Melayari Samudera Maya, Mencari Mutiara Kata: Suatu Metodologi Pemerolehan Kata Baru Berdasarkan Korpus*

Rusli Abdul Ghani
(rusli@dbp.gov.my)
Nur Hafizah Mohamed Husin
(hafizah@dbp.gov.my)
Bahagian Penyelidikan Bahasa
Dewan Bahasa dan Pustaka
(http://www.dbp.gov.my)

Abstrak

Perkembangan pesat dalam bidang Teknologi Maklumat dan Komunikasi (TMK) turut menghasilkan kata dan istilah baru yang belum lagi dikutip, digilap dan diuntaikan menjadi entri kamus yang bertakrif. Kajian ini menghuraikan suatu metodologi untuk mendapatkan kata baru dengan memanfaatkan korpus teks akhbar.

Abstract

The rapid development in Information and Communication Technology (ICT) has given rise to many new words and terms that have yet to be described and defined. This study describes a methodology for the acquisition of new words based on a corpus of journalistic text.

Penghargaan

Kami dengan takzimnya merakamkan setinggi-tinggi penghargaan kepada **Kumpulan The New Straits Times Press (Malaysia) Berhad** kerana pembekalan data NSTP e-Media dalam bentuk digital. Sumbangan ini amat besar ertinya dalam pembinaan dan pengembangan bahasa Melayu.

1.0 Pendahuluan

Perkembangan pesat dunia Teknologi Maklumat dan Komunikasi,[1] baik yang maujud (perkakasan WAP, PIM dan PDA) mahupun yang maya (di WWW, IRC, Usenet dan sebagainya), turut mewarnai alam bahasa Melayu melalui penggunaan laras *technospeak* atau *geekspeak*[2] serta istilah baru yang pelbagai bentuknya (sama ada yang tergubal secara pengakroniman, pemajmukan atau melalui pengitaran semula kata lama dengan pengertian baru[3]).

Web sebagai wadah terbitan jauh lebih luwes dan global sifatnya ketimbang media tradisional. Sebagai wahana ilmu pula, Web lebih pantas dan praktis berbanding mesin cetak. Justeru itu, alam mayalah tempat yang sesuai untuk mencari kata-kata baru. Lagipun, maklumat leksikal pada Web semakin diyakini pengguna; seperti kata Constance Hale, editor *Wired Style* (dikutip dalam Long 2000) "*When it comes down to a choice between what's on the Web and what's in* Webster's*, we tend to go with the Web*."

[1] Sesudah ini disingkatkan kepada TMK (sebagai padanan singkatan Inggeris ICT).
[2] Untuk jargon dan istilah mutakhir TMK dalam bahasa Inggeris lihat sebagai contoh laman Web http://foldoc.doc.ic.ac.uk/foldoc/index.html, http://www.jargon.net dan http://www.webopedia.com.
[3] Untuk kaedah pembentukan kata dan istilah lihat sebagai contoh Bauer (1983).

2.0 Kerangka Analisis

Fokus kami adalah terhadap kata dan penggunaan kata dalam industri TMK berdasarkan andaian bahawa banyak kata atau istilah baru digunakan dalam bidang yang sedang pesat berkembang ini.

TMK secara umumnya dapat ditakrifkan sebagai segmen industri yang merangkumi bidang penghasilan komputer dan perkakasan sampingannya, khidmat perisian dan pengaturcaraan, khidmat komunikasi dan dagangan elektronik yang melibatkan Internet (termasuk WWW) ataupun Intranet swasta dan awam (Henry *et al.* 1999).

Namun demikian, kami tidak pula mahu mencari istilah yang terlalu teknikal[4] kerana apa yang dihajati sebenarnya ialah kata atau istilah baru yang popular sifatnya yang dapat dipertimbangkan untuk menjadi entri kamus umum bahasa Melayu. Lantaran itu, kajian ini tidak memanfaatkan wacana TMK yang teknikal (yang tentunya akan menghasilkan banyak istilah baru) tetapi sebaliknya mengandalkan teks akhbar, jenis wacana yang ternyata popular, umum dan meluas pengaruh dan sebarannya.[5]

Kami juga menyedari hakikat bahawa baru tidaknya sesuatu kata itu relatif sifatnya kerana penentuannya bergantung pada titik rujukan temporal yang diterima pakai. Ada penelitian dibuat berlandaskan dikotomi kuno–moden dengan sela yang berabad lamanya dan ada juga kajian kata baru dibuat berdasarkan sesuatu batas kala dasawarsaan. Misalnya, untuk tujuan penyusunan *The Oxford Dictionary of New Words*, editornya menganggap kata baru itu sebagai " … *any word, phrase, or meaning that came into popular use in English or enjoyed a vogue during the eighties or early nineties.*" (Tulloch 1992: v).

Kami pula menerima pakai takrifan 'kata baru' yang praktis bersesuaian dengan matlamat kami untuk mencari calon entri baru untuk kamus umum. Dalam konteks ini, kata baru ialah mana-mana bentuk kata yang belum lagi terakam dalam *Kamus Dewan* Edisi Ketiga[6] (DBP:1994) atau sebarang makna yang belum terungkap oleh kata entri, subentri atau frasa dalam KD3. Dalam konteks ini 'kata baru' bererti bentuk-bentuk kata dan penggunaan kata yang muncul sesudah terbitnya KD3 (dengan andaian bahawa sebahagian besar perkataan yang digunakan dalam bahasa Melayu sebelum 1994 sudah terakam).

2.1 Dorongan

Penelitian ini merupakan sebuah projek tunasan daripada projek hakiki pembinaan pangkalan data korpus yang ditangani Bahagian Penyelidikan Bahasa, Dewan Bahasa dan Pustaka. Sebagai amalan rutin perakaman dan penginputan data korpus, sesebuah teks yang hendak dimasukkan ke dalam sistem korpus akan dianalisis terlebih dahulu dengan perisian *Malay Text Analysis* (MATA).[7]

Pemprosesan ini akan menghasilkan maklumat perkataan yang terkandung dalam teks tersebut. Sebagai contoh, analisis sebuah buku sosiologi menghasilkan maklumat seperti yang berikut;

```
RECORD NUMBER      : 1
TEXT : Sosiologi

                    **********************************
                    WORD ANALYSIS - ROOT WORD FREQUENCY
                    **********************************

          Options :
          ---------

          Number of words               :          All
          Frequency                     :          All
```

[4] Untuk metodologi pemprosesan istilah lihat sebagai contoh Sager (1990) dan Felber (1984:117-126).
[5] Lihat sebagai contoh Rademan 1996 untuk manfaat dan mudarat menggunakan teks akhbar dalam talian sebagai bahan korpus.
[6] Diringkaskan kepada KD3 sesudah ini.
[7] Sistem analisis teks yang tergabung dalam sistem korpus yang dibangunkan oleh Unit Terjemahan Melalui Komputer, USM.

Total Number of Words	19279	100.00000
Root Words	12779	66.28456
Affixed Words	5551	28.79299
New Words	912	4.73054
Numbers	25	0.12967
Invalid Words	12	0.06224

Root Word		Count	(%)
yang	:	861	4.46600
dan	:	553	2.86841
dalam	:	330	1.71171
tidak	:	210	1.08927
dengan	:	208	1.07889
masyarakat	:	208	1.07889
orang	:	192	0.99590
...			
variasi	:	1	0.00519
virus	:	1	0.00519
wajib	:	1	0.00519
wakil	:	1	0.00519
warganegara	:	1	0.00519
watak	:	1	0.00519
yuran	:	1	0.00519
zat	:	1	0.00519
1295		18330	95.07755

Unknown Word		Count	(%)
2	:	33	0.17117
1	:	31	0.16080
kanak	:	18	0.09337
4	:	17	0.08818
marx	:	16	0.08299
6	:	14	0.07262
china	:	7	0.03631
ketidakpuasan	:	7	0.03631
mengkaji	:	7	0.03631
15	:	6	0.03112
...			
totem	:	1	0.00519
turner	:	1	0.00519
universti	:	1	0.00519
york	:	1	0.00519
421		912	4.73054

Invalid Word		Count	(%)
termasuklah	:	4	0.02075
pembelajaran	:	3	0.01556
organisme	:	1	0.00519
pemimpinnya	:	1	0.00519
peratuskah	:	1	0.00519
persekitarannya	:	1	0.00519
terbentuklah	:	1	0.00519
7		12	0.06224

Tiga senarai lengkap bentuk kata terhasil daripada pemprosesan ini, iaitu 'Known Word,' 'Unknown Word' dan 'Invalid Word.' 'Known Word' merupakan senarai kata dasar yang digunakan dalam teks tersebut dan sepadan dengan entri kamus sistem MATA (yang disediakan berpandukan kata dasar yang terdapat dalam KD3 serta rumus umum morfologi bahasa Melayu). Senarai 'Unknown Word' merupakan gabungan 'New Words,[8] dan 'Numbers' manakala 'Invalid Word' merupakan senarai perkataan yang tersalah eja dan perkataan yang tidak dapat dihuraikan oleh rumus morfologi sistem.

Pada asalnya kami melakukan pemprosesan ini untuk memastikan teks digital yang diinput ke dalam sistem korpus itu sama dengan versi cetakannya. Bagaimanapun, output pemprosesan ini mencadangkan kemungkinan lain yang bermanfaat selain berperanan sebagai penyemak ejaan dan penyemak keutuhan wacana. Keadaan ini mendorong kami untuk meneliti dua kemungkinan;

i. Penggunaan sistem yang sedia ada untuk mendapatkan kata baru; dan

ii. Penghuraian suatu metodologi pemerolehan kata baru dengan tujuan untuk membina sebuah sistem kata baru yang lebih canggih yang dapat menunjangi kerja-kerja pengumpulan entri baru dalam leksikografi dan pengumpulan istilah sumber untuk terminografi.

2.2 Ruang Lingkup dan Data Penelitian

Penelitian ini dibuat dengan menggunakan teks akhbar dalam talian. Kebanyakan akhbar berbahasa Melayu ada laman Web masing-masing; antara yang tersohor ialah Berita Harian,[9] Harakah Daily[10] dan Utusan Online.[11]

Bagaimanapun, atas alasan kepraktisan, kami sekadar menggunakan data digital NSTP e-Media yang dibekal secara pukal oleh pihak NSTP dari semasa ke semasa. Dengan cara ini, kami tidak perlu memuatturunkan, dokumen demi dokumen, data akhbar NSTP daripada Web.

Data NSTP e-Media yang kami gunakan merupakan data Januari-Julai 2000 (dilabelkan NSTP00) daripada *Berita Harian, Berita Minggu, Harian Metro* dan *Metro Ahad*. Saiz data ini ialah sekitar 3 juta perkataan (atau token).

Kami abaikan data akhbar lain pada Web kerana penelitian ini lebih tertumpu pada metodologi pemerolehan kata baru dan bukannya suatu usaha tuntas untuk mengutip sebanyak mungkin kata baru.

2.3 Kajian yang Berkaitan

Beberapa usaha telah dan sedang dilakukan dalam ranah neologisme ini, termasuklah projek CORDON,[12] Neolosearch,[13] AVIATOR[14] dan kajian yang dilakukan Roche (1998a dan 1998b).

Projek CORDON bermatlamat untuk membina perisian yang dapat mengesan neologisme (kata baru) yang terdapat dalam sesebuah teks berdasarkan kaedah statistik kekerapan serta taburan perkataan dan kolokasi perkataan tersebut di dalam korpus pantauan (*monitor corpus*).

AVIATOR yang digarap oleh kumpulan penyelidik Universiti Liverpool pada 1990-1993 menggunakan empat perisian (yang bertindak sebagai turas) untuk mengesan dan mengasingkan kata-kata baru daripada teks akhbar (*The Independent*).[15] Demikian juga dengan penyelidikan yang dibuat oleh Roche (ibid.). Beliau membahagikan proses pencarian dan pengasingan kata baru daripada sesuatu senarai kata kepada beberapa tahap, termasuklah tahap pengasingan kata yang diketahui, tahap pengasingan kata nama khas dan tahap pengasingan akronim supaya baki daripada senarai asal itu nanti merupakan calon kata baru.

Neolosearch (Janicijevic dan Walker 1997) pula merupakan usaha untuk membangunkan suatu sistem automatik pencarian kata baru dalam dokumen berbahasa Perancis yang terdapat pada Internet.

---

[8] Dalam senarai 'unknown word' kata 'totem' merupakan calon kata baru kerana belum terakamkan dalam KD3.
[9] http://www.emedia.com.my/ (The New Straits Times Press (M) Bhd.)
[10] http://www.harakahdaily.com/ (HarakahDaily.com)
[11] http://www.utusan.com.my/ (Utusan Melayu (M) Bhd.)
[12] http://www.ids.-mannheim.de/telri/seminar/jv-cordon.html .
[13] http://www.qucis.queensu.ca/achallc97/papers/a009.html .
[14] http://radar.rdues.liv.ac.uk/newwds.html .
[15] http://www.independent.co.uk/news

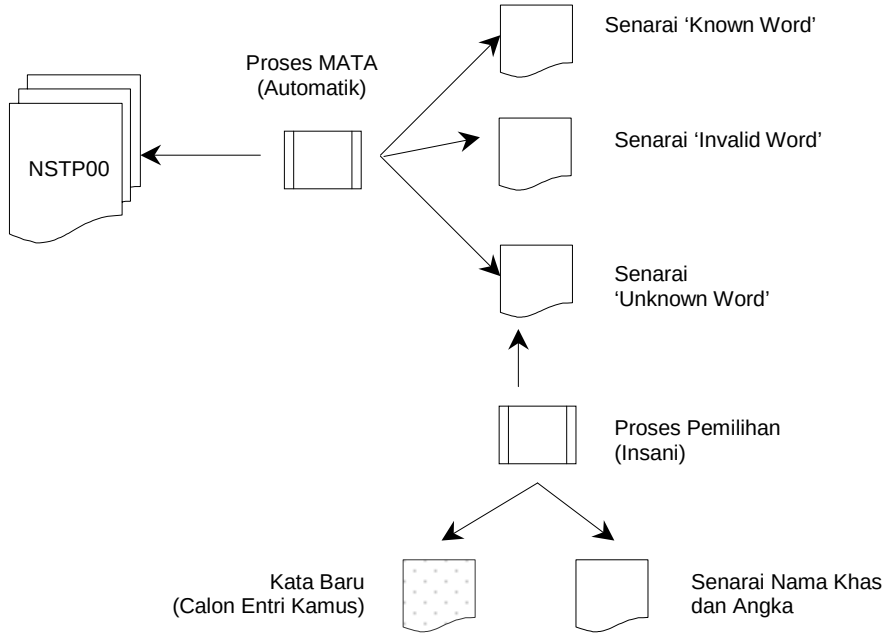
3.0 Metodologi

Dalam bahagian ini kami menghuraikan dua metodologi umum. Yang pertama ialah kaedah yang melibatkan penggunaan sistem sedia ada manakala yang kedua merupakan metodologi cadangan sebagai pembaikan kepada yang pertama.

3.1 Sistem Sedia Ada

Dalam kaedah ini data NSTP00 diproses dengan menggunakan sistem MATA untuk menghasilkan tiga senarai bentuk perkataan yang jumlahnya sekitar 56,100 dan senarai output ini diteliti berpandukan kerelevanan setiap bentuk kata dengan domain kajian ini (Rajah 1).

Rajah 1: Carta Alir Sistem Sedia Ada



3.1.1 Penelitian Output

Senarai 'Known Word' boleh diabaikan daripada penelitian selanjutnya kerana senarai ini mengandungi kata yang sedia terakam dalam KD3.

Senarai 'Invalid Word' juga boleh dikesampingkan kerana mengandungi bentuk kata yang tersalah eja,[16] dan juga bentuk kata terbitan[17] serta bentuk berklitik dan berpartikel[18] yang masing-masing tidak terdaya dileraikan oleh penganalisis morfologi yang sedia ada. Bentuk berklitik dan berpartikel boleh juga dilurutkan klitik dan partikel masing-masing secara insani untuk penelitian selanjutnya kerana ada kemungkinan bentuk kata baru itu hadir dalam bentuk berklitik dan berpartikel. Namun begitu, dalam kajian ini kami mengabaikan senarai ini. Bagaimanapun, senarai ini berguna sekali untuk kajian pembaikan rumus morfologi dan seterusnya pembaikan sistem MATA.

Yang menjadi teras telitian ialah senarai 'Unknown Word.' Penelitian untuk mendapatkan kata baru dalam bidang TMK ini dilakukan secara manual kerana belum ada cara yang dapat mengkategorikan secara

[16] Misalnya 'organisme' dalam contoh teks sosiologi.
[17] Umpamanya 'pembelajaran' dalam contoh teks sosiologi.
[18] Seperti 'termasuklah,' 'pembelajaran' dan 'peratuskah' dalam contoh teks sosiologi.

automatik kata atau istilah mengikut domain pemakaian masing-masing. Senarai ini agak panjang dan yang dapat kami paparkan dan bincangkan di sini hanyalah sebahagian kecil daripada kata yang berkemungkinan layak ditangani sebagai entri kamus umum. Perhatikan senarai yang berikut:

Calon Kata Baru	Kekerapan	Calon Kata Baru	Kekerapan
ADSL	1	html	4
dotcom	142	ICT	99
e-beli	1	intranet	4
e-beli-belah	2	ISDN	23
e-dagang	95	k-ekonomi	16
e-perdagangan	1	pengimbas	3
e-komuniti	8	portal	19
e-niaga	8	webzine	1
e-pembelajaran	4	telko	7
e-perbankan	1	telco	1
e-perniagaan	2	WAP	91
		virtual	19

Senarai ini memperlihatkan beberapa contoh bentuk kata dengan kekerapan masing-masing di dalam korpus NSTP00. Setiap kata ini kemudiannya dikeluarkan baris konkordans dan dicari kolokasi masing-masing untuk menentukan konteks pemakaian.

3.1.1.1 e-Tu, e-Ni

Antara bentuk kata baru yang terpampang jelas dalam senarai di atas ialah bentuk yang berpangkalkan 'e-.' Nampaknya pangkal 'e-' ini sangat produktif dan hampir semua kegiatan manusiawi yang boleh dielektronikkan boleh dilekapkan dengan e-. Keadaan ini tentunya akan menghasilkan bermacam *e*-tu dan *e*-ni. Maka terpulanglah kepada kebijaksanaan ahli perkamusan untuk menangani bentuk begini sebagai entri yang berasingan atau sebagai bentuk bergabung seperti 'anti-,' 'pro-' dan sebagainya. Fenomena ini juga diramalkan akan berulang dengan 'k-' yang berupa kependekan bagi *knowledge* dan buat masa ini jelas tertempel pada ekonomi sahaja.

Kekerapan yang tinggi mencadangkan penggunaan yang popular dan ini perlu diambil kira sebagai salah satu ciri calon entri kamus, meskipun diakui bahawa ada faktor lain yang perlu dicongakkan juga sebelum sesuatu kata itu diterima sebagai entri.

Sehubungan dengan ini, e-dagang nampaknya jauh lebih popular penggunaannya daripada e-perdagangan sedangkan kedua-duanya merupakan bentuk padanan bagi '*e-commerce*.' Hakikat ini memperlihatkan keutamaan pemakaian yang ketara. Demikian juga dengan e-niaga ketimbang e-perniagaan.

3.1.1.2 Singkatan

Ada juga bentuk kata dalam senarai di atas yang berupa singkatan istilah Inggeris; umpamanya ADSL, html, ICT, ISDN dan WAP. Kesemua istilah ini belum terakam dalam kamus mahupun dalam *Glosari Teknologi Maklumat* (DBP: 1996). Justeru itu, output penelitian bukan sahaja boleh dimanfaatkan oleh penyusun kamus tetapi juga oleh penggubal istilah.

Satu contoh singkatan yang kerap diperkatakan dalam akhbar ialah WAP, sama ada berkolokasi akrab dengan portal, telefon dan perkhidmatan (perkhidmatan WAP, portal WAP, telefon WAP) atau terkurung selepas rangkai kata 'protokol aplikasi tanpa wayar.'

Perhatikan kolokasi dan baris konkordans WAP di bawah;

Kolokasi Kiri		Kata Kunci	Kolokasi Kanan	
Bentuk Kata	**Kekerapan**		**Bentuk Kata**	**Kekerapan**
perkhidmatan	13	**WAP**	portal	10
mandi	8		pengguna	5
telefon	7		perkhidmatan	7

untuk memastikan perkhidmatan **WAP** rakan niaga globalnya termasuk Maxis
engecualikan caj perkhidmatan **WAP** sebanyak RM10 selama 12 bulan. Baga
  rti WAP. Untuk perkhidmatan **WAP** itu, Maxis bekerjasama dengan beberap
etika pelancaran perkhidmatan **WAP** pertama di China, di Beijing, baru-ba
Untuk pelancaran perkhidmatan **WAP** di China, Motorola China bekerjasama
yarikat penyedia perkhidmatan **WAP**," katanya. Sementara itu, Bistamam
u ini menawarkan perkhidmatan **WAP**. Produk seterusnya Perkhidmatan Radio
vendor menawarkan perkhimatan **WAP** ke seluruh dunia. Meskipun mengunju
ilai tambah daripada platform **WAP** dan SMS keluaran MTech dikenali eBuzz
hidmatan menggunakan platform **WAP** itu secepat mungkin. "Selain dengan M
erkhidmatan menerusi platform **WAP**. Perkhidmatan itu termasuk perdagan
n sejak Februari lalu, portal **WAP** Maxis boleh dicapai sepenuhnya menggu
  t (GMI), Tim Copper. Portal **WAP** sistem telefon mudah alih merupakan e
  354 wds. Maxis tawar portal **WAP** paling meluas
ian selepas pelancaran portal **WAP** Maxis di Kuala Lumpur, semalam. Hadir
ungi kadar lemak yang rendah. **Wap** Wap yang dihasilkan air panas juga
 Razlan ... standard sejagat. **WAP** Portal Sdn Bhd (WAP Portal), pembekal
 menggunakan telefon seluluar **WAP** mulai semalam. "Selain menawarkan b
  zink. Bahang dilihat semacam **wap** air yang lepas bebas secara rawak. Si
 pula, melalui khidmat sistem **WAP**, pengguna kini boleh menikmati pelbag
uga digalakmencuba mandi susu **wap** herba bagi memberi vitamin pada kulit
an. Donlan berkata, teknologi **WAP** dijangka membuka dimensi baru dalam p
ru `telefon media'. Teknologi **WAP** membolehkan maklumat daripada Interne
amat besar menerusi teknologi **WAP**, Net-Linx tidak mengetepikan perniaga
n ini berlaku sebelum telefon **WAP** muncul. Ramai pemerhati, termasuk s
pai menggunakan telefon upaya **WAP**, Motorola Acoompli A16188 yang dibang
 kadar lemak yang rendah. Wap **Wap** yang dihasilkan air panas juga berper
otokol aplikasi tanpa wayar (**WAP**) dan perkhidmatan mesej pendek (SMS)
nologi aplikasi tanpa wayar (**WAP**). "Malah, Motorola akan membangun d
otokol aplikasi tanpa wayar (**WAP**) bebas, akan melabur sehingga RM25 ju
otokol Aplikasi Tanpa Wayar (**WAP**) percuma untuk langganan baru. DiGi
otokol Aplikasi Tanpa Wayar (**WAP**) turut mempengaruhi reka bentuk model
otokol Aplikasi Tanpa Wayar (**WAP**) mendorong telefon bimbit klon dipasa
otokol aplikasi tanpa wayar (**WAP**) bagi aplikasi Internet mudah alih me
otokol Aplikasi Tanpa Wayar (**WAP**). Naib Presiden Eksekutif Kumpulan
otokol Aplikasi Tanpa Wayar (**WAP**) paling meluas di Malaysia dengan leb
 tokol Aplikasi Tanpa Wayar (**WAP**) dan telefon selular Generasi Ketiga

WAP merupakan kependekan bagi *wireless application protocol*. Istilah ini belum ada dalam senarai istilah terbitan DBP terkini, namun istilah yang telah digunakan dalam terbitan NSTP ialah protokol aplikasi tanpa wayar. Ini menyarankan kepada penggubal istilah suatu bentuk kata yang sudah dipakai dan hanya memerlukan pentauliahan rasmi DBP.

Demikian juga dengan bentuk kata seperti ‘dotcom,’ ‘intranet,’ ‘portal,’ ‘webzine’ dan ‘virtual.’ Bentuk-bentuk ini boleh menjadi bahan asas untuk pembentukan istilah dan terminografi.

3.1.2 Kelemahan Sistem Sedia Ada

Sistem yang sedia ada mempunyai kelemahan tersendiri kerana banyak bentuk kata yang digunakan dalam teks tidak dapat dikesan. Barang diingat bahawa pemproses sistem ini sebenarnya penganalisis teks dan bukannya sistem yang dibangunkan khusus untuk mencari kata baru. Kami hanya membincangkan dua bentuk yang terlepas daripada penurasan sistem ini.

3.1.2.1 Kata Majmuk

Bentuk kata majmuk tidak dapat dikesan oleh sistem ini kerana sistem analisis hanya memproses kata tunggal dan mengeluarkan output yang terdiri daripada senarai kata tunggal. Oleh itu, istilah seperti ‘capaian tanpa wayar’ tidak dapat dikesan kerana yang diproses dan dioutputkan ialah ‘capaian,’ ‘tanpa’ dan ‘wayar’ yang masing-masing sudah terdapat dalam KD3.

Demikian juga dengan pelbagai jenis telefon (telefon bimbit, telefon selular, telefon mudah alih, telefon bergerak, telefon tetap dan telefon awam) dan televisyen (televisyen kabel dan televisyen litar tertutup).

3.1.2.2 Kata Lama Makna Baru

Semua bentuk kata yang memperoleh makna baru tidak dapat dikesan dengan menggunakan metodologi mudah ini kerana sistem hanya berupaya membandingkan bentuk tetapi tidak mampu mengesan makna atau perubahan makna. Oleh itu kata yang sudah mendapat makna baru atau yang sudah diperluas penggunaannya tidak diasingkan sebagai 'unknown word.'

Kata 'maya,' 'laman' dan 'gerbang' sudah memperoleh makna baru. Laman yang dahulunya digarap sebagai;

laman bp kep halaman: *jamuan* ~. (KD3 *q.v*. **laman**)

sudah memperlihatkan kolokasi baru dan pemakaian yang sangat kerap dalam dunia TMK seperti yang dipaparkan dalam jadual di bawah:

Kolokasi Kiri		Kata Kunci	Kolokasi Kanan	
Bentuk Kata	**Kekerapan**		**Bentuk Kata**	**Kekerapan**
melalui	23	**laman**	Web/Webnya	229
menerusi	21		Internet	21
melayari	15		budaya	11
internet	12		WWW	10

Demikian juga dengan 'gerbang' yang dahulunya merupakan binaan maujud yang menjadi pintu masuk utama ke kota, istana atau negeri. Kini 'gerbang' lebih kerap berperanan sebagai bukaan maya ke alam Internet seperti yang ditunjukkan oleh kolokasi dan baris konkordans di bawah:

Kolokasi Kiri		Kata Kunci	Kolokasi Kanan	
Bentuk Kata	**Kekerapan**		**Bentuk Kata**	**Kekerapan**
perkhidmatan	3	**gerbang**	Internet	8
menubuhkan	2		Web	6
			kewangan	2
			utama	2

, restoran terbuka, tempat duduk, **gerbang** laluan dan plaza. "Bagi mengekalk
ura adalah kecil untuk perniagaan **gerbang** maklumat tetapi ia sesuai untuk mem
-dagang. Secara keseluruhannya, **gerbang** Netvigator.com ini mampu menawarkan
ereka yang ingin menjejak kaki ke **gerbang** perkahwinan, modul keibu-bapaan dan
Kampung Lan Kok dan melalui pintu **gerbang** sempadan Terenggnu dan Kelantan.
u banyak syarikat yang menubuhkan **gerbang** sendiri dan keadaan ini akan memung
matlamatnya, iaitu tiba ke pintu **gerbang** Simurgh, dan melihat cahaya yang be
n Alpha Intercontinental Sdn Bhd, **Gerbang** Sutera Sdn Bhd dan Infostas Enginee
iwujudkan yang berperanan sebagai **gerbang** utama bagi pemasaran barangan tempa
a, Celcom akan bekerjasama dengan **gerbang** utama tempatan, Skali.com untuk men
10 rakan kongsi untuk menjayakan **gerbang** web CikguNet, iaitu Berita Harian S
negara ini dengan memperkenalkan **gerbang** web CikguNet yang menggunakan alama
am, statistik terkini menunjukkan **gerbang** web CikguNet sudah dikunjungi kira-
lu. Lebih menarik lagi, pelawat **gerbang** web hasil ilham khas Mimos Berhad i
ira-kira 68 peratus pengunjung ke **gerbang** web berkenaan adalah pelawat tempat
Zealand dan Australia. Walaupun **gerbang** web itu dibina dalam bahasa Melayu
atan menarik yang disediakan oleh **gerbang** web berkenaan ialah rancangan menga
alan perkhidmatan Internet (ISP), **gerbang** web, enjin carian dan kemudahan mel
arakat lain, guru juga menghadapi **gerbang** yang mencabar, iaitu era globalisas

'Melayari' pun tidak semestinya di laut nyata; malah yang lebih kerap dilayari buat masa ini ialah samudera maya.

Kolokasi Kiri		Kata Kunci	Kolokasi Kanan	
Bentuk Kata	**Kekerapan**		**Bentuk Kata**	**Kekerapan**
boleh	8	**melayari**	Web	16
ekonomi	4		laman	15
pengguna	3		Internet	7
pelanggan	2		lebuh	4

```
lelaki dan seorang wanita dalam melayari hidup. Manusia itu tidak dapat lar
desire-academy@hotmail.com atau melayari homepage :http://members.zoom.com/
 ira 400,000 pencari kerja akan melayari internet bagi mencari kerja manaka
 pemilik telefon bimbit sanggup melayari Internet bagi mencari melodi berla
dak boleh memegang komputer dan melayari Internet dan seterusnya berurus ni
ini, lebih ramai penduduk dunia melayari Internet berbanding mereka yang me
olah semakin lalai kerana asyik melayari Internet," katanya. Menjawab soa
nak menonton televisyen mahupun melayari Internet, pastikan anda bersama me
yang sama boleh digunakan untuk melayari Internet. Pengurus Besar Kumpula
, Inc yang membolehkan pengguna melayari kandungan web interaktif dan perib
mur boleh disatukan serta boleh melayari kehidupan sebagai suami isteri ber
banyakan remaja lebih cenderung melayari laman web bersifat negatif. Misa
ap dengan pegawai jualan ketika melayari laman web sesebuah syarikat berken
nggannya yang ketika itu sedang melayari laman web syarikat. "Maklumat ta
ya amat mudah, anda hanya perlu melayari laman web www.lycosasia.com.my dan
an, malah membenarkan pelanggan melayari laman web kasino dengan membuat ba
 Bakal pembeli, umpamanya boleh melayari laman web Bukit Rimau untuk mencar
 borang permohonan dengan hanya melayari laman web K & N Kenanga Holdings B
talian bebas tol 1-800-88-3117, melayari laman web http://www.ygmb.com.my/y
```

3.2 Sistem Cadangan

Metodologi yang digunakan dalam sistem kata baru yang lebih canggih seharusnya dapat mengesan bentuk kata majmuk dan juga perubahan makna. Salah satu cara untuk mengesan perkembangan ini adalah dengan mengambil kira maklumat statistik (kekerapan dan kekerapan berkolokasi dengan kata lain) sesuatu bentuk kata itu di dalam teks tertentu berbanding dengan korpus pantauan yang menyeluruh, representatif dan seimbang (lihat sebagai contoh Church dan Hanks 1990; Daille et al. 1994 dan Smadja dan McKeown 1990 untuk perbincangan dan huraian lanjut tentang penggunaan maklumat statistik leksikal dalam pemprosesan bahasa tabii).

Sebagai contoh, kata 'gerbang' hadir dengan kekerapan dan kolokasi yang berbeza dalam tiga korpus yang berbeza (korpus teks lama,[19] korpus NSTP tahun 1999 dan korpus NSTP 2000) seperti yang ditunjukkan di bawah.

KORPUS	**TEKS LAMA (1.4 JUTA)**		**NSTP99 (8 JUTA)**		**NSTP00 (4 JUTA)**	
Kekerapan	1 dalam 27,000		1 dalam 230,000		1 dalam 87,000	
Kolokasi Kiri	Kata	Kekerapan	Kata	Kekerapan	Kata	Kekerapan
	pintu	39	pintu	18	perkhidmatan menubuhkan	3 2
Kolokasi Kanan	Kata	Kekerapan	Kata	Kekerapan	Kata	Kekerapan
	raja bandar negeri	8 7 3	perkahwinan perdana kota alaf	4 4 3 3	Internet Web kewangan utama	8 6 2 2

[19] Korpus ini terdiri daripada teks hikayat dan teks tradisional lain.

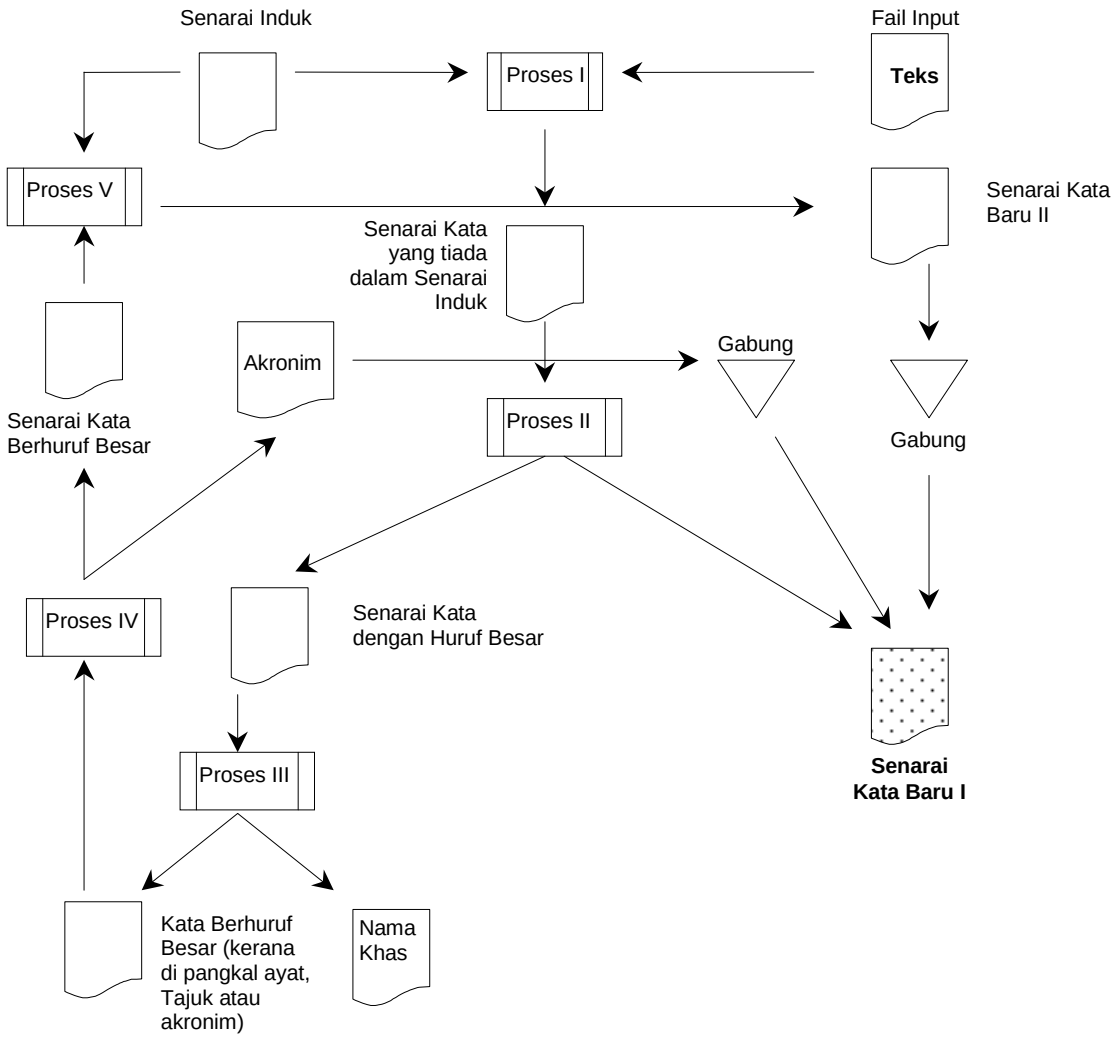
Secara umum dapat dilihat bahawa kekerapan ‘gerbang’ berubah daripada sangat kerap (dalam teks lama) kepada kurang kerap (dalam NSTP99) dan kemudian kembali menjadi agak kerap (NSTP00). Namun demikian kekerapan yang berbeza ini adalah disebabkan oleh kolokasi yang berbeza dan ini mencadangkan penggunaan ‘gerbang’ dengan maksud yang berlainan. Dalam teks lama, ‘gerbang’ jelas merujuk pada pintu yang konkrit manakala dalam NSTP99 ‘gerbang’ merujuk kepada yang nyata (pintu, perdana, kota, alaf[20]) dan yang metaforik (perkahwinan) tetapi dengan kekerapan yang jauh berkurangan. Bagaimanapun, dalam NSTP00 ‘gerbang’ bingkas semula tetapi bukan kerana ‘pintu’ mahupun ‘perkahwinan’ tetapi akibat perkembangan dalam dunia TMK, khususnya Web dan WAP.

Kekerapan relatif ini dapat dimanfaatkan untuk mengesan perubahan makna secara automatik. Namun, sistem sebenar hanya dapat dibangunkan setelah kajian tipologi korpus dan kajian leksikal lain dilakukan terlebih dahulu. Justeru itu, apa yang kami huraikan di sini sebagai sistem cadangan merupakan suatu sistem hipotetis yang akan dibangunkan kemudian.

3.2.1 Carta Alir Sistem

Sistem cadangan ini dinamakan Sistem PEKA (Pencari Kata) dan carta alir sistem adalah seperti dalam Rajah 2.

Rajah 2: Carta Alir Sistem PEKA



[20] Merujuk pada *Millennium Dome* di London.

3.2.2 Sistem PEKA

Sistem PEKA menggunakan senarai induk sebagai pangsi rujukan. Senarai induk ini disediakan berdasarkan kata entri, kata sub-entri dan frasa (kata majmuk) yang terdapat dalam KD3. Seperti yang ditunjukkan dalam Rajah 2 di atas, setiap teks baru akan dibandingkan dengan senarai induk untuk mendapatkan fail output. Proses pencarian kata baru dihuraikan selanjutnya di bawah;

Proses I:

Fail input (teks) baru dibandingkan dengan senarai induk untuk mendapatkan fail output. Proses ini dilakukan secara automatik.

Fail output terdiri daripada semua bentuk kata yang tiada dalam senarai induk, termasuk token berhuruf besar. Fail ini akan menjadi input kepada proses yang berikut.

Proses II:

Kata dengan huruf besar diasingkan daripada kata berhuruf kecil. Proses ini automatik dan menghasilkan dua fail. Fail pertama ialah senarai kata dengan huruf kecil (Senarai Kata Baru I) dan fail kedua ialah senarai kata berhuruf besar (fail ini menjadi input kepada proses selanjutnya).

Proses III:

Dalam langkah ini, kata nama khas diasingkan daripada akronim atau bentuk kata berhuruf besar yang terdapat pada pangkal ayat. Langkah ini boleh berlaku separa automatik atau pun secara insani. Fail dengan senarai nama khas diabaikan manakala yang selebihnya menjalani proses seterusnya.

Proses IV:

Proses IVa: Akronim diasingkan daripada kata dengan huruf besar dan fail akronim digabungkan dengan fail calon kata baru. Proses ini boleh separa automatik atau secara insani.

Proses IVb: Kata berhuruf besar ditukarkan menjadi huruf kecil secara automatik dan senarai ini diproses selanjutnya

Proses V:

Senarai kata daripada proses IVb dibandingkan dengan senarai induk untuk mendapatkan Senarai Kata Baru II yang kemudian digabungkan ke dalam fail Senarai Kata Baru I.

4.0 Kesimpulan

Kajian ini merupakan suatu usaha awal dalam pembangunan sebuah sistem yang dapat mengesan kata-kata baru daripada teks digital secara automatik.

Dengan adanya sistem ini setiap teks yang akan diinput ke dalam pangkalan data korpus akan diproses terlebih dahulu sebagai amalan rutin untuk memperoleh calon kata baru dan istilah sumber yang akan disalurkan kepada penyusun kamus dan penggubal istilah.

Rujukan

AVIATOR. Neologisms in Journalistic Text. http://radar.rdues.liv.ac.uk/newwds.html .

Bauer, L. 1983. *English Word-formation*. Cambridge University Press: Cambridge.

Church, K.W. dan Hanks, P. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16/1.

CORDON. Corpus-oriented Detection of Neologisms. http://www.ids.-mannheim.de/telri/seminar/jv-cordon.html .

Daille, B., Gaussier E. dan Langé, J-M. 1994. "Towards Automatic Extraction of Monolingual and Bilingual Terminology." *Proceedings of COLING 94*. Vol. 1: 515-521.

DBP 1994. *Kamus Dewan* Edisi Ketiga. Dewan Bahasa dan Pustaka: Kuala Lumpur.

DBP 1996. *Glosari Teknologi Maklumat*. Dewan Bahasa dan Pustaka: Kuala Lumpur.

Felber, H. 1984. *Terminology Manual*. Unesco/Infoterm: Paris.

Henry, D., Cooke, S., Buckley, P., Dumagan, J., Gill, G., Pastore, D. dan LaPorte, S., 1999. *The Emerging Digital Economy* II, U.S. Department of Commerce, Economics And Statistics Administration, Office of Policy Development, Washington, DC.

Janicijevic, T dan Walker, D. 1997. NeoloSearch: Automatic detection of neologisms in French Internet documents. http://www.qucis.queensu.ca/achallc97/papers/a009.html.

Long, T. 2000. *A Matter of (Wired News) Style.* http://www.wired.com/news/print/0,1294,39450,00.html .

Rademan, T. 1996. "Using online electronic newspapers in modern English-Language press corpora: Benefits and pitfalls." *ICAME Journal* 22:49-72. http://www.hd.uib.no/icame/ij22/

Roche, S. 1998a. Identifying Neologisms in General Language Corpora. (Kajian Latar yang terdapat di http://www.compapp.dcu.ie/Projects/1998/CL/sroche.cl4/study.html

Roche, S. 1998b. Corpus based English Neologism: Identifier Tool. (Manual Teknikal yang terdapat di http://www.compapp.dcu.ie/Projects/1998/CL/sroche.cl4/

Sager, J.C. 1990. *A Practical Course in Terminology Processing*. John Benjamins Publishing Company: Amsterdam/Philadelphia.

Smadja, F.A. dan McKeown, K.R. 1990. "Automatically Extracting and Representing Collocations for Language Generation." *Proceedings of the 28th Annual Meeting of ACL.*

Tulloch, S. 1992. *The Oxford Dictionary of New Words: A Popular Guide to Words in the News.* Oxford University Press: Oxford.

* Kertas Kerja yang dibentangkan dalam:
Persidangan Linguistik Asean 1
14-16 November 2000
Universiti Kebangsaan Malaysia