Suara Sasterawan, Suara Awam

Rusli Abdul Ghani (rusli@dbp.gov.my)

Bahagian Penyelidikan Bahasa
Dewan Bahasa dan Pustaka Malaysia (http://www.dbp.gov.my)

ABSTRAK

Kajian korpus dilaksanakan berdasarkan kumpulan teks yang terhimpun secara digital dalam pangkalan data untuk tujuan pengolahan dan pencapaian yang mudah dan pantas. Dari segi sistem, pangkalan data korpus tidak dilihat sebagai bermasalah sangat kerana isunya sering kali berlegar di sekitar soal megabait RAM dan gigabait storan sahaja. Namun dari segi data, ada isu yang kerap kali dibangkitkan sehubungan dengan dapatan kajian korpus yang menggunakan data DBP. Ada pihak yang berpendapat bahawa data korpus DBP banyak mengandungi teks yang 'tidak baik' yang ditulis oleh penulis yang kurang tulen dan asli kemelayuannya kerana terpengaruh dengan bahasa lain (terutamanya Inggeris) atau kerana tidak pandai menulis dan sebagainya. Dengan demikian, hasil kajian yang disari dan disaring daripada korpus DBP tidak dapat dijadikan pegangan dan panduan kerana tidak bersandarkan penulis dan penulisan yang 'baik'. Makalah ini akan meneliti persoalan ini dari sudut falsafah dan metodologi kajian korpus dan meleraikan isu tentang golongan yang manakah yang lebih berhak dianggap sebagai mewakili suara umat berbahasa Melayu: bahasawan, sasterawan atau orang awam?

> *Language is not an abstract construction of the learned, or of dictionary makers, but is something arising out of the work, needs, ties, joys, affections, tastes, of long generations of humanity, and has its bases broad and low, close to the ground.*
>
> **Noah Webster**

1.0 PENDAHULUAN

Kajian korpus, atau linguistik korpus, sebagai suatu pendekatan penelitian bahasa dapat diperbahas dari pelbagai sudut. Leech (1992:105) melihat kajian ini sebagai suatu asas metodologi untuk meneliti bahasa, bukan suatu ranah penyelidikan bahasa yang tersendiri. Namun demikian, sebahagian besar ahli bahasa yang memanfaatkan korpus dalam kajian masing-masing berpendapat bahawa Linguistik Korpus bukan perkaedahan semata-mata. Sebagai contoh, Halliday (1993:24)

berpandangan bahawa kajian kuantitatif bidang ini telah membuka ruang untuk perubahan kualitatif dalam pemahaman kita tentang bahasa. Bagi Sinclair (1970 dan ff.) pula bukti korpusan telah memberikan wawasan dan intuisi baru tentang bahasa. Biar apa pun sudut pandang mereka, yang penting dihargai adalah bahawa kajian korpus ini memungkinkan "*a new way of thinking about language*" (Leech 1992:106).

Namun begitu, cara baru untuk meneliti bahasa ini juga memerlukan penghuraian baru falsafah dan metodologinya serta pemerian konsep, istilah dan takrifan baru yang berkaitan dengan bidang ini. Jika tidak, sebarang itlakan atau rumusan daripada kajian ini akan sentiasa dipandang serong.

Salah satu aspek yang kerap dipertikaikan dalam kajian korpus bahasa Melayu ialah datanya. Ada pihak yang mengatakan bahawa data korpus Dewan Bahasa dan Pustaka 'tidak baik' kerana memasukkan 'apa sahaja' karya ke dalam pangkalannya. Ini tentunya soal dasar kerana jatuh bangunnya sesuatu kajian itu berpangsi pada kebenaran data dan keutuhan metodologinya.

Dalam kertas ini, sebagai penjelasan kepada praandaian dan wasangka terhadap data DBP, kami akan menghuraikan rasional dan kriteria yang digunakan dalam mereka bentuk pangkalan data korpus DBP. Dengan demikian suara sasterawan, suara wartawan, suara bangsawan, suara awam atau suara siapa-siapa sahaja yang terkandung dalam pangkalan data korpus DBP dapat dihargai dalam konteks kajian yang ilmiah.

2.0 KORPUS 101

Istilah seperti 'korpus', 'arkib', 'kutipan' serta takrifan masing-masing yang dipakai dalam konteks Linguistik Korpus dan Linguistik Komputeran telah banyak diperkatakan dan dihuraikan (sebagai contoh Atkins, Clear dan Ostler 1992; McEnery

dan Wilson 1996). Sinclair (1996) menggariskan takrifan konsep tertentu Tipologi Korpus dalam "*Preliminary Recommendations on Corpus Typology*" EAG–TCWG–CTYP/P seperti yang berikut:

- ***Korpus*** ialah kumpulan cebisan bahasa (atau teks lengkap) yang dipilih dan disusun mengikut kriteria linguistik[1] yang eksplisit untuk digunakan sebagai sampel sesuatu bahasa;
- ***Korpus komputer*** ialah korpus yang diberi penanda, kod dan diformatkan secara piawai 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.

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

Dengan demikian, dalam linguistik mutakhir, korpus harus dilihat bukan sekadar kumpulan teks semata-mata tetapi sebagai "…*a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration.*" (McEnery and Wilson ibid: 21-22). Biber et al. (1988: 246) pula menganggap 'korpus' harus dapat "… *represent a language or some part of language.*" Dalam kedua-dua pendapat ini kata kuncinya ialah "*representativeness*" .

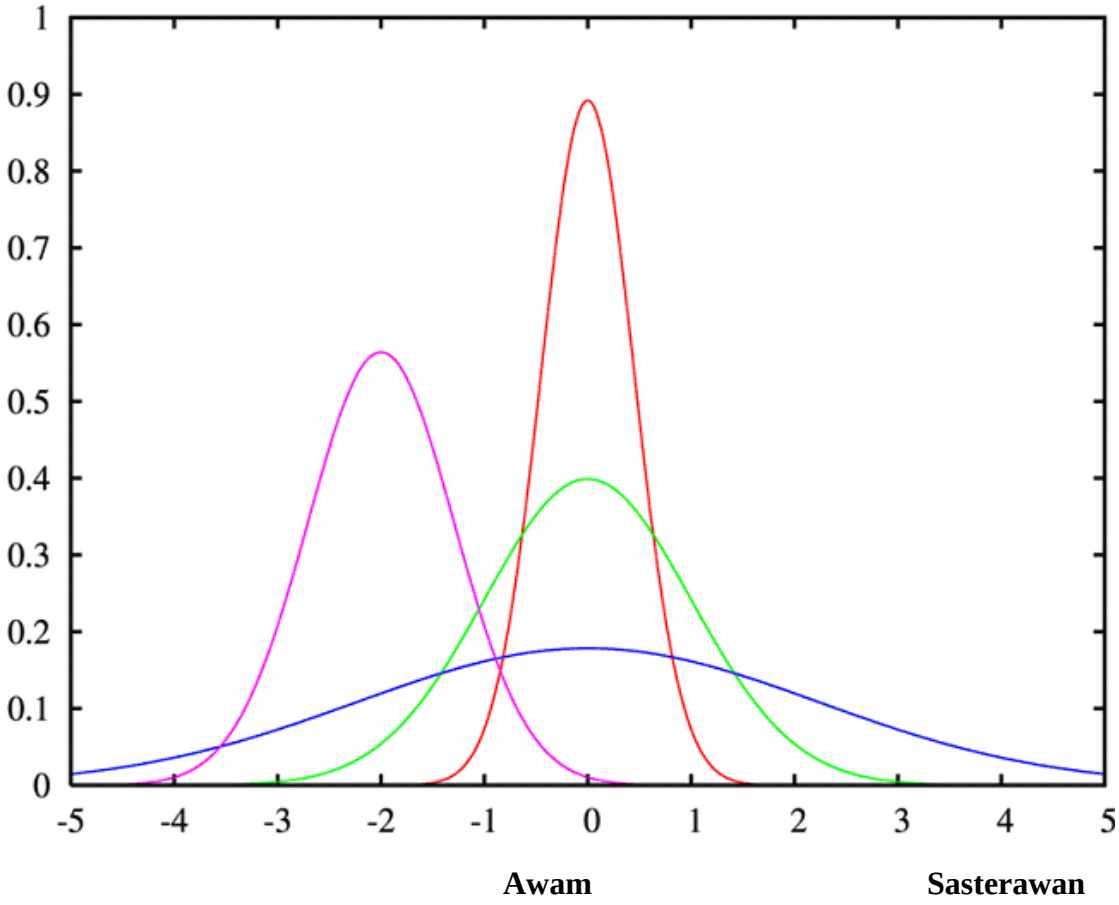
Siapakah yang wajar dianggap 'wakil' suara Melayu? — Tun Bambang, Tun Seri Lanang dan perawi-perawi terdahulu? Raja Ali Haji? Ahmad Rashid Talu dan rakan-rakan yang seangkatannya dengan beliau? Wartawan? Penulis makalah dan buku? Editor DBP? Sasterawan, baik Negara mahupun Negeri? Atau kita-kita yang dianggap penutur asli bahasa Melayu?

Sebetulnya, ujaran dan tulisan mereka ini semuanya wajar dijadikan data, namun penentuan bahan apa dan sebanyak mana yang perlu masuk merupakan isu yang sangat ruwet, yang hanya dapat dirungkai dengan teori statistik, kaedah persampelan dan kajian tipologi.

2.1 Sudut Statistik

Dalam konteks pemprosesan bahasa kami menerima pakai takrif 'representatif' Manning dan Schütze (1999) yang berkata bahawa "… *a sample is representative if what we find for the sample also holds for the general population*."

Oleh sebab kita meneliti fenomena bahasa yang berlaku bagi populasi umum, maka sampel teks haruslah daripada setiap kelompok populasi, dalam spektrum pengguna bahasa yang luas – daripada orang awam kepada sasterawan. Apa yang perlu diingat, sebahagian besar fenomena yang melibatkan populasi yang besar lazimnya mengikut taburan normal seperti rajah di bawah:



Jumlah orang awam tentunya jauh lebih besar daripada sasterawan. Bagi menjamin data kita tidak terpencong (*skew*) ke arah yang gemilang dan cemerlang sahaja (biarpun jauh di sudut hati kita mahu semua penutur asli bahasa Melayu berketerampilan seperti Usman dan Shahnon) maka lebih banyak teks yang dihasilkan oleh orang kebanyakan terkumpul dalam pangkalan data DBP berbanding dengan sasterawan terbilang.

Bagaimanapun, seyogia diingat, kami berusaha untuk mengumpulkan semua karya sasterawan negara dan karya agung lain kerana bilangannya tidak banyak. Dengan demikian, peneliti bahasa yang berhasrat untuk meneliti bahasa Shahnon atau bahasa Tun Seri Lanang masih boleh berbuat demikian tanpa diganggu oleh kepincangan data lain. Kajian terhadap karya sasterawan negara tertentu, dalam

peristilahan de Saussurean, hanya akan memperlihatkan '*parole'* sedangkan kajian korpus sebenarnya berhasrat untuk menyelami '*langue'* orang Melayu.

Untuk mengkaji 'langue' sesuatu populasi, data sampel perlu dikutip kerana tidak mungkin keseluruhan 'bahasa' yang terhasil oleh populasi tersebut dapat dirakamkan. Kaedah persampelan yang boleh digunakan termasuklah persampelan berkadaran, persampelan berstrata atau persampelan rawak berstrata dan beberapa teknik lain (sila rujuk Biber 1988, 1993; Biber et al. 1998; Manning dan Schütze 1999 untuk perbincangan yang lebih terperinci tentang teknik persampelan).

3.0 PANGKALAN DATA KORPUS DBP

Dalam penelitian bahasa kita lazimnya berminat untuk meneliti pelbagai varieti bahasa, bukannya teks invidu. Dengan demikian kita mempunyai dua pilihan untuk membina pangkalan data teks yang autentik:

- Menganalisis setiap ujaran dalam varieti itu; atau
- Membina sampel yang lebih kecil dan munasabah untuk varieti tersebut.

Pilihan pertama tentunya tidak praktis, malah tak tercapaikan kalau pun ada orang yang mencubanya, melainkan dalam kes bahasa yang sudah 'mati' dengan sebahagian teksnya masih wujud. Bagi bahasa yang masih mekar segar bugar, seperti bahasa Melayu, ujaran yang terhasil saban detik bertambah (sekurang-kurangnya output lisannya, biarpun tulisannya agak lamban) dan dari segi teorinya, tak terhinggakan selagi ada penutur asli.

Dengan demikian, kita mengambil laluan kedua dalam menggagaskan reka bentuk pangkalan data korpus DBP. Untuk itu, kita perlu memastikan bahawa sampel kita itu semaksimum mungkin representatif bagi jenis bahasa yang diteliti supaya data

yang terhimpun nanti dapat memberikan kita gambaran yang sebenar mungkin tentang fenomena dan kecenderungan bahasa dalam varieti tersebut.

Apa yang kita tidak mahu ialah korpus kita itu hanya terdiri daripada karya penerima Anugerah Sastera Negara. Kita juga tidak mahu teks kita terdiri daripada genre novel semata-mata kerana tindakan demikian nescaya akan memberikan gambaran yang terpencong dan terpesong tentang bahasa Melayu yang umum.

Secara hipotetisnya, apa yang perlu dilakukan ialah mengumpulkan sampel teks daripada julat pengguna bahasa Melayu yang seluas mungkin dan daripada pelbagai genre, yang apabila dilihat secara menyeluruh dan dalam kesatuan yang utuh dapat memberikan gambaran yang relatif benar bagi keseluruhan populasi pengguna bahasa tersebut. Bagaimanapun, pangkalan data korpus yang sebegini masih merupakan suatu utopia.

Oleh sebab pangkalan data korpus DBP ini juga akan dimanfaatkan untuk pelbagai kajian, seperti kajian leksikal dan leksikografi, analisis wacana, analisis kesalahan bahasa, kajian stilistika dan sebagainya maka pangkalan data korpus DBP perlu direka bentuk seluwes mungkin supaya keperluan pelbagai penyelidik bahasa dan sastera dapat diladeni.

Amalan kita buat masa ini masih lebih berteraskan konsep ‘pilih dan pakai’. Penyelidik perlu memilih teks yang ingin diteliti dan hanya memproses teks yang relevan dengan kriteria yang ditentukannya sendiri.

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 penelitian.

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 percuma atau dibeli (seperti data akhbar) secara pukal. Data selebihnya ditaip semula atau diimbas dan disemak untuk menjamin keandalan teks. 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 mini ini diberikan nama berdasarkan jenis terbitan (buku, majalah, akhbar, efemera), jenis teks (teks lama atau tradisional, terjemahan) atau genre (drama, puisi).

Maklumat mutakhir tentang data yang ada dalam pangkalan data DBP adalah seperti dalam jadual yang berikut:

BIL.	SUB-KORPUS	21-Feb-05
1	AKHBAR	10,111,504
2	AKHBAR97	3,443,849
3	AKHBAR99	6,055,096
4	AKHBAR00	6,800,502
5	AKHBAR01	4,825,314
6	AKHBAR01-EKONOMI	147,924
7	AKHBAR01-HIBURAN	239,035
8	AKHBAR01-SUKAN	926,910
9	AKHBAR02	4,586,869
10	AKHBAR02-EKONOMI	227,605
11	AKHBAR02-HIBURAN	420,438
12	AKHBAR02-SUKAN	1,101,196
13	AKHBAR03	5,114,146
14	AKHBAR03-EKONOMI	74,690
15	AKHBAR03-HIBURAN	676,615
16	AKHBAR03-SUKAN	1,163,734
17	AKHBAR04	6,120,096
18	AKHBAR04-EKONOMI	212,854
19	AKHBAR04-HIBURAN	1,088,132
20	AKHBAR04-SUKAN	986,866
21	DB3	11,137,717
22	DB2	9,739,899
23	DB1	2,759,585
24	DB4	2,316,239

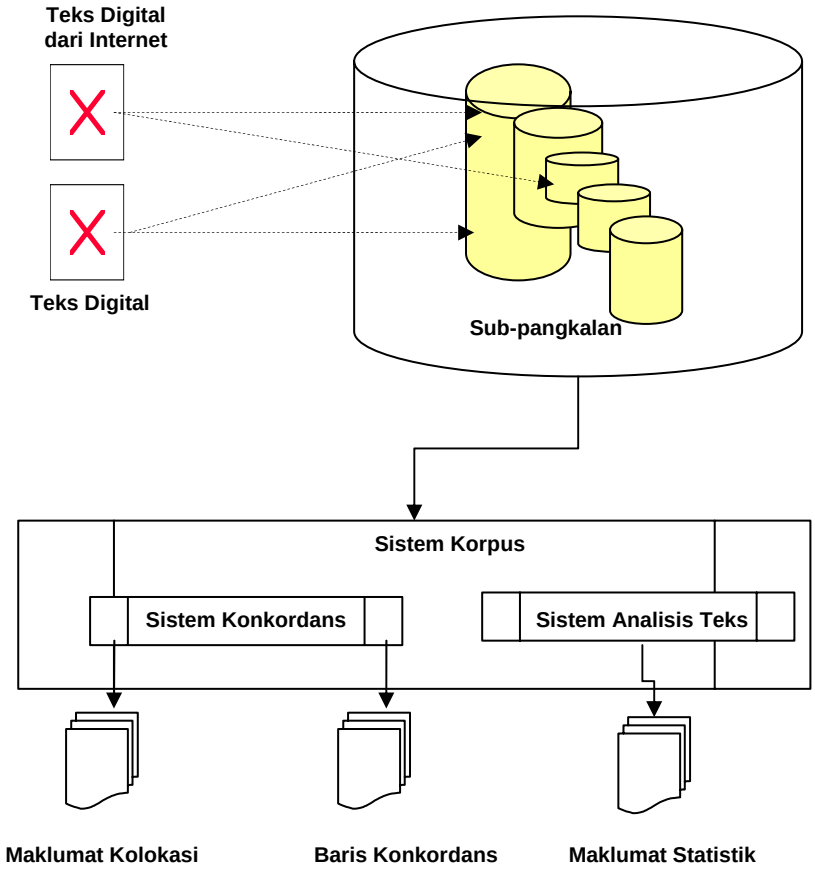
25	DB5	53,078
26	DRAMA	330,712
27	EFEMERAL	179,881
28	KLASIK	2,651,222
29	MAJALAH	4,861,827
30	MAJALAH1	3,361,029
31	MAJALAH ILMIAH	1,887,516
32	MAJALAH BUKAN ILMIAH	2,209,554
33	SUKUAN	3,407,692
34	PUISI	2,348
35	TERJEMAHAN	2,021,191
36	UTUSAN	6,448,577
37	HARAKAH	624,699
38	BUKU TEKS	1,095,726
39	NUSANTARA	944,353
40	KAD BAHAN	3,130,641
	JUMLAH KATA	113,486,861

Pemecahan pangkalan ini sebenarnya didorong oleh batas perkakasan dan keperluan untuk mengasingkan teks berdasarkan wadah terbitan, bukannya disebabkan pertimbangan linguistik.

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 'terkedu'. Atas sebab kelemahan ini dan kekurangan lain maka DBP mengusahakan pembinaan Sistem Bahasa Melayu Bersepadu dengan 'Sistem Korpus' sebagai salah satu komponen yang tersepadu di dalamnya. Sistem Korpus ini nanti akan mampu menyaring teks daripada sub-pangkalan yang ada bagi membentuk pangkalan data korpus mengikut ketentuan ahli bahasa.

Sebab yang kedua ialah kekusutan konsep dan kriteria korpus bahasa Melayu yang seimbang dan representatif belum dapat diuraikan sepenuhnya: 'Seimbang' yang bagaimana dan 'representatif' bagi apa? Lantaran itu, sebagai dasar kami memberikan

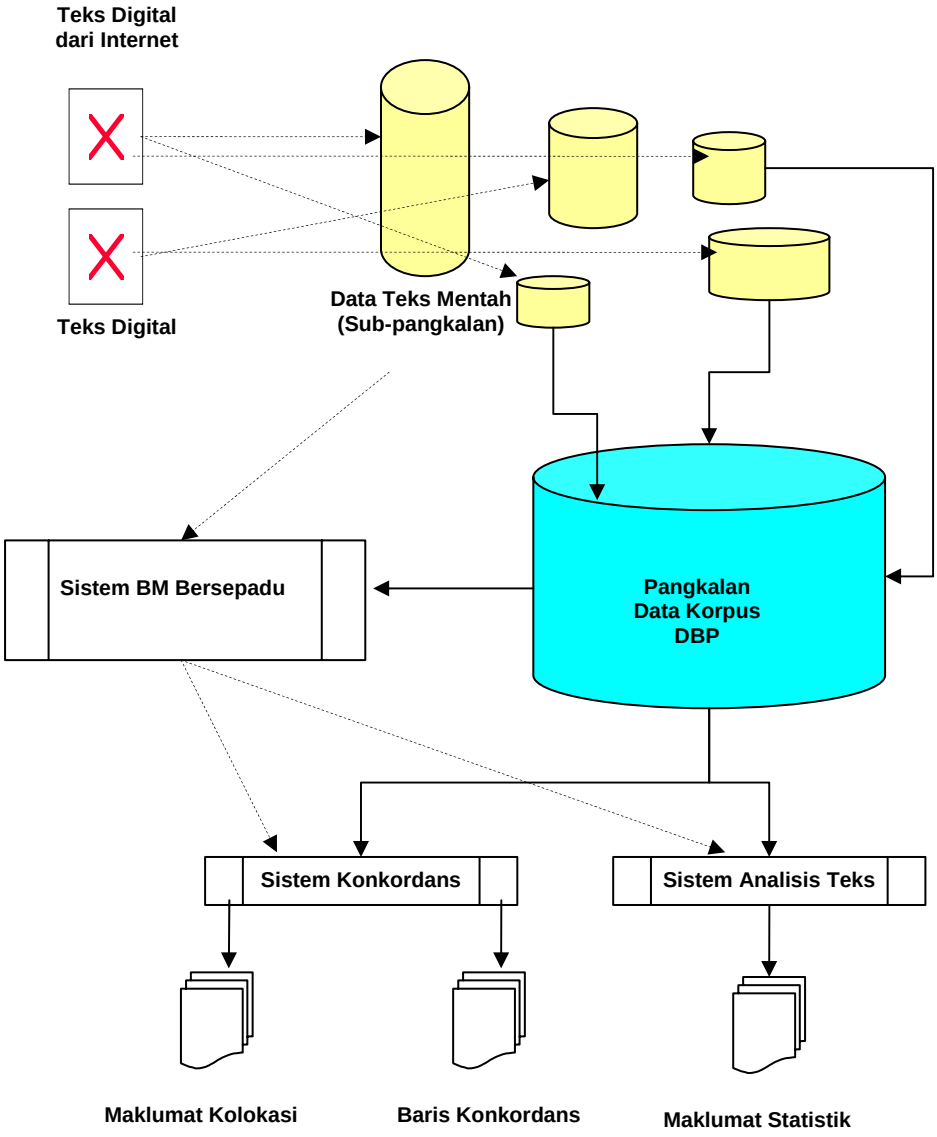
pengguna pangkalan data teks itu kebebasan untuk mentakrifkan sendiri kriteria berpandukan skop kajian masing-masing.



Rajah 1: Pangkalan Data Teks DBP 2005

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

Gambaran menyeluruh reka bentuk pembinaan pangkalan data korpus DBP sedia ada dapat dilihat pada Rajah 1 manakala reka bentuk pangkalan data korpus DBP dalam Sistem Bahasa Melayu Bersepadu dapat dilihat dalam Rajah 2.



Rajah 2: Pangkalan Data Korpus DBP 2005/2006?

4.0 KESIMPULAN

Untuk mengetahui '*la langue*' bahasa Melayu kita haruslah meneliti penggunaan bahasa dalam kalangan suara banyak, suara rakyat. Suara sasterawan itu hanyalah titisan kecil daripada suara awam yang melaut itu. Daripada jumlah yang sekian banyak tentu sahaja ada teks yang kurang baik, sama ada dari segi ejaan, gaya bahasa,

kesesuaian kata dan sebagainya. Namun kelemahan dan kekurangan yang ada pada mereka ini haruslah dilihat sebagai isu '*parole*' seperti juga halnya dengan kekuatan bahasa sasterawan.

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.

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: Edinburgh University Press.

Biber, D. 1988. *Variation Across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1993. Representativeness in Corpus Design. *Literary & Linguistic Computing 8*. hlm 243 - 257.

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

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.

Granger, S. 1993. International Corpus of Learner English dalam Aarts J., de Haan P. dan Oostdijk, N. (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi. 57.72.

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

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

Leech, G. 1992. Corpora and Theories of Linguistic Performance dalam *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82. Stockholm, 4.8 August 1991 edited by Jan Svartvik. Mouton de Gruyter 1992 (105.122).

Manning, C. D., dan Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge Massachusetts: The MIT Press.

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

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

Pearson J. 1998. *Terms in context*. SCL n. 1. Amsterdam and Philadelphia: Benjamins.

Rusli Abdul Ghani, Norhafizah Mohamed Husin, Chin Lee Yim 2004. *Pangkalan Data Korpus DBP: Perancangan, Pembinaan dan Pemanfaatan.* Seminar Sehari UKM, Bangi, 2004.

Sinclair J. dan Renouf A. 1987. A Lexical Syllabus for Language Learning dalam R.A. Carter dan M. McCarthy (eds.) *Vocabulary in Language Teaching*. London: Longman.

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

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

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*. Oxford. Blackwell.

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

Tognini Bonelli, E. 1996. *Corpus Theory and Practice*. Birmingham: TWC.

Tognini Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: Benjamins

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