PENCANTAS PERKATAAN MELAYU UNTUK AKSARA JAWI BERASASKAN PETUA

SULIANA SULAIMAN

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT

2013

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

1 Ogos 2013 SULIANA SULAIMAN

P 47840

PENGHARGAAN

Dengan nama Allah yang Maha Pengasih lagi Maha Penyayang.

Segala puji bagi Allah S.W.T tuhan sekalian alam. Selawat dan salam ke atas junjungan besar Nabi Mujammad S.A.W. Alhamdulillah, setinggi kesykuran ke hadrat Allah S.W.T kerana limpah kurnia dan rahmatNya, dapat saya meyempurnakan tesis ini.

Ucapan penghargaan dan jutaan terima kasih ditujukan kepada barisan peyelia saya iaitu, Prof. Khairuddin Omar, Dr Nazlia Omar dan Encik Zamri Murah yang banyak memberi tunjuk ajar, saranan dan bantuan dari awal sehingga ke peringkat akhir tesis ini disiapkan. Penghargaan ini juga ditujukan buat sarjana tamu Tuan Haji Hamdan Abdul Rahman yang banyak membantu dan mencurahkan ilmu dalam menjayakan tesis ini. Semoga segala bentuk bantuan yang dicurahkan akan diberi ganjaran yang berlipat kali ganda oleh Allah S.W.T.

Penghargaan yang tidak terhingga juga diberikan kepada suami tercinta Mohd Nazir yang telah banyak mendoakan kejayaan dan memberikan dorongan serta galakan buat diri ini. Juga buat anakanda Harith Hakimi yang selalu memahami keadaan mama sebagai seorang pelajar. Tidak lupa juga buat emak dan ayah yang sentiasa mendoakan kejayaan saya. Semoga Allah sentiasa merahmati kalian di dunia dan di akhirat.

Selain itu sekalung penghargaan ditujukan buat Kementerian Pengajian Tinggi dan juga Universiti Pendidikan Sultan Idris kerana telah membiayai segala kos pengajian sepanjang menyempurnakan tesis ini melalui Skim Latihan Bumiputera (SLAB).

Akhir sekali, ucapan terima kasih buat rakan-rakan dan pensyarah Kumpulan Penyelidikan Pengecaman Pola atas segala bantuan dan kerjasama yang diberikan.

ABSTRAK

Pencantas perkataan merupakan proses membuang imbuhan pada perkataan dan menghasilkan perkataan tercantas ataupun kata dasar. Pencantas perkataan boleh digunakan dalam capaian dokumen, transliterasi, pengkelasan teks dan penterjemahan mesin. Pencantas perkataan yang dihasilkan dalam kajian yang lepas bagi Bahasa Melayu lebih tertumpu kepada tulisan Rumi. Set petua yang dihasilkan untuk mencantas imbuhan tidak sesuai untuk kata terbitan Jawi. Perbezaan ketara boleh dilihat pada petua pembuangan akhiran '-an'. Contohnya, untuk tulisan Rumi, akhiran '-an' dieja dengan menggunakan satu cara manakala untuk tulisan Jawi ianya boleh dieja sebagai 'أن', 'ان', 'ءن' dan 'ن'. Oleh yang demikian, set petua Jawi diperlukan untuk membuang imbuhan pada kata terbitan Jawi. Selain itu pencantas perkataan Bahasa Melayu yang menggunakan kamus kata dasar perlu sentiasa dikemas kini untuk memastikan setiap perkataan yang dicantas sama dengan perkataan di dalam kamus untuk mengurangkan ralat. Objektif bagi tesis ini adalah untuk menghasilkan petua cantasan serta membangun dan menilai pencantas perkataan Jawi yang digunakan untuk mencantas kata terbitan dan menghasilkan kata dasar yang merangkumi kata jati melibatkan satu, dua dan tiga suku kata. Set data yang digunakan dalam kajian ini telah ditransliterasi ke dalam Jawi dan dibahagi kepada dua set, iaitu artikel-artikel daripada Utusan Melayu dan Berita Harian yang dipilih secara rawak di antara September 2009 - November 2010. Pangkalan data yang digunakan juga termasuklah Al-Quran terjemahan Sheikh Abdullah Basmeih yang telah digunakan dalam kajian yang lepas. Dalam penghasilan algoritma pencantas perkataan ini, terdapat dua komponen penting telah dihasilkan iaitu petua *nyah-imbuhan* untuk mencantas imbuhan dan petua *pengesanan kesalahan ejaan Jawi (SEDR)* yang digunakan untuk menyemak perkataan yang dicantas. Petua *nyah-imbuhan* melibatkan beberapa proses yang memerlukan pembuangan, penggantian dan penambahan aksara dalam setiap kata terbitan Jawi. Petua *SEDR* pula melibatkan susunan corak ejaan untuk membentuk suku kata Jawi. Sebanyak enam eksperimen telah dilakukan bermula dengan pengiraan ketepatan petua *SEDR*, pengiraan ketepatan petua *Rule Application Order* dan *Rule Frequency Order* menggunakan data Jawi, turutan pembuangan imbuhan, pengiraan ketepatan pencantas berasaskan penilaian Frakes dan Paice serta penilaian algoritma signifikasi berasaskan statistik. Hasil keseluruhan daripada kajian mendapati bagi nilai *min purata ketepatan* (*MPK*) dokumen Jawi yang dicantas adalah 8.43% manakala nilai *MPK* dokumen Jawi yang tidak dicantas adalah 5.14%. Pencantas perkataan Melayu untuk aksara Jawi ini dapat membantu meningkatkan ketepatan dalam capaian dokumen Jawi.

A MALAY STEMMER FOR JAWI CHARACTERS BASED ON RULE

ABSTRACT

Stemming is the process used to remove affixes from words to produce stemmed words; or root words. Stemmers can be used in document retrieval, transliteration, text classification and machine translation. Malay stemmers produced in previous studies focus more on Roman (Rumi) script. The rule set to produce stem affixes is unsuitable for derived Jawi words. Significant differences can be seen in the deaffixation of suffix rule '-an'. For example, for Roman (Rumi) script, the suffix '-an' is spelt using one technique; whilst for Jawi script, it can be spelled as 'أن', 'ان', 'ءن' and 'ن'. Therefore, a new Jawi rule set is needed to remove affixes from Jawi derived words. Additionally, Malay stemmers that use the root word dictionary need to be updated constantly, to ensure every stem word matches with the corresponding dictionary word to reduce the errors. The objective of this thesis is to produce stemmer rule and also develop and evaluate of a new Jawi stemmer used to stem derived words and to produce root words that encompass pure words involving one, two, and three syllables. The data set used in this work, has been transliterated into Jawi and has been divided into two sets. This study use articles from Utusan Melayu and Berita Harian which had been randomly picked between September 2009 – November 2010. The database also includes the translation of Al-Quran by Sheikh Abdullah Basmeih which has been used in one of the previous studies. In developing the proposed stemmer algorithm, two important components had been presented, namely a deaffixation rule to stem affixes and a Spelling Error Detector Rule (SEDR), to validate the stemmed word. The deaffixation rule involves several processes that require the removal, replacement, and character addition, within the Jawi derived word. SEDR involves the arrangement of a spelling pattern to form Jawi syllables. Several experiments were performed to prove that the Jawi stemmer could assist in Jawi document retrieval. A total of six experiments have been carried out starting with accuracy calculation of the SEDR rule, accuracy calculation of *Rule Application Order* and *Rule Frequency Order* using Jawi word, sequencing of the deaffixation, calculation of the stemmer accuracy based on Frakes and Paice assesments, and significant assessment of algorithms based on statistics. The overall result of this study found that the Mean Average Precision (MAP) of the stemmed Jawi documents is 8.43%, while the MAP value of the non-stemmed Jawi documents is 5.14%. Therefore, the Malay stemmer for Jawi characters could help to improve the accuracy of Jawi document retrieval.

KANDUNGAN

BAB 8 KESIMPULAN DAN SUMBANGAN

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENGENALAN

1.1 PENDAHULUAN

Pencantas perkataan digunakan untuk mencantas kata terbitan dan menghasilkan kata dasar. Pencantas perkataan yang terawal dihasilkan oleh Julie Beth Lovins pada tahun 1968 (Lovins, 1968). Kegunaan pencantas perkataan tidak terhad kepada bidang capaian dokumen sahaja tetapi ia juga amat penting dalam bidang transliterasi, serta penyemakan ejaan. Pencantas perkataan yang pertama bagi Bahasa Melayu telah dihasilkan oleh Asim Othman pada tahun 1993 (Asim, 1993). Seterusnya pencantas perkataan yang ada telah dilakukan penambahbaikan dan dapat mencantas perkataan Bahasa Melayu dengan baik. Kajian yang dilakukan telah membuktikan bahasa pencantas perkataan Bahasa Melayu ini dapat membantu dalam capaian dokumen. Walau bagaimanapun tumpuan utama penghasilan pencantas perkataan Bahasa Melayu tertumpu kepada aksara Rumi dan tidak merangkumi ejaan bagi aksara Jawi.

1.2 PERNYATAAN MASALAH

Bahasa Melayu boleh ditulis dengan menggunakan dua jenis aksara yang berbeza iaitu Rumi dan Jawi. Tulisan Jawi pernah digunakan sebagai tulisan utama untuk berkomunikasi dan kemudiannya tulisan ini telah digantikan dengan tulisan Rumi.

Morfologi bagi Bahasa Melayu lebih rumit jika dibandingkan dengan morfologi Bahasa Inggeris (Fatimah, 1995). Kajian mengenai pencantas perkataan dalam Bahasa Melayu telah dilakukan oleh beberapa orang penyelidik seperti Asim (1993), Fatimah (1995), Sock (2000), Idris (2001) dan Taufik (Muhammad Taufik,

2009). Namun hanya Sock (2000) sahaja yang menggunakan kaedah N-gram dalam penghasilan pencantas perkataannya manakala penyelidik yang lain menggunakan kaedah petua morfologi.

Dalam kajian Asim (1993), penyelidik telah menggunakan kamus untuk menyemak imbuhan yang telah dicantas. Walau bagaimanapun dalam kajian Fatimah (1995), telah menggantikan penggunaan kamus yang digunakan oleh Asim dengan 'kamus kata dasar'. Penggunaan kamus kata dasar dapat meningkatkan lagi ketepatan cantasan (Fatimah, 1995). Hasil kajian Idris (2001), penyelidik telah menambah satu lagi kamus khas iaitu 'kamus tempatan' yang dapat mengurangkan ralat dalam cantasan. Fungsi kamus khas adalah untuk mengelakkan ralat terlebih cantas dalam sesuatu subjek tertentu. Dalam kajian Idris (2001), kamus tempatan yang digunakan adalah kamus bagi mata pelajaran Sejarah.

Kamus perlu dikemaskini untuk memastikan perkataan yang baru dicantas dapat ditemui sekaligus mengelak daripada berlakunya ralat terlebih cantas dan ralat terkurang cantas (Kazem, 2005; Somayye Estahbani & Reza Javidan, 2011). Selain itu kajian yang dilakukan terhadap pencantas perkataan Bahasa Melayu lebih tertumpu kepada aksara Rumi (Asim, 1993; Fatimah, 1995; Sock, 2000; Idris, 2001, Muhammad Taufik, 2009).

Walaupun tulisan Rumi dan Jawi digunakan untuk mewakili satu bahasa yang sama tetapi terdapat perbezaan dalam aksara dan gaya ejaan. Aksara Rumi sama seperti aksara yang digunakan dalam Bahasa Inggeris manakala aksara Jawi pula sebahagiannya sama seperti aksara Arab. Terdapat enam penambahan aksara lain dalam Jawi menyebabkan hanya sebahagian aksara Jawi menyerupai aksara Arab.

Bentuk ejaan bagi Rumi dan Jawi juga berbeza. Dalam Rumi, vokal diwakili oleh lima aksara yang berbeza iaitu a, e, i, o dan u manakala bagi Jawi pula lima bunyi vokal hanya diwakili oleh tiga aksara yang berbeza iaitu ا, و dan ي. Setiap suku kata dalam ejaan Rumi mempunyai aksara vokalnya yang tersendiri. Tetapi bagi ejaan Jawi ada suku kata yang dieja tanpa menggunakan sebarang aksara vokal.

Selain itu terdapat perbezaan di antara petua yang digunakan untuk mencantas imbuhan. Perbezaan yang ketara boleh dilihat pada imbuhan awalan, akhiran dan apitan. Contohnya untuk mencantas perkataan *binaan* dan *bukaan* bagi aksara Rumi, akhiran *-an* perlu dipanggil untuk menghasilkan kata dasar *bina* dan *buka*. Tetapi untuk mencantas kata terbitan بينأن dalam aksara Jawi, akhiran ءن- perlu dipanggil untuk menghasilkan kata dasar بينا. Dalam sesetengah situasi, untuk mencantas kata terbitan بوكأن akhiran أن- perlu dipanggil untuk menghasilkan kata dasar بوک. Oleh yang demikian untuk mencantas akhiran *-an* dalam Jawi memerlukan perhatian sama ada untuk memanggil petua أن- atau ءن-.

Perbezaan juga wujud dalam kata serapan Bahasa Inggeris. Berdasarkan buku Daftar Kata Rumi-Sebutan-Jawi terbitan Dewan Bahasa dan Pustaka edisi ke-2 dinyatakan, ejaan kata serapan Bahasa Inggeris yang mempunyai dua aksara Jawi di akhir perkataan yang membentuk kelompok konsonan tidak perlu ditulis secara bersambung. Contohnya perkataan ”golf” perlu ditulis sebagai كولف dan bukannya كولف. Aksara ل dan ف dianggap sebagai dua aksara konsonan dan perlu ditulis secara berasingan tetapi sekiranya aksara konsonan tadi ditambah dengan aksara vokal di akhir perkataan seperti perkataan ‘fakta’ maka suku kata akhir perlu dieja sebagai suku kata terbuka seperti فكتا. Kesemua perkataan Bahasa Inggeris dan Eropah yang memerlukan penggunaan aksara [g] dieja dengan menggunakan aksara ‘ga’ dalam Jawi. Contohnya perkataan seperti agenda dan gimnasium. Untuk aksara [k] pula dieja dengan menggunakan aksara ك dan bukannya ق .

Bagi kata serapan Bahasa Arab pula, aksara vokalnya merujuk kepada lambang kepanjangan atau dikenali juga sebagai mad. Aksara vokal yang sama boleh dijumpai dalam dua perkataan yang berbeza tetapi mempunyai jenis kepanjangan yang berbeza. Perkataan yang mempunyai vokal yang panjang akan dieja dengan menggunakan salah satu aksara vokal ا,و,ي manakala perkataan yang mempunyai aksara vokal pendek akan dieja dengan tanpa menggunakan aksara vokal contohnya ’dam’ (vokal pendek) dieja sebagai دم dan ’am’ (vokal panjang) dieja sebagai عام (Hamdan Abdul Rahman, 1999).

Untuk memastikan kata terbitan dapat dicantas dengan tepat, pencantas perkataan Melayu untuk aksara Jawi telah dibangunkan. Objektif bagi kajian ini akan dibincangkan seperti berikut.

1.3 OBJEKTIF KAJIAN

Objektif bagi kajian ini adalah seperti berikut:

1. Menghasilkan petua penyemakan kesalahan ejaan (SEDR) yang dapat menyemak ejaan kata dasar bagi aksara Jawi.
2. Menghasilkan petua nyah-imbuhan yang digunakan untuk mencantas kata terbitan Jawi.
3. Membangunkan algoritma pencantas perkataan Bahasa Melayu bagi Aksara Jawi yang dapat mencantas kata terbitan Jawi kepada kata dasar.

1.4 KEPENTINGAN KAJIAN

Pencantas perkataan diperlukan sebagai asas untuk menyokong capaian maklumat sesuatu bahasa dan digunakan dalam aplikasi terjemahan dokumen dan carian sesawang (Jelita Asian, 2005). Berdasarkan Van Rijsbergen (1979), teknik pencantas perkataan dapat mengurangkan saiz indeks dan membantu capaian maklumat yang lebih relevan. Menurut beliau lagi proses pencantas perkataan dapat mengurangkan saiz perwakilan dokumen kepada 20%-50% jika dibandingkan dengan perwakilan penuh perkataan. Dalam tesis Belal (2001), penyelidik menyatakan kepentingan pengakar perkataan sebagai fungsi utama dalam sistem capaian dokumen kerana ianya berupaya mengurangkan bilangan perkataan yang berbeza dan sekaligus mengurangkan saiz kamus.

Kepentingan kajian yang dilakukan boleh dilihat dalam beberapa faktor iaitu pencantas perkataan yang dihasilkan merupakan pencantas perkataan untuk mencantas kata terbitan Melayu bagi aksara Jawi. Petua nyah-imbuhan yang dihasilkan bersesuaian untuk digunakan dalam mencantas imbuhan Jawi dan kajian ini juga memperkenalkan pengunaan petua SEDR iaitu petua yang digunakan untuk menyemak kembali ejaan kata dasar dalam Jawi setelah proses pembuangan imbuhan

dilakukan. Dalam kajian pencantas perkataan yang lepas, kamus digunakan sebagai proses penyemakan setelah cantasan dilakukan. Perkataan yang dijumpai di dalam kamus akan dijadikan sebagai kata dasar. Terdapat perbagai jenis kamus yang digunakan seperti kamus kata dasar yang mengandungi senarai lengkap semua kata dasar dan juga kamus tempatan yang mengandungi perkataan khusus bagi sesuatu subjek contohnya kamus sejarah. Peratus ketepatan pencantas bergantung kepada masukan kamus. Lebih tepat kamus membuat padanan ke atas kata cantasan bermakna lebih tinggi ketepatan yang akan dihasilkan oleh pencantas. Walau bagaimanapun untuk memastikan cantasan adalah tepat, kamus perlulah sentiasa dikemaskini untuk memastikan perkataan yang dicantas dapat ditemui dalam kamus (Kazem Taghva, Russek Beckley dan Mohammad Sadeh, 2005; Somayye Estahbani dan Reza Javi, 2011).

Selain itu, kajian ini juga telah membuktikan bahawa penggunaan pencantas perkataan Jawi mampu untuk membantu capaian dokumen Jawi. Dengan adanya pencantas Jawi carian ke atas dokumen Jawi boleh dilakukan dengan mudah tanpa perlu ditukarkan ke dalam ejaan Rumi. Selain itu pencantas Jawi juga penting untuk digunakan dalam proses transliterasi disebabkan ejaan Jawi yang semakin kurang dikuasai oleh generasi muda. Proses penyemakan ejaan bagi skrip Jawi juga dapat dilakukan secara automatik.

1.5 SKOP KAJIAN

Skop kajian yang dijalankan merangkumi kesemua imbuhan-imbuhan Bahasa Melayu dalam tulisan Jawi termasuk imbuhan awalan, apitan, akhiran dan sisipan yang digabungkan dengan kata dasar satu suku kata, dua suku kata dan tiga suku kata untuk membentuk kata terbitan yang ditransliterasi ke dalam skrip Jawi. Kesemua set data yang akan diuji diambil daripada artikel-artikel yang terkandung dalam Utusan Melayu dan ditukar dalam ejaan Jawi melalui proses transliterasi Rumi-Jawi menggunakan sistem TERUJA (2011) dan e-Jawi Converter (2011) untuk menghasilkan 1200 kata terbitan dalam Jawi. Berdasarkan kajian Yon Hendri (2009) TERUJA menghasilkan ketepatan sebanyak 70.7%. Seterusnya untuk melihat sejauh mana pencantas perkataan Jawi ini dapat membantu capaian dokumen, tafsir Al-Quran

seperti yang digunakan dalam kajian Fatimah (Fatimah Ahmad, 1995) dan Taufik (Muhammad Taufik, 2006) telah digunakan.

1.6 STRUKTUR ORGANISASI TESIS

Tesis ini telah ditulis dan disusun dalam lapan bab yang utama. BAB I dimulai dengan perbincangan mengenai latar belakang kajian yang dijalankan. Seterusnya bab ini juga membincangkan mengenai pernyataan masalah, objektif kajian, kepentingan kajian dan juga skop. Struktur tesis yang berikutnya dibincangkan dalam bab yang berikut.

Bab II menerangkan kepentingan pencantas perkataan serta jenis pencantas perkataan yang biasa digunakan. Pencantas-pencantas perkataan untuk bahasa lain seperti Bahasa Inggeris, Bahasa Perancis, Bahasa Arab, Bahasa Indonesia dan Bahasa Melayu turut dibincangkan dengan teliti. Selain itu jenis ralat yang ditemui dalam pencantas perkataan serta pengujian pencantas perkataan juga dimuatkan dalam bahagian ini.

Bab III membincangkan perbezaan yang wujud antara ejaan Rumi dan Jawi. Pembentukan perkataan melalui kata tunggal, kata terbitan, kata ganda dan juga partikel diterangkan dengan lengkap dalam bab ini.

Bab IV menerangkan metodologi tesis yang dijalankan. Perkara yang dibincangkan adalah berkaitan dengan reka bentuk kajian, kerangka kerja kajian dan reka bentuk eksperimen. Bab ini juga turut memuatkan sukatan prestasi, jenis ujian signifikasi dan alatan kajian yang digunakan sepanjang membangunkan tesis ini.

Bab V membincangkan proses yang terlibat dalam pembangunan petua pencantas perkataan melayu bagi aksara Jawi. Penghasilan petua pengesanan kesalahan ejaan (SEDR) dan petua nyah-imbuhan Jawi diterangkan dengan lebih lanjut. dalam bab ini. Eksperimen I dijalankan untuk menentukan turutan cantasan imbuhan yang bersesuaian dengan ejaan Jawi. Selain itu juga petua mengenai SEDR yang dihasilkan turut dinilai untuk menentukan peratus ketepatan petua SEDR. Hasil

dapatan daripada eksperimen dibincangkan dan digunakan untuk membangunkan algoritma cantasan seperti yang diterangkan dalam bab VI.

Bab VI menerangkan mengenai pembangunan algoritma pencantas perkataan Melayu bagi aksara Jawi. Algoritma pencantas yang telah dibangunkan kemudiannya dinilai dengan menentukan ketepatan cantasan yang dilakukan. Selain itu, penilaian menggunakan kaedah Paice (1990) dan Frakes & Fox (2003) turut dilakukan untuk melihat keberkesanan algoritma cantasan yang dihasilkan. Hasil eksperimen yang dijalankan turut dibincangkan.

Bab VII menerangkan eksperimen sistem capaian maklumat yang dijalankan untuk menguji sama ada algoritma cantasan yang dihasilkan dapat membantu capaian dokumen Jawi atau sebaliknya. Dalam ujian ini, set dokumen daripada tafsir Al-Quran yang sama seperti kajian Fatimah (Fatimah Ahmad, 1995) dan Taufik (Muhammad Taufik, 2006) telah digunakan. Ujian pasangan sampel-t juga turut dilakukan untuk menentukan sama ada terdapat perbezaan yang signifikan antara min purata ketepatan (MPK) bagi dokumen yang dicantas dengan dengan dokumen yang tidak dicantas dan hasil eksperimen turut dibincangkan dalam bab ini.

Bab VIII menerangkan mengenai kesimpulan bagi kajian yang telah dijalankan. Bab ini juga turut membincangkan dapatan kajian, sumbangan kajian serta cadangan perluasan kerja kajian.

1.7 KESIMPULAN

Bab ini menyentuh secara ringkas dan padat mengenai objektif, pernyataan masalah, kepentingan kajian, skop dan struktur organisasi mengenai kajian. Setiap aspek penting dibincangkan dalam bab ini. Bab ini mengemukakan gambaran menyeluruh mengenai kajian yang dijalankan dan membantu para penyelidik yang ingin memahami kajian ini.

Objektif bagi kajian ini adalah untuk menghasilkan petua penyemakan kesalahan ejaan (SEDR) yang dapat menyemak ejaan kata dasar bagi aksara Jawi,

menghasilkan petua nyah-imbuhan yang digunakan untuk mencantas kata terbitan Jawi dan juga membangunkan algoritma pencantas perkataan Bahasa Melayu bagi Aksara Jawi yang dapat mencantas kata terbitan Jawi kepada kata dasar. Untuk memastikan kesemua objektif dapat dicapai, kajian untuk menghasilkan petua SEDR serta petua nyah imbuhan yang baru telah dilakukan dengan melihat penggunaan setiap petua menerusi buku-buku seperti "Panduan Membaca dan Menulis Jawi", "Tatabahasa Dewan", ditulis oleh Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa dan Abdul Hamid Mahmood terbitan Dewan Bahasa dan Pustaka, "The Morphology of Malay", "Daftar Ejaan Rumi Jawi" dan "Daftar Kata Bahasa Melayu Rumi-Sebutan-Jawi" terbitan Dewan Bahasa dan Pustaka. Seterusnya enam eksperimen telah dilakukan untuk menentukan objektif telah dicapai.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Algoritma pencantas perkataan merupakan proses untuk membuang perkataan berimbuhan dan menghasilkan perkataan yang dicantas ataupun dikenali sebagai kata dasar. Contohnya perkataan 'bermain', 'permainan' dan 'mainan' akan dicantas dan menghasilkan kata dasar 'main'. Berdasarkan Savoy (1993), kata dasar diperoleh dengan membuang awalan serta akhiran atau kedua-duanya.

Algoritma pencantas perkataan mempunyai fungsi yang penting dalam sistem capaian maklumat seperti mengurangkan saiz perkataan. Selain itu algoritma pencantas juga berkebolehan untuk mengenal perkataan yang sama dari segi semantik dan sekaligus meningkatkan nilai dapatan (Willet, 1988).

Kebanyakan kajian dalam bidang capaian maklumat lebih mementingkan penambahbaikan prestasi berbanding dengan pengurangan storan (Harman, 1987). Algoritma pencantas tidak menjamin penambah baikan untuk capaian yang lebih efektif dalam sebarang keadaan (Harman, 1991).

Harman (1991) menguji tiga jenis algoritma pencantas yang berbeza iaitu S-stemmer (Harman, 1991), Lovins (1968) dan Porter (1980). Harman (1991) membandingkan carian yang menggunakan pencantas dengan carian yang tidak menggunakan pencantas. Selepas penilaian terperinci dilakukan, Harman (1991) menyimpulkan tiada pencantas yang dapat membantu meningkatkan penilaian secara konsisten. Popovic dan Willet (1992) menguji sama ada pembuangan akhiran lebih

efektif terhadap bahasa yang lebih komplek dalam aspek morfologi seperti Serbia. Algoritma yang sama seperti algoritma Porter (1980) telah dihasilkan dalam Bahasa Serbia dan algoritma ini diuji menggunakan koleksi set ujian yang kecil. Daripada ujikaji didapati terdapat perkembangan yang signifikan dalam ketepatan pada 10 dokumen yang teratas. Dalam ujikaji ini, eksperimen kawalan telah dilakukan. Teks dalam Bahasa Serbia telah diterjemah ke dalam Bahasa Inggeris dan teks berkenaan diuji sekali lagi. Hasil daripada eksperimen kawalan ini mengukuhkan lagi pendapat yang dibentangkan oleh Harman (1991) iaitu algoritma Porter tidak dapat membantu meningkatkan capaian dokumen Inggeris. Popovic dan Willet (1992) menyatakan keberkesanan sesuatu algoritma itu dipengaruhi oleh bahasa yang mempunyai morfologi yang kompleks.

Dalam capaian dokumen, perkataan yang mempunyai maksud yang sama akan dikumpulkan dalam satu kumpulan atau dikenali sebagai indeks cantas yang akan meningkatkan padanan dokumen bagi pertanyaan (Van Rijsbergen, 1979). Perkataan yang dicantas tidak mengandungi elemen-elemen linguistik. Bagi ahli linguistik mereka lebih tertarik untuk mencari lemma yang tepat bagi perkataan berimbuhan dan bukannya hasil cantasan.

Pencantas perkataan mampu untuk meningkatkan lagi capaian dokumen bagi sesetengah bahasa. Terdapat pencantas perkataan yang tidak mendatangkan sebarang perubahan terhadap capaian dokumen. Oleh yang demikian penting untuk menguji pencantas perkataan yang dihasilkan untuk menentukan sama ada pencantas berkenaan mampu membantu meningkatkan lagi capaian dokumen.

2.2 PENCANTAS PERKATAAN

Mencantas perkataan (*Stemming*) merupakan teknik untuk mengumpul kesemua perkataan ke dalam satu kelompok morfologi yang sama (McNamee, 2008). Konsep pencantas perkataan telah lama digunakan pada sekitar tahun 1960-an. Matlamat utama pencantas perkataan adalah mengurangkan bilangan perkataan yang terkandung dalam suatu senarai contohnya kamus. Algoritma pencantas perkataan diguna untuk memperbaiki tahap kecekapan dan capaian sistem maklumat. Ianya juga diguna

sebagai salah satu cara meningkatkan capaian dan juga berkeupayaan untuk menambah ketepatan dalam sistem capaian maklumat (Kowalski, 1997).

Pencantas perkataan dapat membantu meningkatkan mutu capaian maklumat serta mengurangkan saiz kamus pengindeksan dan sekaligus menjimatkan ruang storan (Paice, 1996; Argraw & Askar, 2007; Felipe et. al, 2010). Pencantas perkataan juga berguna untuk aplikasi yang memerlukan perkataan mempunyai bentuk morfologi yang sama diproses serta dikumpulkan dalam satu kumpulan seperti pengkelasan teks, sistem capaian maklumat dan carian kamus (Lily Suryana Indradjaja & Bressan, 2003). Pengumpulan perkataan yang mempunyai maksud yang sama akan meningkatkan nilai purata panggilan dan dapat meningkatkan capaian (Carmen et.al., 2005; Tesfaye & Abebe, 2010; Mohammad N.Al-Kabi et. al. 2006).

Untuk melakukan carian maklumat menggunakan pangkalan data, indeks perlu diwujudkan untuk memudahkan proses carian. Carian boleh dilakukan dengan menggunakan satu perkataan atau pelbagai perkataan yang dikenali sebagai varian kata (*terms variants*). Terdapat tiga jenis variasi bagi varian kata iaitu:

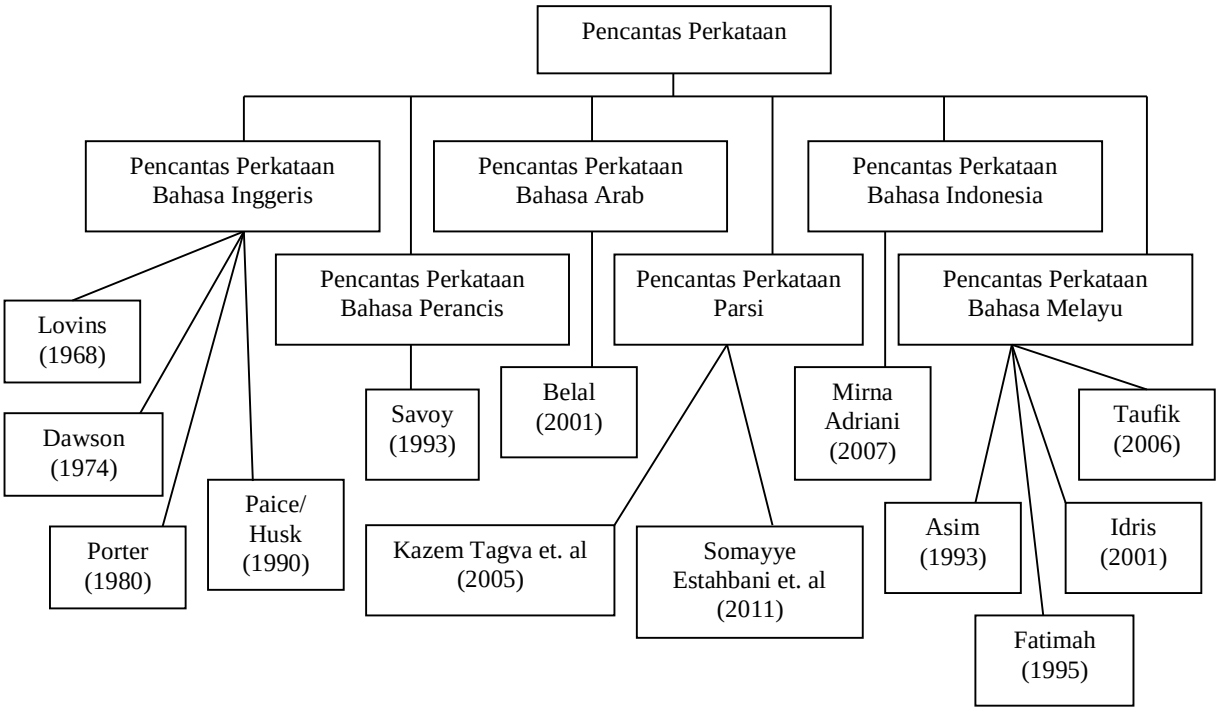
- Variasi morfologi (*morphological variation*): perkataan yang sama konsep dan morfologi tetapi muncul dalam bentuk yang berbeza. Contohnya *'connecting'* dan *'connected'*.
- Variasi lexico-semantik (*lexico-semantic variation*): perkataan yang berlainan tetapi mewakili maksud yang sama. Contohnya *'aoxemia', 'anoxaemia*' dan *'breathing problem'* dimasukkan ke dalam maksud *'breathing disorder'*.
- Variasi sintaktik (*syntactic variation*): Berkaitan dengan perkataan yang lebih daripada satu. Contohnya *'consideration of these domain properties*' dan *'considering certain domain properties'* dirujuk sebagai *'considering domain properties'*.

Daripada kaedah yang ada, algoritma pencantas sering digunakan untuk mencari kata dasar yang terdapat dalam perkataan berimbuhan. Variasi lexico-semantik pula sering kali menggunakan carian jadual leksikal atau carian *thesaurus* (Paice, 1996).

Bagi perubahan semantik penggunaan pertanyaan yang berkaitan dengan terma semantik diperlukan (Galvez et al., 2005).

Proses padanan carian dengan teks pangkalan data akan lebih mudah dengan menggunakan kaedah gabungan (*conflation method*) (Galvez et al., 2005). Kaedah gabungan (*conflation*) boleh dilihat dalam dua sudut yang berbeza iaitu teknik linguistik dan teknik bukan linguistik. Teknik linguistik menggunakan pendekatan pemprosesan bahasa tabii. Kaedah gabungan (*conflation*) yang boleh dilihat dalam teknik linguistik adalah seperti *lemmatization* dan analisis morfologi. Penghasilan kata cantasan akan bergantung kepada maklumat leksikal yang disimpan dalam kamus elektronik atau leksikon (Galvez et al., 2005). Contoh yang boleh dilihat adalah seperti penganalisis morfologi yang dihasilkan oleh Karttunen (1983).

Teknik bukan linguistik pula tidak menggunakan sebarang pendekatan pemprosesan bahasa tabii. Perkataan yang mengandungi imbuhan akan dikurangkan dalam satu bentuk yang sama. Untuk teknik bukan linguistik biasanya algoritma pencantas dan pelucutan akhiran (*suffix stripping*) digunakan (Galvez et al., 2005). Contoh pencantas perkataan yang sering digunakan adalah pencantas perkataan Lovins (1968) dan Porter (1980) yang digunakan untuk mencantas perkataan Bahasa Inggeris. Banyak kajian telah melaporkan keberkesanan pencantas dalam sistem capaian maklumat terutamanya untuk Bahasa Inggeris (Harman , 1991; Hull, 1996; Krovetz, 1993). Rajah 2.1 di sebelah menunjukkan pencantas perkataan yang berbeza mengikut bahasa.



Rajah 2.1 Pencantas perkataan yang berbeza mengikut bahasa

2.2.1 Pencantas Perkataan Pembuangan Imbuhan

Pencantas perkataan pembuangan imbuhan menggunakan teknik pembuangan akhiran atau awalan daripada perkataan dan menghasilkan perkataan yang dicantas. Walau bagaimanapun ada di antara hasil cantasan yang perlu diubah. Contoh pencantas perkataan pembuangan imbuhan adalah pembuangan kata jamak dalam Bahasa Inggeris. Pencantas perkataan pertama yang menggunakan teknik ini adalah pencantas perkataan Lovins (1968). Tetapi kebanyakan teknik menggunakan lelaran (*iterative*) padanan terpanjang. Dalam pencantas lelaran ini, rentetan aksara yang mempunyai padanan terpanjang akan dibuang daripada perkataan berdasarkan petua khas. Hasil perkataan yang dicantas tidak semestinya tepat dari segi linguistik.

Salah satu pencantas perkataan pembuangan imbuhan Bahasa Inggeris yang efisyen adalah pencantas perkataan Porter (1980). Pencantas perkataan lain yang menggunakan teknik ini adalah seperti Salton (1968), Paice (1990) dan MARS (Niedermair et. al., 1985).

2.2.2 Pencantas Perkataan Varieti Pengganti (*Sucessor variety stemmer*)

Pencantas perkataan varieti pengganti (Haffer & Weiss, 1974) adalah berdasarkan kajian dalam struktur linguistik yang diguna untuk menentukan perkataan dan sempadan morfem berdasarkan agihan fonem dalam sebutan yang besar (Frakes, 1992).

Pengganti varieti boleh dinyatakan sebagai bilangan aksara yang berbeza mengikut perkataan dalam teks. Setelah pengganti perkataan dijumpai, perkataaan akan dipecah berdasarkan varieti ini. Proses pemecahan boleh dilaksanakan seperti yang dicadang oleh Hafer & Weiss (1971) dengan menggunakan satu kaedah seperti berikut: *cutoff method*, *peak and plateu method*, *complete word method* dan *entropy method.*

Setelah selesai proses segmentasi, segmen akan dipilih sebagai hasil cantasan dengan menggunakan petua berikut (Hafer & Weiss, 1974):

jika (segmen pertama muncul dalam ≤ 12 perkataan dalam korpus)
perkataan pertama adalah perkataan yang dicantas
atau (segmen kedua adalah perkataan yang dicantas)

2.2.3 Pencantas Perkataan Carian Jadual

Pencantas perkataan carian jadual merupakan pencantas perkataan yang paling mudah di antara kesemua pencantas yang lain. Kesemua indeks kata dan perkataan yang dicantas diletak dalam satu jadual. Rajah 2.2 menunjukkan contoh jadual yang mengandungi indeks kata dan perkataan yang dicantas.

Perkataan	Perkataan Yang dicantas
pembangunan	bangun
bangunan	bangun
bangunkan	bangun
bangunnya	bangun
membangunkan	bangun
membangun	bangun

Rajah 2.2 Contoh jadual yang mengandungi indeks kata dan perkataan yang dicantas bagi pencantas perkataan carian jadual

Carian jadual boleh diguna untuk mencantas pertanyaan dan perkataan yang terkandung dalam dokumen setelah jadual lengkap diisi dengan maklumat. Carian yang pantas boleh dilakukan dengan menggunakan teknik *hash* dan *B-tree*. Pencantas perkataan ini berkeupayaan menghasilkan ketepatan yang tinggi.

2.2.4 Pencantas Perkataan N-Gram

Pencantas perkataan N-Gram telah digunakan oleh Adamson & Boreham (1971). Rajah 2.3 menunjukkan contoh perkataan '*engineering*' dan '*engineers*' yang dipecahkan ke dalam bentuk diagram.

Engineering =>

1	2	3	4	5	6	7	8	9	10
en	ng	gi	in	ne	ee	er	ri	in	ng

Diagram Unik =>

1	2	3	4	5	6	7	8
en	ng	gi	in	ne	ee	er	ri

Engineers =>

1	2	3	4	5	6	7	8
en	ng	gi	in	ne	ee	er	rs

Diagram unik=>

1	2	3	4	5	6	7	8
en	ng	gi	in	ne	ee	er	rs

Rajah 2.3 Perkataan *engineering* dan *engineers* yang dipecahkan dalam bentuk diagram

Rajah 2.3 menunjukkan perkataan '*engineering*' mempunyai sepuluh diagram dan lapan daripadanya adalah unik. Manakala perkataan '*engineers*' pula mempunyai lapan diagram dan kesemuanya adalah unik. Pencantas akan mengira ukuran yang berkaitan di antara pasangan perkataan berdasarkan perkongsian diagram unik. Walau bagaimanapun terdapat kekangan untuk menyediakan jadual yang lengkap bagi sesuatu bahasa (Frakes, 1992). Selain itu simpanan yang besar diperlukan untuk menyimpan kesemua jadual. Diagram boleh didefinasi sebagai pasangan aksara yang berurutan. Kedua-dua perkataan *engineering* dan *engineers* berkongsi sepuluh diagram unik iaitu: en ng gi in ne ee er. Selepas pengiraan dilakukan, pengiraan yang sama akan dikira. Kebarangkalian digunakan untuk membuat ukuran persamaan seperti berikut:

$$S = 2C / (A+B)$$

Dengan A merupakan bilangan diagram unik dalam perkataan pertama dan B merupakan diagram unik dalam perkataan kedua. C pula merupakan bilangan diagram unik yang dikongsi dalam A dan B.

2.3 PENCANTAS PERKATAAN BAHASA INGGERIS

Pencantas perkataan Inggeris yang pertama telah dihasilkan oleh Julie Beth Lovins pada tahun 1968. Tujuan utama pencantas perkataan Inggeris diwujudkan bertujuan untuk memproses setiap perkataan Inggeris dan menghasilkan bentuk kata dasar. Pencantas perkataan ini banyak digunakan untuk mencapai dokumen yang relevan dalam sistem capaian maklumat dokumen.

2.3.1 Algoritma Lovins

Algoritma Lovins telah dihasilkan oleh Julie Beth Lovins pada tahun 1968 (Lovins, 1968). Algoritma ini berasaskan padanan lelaran terpanjang (*interative longest match*) yang merupakan salah satu daripada teknik pencantas perkataan pembuangan imbuhan.

Pencantas perkataan Lovins mengandungi 294 senarai akhiran yang dibahagi kepada 11 subset. Subset ini disusun secara menurun berdasarkan panjang akhiran dan disimpan mengikut susunan abjad supaya mudah diuruskan. Setiap subset diberi awalan khas yang mengandungi panjang akhiran di dalamnya. Setiap akhiran mempunyai kod keadaan dan pulangan pembawa sebagai penentu had. Kod keadaan mengandungi aksara yang membawa maklumat mengenai sekatan kontektual untuk mencantas (Lovins, 1968).

Berdasarkan pencantas perkataan Lovins, apabila akhiran telah dijumpai, ia akan dibuang berdasarkan set petua. Dalam kes ini sebanyak 29 set petua telah dihasilkan. Proses ini diulang sehingga tiada lagi akhiran untuk dibuang. Perkataan tercantas berkemungkinan tidak dapat dikumpulkan dalam kumpulan yang sama walaupun setelah kesemua akhiran dibuang. Oleh itu teknik kod semula (recoding) digunakan. Sebanyak 34 petua kod semula dihasilkan untuk menukarkan perkataan tercantas yang tidak tepat kepada kata dasar.

2.3.2 Algoritma Dawson

Algoritma pencantas perkataan Dawson (Dawson 1974) dihasilkan berdasarkan algoritma yang dihasilkan oleh Lovin (1968). Algoritma ini mengunakan pendekatan lelaran padanan terpanjang.

Dawson menggunakan senarai yang mengandungi 260 akhiran dalam Bahasa Inggeris dan kod keadaan seperti yang digunakan oleh Lovins. Walau bagaimanapun Dawson mendapati senarai yang diperolehinya mempunyai kekurangan dalam aspek akhiran dan perkataan yang berbentuk jamak (*plural*). Dawson mengemas kini senarai daripada Lovin dan menghasilkan senarai baru yang mengandungi 1200 akhiran. Untuk mengelakkan masalah dari segi pemprosesan dan penyimpanan, Dawson menyimpan kesemua akhiran dan kod kawalan secara terbalik dan menghasilkan indeks berdasarkan bilangan dan aksara terakhir bagi setiap akhiran dan kod kawalan.

Dalam algoritma Dawson, kod semula tidak dilakukan tetapi Dawson menggunakan teknik '*partial matching*' yang memadankan perkataan jika

cantasannya hampir sama. Oleh yang demikian padanan perkataan diberi sedikit peruntukan untuk dilakukan.

Asas bagi algoritma Dawson adalah seperti berikut. Sekiranya cantasan berpadanan dengan dua jenis bilangan aksara dan aksara berikutnya adalah kepunyaan pencantas yang mempunyai akhiran yang sama maka kedua-dua hasil cantasan akan digabungkan dalam satu bentuk yang sama (Popovic, 1991)

2.3.3 Algoritma Porter

Tujuan utama algoritma Porter (Porter, 1980) dicipta adalah untuk memproses perkataan Inggeris dan untuk mendapatkan kata dasar bagi setiap perkataan berimbuhan dan sekaligus menyumbang kepada keberkesanan capaian maklumat. Secara asasnya algoritma ini berasaskan pembuangan imbuhan akhiran atau dikenali sebagai *suffix stripping*. Algoritma ini digunakan dengan meluas dalam sistem capaian maklumat dokumen bagi bahasa Inggeris.

Algoritma Porter mengandungi set petua bagi setiap keadaan. Untuk kes ini keadaan yang dimaksudkan boleh dibahagi kepada tiga kelas iaitu: keadaan pada pencantas, keadaan pada akhiran dan keadaan pada petua. Jenis keadaan pencantas boleh disenaraikan seperti berikut:

1. Pengukuran: ditanda sebagai *m* bagi pencantas berasaskan turutan berselang-seli vokal-konsonan. Vokel adalah *A,E,I,O, U* dan *Y* merupakan konsonan manakala C adalah jujukan konsonan dan V adalah jujukan vokal. Persamaan bagi *m* adalah seperti berikut:

   $$[C]\ (VC)^{m}[V]$$

   Superskrip *m* dalam persamaan di atas menunjukkan bilangan jujukan VC dan tanda kurungan menunjukan pilihan yang wujud bagi kandungannya.
2. *S - pencantas yang berakhir dengan S (dan yang sama dengan aksara yang lain)
3. *V* - pencantas yang mengandungi vokal
4. *d - pencantas yang berakhir dengan dua konsonan

5. *o - pencantas yang berakhir dengan jujukan konsonan-vokal-konsonan di mana konsonan tidak diakhiri dengan jujukan W, X atau Y.

Bentuk keadaan akhiran adalah seperti berikut:

((akhiran_semasa) == corak)

Petua pembuangan akhiran adalah seperti berikut:

(keadaan) S1 → S2

Petua di atas bermaksud sekiranya perkataan berakhir dengan akhiran S1, dan cantasan sebelum S1 memenuhi keadaan yang diberikan maka S1 digantikan dengan S2. Keadaan biasanya diberi dalam bentuk *m* dan boleh juga mengandungi ungkapan seperti *dan, atau* dan *tidak.*

Petua dibahagikan kepada beberapa langkah. Petua dalam langkah yang tertentu diperiksa dalam jujukan dan hanya satu petua sahaja yang boleh digunakan dan petua ini berkemungkinan merupakan padanan terpanjang S1 bagi perkataan yang diberikan. Algoritma Porter menggunakan kamus yang mengandungi 60 akhiran dan sedikit sensetif-konteks serta petua *recoding*. Ini menjadikan algoritma Porter lebih ekonomikal dari segi masa dan simpanan. Disebabkan algoritma ini lebih ringkas, ujian capaian (Porter, 1980) menunjukkan algoritma ini lebih baik daripada kebanyakan algoritma yang lebih rumit seperti yang dinyatakan oleh Dawson (1974).

Antara kelebihan algoritma Porter adalah ia lebih ringkas dan efisyen dalam aspek pemprosesan, mudah dilaksana menggunakan mana-mana bahasa pengaturcaraan tinggi serta mempunyai nilai yang tinggi apabila diuji ke atas ujian capaian (lennon et. al., 1981). Algoritma Paice/Husk lebih padat berbanding Lovins dan Dawson tetapi lebih padat seperti Porter.

2.3.4 Algoritma Paice/Husk

Algoritma Paice/Husk telah dihasilkan oleh Chris Paice sekitar tahun 1990 dan dibantu oleh Gareth Husk (Paice, 1990). Walaupun algoritma ini senang untuk dilaksanakan dan efisyen, ianya dikenali juga dengan sifat yang agresif. Algoritma ini

menggunakan jadual petua tunggal yang menentukan sama ada pembuangan atau penukaran perlu dilakukan bagi bahagian akhir perkataan. Teknik penukaran ini bertujuan mengelak masalah kesilapan ejaan dengan menggantikan hujung sesuatu perkataan tanpa perlu melakukan proses berasingan semasa melakukan cantasan berbanding dengan membuang terus hujung perkataan berkenaan (Paice, 1990). Petua dikumpulkan dalam satu bahagian kumpulan yang sama dengan aksara terakhir untuk akhiran dan menyebabkan petua dapat dicapai dengan pantas memandangkan aksara terakhir dapat dilihat pada perkataan semasa atau perkataan terpenggal. Arahan petua dalam setiap kumpulan adalah signifikan dan sesetengah petua dihadkan bagi sesuatu perkataan. Sebelum padanan petua dilarikan, satu ujian ringkas telah dilakukan. Selepas petua tertentu digunakan, proses diikuti dengan lelaran atau mungkin juga ditamatkan.

Setiap petua mempunyai lima komponen yang mana dua daripadanya merupakan pilihan. Komponen yang terlibat adalah seperti berikut:

- Akhiran bagi satu atau lebih aksara, lakukan dalam arahan terbalik
- Pilihan yang ditanda sebagai ”*”
- Digit yang menentukan jumlah pembuangan
- Pilihan jujukan tambahan satu atau lebih aksara
- Simbol sambungan , ’>’ atau ’.’

2.4 PENCANTAS PERKATAAN BAHASA PERANCIS

Algoritma ini telah dihasilkan oleh Savoy (1993) yang mengandungi sebanyak 52,627 masukan. Terdapat enam medan yang disambung kepada setiap masukan dan setiap medan berupa kata dasar atau jenis cantasan (kata kerja, kata nama, adjektif dan sebagainya), kunci yang menunjuk kepada fail, maklumat maskulin, maklumat jantina dan pilihan terjemahan perkataan dalam Bahasa Inggeris.

Fail deklensi menyimpan 100 masukan untuk kata nama, adjektif, kata ganti nama dan sebagainya. Fail ini juga turut menyimpan 132 masukan untuk kata kerja dan dilaksanakan menggunakan carian pohon gelintar pelbagai jalan. Terdapat enam

medan utama untuk setiap masukan (Savoy, 1993). Medan yang dimaksudkan adalah seperti berikut:

- Kekunci
- Akhiran
- Maklumat feminin bilangan tunggal
- Maklumat maskulin bilangan tunggal
- Maklumat feminin bilangan yang lebih daripada satu
- Maklumat maskulin bilangan yang lebih daripada satu

Masukan bagi kata kerja pula adalah seperti berikut:

- Kekunci
- Akhiran
- Jenis kala (tense)
- Bilangan tunggal pertama
- Bilangan tunggal kedua
- Bilangan tunggal ketiga
- Bilangan yang lebih daripada satu pertama
- Bilangan yang lebih daripada satu kedua
- Bilangan yang lebih daripada satu ketiga

Algoritma Savoy bekerja dalam dua fasa yang berbeza. Fasa yang pertama akan membuang *inflectional suffixes* dan fasa kedua akan membuang kesemua akhiran kata terbitan. Fasa pertama dimulakan dengan membuang hujung aksara satu persatu dan setelah semuanya dibuang perkataan yang tertinggal dibanding dengan yang ada di dalam kamus. Sekiranya perkataan dijumpai maka perkataan tadi akan dijadikan sebagai kata dasar dan disemak dalam fail diklensi untuk akhiran yang sepadan. Jika tidak proses akan diulang sehinggalah perkataan dijumpai di dalam kamus.

Fasa kedua lebih tertumpu kepada pembuangan akhiran bagi kata terbitan. Empat jenis jadual yang berbeza telah dihasilkan dan kesemuanya berkaitan dengan kata kerja, kata nama, adjektif dan adverba. Algoritma menggunakan petua