Manuscript Submitted	20.4.2021
Accepted	16.6.2021
Published	30.6.2021

Pencantas Perkataan Jawi Lama (Old Jawi Stemmer) dalam Bahasa Melayu berasaskan Petua

Che Wan Shamsul Bahri C.W.Ahmad
Fakulti Sains dan Teknologi Maklumat
Kolej Universiti Islam Antarabangsa Selangor (KUIS), Malaysia
cwshamsul@kuis.edu.my

Khairuddin Omar[1], Mohammad Faidzul Nasruddin[2] & Mohd Zamri Murah[3]
[1, 2,3] Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia(UKM), Malaysia
{[1]ko, [2]mfn, [3]zamri }@ukm.edu.my

Abstract

The word stemming works to remove the affix of a word by generating the base word for that word. Stemming is widely used in natural language processing (NLP) such as machine transliteration, machine translation and document access. With the use of stemming, the dictionary size can be reduced because words in the same morphology do not need to be entered repeatedly. Instead the words are included in the same group. There are two types of script writing in Malay language, either use the Latin spelling system or Jawi spelling system. Many studies on the Malay word grapple more focused on Latin spelling system compared with Jawi spelling system. This paper proposes Malay stemmer for old Jawi characters by using a set of rules in old Jawi (a set of rules used to constrain various forms of words derived from old Jawi). There are 187 rules was developed for the stemmer, called as PEJAL. There are 2500 derived Jawi words consisting of prefixes, suffixes, suffixes, insertions and tested using this stemmer. The experimental results showed that 88.5% of the Jawi words were successfully stemmed correctly.

***Keywords**:word stemmer, Jawi, information access, transliteration*

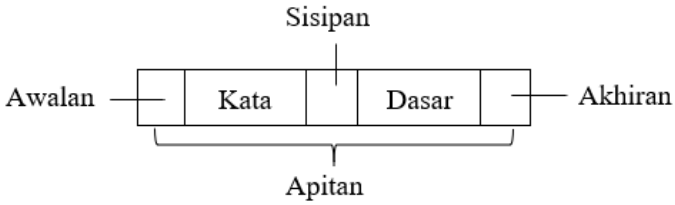
Abstrak

Pencantas perkataan berfungsi untuk membuang imbuhan sesuatu perkataan dengan menghasilkan kata dasar bagi perkataan tersebut. Cantasan banyak digunakan dalam bidang pemprosesan bahasa tabii(PBT) seperti transliterasi mesin, penterjemahan mesin dan capaian dokumen. Dengan penggunaan cantasan, saiz kamus dapat dikurangkan kerana perkataan dalam morfologi yang sama tidak perlu dimasukkan berulang kali. Sebaliknya perkataan-perkataan tersebut dimasukkan dalam kumpulan yang sama. Dalam Bahasa Melayu, terdapat dua jenis skrip penulisan sama ada menggunakan iaitu sistem ejaan Rumi atau sistem ejaan Jawi. Banyak kajian berkaitan pencantas perkataan Bahasa Melayu lebih tertumpu kepada sistem ejaan Rumi berbanding dengan sistem ejaan Jawi. Kertas ini mencadangkan pencantas perkataan Melayu bagi aksara Jawi dengan menggunakan satu set peraturan dalam Jawi lama (satu set peraturan yang digunakan untuk mengekang pelbagai bentuk perkataan terbitan Jawi lama). Terdapat 187 petua yang digunakan untuk pencantas Jawi ini yang dipanggil sebagai PEJAL. Terdapat 2500 perkataan Jawi terbitan terdiri daripada awalan, apitan, akhiran, sisipan dan diuji menggunakan pencantas ini. Hasil uji kaji menunjukkan 88.5% daripada perkataan Jawi berjaya dicantas dengan betul.

***Kata kunci:** pencantas perkataan, Jawi, capaian maklumat, transliterasi*

1. Pendahuluan

Setiap perkataan boleh dikategorikan sama ada kata akar atau kata terbitan. Kata akar ialah kata yang belum diberi apa-apa imbuhan. Manakala kata dasar pula ialah kata akar atau kata terbitan yang dapat diberi imbuhan. Apa yang dapat disimpulkan ialah kedua-dua kata tersebut boleh menjadi kata terbitan. Kata terbitan ialah bentuk kata yang mengandungi kata dasar yang mendapat atau menerima imbuhan. Dengar erti kata lain, kata terbitan adalah perkataan yang terjadi atau terdiri daripada kata akar dan imbuhan yang disatukan sebagaimana yang ditunjukkan Rajah 1.



RAJAH 1. Kedudukan setiap kumpulan imbuhan apabila dicantumkan dengan kata dasar

Pencantas perkataan berfungsi untuk membuang imbuhan sesuatu kata terbitan dengan menghasilkan kata dasar bagi perkataan tersebut. Ia banyak digunakan dalam pengkelasan teks, transliterasi, capaian dokumen, dan penterjemahan mesin. Dalam bidang pemproses bahasa tabii (PBT), algoritma pencantas perkataan digunakan untuk mencari kata dasar sesuatu perkataan daripada perkataan yang mengandungi imbuhan atau kata terbitan (Melucci, 2008). Pencantas dapat digunakan bukan hanya untuk mengindeks dan mengurangkan saiz perbendaharaan kata tetapi juga untuk meningkatkan prestasi pencarian maklumat (Muhamad Taufik Abdullah et al., 2009). Pencantas sangat penting dalam kebanyakan bahasa dan begitu juga dalam bahasa Melayu. Bahasa Melayu mempunyai dua jenis tulisan yang berbeza iaitu tulisan Jawi dan tulisan Rumi. Tulisan Jawi diadaptasi daripada aksara Arab dengan penambahan enam aksara baharu untuk disesuaikan penggunaannya dalam bahasa Melayu. Manakala tulisan Rumi pula menggunakan abjad Roman atau Latin. Bahasa Melayu dituturkan di banyak negara nusantara seperti Indonesia, Malaysia, Singapura dan Brunei. Tulisan Jawi digunakan dalam buku, manuskrip, surat antara raja dan lain-lain (Nasrudin et al., 2008).

Antara pengkaji utama dalam cantasan perkataan Bahasa Melayu adalah Asim (1993), Fatimah Ahmad, (1995), Idris dan Syed Mustapha, (2001), Muhamad Taufik et al. (2009), Tai et al. (2000) dan (Suliana Sulaiman, 2013). Sebilangan besar kajian menggunakan peraturan morfologi dalam algoritma cantasan mereka, namun (Tai et al., 2000) menggunakan kaedah gabungan N-gram. Idris dan Syed Mustapha (2001) mengenal pasti perkataan untuk dicantas dengan penggunaan kamus tambahan, yang dikenali sebagai kamus tempatan, yang dapat mengurangkan ralat pencantas. Fatimah Ahmad (1995) dan Muhamad Taufik Abdullah et al. (2009) juga menggunakan kamus kata dasar untuk memastikan bahawa kata dasarnya betul. Walau bagaimanapun, mereka menggunakan kamus kata dasar dan hanya boleh digunakan untuk perkataan yang berasal dari bahasa Melayu yang ditulis dalam tulisan Rumi sahaja. Kertas ini mencadangkan algoritma cantasan Jawi yang boleh digunakan untuk mendapatkan kata dasar sesuatu perkataan Melayu. Kertas ini disusun menjadi enam bahagian. Bahagian 2 membincangkan kajian-kajian lepas yang berkaitan cantasan perkataan Bahasa Melayu. Bahagian 3 membincangkan kajian berkaitan petua ejaan Bahasa Melayu untuk perkataan Jawi lama. Bahagian 4 menerangkan petua cantasan perkataan Jawi lama. Bahagian 5 membincangkan eksperimen dan hasilnya. Akhirnya, bahagian 6 membentangkan kesimpulan kajian.

2. Kajian-Kajian Lepas

Kajian-kajian lepas yang berkaitan cantasan perkataan dalam Bahasa Melayu kebanyakannya hanya memfokuskan cantasan untuk perkataan Rumi sahaja. Contoh kajian tersebut adalah seperti kajian Tai et al. (2000), Idris (2001), Leong et al. (2012), Muhamad Taufik et al. (2009) dan Fadzli et al. (2012).

Kajian yang melibatkan cantasan perkataan dalam domain Jawi hanya dijalankan oleh Suliana (2013). Suliana telah membangunkan dua set petua iaitu **petua pengesanan kesalahan ejaan Jawi** (SEDR) dan **petua nyah-imbuhan Jawi** dalam menghasilkan algoritma pencantas perkataan Jawi. Petua SEDR berfungsi untuk menyemak pola ejaan Jawi yang dicantas supaya tidak berlaku ralat, manakala petua nyah-imbuhan pula berfungsi untuk membuang atau mencantas imbuhan yang ada pada sesuatu perkataaan Jawi. Suliana (2013) mencadangkan 197 petua cantasan untuk perkataan Jawi. Untuk menyemak sama ada hasil cantasan tersebut betul atau sebaliknya, Suliana menggunakan SEDR. SEDR yang dihasilkan mempunyai ketepatan sehingga 97.8% yang diuji terhadap 3018 perkataan Jawi, hanya 67 patah perkataan sahaja mempunyai ralat.

Yonhendri et al. (2009) pula menggunakan petua cantasan untuk mencantas perkataan Rumi bagi dipadankan kata dasar Jawi. Yonhendri menggunakan algoritma cantasan berasaskan suku kata untuk transliterasi mesin Rumi-Jawi.

Hasil cantasan boleh dibahagikan kepada dua kategori utama iaitu tepat atau ralat. Tepat adalah apabila hasil cantasan sama seperti kata dasar perkataan tersebut. Manakala ralat pula apabila hasil cantasan bukannya perkataan dasar bagi perkataan tersebut. Ralat dibahagikan kepada beberapa jenis iaitu ralat terkurang cantas, ralat terlebih cantas, ralat tidak berubah dan ralat lain-lain. Ralat lain-lain adalah seperti hasil cantasan tidak termasuk dalam mana-mana kategor ralat yang lain. Contohnya perkataan دأبايكن (diabaikan) sepatutnya menghasilkan أباي (abai) tetapi menghasilkan output أابي yang silap kerana disebabkan petua silap yang dihasilkan. Jadual 1 di bawah menunjukkan contoh hasil cantasan.

JADUAL 1. Contoh Hasil Cantasan

Hasil Cantasan	Sebelum cantas	Selepas Cantas
Terkurang cantas	ڤمرداڬغن (perdagangan)	رداڬغ
	بوروانڽ (buruannya)	بوروان
Terlebih cantas	ڤڠوات (penguat)	وات
	مڠرنياكن (mengurniakan)	رنيا
Cantasan tepat	منجالينكن (menjalinkan)	جالين
	منجالنكن (menjalankan)	جالن

3. Petua Ejaan Jawi Lama

Dalam ejaan Jawi lama terdapat dua perkataan dieja secara berangkai atau bersambung. Ia tujuan untuk memendekkan atau mempercepatkan penulisan. Di antara perkataan yang kerap digabungkan adalah perkataan يڬ (yang). Contohnya perkataan 'barang yang' (بارڬيڬ) dan 'yang amat' (يڬامت). Antara contoh lain ialah perkataan 'barangsiapa' (بارڬسياڤ) yang mana ditulis dalam dua perkataan yang berasingan dalam Jawi baru iaitu بارڬ سياڤا. Contohnya sebagaimana pada Jadual 2 di bawah :

JADUAL 2. Ejaan Jawi Lama dan Baru

Rumi	***Jawi baru***	***Jawi lama***
yang tersebut	يڬ ترسبوت	يڬترسبت
yang dipertua	يڬ دڤرتوا	يڬدڤرتوا
yang berkuasa	يڬ بركواسا	يڬبركواس
mereka itu	مريك ايت	مريكئيت

Dalam ejaan Jawi lama terdapat beberapa hukum yang perlu dipatuhi atau dijadikan panduan dalam penulisan.

a. Hukum sisipan alif

Hukum sisipan alif digunakan dalam Jawi lama apabila kata dasar mempunyai imbuhan akhiran sebagaimana contoh pada Jadual 3 di bawah.

JADUAL 3. Kata terbitan Jawi lama dengan sisipan alif

Kata dasar	***Kata terbitan (dengan sisipan alif) dalam Jawi lama***
بيلغ	بيلاغن (bilangan)
اوچڤ	اوچاڤن (ucapan)
ريحت	كريحاتن (kerehatan)

Nota: Apabila mencantas kata terbitan ini, huruf *alif* perlu digugurkan.

b. Hukum deranglu

Hukum deranglu ialah semua bunyi ke atas bagi dua suku kata terakhir sesuatu perkataan apabila berakhirkan bunyi da-ra-nga-la-wa (درغلو) memakai huruf saksi *alif*.

ڤدادا، بيچارا، مڠغا، ڤنجالا، ترتاوا
pedada, bicara, menganga, penjala, tertawa

Hukum luar deranglu (HLD) ialah kaedah mengeja perkataan yang mengandungi dua suku kata akhir berbentuk a-a, akan dieja tanpa menggunakan huruf saksi *alif* pada suku kata terakhir. Misalnya perkataan kata dieja (كات) *kaf- alif-ta* saja tanpa huruf saksi *alif* pada suku kata terakhir.

ساي، منجاج، نراچ، ببراڤ، مناڽ
saya, menjaja, neraca, beberapa, menanya

Nota: Apabila mencantas kata terbitan yang mengandungi perkataan yang tertakluk kepada *hukum deranglu*, huruf *alif* perlu dikekalkan pada ahkir perkataan, manakala tiada huruf *alif* diakhir perkataan bagi hukum luar deranglu. Kaedah yang mudah untuk ingat adalah selepas berakhir huruf (بتاڤ سكنڽ چيك جڬكوت) tidak perlu dimasukkan *alif*.

c. Hukum ka-ga

Hukum *ka-ga* ialah semua bunyi *ka* dan *ga* di akhir sesuatu perkataan tanpa mengira apa bunyi di depannya hanya memakai huruf *ka* dan *ga*, tanpa huruf saksi *alif* di akhir.

بلاك، جناك، ڤوساك، درماك، ڤوجڠك، تتڠك
belaka, jenaka, pusaka, dermaga, pujangga, tetangga

Nota: Hukum *ka-ga* adalah sama sebagaimana hukum luar deranglu.

d. Hukum ra-ma

Hukum *ra-ma* pula ialah peraturan yang menentukan semua perkataan yang mempunyai dua suku kata ra-ma di akhir hendaklah memakai huruf saksi *alif*. Hukum ini adalah pengecualian daripada hukum deranglu di atas.

اسراما، انيكاراما، ايراما، دراما، چڠكراما، ڤانوراما
asrama, anekarama, irama, drama, cengkerama, panorama

Nota: Apabila mencantas kata terbitan yang mengandungi perkataan dua suku kata *ra-ma* , huruf *alif* perlu dikekalkan pada kedua-dua suku kata tersebut.

e. Hukum ha - pertama

Hukum *ha* ialah peraturan yang melibatkan huruf *ha* dalam ejaan yang berbunyi ke atas. Hukum ini terbahagi kepada dua, iaitu *ha* pertama dan *ha* kedua. (*a- ha*)

Apabila bunyi *ha* (ke atas terbuka) terdapat pada suku kata pertama sesuatu perkataan yang mengandungi lebih daripada dua suku kata (kata dasar), suku kata ha tersebut tidak memerlukan huruf saksi *alif*.

هلوبا، هروان، هريماو، هلوان
haloba, haruan, harimau, haluan

Namun ada perkataan yang tidak terlibat dengan hukum ini iaitu :

هالا، هاري، هادڤن، هالاجو
hala, hari, hadapan, halaju

Nota: Tiada masalah dalam cantasan, namun perlu diperhalusi semasa transliterasi kerana hukum ha-pertama tidak menggunakan huruf saksi *alif* pada suku kata pertama.

f. Hukum ha - kedua

Hukum *ha* kedua ialah suku kata bunyi *ha* terletak pada suku kata kedua dalam sesuatu perkataan, sama ada perkataan itu sendiri daripada dua suku kata atau lebih. Perkataan itu, didahului oleh suku kata terbuka ke atas *a.ha* seperti dalam perkataan .

مها، بهارو، سهاج، ڤهالا
maha, baharu, sahaja, pahala

مهاديوا، مهاڬورو، مهاسيسوا، مهاميرو
mahadewa, mahaguru, mahasiswa, mahameru

بهاس، بهاڬيا، بهاوا، چهاي، بهاڬي، بهاس، ڬهارو
bahasa, bahagia, bahawa, cahaya, bahagi, bahasa, gaharu

Nota: Tiada masalah dalam cantasan, namun perlu diperhalusi semasa transliterasi kerana hukum ha-kedua menggunakan huruf saksi *alif* pada suku kata pertama dan kedua.

4. Algoritma Petua Cantasan Jawi Lama (PEJAL)

Petua cantasan Jawi lama (PEJAL) sebenarnya adalah penambahbaikkan daripada petua Suliana et. al (2013) yang hanya khusus untuk Jawi moden sahaja. Rajah 2 adalah algoritma petua cantasan kata terbitan bagi Jawi lama.

```
Input : TokenKata
Panjang_perkataan = panjang_TokenKata

TokenKata == AksaraJawi?
Jika YA
     Lakukan pra pemprosesan,  Fungsi PraPemproses
     Kemudian ke A.
sebaliknya
     Cetak "Bukan aksara Jawi"

A. Tentukan Panjang_perkataan , L
Jika L >=5

     Semak pola
     Fungsi Petua Imbuhan Apitan
     Semak petua apitan,  jika sepadan
          cantas imbuhan apitan
          kembalikan kata akar


     Fungsi Petua Imbuhan Awalan
     Semak petua imbuhan awalan, jika sepadan
          cantas imbuhan awalan
          kembalikan kata akar


     Fungsi Petua Imbuhan Akhiran
     Semak petua imbuhan akhiran, jika sepadan
          cantas imbuhan awalan
          kembalikan kata akar

     Cetak " Tiada dalam Petua"
sebaliknya
     Cetak " Tidak dicantas"

     Semak kata akar dan buat jajaran jika perlu

Fungsi PraPemproses

     Buang kata henti
     Buang tanda kata ganda (٢)
     Buang diakritik
     Buang perkataan berangkai (yang)
     Buang kata akhir (nya)
     Buang Partikel (lah, kah)
```

RAJAH 2. Kod Pseudo bagi Algoritma Petua cantasan Jawi lama (PEJAL)

Perisian yang digunakan dalam pembangunan ini adalah PHP sebagai bahasa pengaturcaraan utama untuk pembangunan sistem, manakala MySQL digunakan untuk pembangunan pangkalan data dan Laragon. Laragon adalah persekitaran pembangunan yang universal untuk PHP, Node.js, Python, Java, Go dan Ruby yang portable, cepat, ringan, dan mudah dipakai. Laragon biasanya digunakan sebagai pengganti XAMPP.

5. Hasil Ujikaji

Ujikaji dijalankan terhadap beberapa set data daripada korpus Jawi lama iaitu Majalah Qalam (MQ) terbitan tahun 1951 – 1960 dan Hikayat Merong Mahawangsa (HMM) tahun 1916. Jadual 4 di bawah menunjukkan hasil ujikaji keberkesanan algoritma petua cantasan terhadap MQ.

JADUAL 4. Ujikaji petua cantasan terhadap MQ

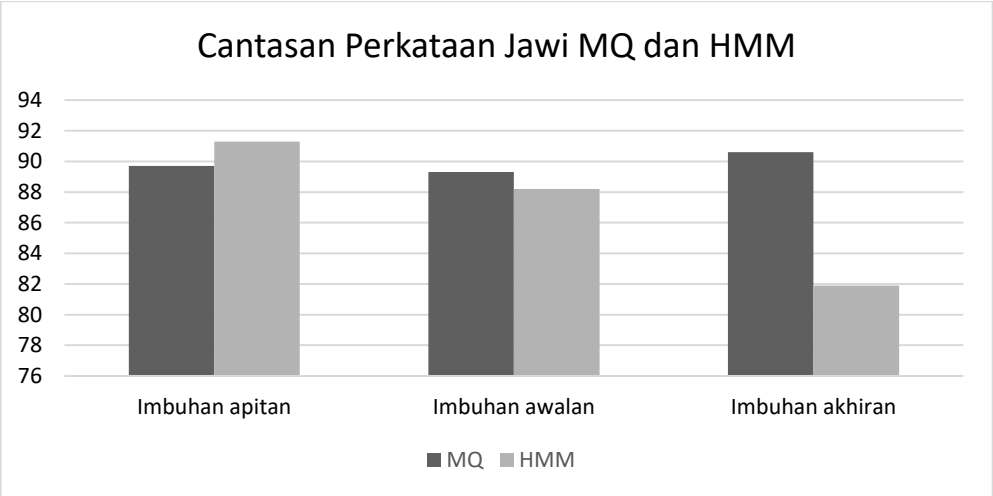
Jenis Perkataan	Purata Ketepatan
Perkataan imbuhan apitan	89.7
Perkataan imbuhan awalan	89.3
Perkataan imbuhan akhiran	90.6
Jumlah keseluruhan	**89.9%**

Jadual 5 di bawah menunjukkan hasil ujikaji petua cantasan terhadap HMM. Peratusan yang paling tinggi adalah bagi perkataan yang mempunyai imbuhan apitan berbanding dengan perkataan imbuhan awalan dan akhiran.

JADUAL 5. Ujikaji petua cantasan terhadap HMM

Jenis Perkataan	Peratusan Ketepatan
Perkataan imbuhan apitan	91.3
Perkataan imbuhan awalan	88.2
Perkataan imbuhan akhiran	81.9
Jumlah keseluruhan	**87.1%**

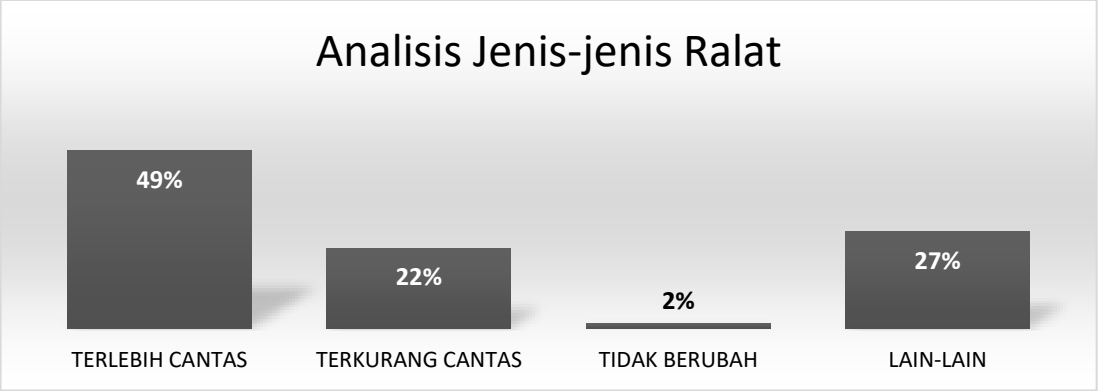
Rajah 3 di bawah menunjukkan perbandingan ketepatan cantasan bagi teks MQ dan HMM. Purata ketepatan cantasan keseluruhannya adalah 88.5%.



RAJAH 3. Prestasi petua cantasan Jawi lama

Analisis Ralat

Ralat cantasan yang berlaku setelah petua cantasan Jawi lama diuji adalah seperti Rajah 4 di bawah. Kebanyakan ralat yang berlaku adalah ralat terlebih cantas. Sebanyak 49 peratus ralat yang berlaku adalah termasuk dalam kategori ini. Ralat ini berlaku apabila kata terbitan yang dicantas menghasilkan perkataan yang terlebih cantas daripada kata dasar. Contohnya, perkataan (فندغن) dicantas menjadi (دغ), sepatutnya menjadi (فندغ).



RAJAH 4. Analisis jenis-jenis ralat yang berlaku berdasarkan ujikaji

Ralat terkurang cantas pula adalah berlawanan dengan ralat terlebih cantas. Ralat jenis ini biasanya perkataan masih lagi tidak dicantas dengan sempurna. Dalam ujikaji ini, hanya 22 peratus ralat yang berlaku dalam kategori ini. Manakala ralat dalam kategori lain-lain pula adalah sebanyak 27 peratus.

Hanya dua peratus ralat yang berlaku dalam kategori tidak berubah. Ralat tidak berubah berlaku apabila petua cantasan tidak menemui mana-mana perkataan yang sepadan untuk dicantas maka sistem tersebut akan mengekalkan perkataan tersebut tanpa melalui cantasan.

6. Kesimpulan

Algorima cantasan yang dihasilkan untuk Bahasa Melayu bagi perkataan Jawi lama berfungsi dengan baik sehingga mencapai ketepatan sehingga 88.5%. Ketepatan yang paling tinggi adalah bagi perkataan yang mempunyai imbuhan apitan dan diikuti dengan imbuhan awalan. Jika petua cantasan ini diperhalusi dan ditambahbaik lagi pada masa akan datang, hasil cantasan akan menjadi dengan lebih baik lagi.

Penghargaan

Penghargaan kepada pihak Kolej Universiti Islam Antarabangsa Selangor (KUIS) kerana tajaan Skim Latihan Kakitangan KUIS (SLAK).

Rujukan

Asim, O. (1993). *Pengakar Perkataan Melayu dan Sistem Capaian Dokumen*. Universiti Kebangsaan Malaysia, Bangi.

Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., & M Satar, S. D. (2012). Simple Rules Malay Stemmer. *The International Conference on Informatics and Applications (ICIA2012), January 2012*, 28–35. http://sdiwc.net/digital-library/download.php?id=00000187.pdf

Fatimah Dato Ahmad. (1995). *Sistem capaian dokumen bahasa melayu: satu pendekatan eksperimen & analisis*. Universiti Kebangsaan Malaysia.

Idris, N., & Syed Mustapha, S. M. F. D. (2001). *Stemming For Term Conflation In Malay Texts*. *September 2016*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.3762

Leong, L. C., Basri, S., & Alfred, R. (2012). Enhancing Malay stemming algorithm with background knowledge. *Pacific Rim International Conference on Artificial Intelligence*, 753–758.

Melucci, M. (2008). A Basis for Information Retrieval in Context. *ACM Transaction on Informa-Tion System (TOIS)*, *26*, 14–41.

Muhamad Taufik Abdullah, Fatimah Ahmad, Ramlan Mahmod, & Sembok, T. M. T. (2009). Rules frequency order stemmer for malay language. *IJCSNS International Journal of Computer Science and Network Security*, *9*(2), 433–438. http://paper.ijcsns.org/07_book/200902/20090258.pdf

Nasrudin, M. F., Omar, K., Zakaria, M. S., & Yeun, L. C. (2008). Handwritten cursive Jawi character recognition: A survey. *2008 Fifth International Conference on Computer Graphics, Imaging and Visualisation*, 247–256.

Suliana Sulaiman. (2013). *Pencantas Perkataan Melayu Untuk Aksara Jawi Berasaskan Petua*. Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaan Malaysia.

Tai, S. Y., Ong, C. S., & Abullah, N. A. (2000). On designing an automated Malaysian stemmer for the Malay language (poster session). *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages.*

Tai, S. Y., Ong, C. S., & Abullah, N. A. (2000). On designing an automated Malaysian stemmer for the Malay language. *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, 207–208.

Yonhendri, Heryanto, A., Omar, K., & Nasrudin, M. F. (2009). *Transliteration Engine Rumi to Jawi (TERUJA)*.