

Pengkomputeran Bahasa Melayu: Kegiatan, Kerjasama dan Kemajuan

Rusli Abdul Ghani (rusli@dbp.gov.my)
& Chin Lee Yim (chin@dbp.gov.my)

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

Abstrak

Kertas ini meninjau upaya Dewan Bahasa dan Pustaka Malaysia selama kira-kira suku abad melaksanakan kegiatan pembinaan dan pengembangan bahasa dengan berlandaskan teori dan amalan perencanaan bahasa dan berpangsiakan teknologi maklumat. Usaha ini ditinjau dari sudut kegiatan pengkomputeran bahasa yang dilaksanakan di DBP, bentuk-bentuk kerjasama dengan pihak lain, termasuk badan swasta dan institusi pengajian tinggi di dalam dan di luar negeri serta menilai kemajuan yang telah dicapai setakat ini.

1.0 PENDAHULUAN

Sekitar dua dekad yang lalu, tatkala demam komputer baru mulai mewabak, ketika komputer peribadi seperti *Spectrum ZX*, *Commodore 64*, *Atari* dan *Amstrad* baru menjadi buah mulut *technogeek* era itu, Dewan Bahasa dan Pustaka Malaysia (DBP) sudah pun menghargai peri pentingnya komputer sebagai alat bantu dalam merealisasikan matlamat perencanaan bahasanya.

Pada mulanya, penggunaan komputer di DBP terhad kepada pemprosesan kata (*word processing*), khususnya dalam penyediaan manuskrip untuk penerbitan. Biarpun ini bukanlah pengkomputeran bahasa dalam erti kata yang sebenar namun permulaan yang sederhana ini menyediakan tapak yang sesuai untuk usaha pengkomputeran bahasa selanjutnya. Makalah ini melakarkan usaha-usaha tersebut

dari mula hingga ke masa ini dan mencadangkan langkah yang perlu diambil untuk mendorong dan menyemarakkan lagi pengkomputeran bahasa Melayu.

2.0 KEGIATAN DAN KERJASAMA AWAL

Penerapan teknologi komputer dalam kerja kebahasaan bermula secara berperingkat-peringkat di DBP. Yang terawal dilaksanakan ialah penggunaan komputer dalam peristilahan.

2.1 Kerjasama Peristilahan

Usaha sulung Dewan Bahasa dan Pustaka dalam pengkomputeran bahasa ialah pengkomputeran kerja peristilahan. Pada 1982, DBP telah menerima bantuan daripada IBM Corporation dan Universiti Sains Malaysia (USM) untuk menyediakan sebuah pangkalan data yang memuatkan istilah yang dibentuk oleh pelbagai jawatankuasa istilah. Pembinaan pangkalan data ini memudahkan pegawai istilah untuk menyediakan bahan mesyuarat, membandingkan istilah yang sudah dibentuk dalam pelbagai bidang, mengemas kini (*updating*) dan mencetak senarai mengikut format yang diperlukan.

Sistem ini kemudian dikembangkan pada 1989 dengan adanya perisian FAIRS (*FACOM Advance Information Retrieval System*). Pada 1992 sistem ini ditatarkan sekali lagi menjadi 'Bank Istilah Berkomputer dalam Talian' (ringkasnya BIKOMTA) yang beroperasi pada ENCORE INFINITY 90.

Sistem ini masih boleh digunakan tetapi tidak begitu dimanfaatkan sekarang kerana tidak 'mesra pengguna' disebabkan tidak berasaskan GUI (*graphical user interface*) (Johari 1996).

2.2 Penterjemahan Automatik

Pada awal 1987, CICC (*Center of the International Cooperation in Computerization*, Tokyo, Jepun) di bawah naungan *Ministry of International Trade and Industry* (MITI) Jepun telah mengutarakan gagasan untuk melaksanakan penelitian dan pembangunan metodologi penterjemahan automatik bagi lima bahasa Asia, iaitu Jepun, bahasa Melayu, bahasa Indonesia, bahasa Cina, dan bahasa Thai.

Matlamat projek ini adalah untuk mewujudkan suatu dasar atau asas kepakaran teknologi dalam kalangan peneliti di negara-negara sedang membangun dengan cara melakukan suatu penelitian bersama dalam bidang pemprosesan bahasa tabii (*natural language processing*). Penterjemahan automatik dipilih kerana adanya matlamat ekonomi (penterjemahan dokumen, buku, majalah), budaya dan teknologi yang jelas. [Lihat sebagai contoh, Tanaka et al. (1989)].

Di bawah payung Kementerian Pendidikan Malaysia, sekumpulan peneliti Malaysia dibentuk yang terdiri daripada pensyarah Universiti Teknologi Malaysia (UTM) dan pegawai Dewan Bahasa dan Pustaka. Mereka ini telah diberikan kursus dan terlibat dalam penelitian awal pada 1987–1988 dalam fasa Kajian Asas perencanaan CICC.

Fasa ini diikuti dengan fasa penelitian dan pembangunan (R&D) dari 1989–1992. DBP pada waktu itu terlibat dalam penelitian rumus bahasa Melayu untuk komponen analisis dan penjanaan serta penyediaan kamus kata umum dan kamus istilah atau teknikal (bidang teknologi maklumat). Fasa R&D diikuti dengan fasa penilaian dari 1993–1994. [Untuk maklumat lanjut tentang pencapaian projek MMT ini sila rujuk laman Web CICC di <<http://www.cicc.or.jp/english/index.html>>]

Fasa berikutnya ialah kajian susulan (*follow-up study*). Bagaimanapun, dengan tertubuhnya Institut Terjemahan Negara Malaysia (<http://www.itnm.com.my>), atau

ringkasnya ITNM, pada 1993 maka tanggungjawab R&D penterjemahan, baik manual mahu pun mesin, beralih ke ITNM dan tidak lagi pada dasarnya menjadi peranan yang perlu dimainkan oleh DBP biarpun kerja-kerja penterjemahan masih dilakukan di Dewan Bahasa dan Pustaka secara *ad hoc*.

Untuk laporan awal tentang pelibatan Malaysia dalam projek *Multilingual Machine Translation* (MMT) yang didanai pelbagai syarikat teknologi maklumat Jepun itu, lihat sebagai contoh Mohd. Amin dan Zaharah (1988) atau Safie dan Zaharah (1989).

Meskipun R&D dalam MMT terhenti pada waktu itu, namun kesan sampingannya ternyata meluas. Projek ini telah menyedarkan Dewan Bahasa dan Pustaka tentang betapa perlunya komputer dan teknologi komunikasi digabungkan dalam kerja-kerja utama DBP agar sebahagian kerja yang selama ini rutin dan mekanikal dapat diautomatiskan, justeru itu dipercepat pelaksanaannya.

Selain itu, projek ini juga telah berjaya menanam bibit teknikal yang diperlukan untuk penelitian linguistik komputeran yang selanjutnya.

2.3 Pangkalan Data Bahasa Melayu

Idea untuk mewujudkan sebuah pangkalan data bahasa Melayu ini mula direncanakan pada awal '90-an dan mendapat dorongan yang kuat dengan termeterainya Memorandum Persefahaman di antara DBP dengan USM untuk sama-sama berusaha membina pangkalan data bahasa Melayu pada 27 Oktober 1993.

Usaha ini telah berjaya menghasilkan Sistem Korpus dan Sistem Kamus yang sehingga kini masih digunakan untuk menempatkan data teks bahasa Melayu dan *Kamus Dewan* dan *Kamus Inggeris–Melayu Dewan*.

Bagaimanapun, Sistem Leksikon yang turut tersenarai dalam kerangka memorandum tersebut hanya terlaksana setakat spesifikasi sistem kerana kendala penelitian dalam aspek-aspek kebahasaan dan perincian data yang diperlukan untuk mewujudkan Bank Leksikon.

3.0 KEGIATAN DAN KERJASAMA SEMASA

Kegiatan semasa dalam pengkomputeran bahasa Melayu tertumpu pada beberapa skop yang utama, iaitu

- pengumpulan teks untuk pangkalan data korpus;
- Pembinaan dan pengisian pangkalan data umum;
- Pembinaan dan pengisian pangkalan data leksikal;
- penataran sistem bahasa yang sedia ada dan pembinaan sistem-sistem baru dalam satu jaringan kerja yang bersepadu dan efisien; dan
- penelitian linguistik korpus (*corpus linguistics*) dan linguistik komputeran (*computational linguistics*).

3.1 Pangkalan Data Korpus

Pangkalan Data Korpus terdiri daripada Sistem Korpus dan Data Korpus. Sistem Korpus merupakan perisian yang dibangunkan oleh kumpulan penyelidik Unit Terjemahan Melalui Komputer di Universiti Sains Malaysia (USM) dalam rangka kerjasama USM-DBP berlandaskan Memorandum Persefahaman yang termeterai pada 1993.

Sistem korpus ini dilengkapi kemudahan untuk mencapai bentuk-bentuk kata, kata terbitan dan rangkai kata dan memaparkan hasil carian tersebut dalam bentuk

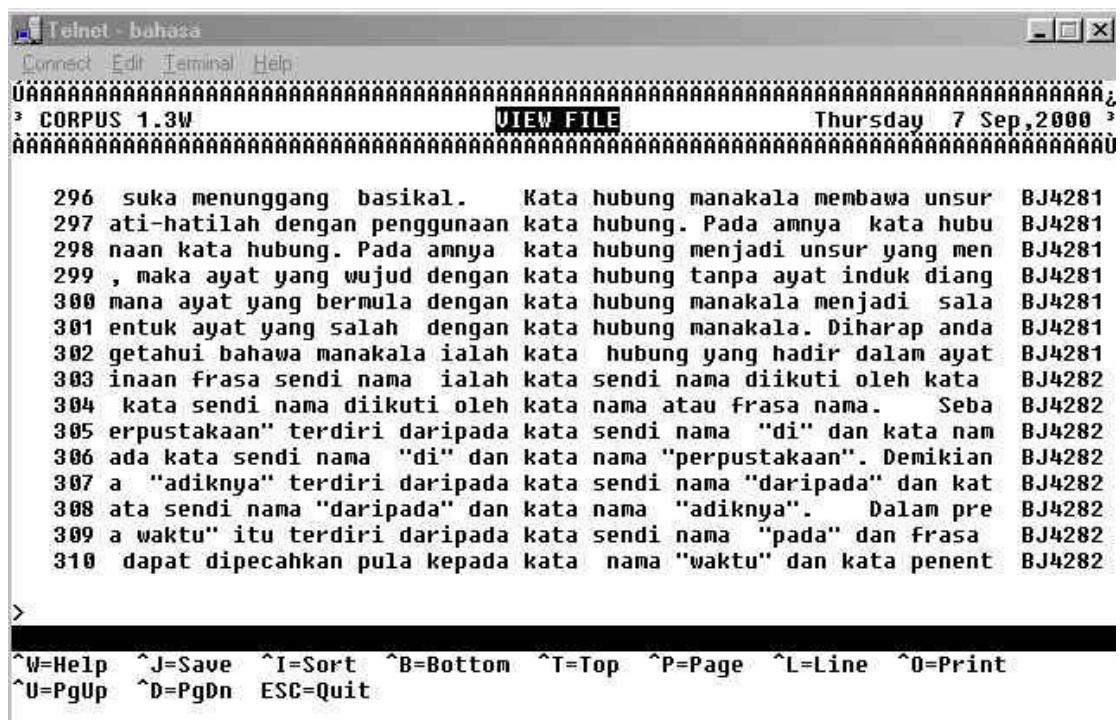
baris-baris konkordans, dengan kata kunci carian tersusun dan terisih di tengah-tengah baris.

Sistem ini juga dapat melaksanakan analisis teks dan memaparkan maklumat seperti kekerapan dan peratusan kata dasar, imbuhan (afiks), kata tak sah (*invalid words*) dan kata tak diketahui (*unknown words*). Fungsi *unknown words* ini sangat bermanfaat untuk mendapatkan kata-kata baru atau bentuk kata yang belum terakam dalam kamus.

Pencarian berdasarkan kata kunci ini boleh dibuat melalui pelbagai teknik carian, sesuai dengan maklumat yang ingin disarikan dan diperagakan. Teknik carian yang lazim adalah seperti yang berikut;

a) *Capaian melalui Kata Kunci (Bentuk Kata)*

Sesuatu bentuk kata boleh dicapai dengan menaipkan kata tersebut. Misalnya, capaian yang menggunakan kata kunci 'kata' akan memaparkan semua kehadiran bentuk ini dalam sesuatu korpus teks, seperti contoh di bawah;



Selain itu, data korpus juga dimanfaatkan oleh peneliti dari institusi pengajian/peguruan tinggi di Malaysia dan juga di luar negeri yang meneliti bahasa Melayu dari pelbagai sudut pandang.

Untuk maklumat lanjut tentang pangkalan data korpus DBP sila ke laman pengkomputeran bahasa di <http://dbp.gov.my/2000/bahasa/frame_utama.htm>.

3.2 Pangkalan Data Leksikal

Pangkalan ini pada asalnya direncanakan sebagai Pangkalan Data Perkamusan (Rusli et al. 1996a) yang berperanan sebagai,

- tempat simpanan data kamus (terbitan Dewan Bahasa dan Pustaka) dalam bentuk digital yang berstruktur;
- sistem penyusunan pelbagai jenis kamus; dan
- penjana kamus-kamus lain (seperti kamus saku) daripada kamus sedia ada.

Pada masa ini data yang termuat dalam Pangkalan Data Perkamusan terbahagi kepada pangkalan:

- Kamus Umum (*Kamus Dewan Edisi Ketiga*)
- Kamus Dwibahasa/Berbilang Bahasa (*Kamus Inggeris-Melayu Dewan*)
- Kamus Istilah (*Kamus Komputer*)

Pangkalan data untuk kamus-kamus yang disenaraikan di atas masih pada peringkat pengujian.

Bagaimanapun, dalam perancangan terbaru kami, pangkalan data perkamusan ini diperluas menjadi pangkalan data leksikal. Ini bererti setiap kamus tidak lagi

disusun sebagai sebuah buku (pangkalan) yang berasingan tetapi dalam sebuah pangkalan utama yang boleh dicapai melalui kata kunci atau kata entri atau pun makna.

3.3 Pangkalan Data Umum

Pangkalan data ini bertujuan untuk,

- menempatkan segala maklumat kebahasaan, khususnya buku dan bahan rujukan yang berkaitan dengan pedoman dan rumus bahasa Melayu;
- menempatkan maklumat kegiatan dan penelitian bahasa Melayu;
- menempatkan maklumat peneliti dan bidang penelitian masing-masing; dan
- menempatkan segala perisian yang berkaitan dengan pemrosesan bahasa tabii untuk kegunaan peneliti bahasa.

(Spesifikasi sistem dirincikan dalam Rusli et al. 1996b).

Sebahagian daripada buku pedoman utama yang dihasilkan oleh DBP dan juga Majlis Bahasa Brunei Darussalam-Indonesia-Malaysia (Mabbim) sudah didigitalkan atau ditukarkan ke dalam bentuk PDF (*portable document format*) dan diletakkan pada laman Web DBP. Antara buku pedoman dan panduan umum yang sedang digarap termasuklah,

- i. *Pedoman Umum Pembentukan Ejaan Bahasa Melayu*
- ii. *Panduan Utamakanlah Kesantunan Berbahasa*
- iii. *Panduan Penyusunan Kamus Istilah*
- iv. *Pedoman Umum Pembentukan Istilah Bahasa Melayu*
- v. *Pedoman Tatanama Kimia*

3.4 Pembinaan Sistem Bahasa Melayu Bersepadu (SBMB)

Usaha mutakhir Dewan Bahasa dan Pustaka dalam pengkomputeran bahasa Melayu ialah penataran sistem-sistem bahasa yang sedia ada dan pembinaan sistem baru untuk menampung keperluan yang sudah dikenal pasti. Sistem-sistem ini digabungkan dalam suatu sistem yang terkait erat dan saling mencapai untuk membentuk suatu sistem yang disebut Sistem Bahasa Melayu Bersepadu (SBMB).

Sistem Bahasa Melayu Bersepadu (SBMB) ini merupakan gabungan enam sistem utama yang akan digunakan dalam melaksanakan kerja-kerja kebahasaan Dewan Bahasa dan Pustaka. Keenam-enam sistem ini ialah;

- i. Sistem pengurusan dan penyusunan kamus
- ii. Sistem pengurusan dan pembinaan ensiklopedia
- iii. Sistem pengurusan dan pembinaan istilah
- iv. Sistem pengurusan dan pembinaan bahasa sukuan
- v. Sistem pengurusan dan pembinaan korpus
- vi. Sistem Khidmat Nasihat Bahasa

SBMB ini dijangka siap pembinaannya dan beroperasi dalam masa 24 bulan. Apabila siap kelak, sistem ini akan mengubah cara kerja kebahasaan Dewan Bahasa dan Pustaka dengan tumpuan kepada pembinaan dan pengembangan bahasa berasaskan Web dan Internet.

3.5 Penelitian Linguistik Korpus

Selain pembinaan pangkalan data, pembelian perkakasan komputer dan penyediaan perisian, yang paling utama ialah penelitian yang dilakukan dengan menggunakan pangkalan, perkakasan dan perisian tersebut.

Salah satu projek yang kami laksanakan dengan sumber kepakaran dari Universiti Lancaster, England dan Universiti Malaya ialah penelitian kelas kata bahasa Melayu dengan berdasarkan korpus. Penelitian ini berjaya menghasilkan leksikon beranotasi, suatu set tanda kelas kata (*tagset*) bahasa Melayu, dan kaedah untuk menandakan teks bahasa Melayu (penghurai atau *parser*) (Knowles dan Zuraidah, Akan terbit). Biarpun korpus yang digunakan dan ditandakan dengan kelas kata itu hanya sebesar sekitar 120,000 perkataan, namun penelitian ini telah meletakkan dasar yang kukuh untuk penelitian korpus yang lebih besar.

Pihak Dewan Bahasa dan Pustaka juga menjalankan kajian-kajian kecil yang berkaitan dengan korpus dan hasilnya disebar dalam bentuk makalah yang dibentangkan dalam seminar atau diterbitkan dalam majalah. Sebahagian hasil penelitian kami ada termuat dalam laman Web DBP di http://dbp.gov.my/2000/bahasa/frame_utama.htm.

4.0 Ke Mana Kita?

Sesudah lebih dua puluh tahun bergiat dalam bidang pengkomputeran bahasa ini, tentunya sudah banyak yang terlaksana seperti yang dihayati dan ada juga yang tidak tercapai seperti yang diidamkan. Daripada penilaian kami, yang sering menjadi kendala dan hambatan dalam pelaksanaan pengkomputeran bahasa bukannya dana tetapi sumber manusia.

Kepakaran dalam bidang pemprosesan bahasa tabii di Malaysia sangat sedikit, apatah lagi di DBP. Yang pasti, upaya pengkomputeran bahasa ini memerlukan input

dan sokongan daripada pelbagai pihak yang berlainan disiplin dan domain kepakaran. Justeru itu, pelaksanaan sesuatu projek pengkomputeran yang rumit harus melibatkan pihak-pihak lain.

Sebagai contoh, kerjasama pada peringkat serantau yang dicetuskan dalam wadah Majlis Bahasa Brunei Darussalam–Indonesia–Malaysia (akronimnya Mabbim) harus ditafsirkan dan dilanjutkan dalam kerjasama teknologi maklumat, dan dengan demikian kegiatan Mabbim tidak lagi tertumpu pada soal peristilahan semata-mata. Gagasan ini sudah pun diterima oleh Sidang Eksekutif Mabbim tatkala merestukan penubuhan Gerbang Bahasa Melayu/Bahasa Indonesia dan memberi izin agar masing-masing negara anggota dapat merencanakan penubuhan gerbang atau portal bahasa masing-masing. Cuma setakat ini, pelaksanaannya belum dapat digiatkan sepertimana yang diharapkan.

Usaha sama dengan institusi, baik swasta mahu pun kerajaan (baca pemerintah di Indonesia), baik di dalam mahu pun di luar negeri, harus dirangsangkan dan digerakkan. Dalam hal ini, DBP menaruh harapan yang besar terhadap kerjasama dengan BPPT dalam beberapa aspek pengkomputeran bahasa. Sebagai contoh, BPPT sudah pun mewujudkan pusat UNL (*Universal Networking Language*) untuk Bahasa Indonesia dan penelitian tentang UNL sudah banyak dilakukan sedangkan DBP baru sahaja mendaftar sebagai ahli Persatuan UNL. Lantaran itu, ada baiknya penelitian UNL bahasa Melayu disekalikan dengan UNL bahasa Indonesia mengingat bahawa kedua-dua bahasa ini berkongsi struktur, tatabahasa, dan sebahagian besar kosa kata. Perbezaan yang ada hanyalah pada beberapa patah kata dan makna sahaja.

5.0 Penutup

Makalah ini melakarkan serba ringkas tentang upaya Dewan Bahasa dan Pustaka Malaysia untuk memberi erti kepada *raison d'être*-nya melalui pengkomputeran agar pembinaan dan pengembangan bahasa Melayu dapat dipergiat, diperluas dan dipercepat sesuai dengan tuntutan alam yang global dan nirsempadan.

BIBLIOGRAFI

Johari Md. Hassan 1996. "Sistem Peristilahan Berkomputer" dalam *Bengkel Reka Bentuk Sistem Bahasa Melayu Berkomputer Bersepadu* (Edaran Terhad).

Mohd. Amin Alias dan Zaharah Abd. Ghafur 1988. "Projek Kerjasama Penyelidikan dan Pembangunan Sistem Terjemahan Automatik antara Malaysia dan Jepun" dalam *Bengkel Kebangsaan Penterjemahan Berkomputer*. DBP.

Rusli Abdul Ghani, Shariffulizan Malek, Diarani Mat Adam dan Zaiton Darois 1996a. "Spesifikasi Fungsian Pangkalan Data Perkamusan" *Bengkel Reka Bentuk Sistem Bahasa Melayu Berkomputer Bersepadu* (Edaran Terhad).

Rusli Abdul Ghani, Shariffulizan Malek, Diarani Mat Adam dan Zaiton Darois 1996b. "Spesifikasi Fungsian Pangkalan Data Umum Bahasa Melayu" *Bengkel Reka Bentuk Sistem Bahasa Melayu Berkomputer Bersepadu* (Edaran Terhad).

Safie Mat Yatim dan Zaharah Abd. Ghafur 1989. "Projek Kerjasama Penyelidikan dan Pembangunan Sistem Penterjemahan Berkomputer di antara Malaysia dan Jepun" dalam *Kumpulan Kertas Kerja Simposium Antarabangsa Penterjemahan Berkomputer 1989, 22-24 Ogos 1989*. Kementerian Pendidikan Malaysia. (hlm 7–13).

Tanaka, H, Ishizaki, S, Uehara, A dan Uchida, H 1989. "Research and Development of Cooperation Project on a Machine Translation System for Japan and its Neighboring Countries" dalam *Kumpulan Kertas Kerja Simposium*

Antarabangsa Penterjemahan Berkomputer 1989, 22-24 Ogos 1989.

Kementerian Pendidikan Malaysia. (hlm 1-6).

Knowles, G. dan Zuraidah Mohd. Dom (Akan terbit). *Word Class in Malay: A Corpus-based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

“Research Cooperation on Machine Translation System.” Dicapai (*accessed*) pada 15

September 2003 daripada Laman CICC di

<<http://www.cicc.or.jp/english/index.html>>.