PENGEKSTRAKAN DATA DARI KAMUS DEWAN BERASASKAN PERATURAN

Pua Chee Wei

Dr Saidah Saad

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Mengenai pembelajaran yang sedia ada, Pembelajaran kemahiran menulis menjadi satu cabaran kepada pelajar kerana kekurangan kemahiran dalam aspek pemilihan kata ataupun istilah. Kebanyakan pelajar kurang memahami penggunaan sesuatu perkataan dalam konteks tertentu dan menimbulkan masalah di dalam pemilihan perkataan atau istilah yang tidak tepat dalam pembinaan ayat, seterusnya menjejaskan makna atau mesej yang ingin disampaikan. Selain itu, laman web Dewan Bahasa dan Pustaka (DBP) sedia ada yang bersifat statik dan dibina secara "hard coded" di dalam pengaturcaraannya, menimbulkan masalah bila data baru perlu dimasukkan atau pengemaskinian perlu dilakukan. Oleh itu, satu laman web kamus yang dinamik perlu dibina bagi membantu dan memudahkan pengguna memahami dan menggunakan kosa kata yang diperlukan serta mengenal perkataan baharu. Kaedahnya adalah melalui proses pengekstrakan bagi membaca data pada e-kamus yang tidak berstruktur dan dipecahkan ke dalam tag atau tandaan yang ditetapkan secara berstruktur mengikut nama tandaan yang ditetapkan. Hasil output adalah dalam bentuk data CSV dan XML, di mana ianya akan memudahkan proses pengemaskinian data kamus pada masa akan datang. E-Kamus yang mempunyai informasi yang lengkap amat diperlukan kerana kepelbagaian makna dan sifat yang sering mengelirukan dan dapat memberi kefahaman yang tepat kepada pengguna mengenai sesuatu istilah pada masa akan datang.

1. PENGENALAN

Perkataan "kamus" berasal daripada bahasa Arab, iaitu qamus. Kamus asalnya merupakan sejenis buku rujukan yang menerangkan makna sesuatu kata dan berfungsi untuk membantu pengguna mengenal perkataan baharu. Selain menerangkan maksud perkataan, sesetengah kamus turut mengandungi panduan sebutan, etimologi dan contoh penggunaannya. Kini, penggunaan komputer dalam proses pengajaran dan pembelajaran telah meluas dengan pesat. Melalui teknologi internet, ianya telah membantu pelajar mengakses laman web ilmiah dan informasi pendidikan terkini secara mudah dan pantas. Nolan dan Martin (1994), menyatakan bahawa pembelajaran dalam suasana baru seperti menggunakan internet di sekolah telah membawa banyak perubahan kepada pelajar.

Mengenai pembelajaran yang sedia ada, pembelajaran kemahiran menulis menjadi satu cabaran kepada pelajar kerana kekurangan kemahiran dalam aspek pemilihan kata ataupun istilah. Kebanyakan pelajar kurang memahami penggunaan sesuatu perkataan dalam konteks tertentu, oleh itu, pemilihan perkataan atau istilah yang tidak tepat dalam pembinaan ayat, akan menjejaskan makna atau mesej yang ingin disampaikan. Jadi, satu laman web kamus yang dapat memaparkan medan tentang sesuatu perkataan seperti Entri, Definisi Entri, Peribahasa Entri, Nama Saintifik Entri dan sebagainya perlu dibangunkan. Informasi yang lengkap amat diperlukan kerana kepelbagaian makna dan sifat yang sering mengelirukan.

2. PENYATAAN MASALAH

Penguasaan kosa kata Bahasa Melayu yang terhad merupakan masalah utama yang dihadapi oleh pelajar terutamanya bukan Melayu dalam kemahiran menulis. Mereka mempunyai idea yang baik untuk disampaikan melalui penulisan tetapi sering kali mengalami masalah memilih dan menggunakan kosa kata yang sesuai. Ini menyukarkan penyampaian dan pengaplikasian bentuk pembinaan ayat dalam bahasa yang lain. Menurut kajian kesilapan Bahasa Melayu dalam kalangan pelajar China Yunnan, didapati bahawa pelajar melakukan kesilapan dalam aspek pengimbuhan Bahasa Melayu kerana mereka keliru dan tidak begitu memahami penggunaan imbuhan dengan betul (Universiti et al. 2016). Sebagai akibat kurangnya pengetahuan terhadap penggunaan imbuhan dalam bahasa Melayu menjadikan para pelajar menghasilkan imbuhan yang tidak tepat, lantas menjejaskan makna yang ingin disampaikan.

Di pihak pembangun kamus (DBP) pula, masalah utama yang dihadapi adalah untuk mengekstrak maklumat yang telah dibukukan (tidak berstruktur) ke bentuk yang lebih berstruktur bagi memudahkan proses pengemaskinian data. Ini kerana proses penandaaan secara manual, memakan masa yang lama dan melibat kos pengambilan pekerja untuk proses pengemaskinian yang melibatkan ratusan ribu perkataan.

3. OBJEKTIF KAJIAN

Matlamat utama adalah untuk mengekstrak maklumat dari kamus Dewan secara automatik ke bentuk yang lebih berstruktur. Objektif projek ini adalah seperti berikut:

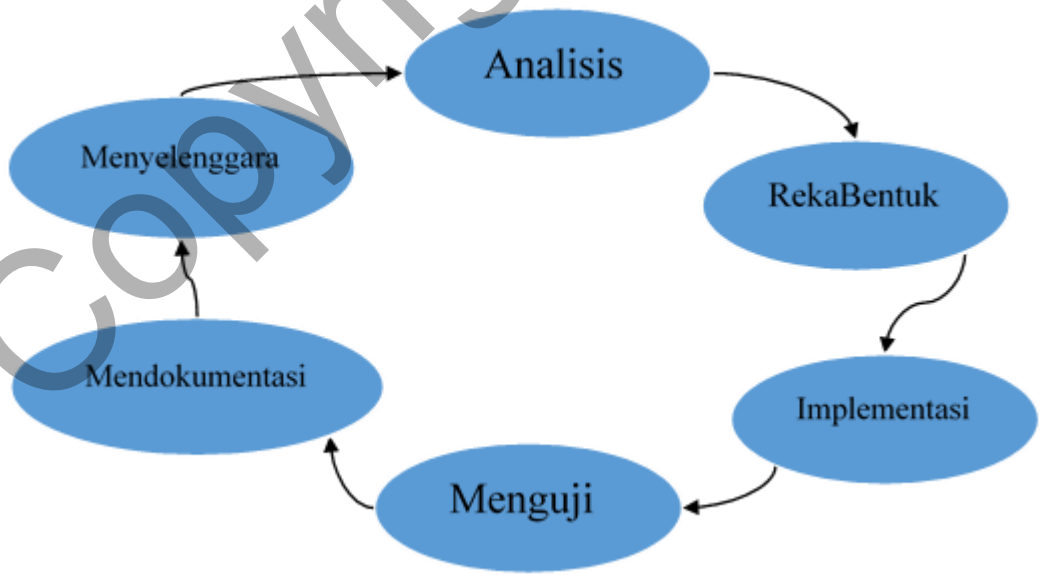
- Membina algoritma bagi proses penandaan (tagging) berasaskan Kamus Dewan.

- Membangun sistem ekamus yang dinamik bagi proses penyimpanan dan capaian perkataan.

4. METOD KAJIAN

Kaedah yang akan digunakan adalah Kitar Hayat Pembangunan Sistem (SDLC) di mana amalan AGILE (Agile) digunakan. Dalam model AGILE, "kegagalan pantas" adalah perkara yang baik. Pendekatan ini menghasilkan kitaran yang berterusan, masing-masing memaparkan perubahan kecil sehingga menyempurnakan sistem terakhir. Pada setiap peringkat, produk akan diuji. Model AGILE membantu pembangun mengenal pasti dan menangani isu-isu kecil mengenai projek sebelum mereka berubah menjadi masalah yang lebih besar, dan perlu melibatkan pihak berkepentingan bagi mendapat maklum balas mereka sepanjang proses pembangunan. Dengan metodologi ini, proses pembangunan kod dapat dipercepat dan mengurangkan kos walaupun terdapat kegagalan semasa implementasi.

Fasa-fasa yang dijalankan dalam kajian adalah seperti Rajah 1 di bawah:



Rajah 1: Fasa-fasa dalam Metodologi

4.1 Fasa Analisis Sistem (System Analysis):

Fasa ini dibahagikan kepada mengumpul data dan menganalisis data. Data akan dikumpul dari Kamus Dewan sedia ada. Kemudian, data akan dianalisis untuk mengenal pasti keperluan penggunaan tanda '*tag*' yang digunakan pada kamus. Berikut merupakan tandaan yang telah dikenalpasti dari pakar domain:

1. Entri
2. Entri Jawi
3. Fonetik Entri
4. Definisi Entri
5. Contoh Ayat Entri
6. Contoh Penggunaan Entri
7. Frasa Entri
8. Definisi Frasa Entri
9. Contoh Ayat Frasa Entri
10. Contoh Penggunaan Frasa Entri
11. Peribahasa Entri
12. Definisi Peribahasa Entri
13. Contoh Ayat Peribahasa Entri
14. SubEntri
15. Jawi SubEntri
16. Fonetik SubEntri
17. Definisi SubEntri
18. Contoh Ayat SubEntri
19. Contoh Penggunaan SubEntri
20. Peribahasa SubEntri
21. Definisi Peribahasa SubEntri
22. Contoh Ayat Peribahasa SubEntri
23. Bentuk Pasif
24. Definisi Bentuk Pasif
25. Contoh Penggunaan Bentuk Pasif
26. Etimologi
27. Nota Tambahan

Terdapat dua lagi tandaan yang dikenalpasti iaitu "Label Variasi Sosial" dimana variasi sosial pengguna bahasa dapat ditinjau dari status sosial dan pendidikan merupakan salah satu bentuk dari status sosial yang keberadaannya terlihat jelas di masyarakat. Satu lagi iaitu "Label Bahasa Dialek" di mana dialek ialah variasi daripada satu bahasa tertentu yang dituturkan oleh sekumpulan penutur dalam satu-satu masyarakat bahasa. Dialek mempunyai bentuk tertentu, dituturkan dalam kawasan tertentu dan berbeza daripada bentuk yang standard/ baku dari segi sebutan, tatabahasa, dan penggunaan kata-kata tertentu, tetapi perbezaannya tidaklah begitu besar untuk dianggap sebagai satu bahasa yang lain. Dialek selalunya digunakan dalam situasi formal atau rasmi, namun terdapat kecenderungan pengguna bahasa yang mencampurkan unsur dialek dalam penggunaan pada situasi formal. Dialek-dialek tempatan di Semenanjung Malaysia sebenarnya ialah

variasi daripada bahasa Melayu. Dalam satu-satu dialek tempatan di Malaysia, terdapat pula ideolek-ideolek, iaitu kelainan pengucapan yang disebabkan oleh perbezaan individu, dan kelainan stilistik yang terjadi sebagai akibat perbezaan konteks.

4.2 Fasa Rekabentuk Sistem (System Design):

Fasa ini melibatkan menyediakan pelbagai rekabentuk sistem dan memilih reka bentuk sistem yang paling sesuai. Antara muka sistem yang bersesuaian akan dipilih bagi memenuhi keperluan kajian.

4.3 Fasa Implementasi Sistem (System Implementation):

Fasa ini akan membincangkan aspek pembangunan dan implementasi sistem pembangunan berdasarkan ciri-ciri yang telah ditentukan. Kod-kod yang dihasilkan akan disemak sama ada mencapai keperluan atau tidak.

4.4 Fasa Menguji (Unit Testing):

Kod-kod akan dikumpul semula bagi tujuan pengujian. Sistem-sistem yang dibahagi akan digabung semula bagi menjadikan sebuah sistem yang lengkap. Jika sebahagian kod mempunyai masalah, penguji dapat mengesan dan membuat pembaikan semula.

4.5 Fasa Dokumentasi (Documentation):

Segala maklumat dan data yang telah digunakan akan didokumentasikan sebagai rujukan. Fasa dokumentasi ini penting kerana ia akan memastikan pembangunan sistem ini selalu menepati keperluan pengguna. Dokumentasi ini merekodkan maklumat tentang keperluan pelanggan, matlamat sistem dibangunkan, cara penggunaan dan maklumat pemaju system.

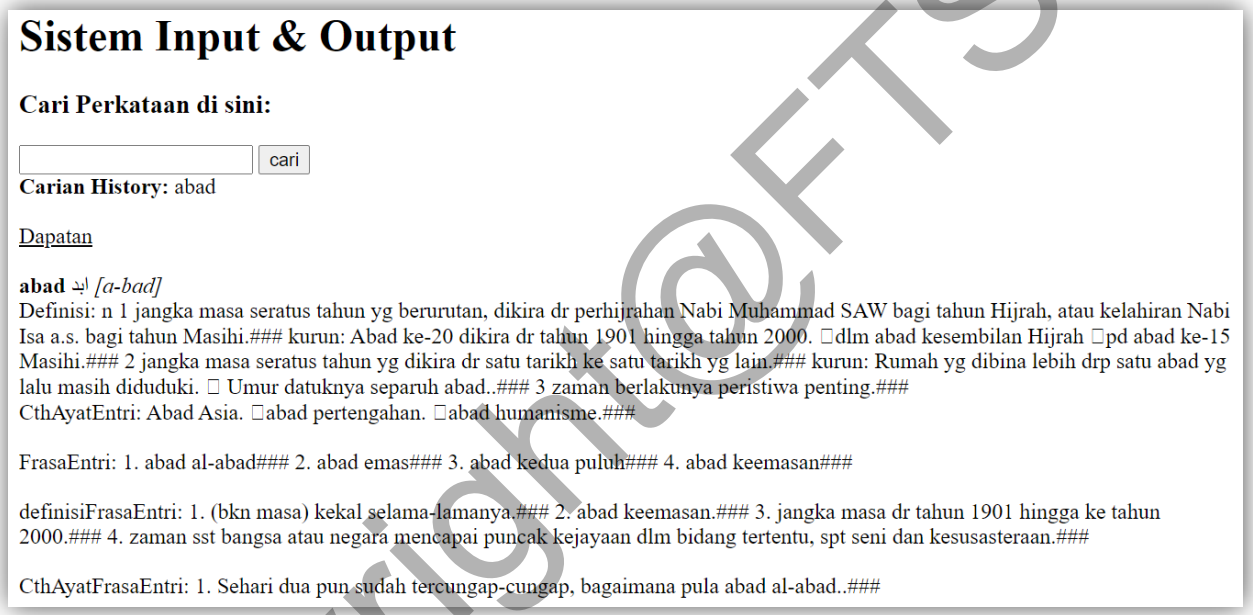
4.6 Fasa Menyelenggara (Maintenance):

Fasa ini merupakan proses terakhir dalam pembangunan sistem maklumat. Organisasi sistem maklumat dan penilaian berjadual akan dibuat.

5. HASIL KAJIAN

Bahagian ini menerangkan hasil dapatan selepas pembangunan algoritma dan sistem ini selesai. Antara perisian yang digunakan untuk pembangunan ialah *Eclipse IDE For Java Developers* untuk pembangunan algoritma, HTML dan PHP untuk pembangunan sistem.

Rajah 2 merupakan antara muka menu utama. Pengguna boleh memasukkan kata laluan yang ingin dicari dan keputusan akan ditunjukkan pada lampiran yang sama. Reka bentuk ini memudahkan pengguna untuk menukar kata laluan dengan lebih senang. Tanda "###" pada Rajah 2 di bawah merupakan tanda pecahan bertujuan untuk memisahkan kepelbagaian makna atau contoh yang ada pada entri.



Rajah 2: Antara Muka Menu Utama

Rajah 3 di bawah menunjukkan alogritma mengambil sumber data daripada fail Kamus. Selepas maklumat data dicapai, maka setiap garisan maklumat dibaca dan dibuangkan garisan kosong, dan selepasnya disimpan dalam ArrayList Pertama.

```
create ArrayList data2 to format data;
for i =0 to data1.size() do
        sentences ← read array content;
        if last sentences not equal to punctuation mark (.) do
                vocabulary ← sentences + one empty space;
        else
                data2 ← vocabulary;
                vocabulary ← empty;
END FOR LOOP
```

Rajah 3: Algoritma Penyusunan Data

Rajah 4 menunjukkan algoritma pengekstrakan data tahap pertama. Selepas maklumat data diformatkan, proses pengekstrakan data akan dijalankan. Namun begitu, oleh sebab maklumat data yang berbeza dan tidak sama, maka data akan dikategorikan kepada tiga bahagian, iaitu Bahagian A, B dan C. Rajah 4 merupakan algoritma untuk mengekstrakan maklumat data bagi Bahagian A. Bahagian A merupakan kumpulan data yang tidak mempunyai perkataan Jawi dan juga perkataan fonetik pada kosa katanya. Contoh data Bahagian A adalah seperti dalam Jadual 1 di bawah:

Jadual 1: Contoh Data Bahagian A

1.	a.n. sing atas nama.
2.	abu2 ابو adj kalah dlm permainan congkak, gasing dsb.
3.	A. sing Abdul. A. digunakan pd pangkal nama lelaki, biasanya bergabung dgn salah satu nama Allah: A. Rahman Hassan.

Contoh data kedua dalam Jadual 1 di atas merupakan data pengecualian yang mempunyai

```
//At here will divide data into three parts, flagCthAyat is close (By default).
for i = 0 to data2.size() do
        words ← read array content;
//Part One: Data without fonetik word and Person's name.
        If words do not contain square bracket && date of month (April) do
                split words into array;
                for each word in words do
                        if first word do
                                entry ← first word;
                        else if word is not alphabet && number do
                                jawi ← word;
//The rest of words would be the meaning.
                        else
                                if word contains semicolon (;) && flagCthAyat is open
                        do
                                        replace semicolon with punctuation;
                                        cthAyat ← add word + newline;
                                        definisi ← add new line (\n); flagCthAyat close;
                                else if flagCthAyat is open do
                                        cthAyat ← word;
                                else if word contains colon (;) do
                                        replace semicolon with punctuation;
                                        definisi ← add word + newline;
                                else if word contains colon (:) do
                                        replace semicolon with punctuation;
                                        definisi ← add word; flagCthAyat open;
                                else
                                        definisi ← word;
END FOR LOOP
save record into excel;
```

perkataan Jawi tetapi tiada fonetik perkataan. Jadi, dalam proses pengekstrakan data Bahagian A, perkataan Jawi juga akan disemak sama ada dalam data atau tidak.

Rajah 4: Algoritma Pengekstrakan Data Tahap Pertama

Rajah 5 merupakan algoritma untuk mengekstra maklumat data bagi Bahagian B. Bahagian B merupakan kumpulan data yang tidak mempunyai perkataan Jawi dan juga perkataan fonetik pada kosa katanya. Tambahan pula, ruangan Entri kumpulan data Bahagian B merupakan nama sejarah atau nama terkenal. Berdasarkan Rajah 5, ruangan Entri akan diinputkan perkataan sehingga sistem mengesan simbol kurungan, maka ruangan definisi akan diteruskan penginputan perkataan sehingga habis kerana nilai y telah menjadi 1.

```
//Part Two: Data contains person name and do not have fonetik word.
        else
                If word contains date of month (April) do
                        split words into array;
                        y ← 0;
                        for word in words do
                                if y!= 0 do
                                        definisi ← word + empty space;
                                if y==0 do
                                        entry ← word + empty space;
                                If word contains bracket do
                                        y ← y + 1;
                        END FOR LOOP
                        save record into excel;
```

Rajah 5: Algoritma Pengekstrakan Data Tahap Kedua

Contoh data Bahagian B adalah seperti kotak di bawah:

A. Samad Said (9 April 1935 –). Sasterawan yg menerima Anugerah Sastera Negara yg keempat pd tahun 1985, banyak menulis dlm genre novel, drama dan puisi yg berkaitan dgn kemanusiaan dan kemasyarakatan, dan antara novelnya yg ter--kenal ialah Salina dan Hujan Pagi, manakala kumpulan puisinya ialah Suara dari Dinding Dewan dan Rindu Ibu, dan dramanya ialah Wira Bukit.

Rajah 6 merupakan algoritma untuk mengekstra maklumat data bagi Bahagian C. Bahagian C merupakan kumpulan data yang mempunyai perkataan Jawi dan juga perkataan fonetik pada kosa katanya. Tambahan pula, maklumat data dalam Bahagian ini juga mempunyai medan-medan yang lain, contohnya Etimologi, Nota Tambahan, Peribahasa dan Entri berbentuk Pasif.

Berdasarkan Rajah 6 di bawah, SubCount digunakan untuk mengirakan kewujudan perkataan Sub Entri. Kewujudan perkataan Sub Entri sentiasa ditemani dengan simbol (~), jadi kiraan SubCount akan ditambah sekiranya sistem mengesan simbol (~). Selain itu, ruangan untuk menginputkan maklumat Entri, Entri Jawi dan juga Fonetik Entri akan dikosongkan sebelum menginputkan SubEntri, SubJawi atau SubFonetik yang baharu. Selain itu, setiap flag yang ada pada algoritma tahap ketiga ini akan digunakan sebagai pemeriksa dalam menyemak simbol tertentu dikesan atau tidak. Contohnya, simbol @ untuk medan Etimologi dan simbol $ untuk medan peribahasa.

Contoh data Bahagian C adalah seperti dalam Jadual 2:

Bil	Contoh Data	Jenis
1.	afdeling افـديلـيغ [af-de-ling] Id n bahagian atau sek-syen. **@ASAL: [drp Bld afdeling].**	Contoh Data bercampur dengan Etimologi.
2.	Afrika افـريـكا [af-ri-ka] n benua kedua terbesar di dunia yg meliputi negara Tunisia di utara dan negara Afrika Selatan di selatan dan di-kelilingi Laut Mediteranean di utara, Lautan Atlantik di barat dan selatan, serta Lautan Hindi di timur dan selatan. Pulaunya yg terbesar ialah Madagascar.; **/Keluasan: 30 370 000 km²; penduduk: 1.216 bilion (2016).**	Contoh Data bercampur dengan Nota Tambahan.
3.	mengacarakan مغاچاراکـن [me-nga-tsa-ra-kan||meng-a-tsa-ra-kan] **kt [psf: diacarakan, acarakan]** 1 menjadikan atau memasukkan (sst) sbg acara dlm majlis, upacara, persembahan dsb: Cadangan Jawatan-kuasa Persembahan utk mengacarakan segmen tsb dlm majlis penutupan persidangan itu telah diterima.	Contoh Data bercampur dengan Entri Bentuk Pasif.
4.	abuk2 ابـوق [a-buk] Id; ~ **mengabuk** مڠ ابـوق **[me-nga-buk||meng-a-buk]** ktt	Contoh Data bercampur

	mengaku barang orang lain sbg miliknya.	dengan Sub Entri.

Jadual 2: Contoh Data Bahagian C

```
//Part Three: Data with sub entry and fonetik word and jawi word and others field of entry.
//flagEntri, flagJawa, flagFonetik, flagDefinisiFrasa, flagCthAyatFrasa, flagDefinisiPrb, flagCthAyatPrb, flagCthAyat, flagCthPenggunaan, flagCthPenggunaanFrasa, flagCthPenggunaanSubEntri, flagCthPenggunaanBentukPasif, flagDefinisiBentukPasif, flagNotaTambahan dan flagEtimologi is close (By Default).
    else:
        split words into array;
        count ← 0; subcount ← 0; //subcount is to alert subentry.
        for word in words do
            if 0<=count<=2 && word match pattern alphabet do
                if subCount==0 do
                    entry ← word + empty space;
                    flagEntri open;
                else
                    SubEntri ← word + empty space;
                    Flag Entri open;
            if 1<=count<=3 && word match pattern Jawi do
                if subCount==0 do
                    jawi ← word + empty space;
                    flagJawi open;
                else
                    JawiSubEntri ← word + empty space;
                    Flag Jawi open;
            if 2<=count<=7 && word match pattern fonetik do
                if subCount==0 do
                    fonetik ← word + empty space;
                    flagFonetik open;
                else
                    fonetikSubEntri ← word + empty space;
                    FlagFonetik open;
```

Rajah 6: Algoritma Pengekstrakan Data Tahap Ketiga

6. KESIMPULAN

Kesimpulannya, sistem kamus ini telah dibangunkan berteraskan objektif kajian, keperluan pengguna dan reka bentuk yang telah dirangka. Selain itu, Hasil pengujian mendapati bahawa proses pengekstrakan data masih perlu dibuat pembaikan kerana terdapat simbol-simbol dan maklumat data yang tidak berkenaan dalam proses pengekstrakan data ini. Secara keseluruhannya, sistem ini dengan harapan mampu memanfaatkan pelajar dan secara tidak langsung dapat memanfaatkan pelajar yang lemah dalam Bahasa Melayu. Dengan ini, kesukaran yang dihadapi oleh pelajar dalam penulisan atau komunikasi dapat diatasi.

7. RUJUKAN

Normawati Binti Abd Rahman. 2006. Kamus Elektronik Bahasa Melayu – Bahasa Inggeris: Pendekatan Carian Kehampiran (Proximity Search). Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.

Universiti, P., Yunnan, K. & Akademi, D. I. 2016. Kesilapan bahasa Melayu dalam kalangan pelajar pelajar universiti kebangsaan Yunnan di akademi Pengajian Melayu. *Jurnal Melayu* 15(2): 196–209.

Azean, N., Atan, B., Bin, Y. & Pendidikan, J. F. 2010. *Pembangunan E-Kamus Bagi Javascript Berdasarkan Rekabentuk Teori Beban Kognitif*.