriteria yang dikenal pasti untuk dijadikan pangkalan korpus bahasa Melayu yang 'seimbang' dan 'representatif' bagi penelitian penggunaan sebenar bahasa Melayu.

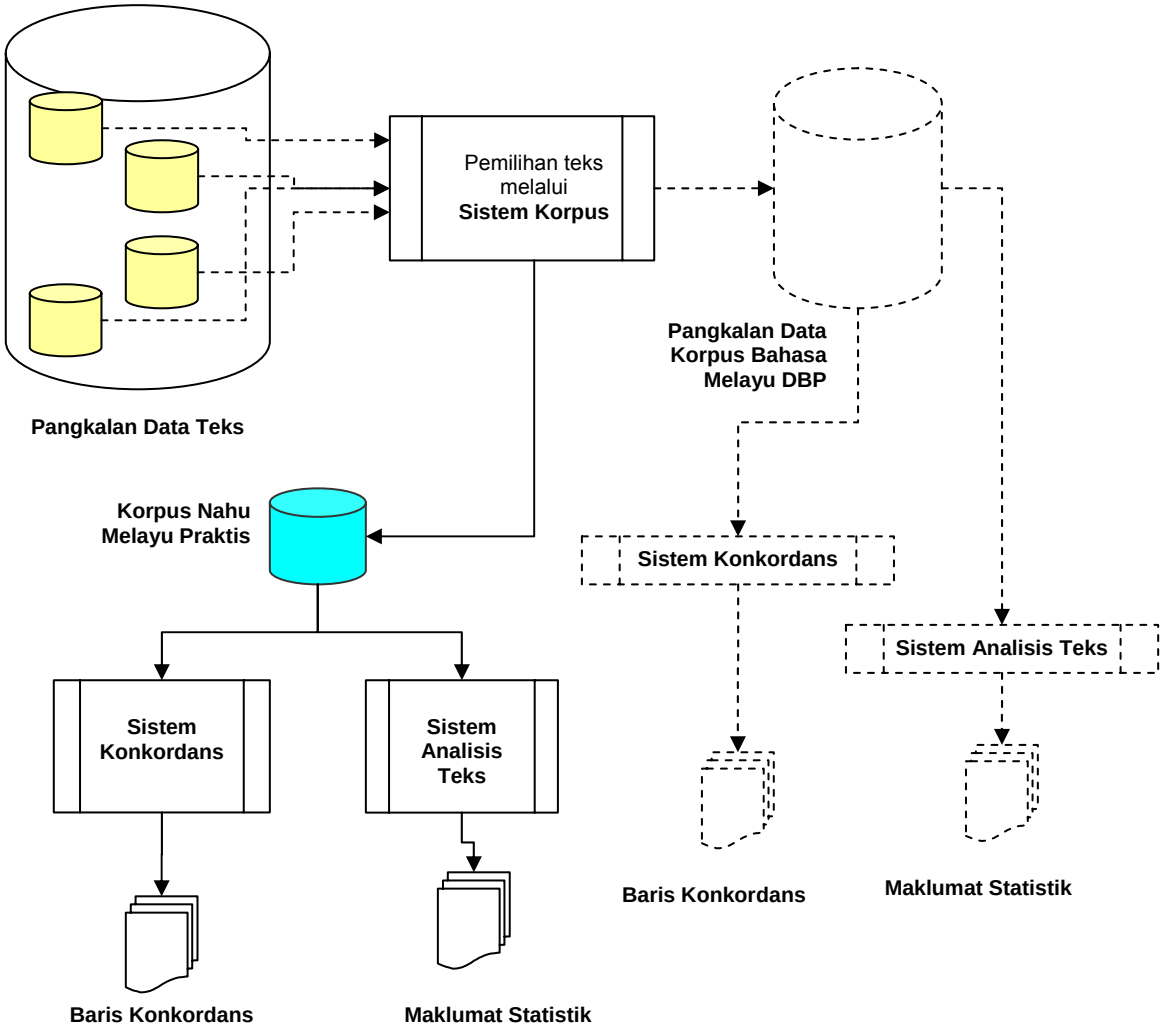
Kajian yang dijalankan di bawah projek Nahu Praktis Bahasa Melayu ini merupakan langkah awal dalam penyediaan suatu himpunan data teks yang pelbagai dan mewakili genre serta sumber utama penghasilan bahasa Melayu tulisan. Kajian ini akan membantu penyelidik untuk memahami profil dan peri laku kata dan bentuk kata dalam korpus yang sederhana besar dan ini akan membawa kepada pembinaan pangkalan data korpus bahasa Melayu DBP yang seimbang dan representatif. Carta alir binaan ini digambarkan dalam Rajah 2.

## 3.0 TIPOLOGI DATA

Dalam bahagian ini kami menghuraikan tipologi teks yang ada dalam pangkalan data teks DBP dan kemudian menghuraikan tipologi data korpus Nahu Melayu Praktis (sebesar 5 juta kata) yang dipilih dan dipetik daripada pangkalan sebesar 100 juta.

### 3.1 Tipologi Teks DBP

Data teks yang terkumpul dalam pangkalan data DBP sekarang ini (sehingga Mac 2004) berjumlah kira-kira 100 juta kata dan tersimpan dalam sub-pangkalan seperti yang dirincikan Jadual 1-11.




**RAJAH 2 PANGKALAN DATA KORPUS BAHASA MELAYU DBP**

**Jadual 1: Data Akhbar**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| AKHBAR | data 94,95,96 dan 98. | 10,111,504 |
| AKHBAR97 | data 97. | 3,443,849 |
| AKHBAR99 | data akhbar NSTP Online tahun 1999 | 6,055,096 |
| AKHBAR00 | data akhbar NSTP Online tahun 2000 | 6,800,502 |
| AKHBAR01 | data akhbar NSTP Online tahun 2001 | 4,825,314 |
| AKHBAR01-EKONOMI | data akhbar NSTP Online tahun 2001-ekonomi | 147,924 |
| AKHBAR01-HIBURAN | data akhbar NSTP Online tahun 2001-hiburan | 239,035 |
| AKHBAR01-SUKAN | data akhbar NSTP Online tahun 2001-sukan | 926,910 |
| AKHBAR02 | data akhbar NSTP Online tahun 2002 | 4,586,869 |
| AKHBAR02-EKONOMI | data akhbar NSTP Online tahun 2002-ekonomi | 227,605 |

| | | |
|---|---|---|
| AKHBAR02-HIBURAN | data akhbar NSTP Online tahun 2002-hiburan | 420,438 |
| AKHBAR02-SUKAN | data akhbar NSTP Online tahun 2002-sukan | 1,101,196 |
| AKHBAR03 | data akhbar NSTP Online tahun 2003 | 5,114,146 |
| AKHBAR03-EKONOMI | data akhbar NSTP Online tahun 2003-ekonomi | 74,690 |
| AKHBAR03-HIBURAN | data akhbar NSTP Online tahun 2003-hiburan | 676,615 |
| AKHBAR03-SUKAN | data akhbar NSTP Online tahun 2003-sukan | 1,163,734 |
| JUMLAH NSTP (Berita Harian, Berita Minggu, Harian Metro) | | 45,915,427 |
| UTUSAN | data Utusan Online. | 6,448,577 |
| | | |
| HARAKAH | data Harakah Edisi Internet | 624,699 |
| | | |
| | **JUMLAH DATA AKHBAR** | **52,988,703** |

**Jadual 2: Data Buku**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| DB3 | data buku 70-an ke atas. | 11,137,717 |
| DB2 | data buku 70-an ke atas. | 9,739,899 |
| DB1 | data buku tahun 60-an ke bawah. | 2,759,585 |
| DB4 | data buku 70-an ke atas. | 1,807,618 |
| | **JUMLAH DATA BUKU** | **25,444,819** |

**Jadual 3: Data Majalah**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| MAJALAH | data majalah. | 4,861,827 |
| MAJALAH1 | data majalah tambahan | 3,361,029 |
| MAJALAH ILMIAH | data majalah ilmiah | 1,887,516 |
| MAJALAH BUKAN ILMIAH | data majalah bukan ilmiah | 2,119,001 |
| | **JUMLAH DATA MAJALAH** | **12,229,373** |

**Jadual 4: Data Teks Melayu Lama/Tradisional**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| KLASIK | data teks Melayu lama atau teks tradisional | 2,440,258 |
| | **JUMLAH DATA TEKS MELAYU LAMA** | **2,440,258** |

**Jadual 5: Data Teks Terjemahan**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| TERJEMAH | data terjemahan ke dalam Bahasa Melayu | 1,886,106 |
| | **JUMLAH DATA TERJEMAHAN** | **1,886,106** |

**Jadual 6: Data Teks Sabah dan Sarawak**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| SUKUAN | data bahasa Melayu terbitan Sabah & Sarawak | 1,038,250 |
| | **JUMLAH DATA BAHASA MELAYU SABAH DAN SARAWAK** | **1,038,250** |

**Jadual 7: Data Buku Teks**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| BUKUTEKS | data buku teks sekolah | 1,095,726 |
| | **JUMLAH DATA BUKU TEKS** | **1,095,726** |

**Jadual 8: Data Teks Drama**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| DRAMA | data drama. | 215,867 |
| | **JUMLAH DATA DRAMA** | **215,867** |

**Jadual 9: Data Efemeral**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| EFEMERAL | data efemeral (brosur, iklan, borang, resepi dsbnya) | 173,131 |
| | **JUMLAH DATA EFEMERAL** | **173,131** |

**Jadual 10: Data Puisi**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| PUISI | data puisi (sajak, syair, pantun) | 2,348 |
| | **JUMLAH DATA PUISI** | **2,348** |

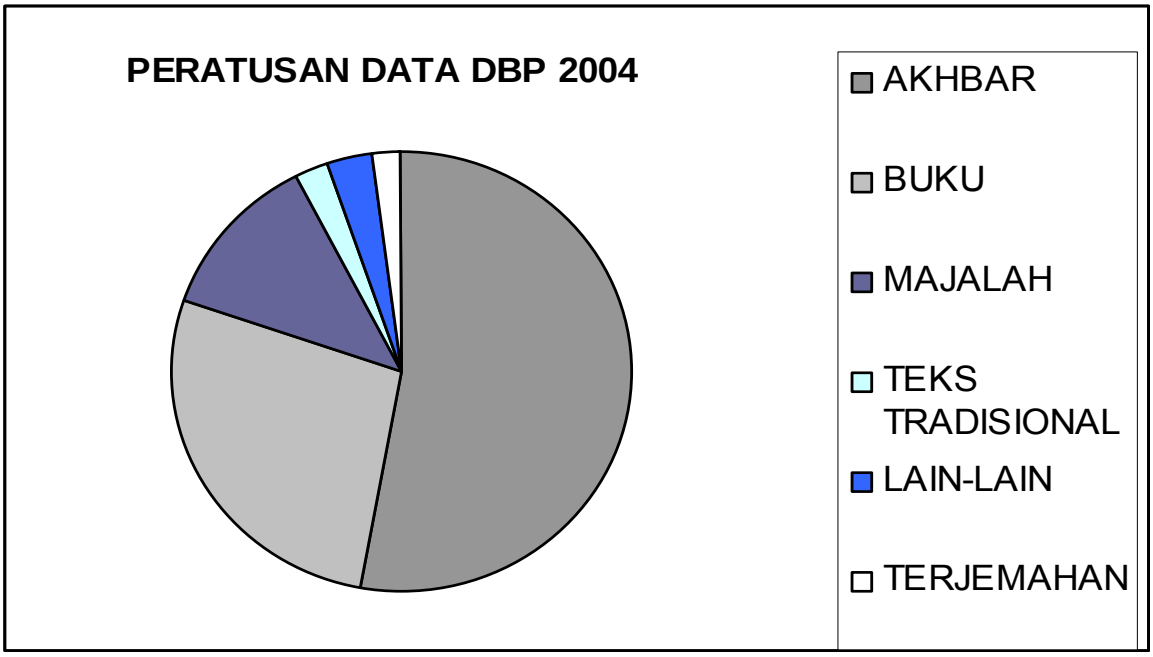
**Jadual 11: Data Kad Bahan**

| SUB-PANGKALAN | KETERANGAN | KATA (Sehingga Mac 2004) |
|---|---|---|
| KAD BAHAN | Rekod Kad Bahan (kutipan kad bahan untuk penyusunan kamus) | 3,130,641 |
| | **JUMLAH DATA KAD BAHAN** | **3,130,641** |

Data ini boleh dikelompok dan diringkaskan seperti dalam Jadual 12 dan Rajah 3.

**Jadual 12: Jumlah Kumulatif Data Teks (Sehingga Mac 2004)**

| SUB-PANGKALAN | JUMLAH KATA | PERATUS |
|---|---|---|
| AKHBAR | 52,988,703 | 52.65% |
| BUKU | 27,797,010 | 27.62% |
| MAJALAH | 12,229,373 | 12.15% |
| TEKS TRADISIONAL | 2,440,258 | 2.43% |
| LAIN-LAIN | 3,303,772 | 3.28% |
| TERJEMAHAN | 1,886,106 | 1.87% |
| JUMLAH | 100,645,222 | 100.00% |




**Rajah 3 Peratusan Data DBP 2004**

Daripada Jadual 12 dan Rajah 3, dapat dilihat bahawa data akhbar menjuzuki lebih daripada 50% saiz pangkalan data teks manakala data buku sesuku daripada jumlah sebenar. Bagi data akhbar, ini mencerminkan jumlah dan isi padu penghasilan bahasa Melayu yang wajar kerana akhbar diterbitkan harian dan mingguan. Penghasilan data akhbar dilihat

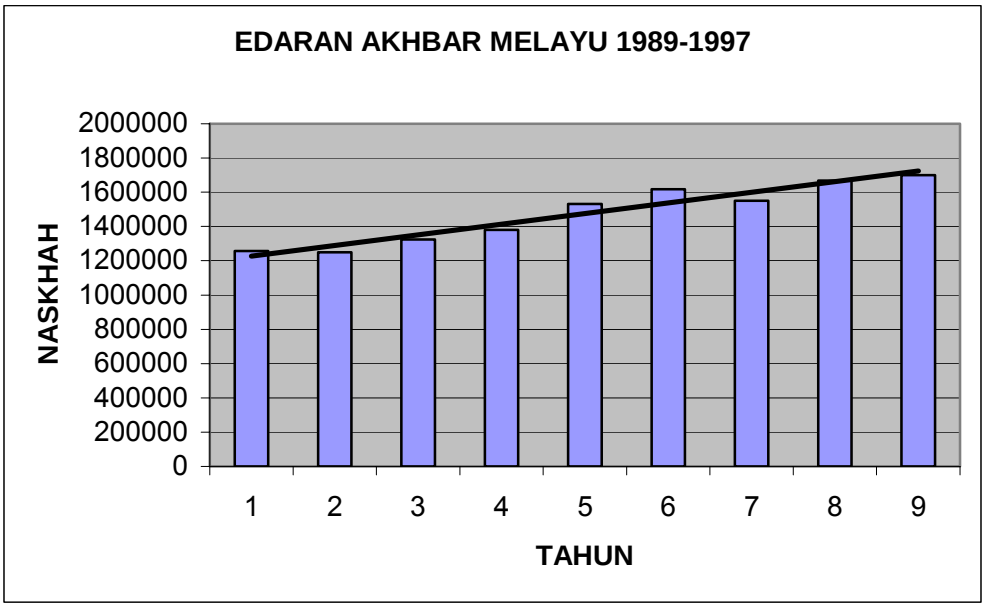
daripada angka edarannya mengikut sumber MediaGuide99 di bawah (Rajah 4) sekitar 1.5 juta naskhah setahun.

Berdasarkan sumber yang sama, data majalah seharusnya lebih banyak daripada data buku kerana majalah lazimnya diterbitkan bulanan dan bilangan majalah berbahasa Melayu banyak di pasaran seperti diperlihatkan dalam Jadual 13 dan Rajah 5.

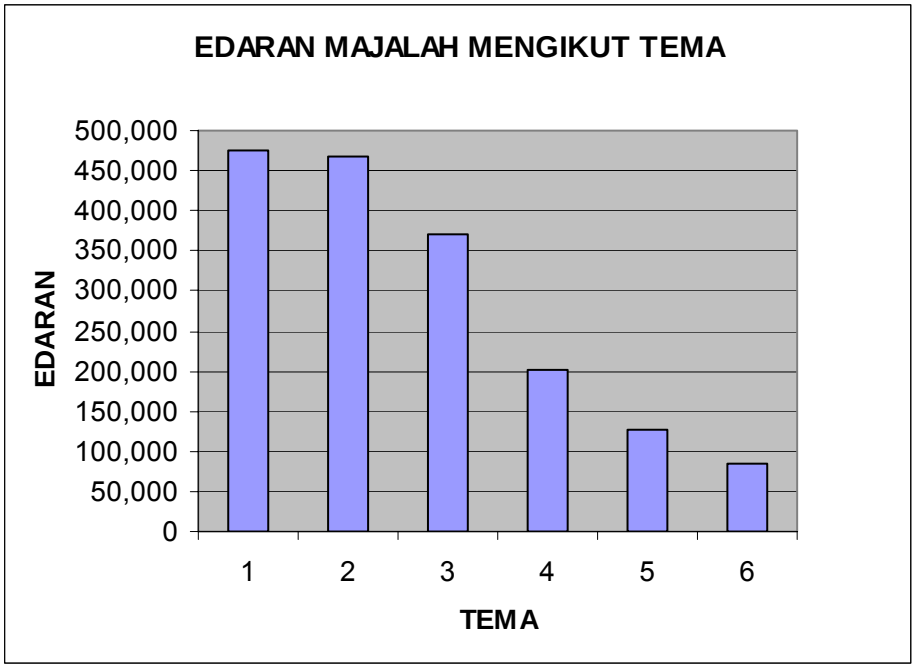
Jadual di bawah memperlihatkan senarai majalah berbahasa Melayu berserta edarannya mengikut tema utama.

**Jadual 13 Edaran Majalah mengikut Tema**

| TEMA | EDARAN |
|---|---|
| 1. Ilmiah | 476 |
| 2. Hiburan | 475 |
| 3. Wanita | 467 |
| 4. Humor | 370 |
| 5. Misteri | 201 |
| 6. Kanak-kanak | 125 |
| 7. Sukan | 85 |
| | |




**Rajah 4 Edaran Akhbar Berbahasa Melayu**




**Rajah 5 Edaran Majalah Berdasarkan Tema (**Sumber: MediaGuide99)

Dengan taburan data dan maklumat penghasilan data di atas kami telah menyediakan suatu garis panduan untuk memetik lima juta kata daripada pangkalan data teks sebesar 100 juta kata berdasarkan perkadaran tertentu yang mencerminkan saiz, liputan, dan pengaruh yang besar terhadap pengguna bahasa Melayu.

### 3.2 Tipologi Korpus Nahu Melayu Praktis

Korpus ialah himpunan teks yang dikumpulkan berdasarkan kriteria reka bentuk tertentu, untuk tujuan dan objektif yang spesifik. Dalam kes Nahu Melayu Praktis (sesudah ini diringkaskan sebagai NMP) ini objektif utama adalah untuk meneliti pelbagai fenomena bahasa (morfologi dan sintaksis) berdasarkan korpus yang representatif dan memadai besarnya.

Untuk dikatakan 'representatif' sesebuah korpus itu perlu mengambil kira tiga aspek utama iaitu, saiz, keberkadaran (*proportionality*) dan keautentikan (Kučera 2002).

Dalam kes korpus NMP saiz permulaan yang dipilih ialah 5 juta kata, yakni 1/5 daripada jumlah keseluruhan yang ada dalam pangkalan data. Pemilihan saiz ini merupakan

keputusan arbitrari dengan andaian bahawa buat permulaan lebih elok dikaji data yang tidak terlalu besar (mengingat bahawa Korpus Brown memberikan dapatan linguistik yang sahih dengan satu juta kata). Saiz ini boleh digandakan dengan mudah untuk kajian susulan.

Data teks hanya dipilih daripada tiga kelompok utama data, iaitu Akhbar, Majalah dan Buku mengikut perkadaran teks yang ada dalam keseluruhan pangkalan data dan berasaskan perkiraan kadar pendedahan (*exposure*) bahasa[7], khususnya dari segi saiz dan kekerapan pendedahan penutur bahasa Melayu kepada pelbagai topik dan jenis tulisan dalam tiga wadah utama penyebaran bahasa Melayu ini.

Perkadaran untuk setiap kelompok juga adalah berdasarkan angka pengedaran kerana kesan dan pengaruh kelompok itu berkadaran langsung dengan jumlah terbitan dan edaran bahan.

Dari segi autentiknya tidaknya data, kami menganggap semua data yang diterbitkan dan kemudian dipilih dan diinputkan ke dalam pangkalan data tanpa sebarang pindaan (tidak termasuk pindaan jenis dan saiz fon serta penandaan TEI yang berasaskan SGML) sebagai data yang autentik kerana mewakili data sebenar.

Komposisi data NMP diringkaskan dan dipaparkan dalam Jadual 14 .

**Jadual 14 Komposisi Data Korpus NMP**

| KELOMPOK DATA | % DATA | KATA (Juta) |
|---|---|---|
| **Akhbar** | **50** | **2.50** |
| Utusan Malaysia | (25) | 1.25 |
| Berita Harian NSTP | (20) | 1.00 |
| Harakah | (5) | 0.25 |
| **Majalah** | **30** | **1.50** |
| Terbitan DBP | (10) | 0.50 |
| Terbitan Luar DBP | (20) | 1.00 |
| **Buku (Terbitan DBP dan Luar DBP)** | **20** | **1.00** |
| Fiksyen | (10) | 0.50 |
| Bukan Fiksyen | (10) | 0.50 |
| **JUMLAH** | **100** | **5.00** |

Dengan perkadaran data yang sedemikian, maka kajian tatabahasa yang dilakukan terhadap korpus NMP ini bolehlah dianggap sebagai mencerminkan fenomena bahasa Melayu secara keseluruhan. Bagaimanapun, kajian susulan perlu dilakukan dengan data yang lebih besar (dengan perkadaran yang serupa) untuk mengesahkan dapatan tersebut. Perlu diingat bahawa pendedahan terhadap sesuatu teks itu hanyalah anggaran berdasarkan jumlah terbitan atau edaran semata-mata kerana kita tahu bahawa sesuatu terbitan itu boleh dibaca oleh ramai orang, lebih-lebih lagi dengan adanya akhbar, majalah dan bahan-bahan lain dalam laman Web. Dengan demikian, jumlah pembaca tidak semestinya sama dengan jumlah pembeli.

## 4.0 KESIMPULAN

Kertas ini menghuraikan secara ringkas perancangan, pembinaan, dan pemanfaatan teks yang sedia terkumpul dan tersimpan dalam pangkalan data Dewan Bahasa dan Pustaka. Usaha kumpulan penyelidik UKM dan DBP untuk memanfaatkan sebahagian daripada data ini dalam kajian tatabahasa merupakan langkah yang wajar dilaksanakan dan diharap akan menjadi perintis kepada kajian-kajian lain. Sebarang kajian yang dilakukan akan memberikan perspektif dan wawasan (*insight*) yang berguna dalam pembinaan dan pengembangan bahasa Melayu.

## Bibliografi


Aarts, J. 1991. 'Intuition-based and observation-based grammars' dalam Aijmer dan Altenburg 1991, hlm 44-62.

Aarts, J. dan Meijs, W. (ed.) 1986. *Corpus Linguistics* II, Amsterdam: Rodopi.

Aijmer, K. dan Altenberg, B. (ed.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.


[7] Berdasarkan jumlah terbitan dan edaran akhbar, buku dan majalah.

Atkins, B. T. S. dan Levin, B. 1995. 'Building on a corpus: a linguistic and lexicographical look at some near-synonyms' dalam *International Journal of Lexicography* 8:2, 85-114.

Atkins, S., Clear, J. dan Ostler, N. 1992. 'Corpus Design Criteria' dalam *Literary and Linguistic Computing* 7(1): 1-16.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh University Press, Edinburgh.

Biber, D., Conrad, S. dan Reppen. R. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, Cambridge, UK.

Francis, N. dan Kučera, H. 1964. *Manual of Information to accompany the a standard corpus of present-day edited American English, for use with digital computers.* Department of Linguistics, Brown University, Providence, Rhode Island.

Garside, R., Leech, G. dan McEnery, A. (ed.). 1997. *Corpus Annotation*. London: Longman.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Longman, London.

Kučera, K 2002. 'the Czech National Corpus: Principles, Design, and Results' dalam *Literary and Linguistic Computing* 17(2): 245-257.

McEnery A., dan Wilson, A. 2001. *Corpus Linguistics* (Edisi Ke-2). Edinburgh University Press, Edinburgh.

McEnery, A. dan Wilson, A. 1993. 'The role of corpora in computer-assisted language learning' dalam *Computer Assisted Language Learning* 6(3): 233-48.

Sinclair, J. (ed.). 1987. *Looking Up.* HarperCollins, London.

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

Sinclair, J. 1996. "*Preliminary Recommendations on Corpus Typology*" EAG–TCWG–CTYP/P di laman Web <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>

Stubbs, M. 1996. *Text and Corpus Analysis*. Blackwell, Oxford.

Summers, D. 1993. 'Longman/Lancaster English Language Corpus – Criteria and Design dalam *International Journal of Lexicography* 6:3, 181-208.

Zaiton Ab. Rahman 1987. Kertas Rancangan Projek Analisis Teks Secara Komputer. Cawangan Penyelidikan, DBP (Tidak diterbitkan).

Zampolli, A. dan Ostler, N. (ed.). 1993. *'Special Section on Corpora', Literary and Linguistic Computing* 8(4).