https://doi.org/10.17576/apjitm-2021-1001-01
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 10 No. 1, June 2021: 1 - 16
e-ISSN: 2289-2192

Received: 7 September 2020
Accepted: 19 February 2021
Published: 1 June 2021

HAPAXES AS PRODUCTIVITY OF WORD VARIATIONS BASED ON CLASSICAL TEXTS

MUHAMAD FADZLLAH ZAINI
MAZURA MASTURA MUHAMMAD
ANIDA SARUDIN
SITI SANIAH ABU BAKAR
ZULKIFLI OSMAN

ABSTRACT

This study aims to examine the productivity of old Malay words that are contained in classic Malay texts. The ability of classical Malay texts was able to be examined through the presence of old Malay word variations based on the appearance of 1.000000 / corpus. In examining this emergence, words that are *hapaxes* were removed from the corpus-generated results. This study utilized the quantitative research design using the corpus linguistic statistical approach. This study used 7 classical Malay texts entitled Tips for Building a House. This classical text is non-literary and delivers architectural knowledge from feudal Malay scholars. The findings show that almost 54% of word are *hapaxes* in this classical text discourse. The stimulation of old Malay lexical used by past scholars showed less words reuse.

Keywords: *Hapaxes*, Lexical Productivity, Computational Linguistics, Classical Text.

HAPAXES SEBAGAI PRODUKTIVITI VARIASI KATA BERDASARKAN TEKS KLASIK

ABSTRAK

Kajian ini bertujuan untuk meneliti produktiviti kata Melayu lama yang terdapat dalam teks klasik Melayu. Keupayaan teks klasik Melayu dapat dilihat dengan kehadiran variasi kata Melayu lama berdasarkan kemunculan 1.000000 / korpus. Dalam meneliti kemunculan ini, kata yang bersifat *hapaxes* dikeluarkan daripada keputusan janaan korpus. Kajian ini menggunakan reka bentuk penyelidikan kuantitatif berdasarkan Pendekatan Statistik Linguistik Korpus. Kajian ini menggunakan 7 teks klasik Melayu berjudul Petua Membina Rumah. Teks klasik ini bersifat non-sastera yang bertindak sebagai penyampaian ilmu seni bina dikalangan sarjana Melayu feudal. Dapatan menunjukkan hampir 54% kata bersifat *hapaxes* dalam wacana teks klasik ini. Rangsangan kata Melayu lama yang digunakan oleh sarjana terdahulu menunjukkan kurang penggunaan semula kata.

Kata kunci; *Hapaxes*, Produktiviti Kata, Linguistik Berkomputer, Teks Klasik

PENGENALAN

Kajian ini merupakan penelitian terhadap produktiviti kata Melayu lama dalam teks klasik Melayu. Produktiviti ini merupakan taburan kata yang berada dalam sesebuah teks. Taburan ini memperlihatkan kekerapan kata digunakan bagi representasi teks. Keupayaan mental

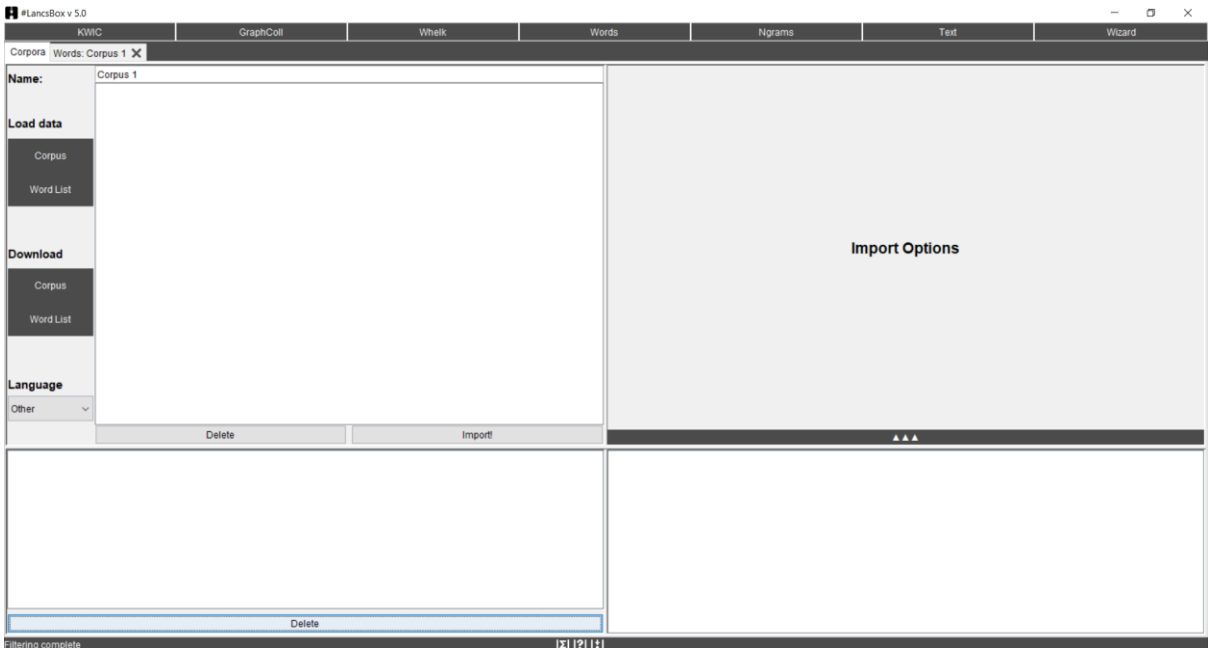
menghasilkan perbendaharaan kata memberikan daya terhadap representasi penulisan sesebuah teks (Bradac, Davies, Courtright, Desmond, & Murdock, 1977). Perbendaharaan kata ini merupakan kumpulan kata yang berada dalam konstruk mental bagi membentuk unjuran komunikasi. Hubungan ini merupakan bentuk komunikasi lisan kepada tulisan yang membentuk perbendaharaan kata (Zaytseva, Miralpeix, & Pérez-Vidal, 2019). Penghasilan konstruk mental daripada lisan kepada tulisan memberikan imej sarjana terdahulu dalam ilmu yang dimiliki (Hassan, 2016). Bagi melihat konstruk mental sarjana Melayu feudal, teks klasik Melayu perlu diamati. Pengamatan ini bagi melihat rangsangan keilmuan tamadun terdahulu dalam kehidupan mereka (Hernández-Campoy & Conde-Silvestre, 2012). Teks klasik merupakan sumber maklumat peribumi Melayu yang mencatatkan soal kemajuan tamadun Melayu feudal (Ding Choo Ming, 2016). Sejarah teks klasik Melayu telah menunjukkan kepenggunaan ilmu terdahulu yang telah dilestarikan bagi keperluan masa kini, seperti kajian (Ahmad, Lukman, & Yusof, 2016; Amer Hudhaifah, 2017; Faisal & Wahidah, 2012; Filzah, Mat, Wan, & Firdaus, 2018; Ghani, 2015; Riswadi & Mustaffa, 2017; Safwan & Zubir, 2018; Wan Mohd Dasuki & Radziah, 2014; Yakob, 2018). Teks klasik Melayu mempunyai kandungan ilmu yang tersendiri dan dicatat dalam tulisan jawi (Yakob, 2018). Merentasi era digital, keupayaan menterjemah teks klasik dalam mode dalam talian sangat diperlukan. Setakat ini hanya satu sahaja korpus yang mengumpulkan rujukan teks klasik Melayu iaitu *Malays Concordance Project* (MCP) yang berpangkalan di Universiti Kebangsaan Australia. Keperluan dalam mendigitalkan kata atau wacana ini penting bagi pelestarian ilmu sebagai khazanah berharga. Kebanyakan teks klasik dunia terutama dalam bahasa Inggeris telah digital bagi penyimpanan dan kegunaan penyelidikan. Dalam kerancakan ini, usaha sarjana berupaya dalam mendigitalkan segala bentuk wacana bahasa Melayu dibangunkan sebagai data korpus (Sukawai & Omar, 2020; Tiun, Abdullah, Kong, & Muhammad, 2013).

Selesai soal pembangunan korpus, penyelidikan atau pengamatan data korpus perlu diberikan fokus. Antara analisis yang dilakukan mengikut prosedur linguistik korpus adalah analisis senarai kata (*wordlist analysis*), analisis senarai kata kunci (*keywords analysis*) dan analisis penghubungan kata (*co-occurrence analysis*) (S. Smith, 2020). Analisis senarai kata kunci (*wordlist analysis*) merupakan bentuk penelitian terhadap kata berbeza (*type*) yang terdapat dalam teks korpus. Dalam penelitian ini, penggunaan prinsip *Zipf Law* digunakan untuk mendapatkan kedudukan *hapaxes* atau *hapax legomena* dengan perincian logaritma 10. Prinsip ini menunjukkan keupayaan janaan kata yang mempunyai aras kekerapan dalam graf lekuk. *Zipf Law* berupaya menjana senarai kata berdasarkan saiz korpus yang digunakan (Delgado, Ángela, & García, 2007). *Hapaxes* merupakan kumpulan tahap jenis kata yang muncul sekali (1.000000) kekerapan menerusi data korpus (Davis, 2019). Dalam penentuan *hapaxes* penggunaan prinsip *Zipf Law* diaplikasikan bagi representasi kata berkedudukan bawah (Brezina, 2018). Selain daripada *Zipf Law* terdapat juga sarjana menggunakan *Heap's Law* (Boytsov, 2017; Davis, 2019; Delgado et al., 2007; Lü, Zhang, & Zhou, 2013) bagi menganalisis kemunculan *hapaxes*. Akan tetapi, menerusi Pendekatan Statistik Linguistik Korpus menggunakan prinsip *Zipf Law* sebagai teras analisis graf lekuk data korpus. *Hapaxes* kurang dibincangkan dalam kajian menggunakan data korpus terutama teks klasik Melayu. *Hapaxes* memiliki pelbagai fungsi iaitu mendedahkan kekayaan kata, kehadiran kata baharu (*neologism*) dan kesilapan ejaan (Lardilleux & Lepage, 2009). *Hapaxes* juga berperanan dalam produktiviti morfologi dengan hadiran kata yang mempunyai penambahan atau imbuhan (Pierrehumbert & Granell, 2019). Oleh hal yang demikian, kajian ini telah melakukan eksperimen terhadap data korpus teks klasik Melayu bagi meneliti produktiviti kata pada aras *hapaxes*. Keupayaan teks klasik mengandungi variasi kata tinggi seperti, 58% *hapaxes* dalam kesusasteraan *shakespeare's* (Lardilleux & Lepage, 2007). Hal yang sama

turut berlaku dalam teks klasik yang lain iaitu lebih daripada 50% *hapaxes* seperti dalam *William Blak e's Poem* (Davis, 2019). Keupayaan *hapaxes* dalam mengekspresikan variasi kata sangat tinggi berbanding ketetapan nilai kekerapan tinggi, sederhana dan rendah. Sebaliknya, *hapaxes* sebagai nilai kemunculan sekali memberikan nafas baharu dalam mencari sumber kata yang unik. Signifikan kajian ini, integrasi antara aplikasi komputer dan bahasa digunakan bagi menganalisis teks (kata) bagi melihat trend taburan kata yang berlaku dalam teks klasik Melayu. Hal ini memberikan pendedahan terhadap teks klasik Melayu dalam eksperimen kata unik yang hadir dengan kekerapan sekali. Di samping itu, kajian ini melibatkan data yang empirikal tidak bias terhadap data yang dikaji secara menyeluruh. Pendedahan ini mampu menunjukkan variasi kata yang jarang digunakan di samping menunjukkan fenomena kata akar divariasikan berdasarkan pembentukan kata.

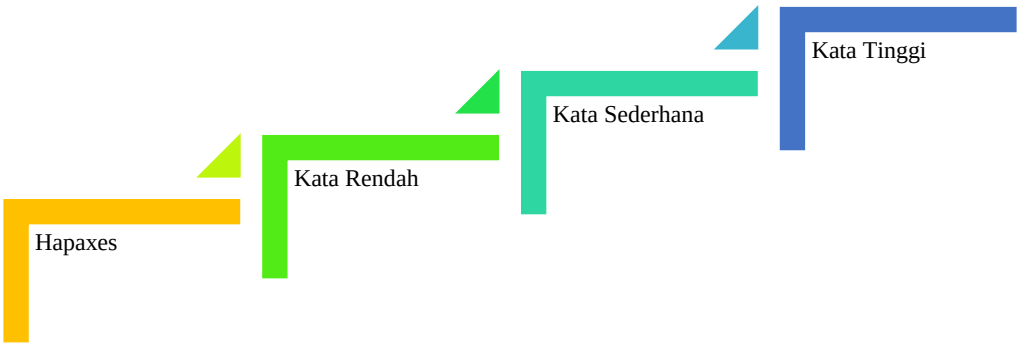
METODOLOGI

Kajian ini menggunakan reka bentuk penyelidikan kuantitatif iaitu statistik linguistik korpus. Statistik Linguistik Korpus ini mengimbangi kehadiran kata berdasarkan nilai kekerapan dan nilai signifikan. Kajian telah mengaplikasikan prosedur berpandukan korpus dengan menggunakan senarai kata yang telah dijana. Senarai kata merupakan himpunan kata dengan pemakaian istilah *type* (Michael P, 1998). Hasil janaan ini membentuk kehadiran *hapaxes* dalam rajah lekuk penyesuaian logaritma 10. Kajian ini telah menggunakan aplikasi *#LancsBox 5.0* (Brezina, 2018). Rajah 1 menunjukkan tingkap perisian #LancsBox 5.0. Aplikasi ini merupakan penilaian bagi eksperimen statistik kata. Pemakaian aplikasi dalam kajian ini memberi fokus kepada sub-tingkap *Words*. Penjanaan data menerusi sub-tingkap ini ialah analisis senarai kata dan analisis kata kunci. Penggunaan analisis senarai kata dijana dengan automatik berpandukan kekerapan kata secara relatif atau mentah.



RAJAH 1. Perisian #LancsBox 5.0

Kehadiran *hapaxes* ditentukan berdasarkan nilai 1.000000 dalam kekerapan mentah (Weeber, Baayen, & Vos, 2000).



RAJAH 2. Aras senarai kata

Rajah 2 menunjukkan aras senarai kata yang asas berdasarkan saiz data korpus. Aras ini berguna untuk melaraskan kedudukan keseluruhan nilai kekerapan yang muncul dalam data korpus. Sebelum ini keperluan hanya dilihat berdasarkan 100, 50, 30 atau 10 kata tertinggi (Crawford & Csomay, 2016). Representasi kata boleh diteliti berdasarkan nilai kekerapan seperti Kekerapan Mentah > 100 atau Kekerapan Relatif > 50.000000. Hal ini dapat menunjukkan indeks kata yang terlibat dalam penelitian nilai kekerapan. Signifikan penggunaan nilai ini membolehkan klasifikasi kata diukur berpandukan kekerapan kata sama ada mentah atau relatif. Berdasarkan nilai senarai kata dalam Korpus Petua Membina Rumah (KPMR) terdapat empat aras kekerapan kata (rujuk Jadual 1).

JADUAL 1. Nilai aras senarai kata KPMR

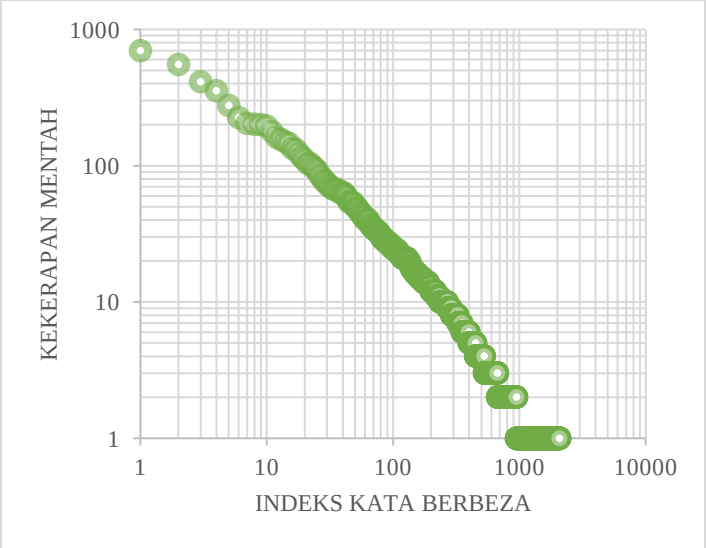
Nilai Aras	Aras
> 100	Kata Tinggi
100 > 50	Kata Sederhana
50 > 10	
10 > 2	Kata Rendah
1	*Hapaxes*

Jadual 1 menunjukkan nilai aras senarai kata KPMR yang kira berdasarkan saiz korpus. Nilai aras ini berpandukan nilai kekerapan mentah senarai kata. KPMR merupakan Korpus Petua Membina Rumah yang telah dihimpunkan menjadi korpus khusus. Petua Membina Rumah merupakan judul teks klasik berdasarkan katalog koleksi Perpustakaan Negara Malaysia. Terdapat tujuh teks klasik dipilih bagi kajian ini. Antara teks klasik yang terlibat MSS741, MSS2001, MSS1849, MSS1521, MSS1415, Fasal Kitab Abu Mahsyar dan Fasal Tajul Muluk. Secara keseluruhan korpus ini mengandungi 14648 *token*, 2079 *type* dan 14.19 nisbah *type token*. Manuskrip ini telah di transliterasi oleh Abdul Rahman Al-Ahmadi merupakan Mualim Tamu Perpustakaan Negara Malaysia pada 1997 hingga 1998. Transliterasi ini telah disemak semula oleh Muhammad Anas Al-Muhsin pada tahun 2019 sebagai keperluan geran FRGS 2019-0036-108-02. Rajah 3 menunjukkan contoh teks klasik Melayu berjudul Petua Membina Rumah dengan kod MSS741. Kod MSS merupakan nombor teks klasik yang ditetapkan oleh Perpustakaan Negara Malaysia. Hal ini menunjukkan pemilikan teks klasik koleksi Perpustakaan Negara Malaysia.

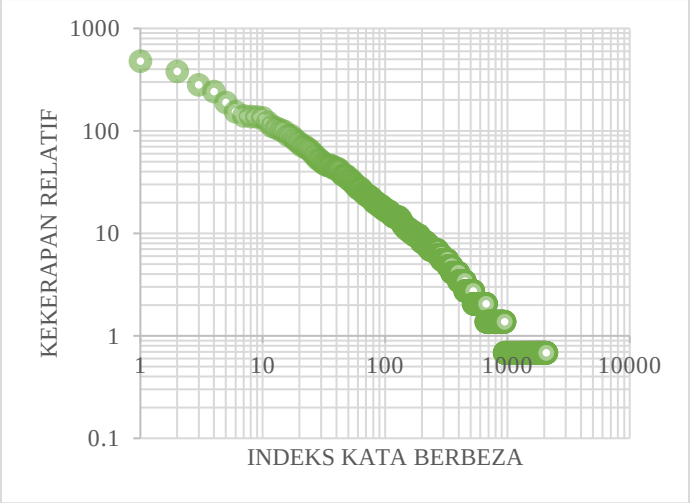
RAJAH 3. Contoh teks klasik Melayu MSS741

KEPUTUSAN

Rajah 4 menunjukkan trend kekerapan mentah teks klasik atau manuskrip petua membina rumah. Manakala Rajah 5 menunjukkan trend kekerapan relatif yang telah dinormalisasikan per 10 000 perkataan (automatik keterbacaan mesin). Kedua-dua rajah ini menggunakan representasi kekerapan signifikan dijana menerusi graf *scatter*. Graf ini menunjukkan taburan kekerapan kata yang berlaku dalam struktur teks petua membina rumah. Kekerapan kata paling tinggi adalah kata hubung *dan* iaitu kekerapan mentah 697.000000 dan kekerapan relatif 475.832878 per perkataan.



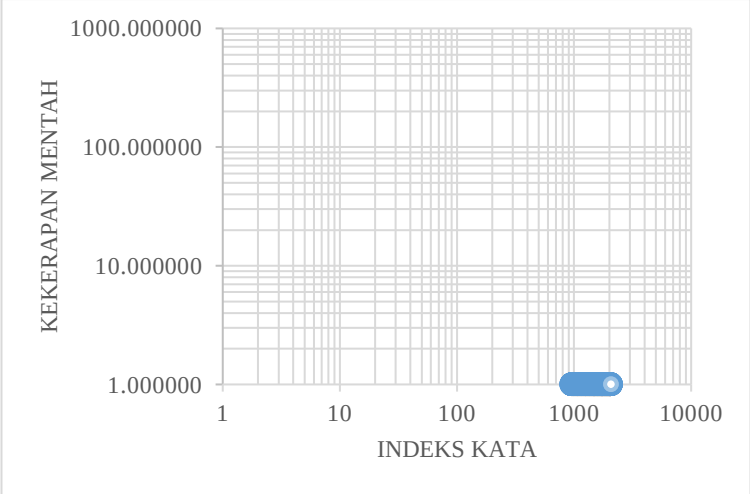
RAJAH 4. Representasi trend kekerapan mentah KPMR



RAJAH 5. Representasi trend kekerapan relatif KPMR

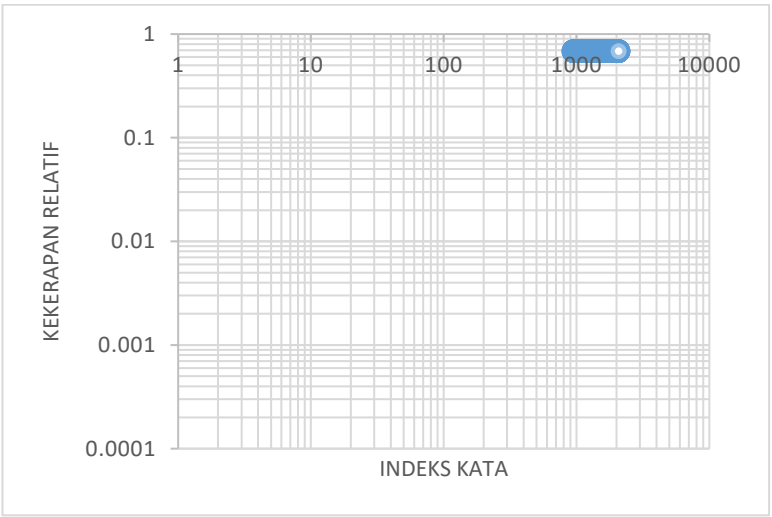
Berdasarkan Rajah 4 dan 5 menunjukkan bentuk pertumbuhan lekuk *type-token* yang dijana bagi memperlihatkan kata yang jarang berlaku dan sering berlaku. Pertumbuhan lekuk ini memberikan satu jangkaan terhadap kekerapan kata dalam teks klasik di mana kewujudan jurang antara indeks kata. Rumusan ini memberi satu tanggapan mudah untuk bacaan kekerapan kata yang dijana menerusi intuitif logaritma = 10 dalam pertumbuhan lekuk *type-token*. Representasi kata secara menyeluruh ini memberikan bentuk kemunculan kata pada indeks tinggi, sederhana dan rendah. Representasi ini merupakan bentuk normal dalam taburan kekerapan kata dalam sesebuah kata. Kata dinilai dengan representasi tertinggi sehingga rendah membentuk satu pertumbuhan graf lekuk yang menjelaskan pemboleh ubah yang berperanan dalam plot representasi kata ini. Nilai kekerapan tertinggi boleh dilihat dengan representasi kekerapan mentah atau relatif sebanyak 100 kali kekerapan ke atas. Kemudian, nilai kekerapan sederhana boleh dilihat menerusi dua penilaian iaitu 100 kali hingga 50 kali kekerapan dan 10 kali hingga 50 kali kekerapan. Akhir sekali, nilai *hapaxes* yang berada pada tahap 1 kali kekerapan kemunculan pada kekerapan mentah. Bagi kekerapan relatif nilai yang telah dinormalisasikan 1 kali kekerapan iaitu 0.682687. Representasi *hapaxes* sangat jelas apabila penelitian terhadap Rajah 4 iaitu kekerapan relatif yang melepasi y = 0 ke bawah. Kekerapan kata pada aras > 100 memerihalkan bentuk kata yang kerap digunakan dalam teks klasik ini. Kekerapan ini dipanggil sebagai kekerapan tertinggi yang melibatkan pelbagai variasi kata sama ada kata hubung dan kata kandungan. Rajah 5 menunjukkan representasi kekerapan relatif bagi > 100 kata. Representasi ini menunjukkan > 100 kata yang terdapat dalam teks klasik.

Rajah 6 dan 7 menunjukkan kedudukan representasi *hapaxes* berdasarkan kekerapan mentah (graf atas) dan kekerapan relatif (graf bawah). Nilai kekerapan mentah ialah 1.000000 dan kekerapan relatif 0.682687 berkedudukan di bawah (akhir).



RAJAH 6. Representasi kekerapan mentah *hapaxes*

Manakala Rajah 6 menunjukkan secara terperinci kedudukan kekerapan mentah yang berada pada nilai y= 1.000000 dan dominan sifatnya kerana merangkumi 54% kata sama ada tersalah eja atau neologisme. Rajah 7 menunjukkan kekerapan relatif yang berlaku pada aras ini 0.682687 di bawah paras 1.000000 setelah dinormalisasi. Walaupun representasi kata ini berada pada aras yang rendah dan kekerapan sekali dalam KPMR, penelitian terhadap perbendaharaan kata dan sintesis boleh dilihat. Hal ini menunjukkan penggunaan bahasa oleh sarjana manuskrip terdahulu bersifat luas dan boleh pelbagaikan variasi kata yang dibentuk.



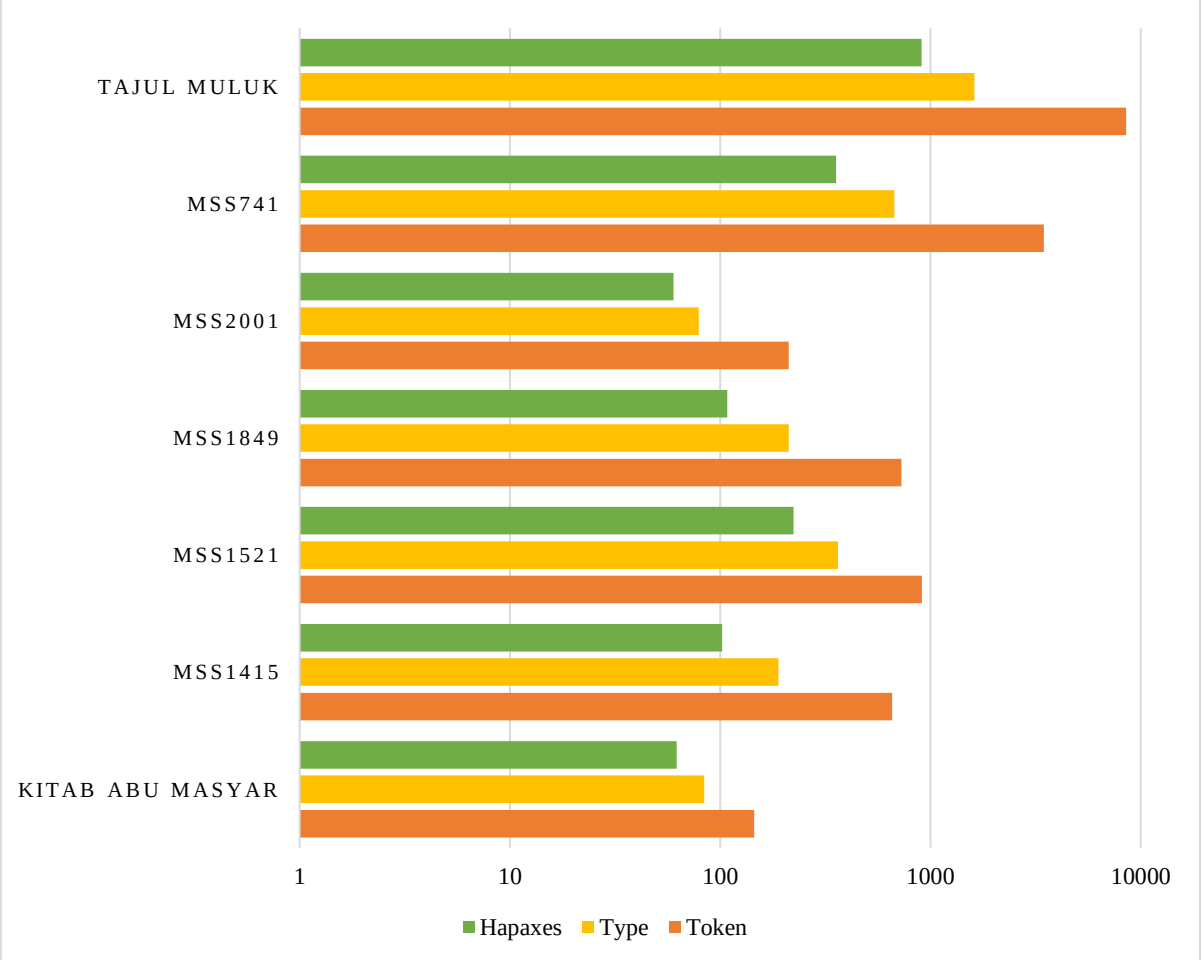
RAJAH 7. Representasi kekerapan relatif *hapaxes*

PERBINCANGAN

Hapaxes atau hapax legomena merupakan bentuk bahagian kata yang berkedudukan 1.000000. Bentuk ini dapat dijana berdasarkan janaan data korpus melalui senarai kata (kekerapan mentah dan kekerapan relatif). Morton (1978) telah menerangkan bagaimana kelas kata terbesar dalam perbendaharaan kata teks adalah merujuk kepada kata yang jarang berlaku atau berlaku sekali (hapax legomena). Kehadiran *hapaxes* merupakan bentuk menarik yang perlu diteliti. Hal ini kerana, *hapaxes* merupakan cerminan beberapa faktor seperti latar

kehidupan, pengalaman, kuasa pemikiran sarjana/penulis. Perkara ini dapat menyampaikan kehalusan dalam makna.

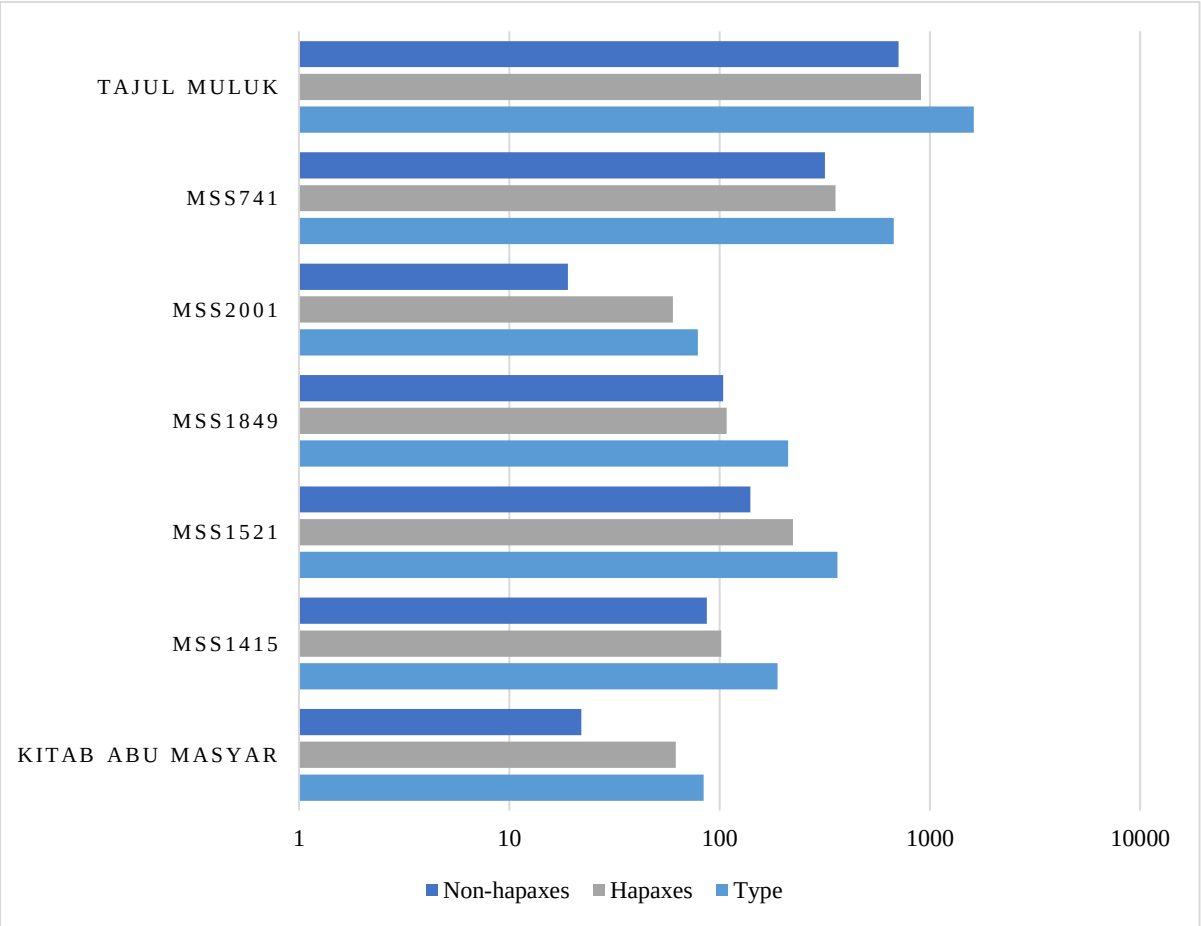
Kebanyakan kajian lepas hanya memerhatikan *rare words* tidak bagi *hapaxes*. Sekilas, kajian teks klasik sama ada di eropah banyak meneliti *hapaxes* bagi memperoleh variasi kata Inggeris lama atau kata natif lama (e.g. Davis, 2019; Domazakis, 2010; Pierrehumbert & Granell, 2019). *Hapaxes* bukan sahaja mengetengahkan neoligsm sebaliknya memerincikan kumpulan kata ini terhadap kesalahan ejaan dan variasi istilah (J. A. Smith & Kelly, 2002). *Hapaxes* merupakan nafas baru bagi meneliti teks klasik Melayu untuk mendapatkan kelompok besar kata yang digunakan dalam sesuatu bidang. Terutama dalam bidang seni bina Melayu lama. Tidak banyak teks klasik Melayu lama dibedah bagi mendapatkan variasi kata. Oleh hal demikian, kajian ini menggunakan *hapaxes* bagi mendapatkan kata Melayu lama yang digunakan oleh sarjana terdahulu. Kajian ini menggunakan tujuh (7) teks klasik Melayu di bawah koleksi simpanan Perpustakaan Negara Malaysia. Berikut merupakan maklumat maklumat bagi data korpus Petua Membina Rumah (KPMR);



RAJAH 8. Maklumat Token, Type dan *Hapaxes*

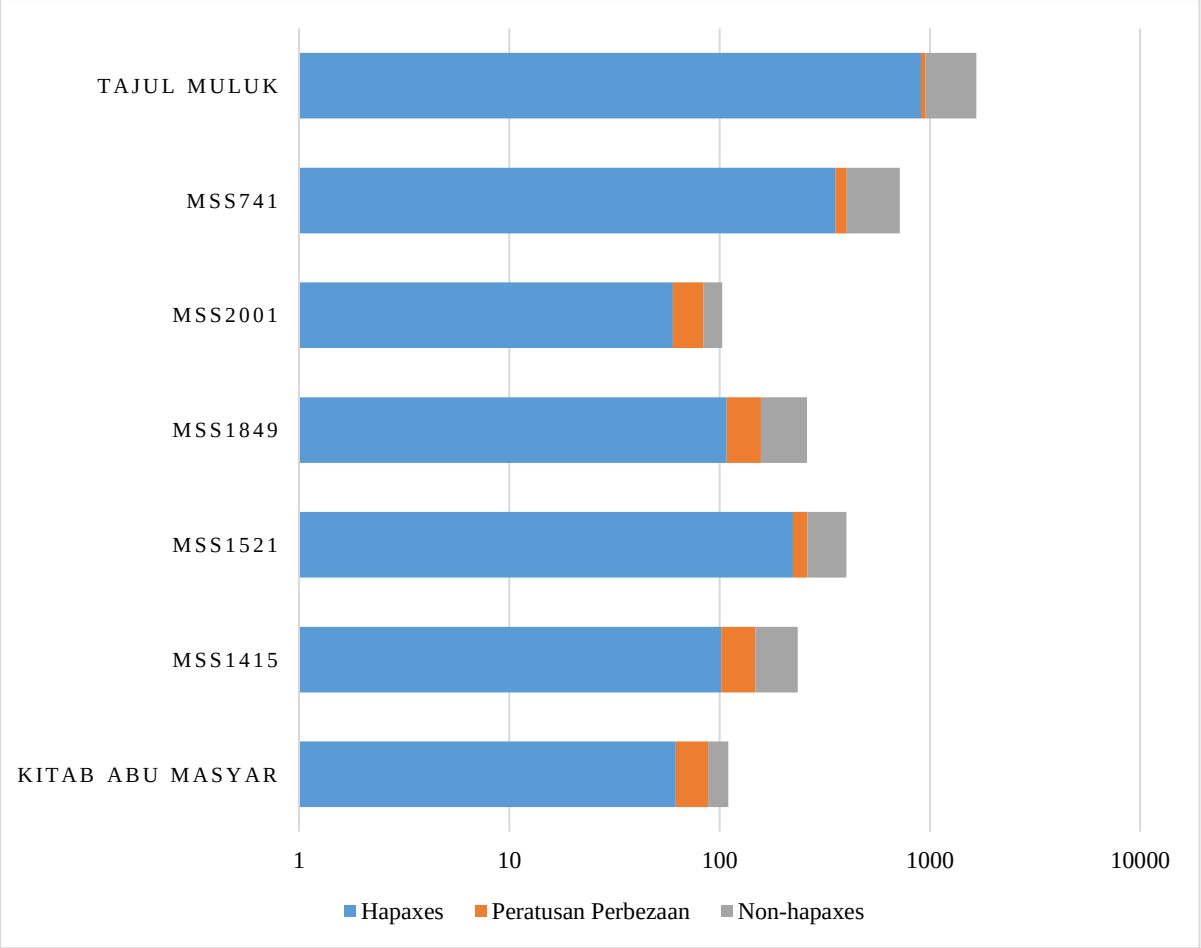
Berdasarkan Rajah 8, menunjukkan maklumat statistik am bagi korpus Petua Membina Rumah (KPMR). Skala graf ini mengikut keselarasan *logarithmic* dengan asas 10. Dalam graf ini menunjukkan bahawa nilai *hapaxes* menghampiri *type* (jenis kata). Nilai

hapaxes yang tinggi adalah 908 kata (Tajul Muluk), 356 kata (MSS741), 223 kata (MSS1521), 108 kata (MSS1849), 102 kata (MSS1415), 62 kata (Kitab Abu Mahsyar), dan 60 kata (MSS2001). Melalui nilai *hapaxes* berdasarkan bahagian-bahagian teks klasik ini dapat diperhatikan jumlah kata non-*hapaxes* adalah sederhana. Kata non-*hapaxes* yang berlaku dalam Tajul Muluk (711), MSS 741(317), MSS2001 (19), MSS1849 (104), MSS1521 (140), MSS1415 (87), dan Kitab Abu Mahsyar (22). Nilai non-*hapaxes* kurang daripada 50% dan bersifat sederhana. Hal ini dapat menunjukkan pola penulis dalam pemakaian kata tidak berulang. Rajah 9 menunjukkan perbandingan *hapaxes* dan non-*hapaxes* berdasarkan Korpus Petua Membina Rumah (KPMR).



RAJAH 9. Perbandingan *Hapaxes* dan Non-*hapaxes*

Perbezaan nilai *hapaxes* dan non-*hapaxes* dapat dilihat berdasarkan jurang yang berlaku dalam *type* (jenis kata) (rujuk Rajah 9).



RAJAH 10. Perbezaan Nilai *Hapaxes* dan Non-*Hapaxes*

Rajah 10 menunjukkan nilai perbezaan antara *hapaxes* dan non-*hapaxes*. Semua nilai jurang ini dibawah aras 50%. Jurang yang tinggi ialah 49.05% (MSS1849) dan jurang yang rendah 24.05% (MSS2001). Jurang ini berperanan bagi mengenal pasti produktiviti kata Melayu lama yang tidak mengalami pengulangan. Hal ini dapat meningkatkan pemerolehan variasi kata Melayu dalam bahagian-bahagian teks klasik Melayu ini. Terdapat empat bahagian teks klasik Melayu yang mempunyai jurang 40% ke atas iaitu Tajul Muluk, MSS741, MSS1849 dan MSS1415. Manakala satu jurang 38.56% dalam MSS1521 dan dua jurang pada aras 26.19% (Kitab Abu Mahsyar) dan 24.05 (MSS2001). Hasil daripada keputusan ini mendapatkan kehadiran *hapaxes* dalam teks klasik Melayu memberi impak positif terhadap penelitian variasi kata.

Hapaxes merupakan kata yang sering dikaji dalam teks klasik atau sastera (Davis, 2019). Dalam kajian Davis (2019) menyatakan hampir 50% *hapaxes* berlaku dalam teks sastera. Di samping itu, peratusan yang sama (50% hingga 80%) juga sering berlaku dalam teks umum (Lardilleux & Lepage, 2007). Penyelidikan kata Inggeris lama, *hapaxes* digunakan bagi mengenal pasti potensi kata dalam wacana (Stiles, 2019). Berdasarkan *hapaxes* dalam KPMR boleh diteliti melalui kata kandungan iaitu kata nama, kata kerja dan kata adjektif dalam bahasa Melayu. Berikut merupakan *hapaxes* kata kandungan yang terdapat dalam KPMR;

JADUAL 2. *Hapaxes* dalam KPMR

Hapaxes	*Teks Klasik/Kekerapan Relatif*	*KWIC*
[pelapit]	MSS1521 (1.000000/10.989012)	*Kain putih maka bubuh diberi bunga tiang* ***pelapit*** *helan maka inilah disuratnya pada kain putih*
[siputkan]	Tajul Muluk (1.000000/1.1724703)	*Lipat maka letak atas pucuk tiang maka* ***siputkan*** *tiang itu dan panjang dan ukuran bendul*
[gentala]	MSS1521 (1.00000/10.989012)	*Tersebut itu pertama lembu tiang tempatnya kedua* ***gentala*** *naga ketiga gajah sepayang yang keempat tulang*
[perkelahian-perkelahian]	MSS1849 (1.000000/13.736264)	*Rumah nescaya empunya rumah itu terlalu jahat* ***perkelahian-perkelahian*** *lagi kehilangan dan pada bulan ramadhan mendirikan*
[sayogialah]	MSS741 (1.000000/2.8860028)	*Jika kena pada belakangnya banyak tawar maka* ***sayogialah*** *kita ingat pada barang suatu yang diperbuat*

Jadual 2 menunjukkan beberapa *hapaxes* yang berlaku secara keseluruhan dalam KPMR. Kata ini merupakan gabungan beberapa kelas kata seperti kata tunggal dan mengalami perubahan kata (kata nama kepada kata kerja). Kehadiran imbuhan sebagai *hapaxes* antara faktor yang berlaku pemakaian kata sekali. Faktor imbuhan memberi penekanan terhadap variasi kata dan fungsi yang berbeza daripada kata tunggal yang berlaku. Faktor ini dikenali sebagai *word-formation pattern* yang berlaku dalam penutur Itali dengan kehadiran *verb-nouns* (Štichauer, 2016). Perkembangan perbendaharaan ini sering berlaku dalam *hapaxes*. Berdasarkan *word-formation* yang berlaku dalam KPMR beberapa imbuhan boleh diperhalusi. Kehadiran bentuk kata sedemikian juga menyumbang kepada variasi kata kerana tidak berdiri dengan fungsi makna atau pemakaian yang sama. Sebaliknya, representasi kata bersifat sedemikian memberikan bentuk variasi kata. Bentuk variasi kata ini ditunjukkan melalui KWIC dengan saiz tingkap 7 kiri dan 7 kanan serta nilai ko-kejadian 0.68. Bentuk variasi kata yang dipengaruhi oleh imbuhan atau penambahan seperti kata akar; *ukur* (kata ini tidak di senarai dalam *hapaxes*). Kata *ukur* mengalami pembentukan kata dengan kehadiran imbuhan iaitu, kata *diukur*, *mengukur*, *mengukurkan*, *seukuran*, dan *ukurlah*. Kelima-lima kata ini merepresentasikan nilai 1.000000 (kekerapan mentah) dan 0.682874 (kekerapan relatif). Jadual 3 menunjukkan KWIC bagi variasi kata *ukur* yang melalui penambahan/imbuhan.

JADUAL 3. KWIC variasi kata *ukur*

Indeks	**Fail**	**Kiri**	**Nod**	**Kanan**
1	TAJUL MULUK.txt	*dipihak lubang maka rasuk maka ambil tali*	*diukur*	*sama panjang tiang itu maka tali itu*
1	TAJUL MULUK.txt	*dan kayu sepayang dan setengah hukama jika*	*mengukur*	*pada bendul di dalam tiang daripada luar*
1	TAJUL MULUK.txt	*tambak cerungnya. adapun seorang al hakim pada*	*mengukurkan*	*tiang rumah hendak dipihak lubang maka rasuk*
1	TAJUL MULUK.txt	*likuran panjang tiang seri serambi basah itu*	*seukuran*	*dengan serambil itu juga tamat. bab ini*
1	MSS1521.txt	*tiga maka semua diambil enam bagi maka*	*ukurlah*	*tujuh lipat itu belitlah pula kedua ukuran*

Seterusnya, terdapat juga variasi kata jauh juga berlaku dalam *hapaxes*. Variasi imbuhan ini berada kedudukan sekali dalam senarai kata dengan nilai ko kejadian bersama

0.68 (bacaan yang sama bagi nilai *hapaxes*). Dalam variasi ini melibatkan imbuhan apitan, akhiran serta kata akar jauh turut berlaku dalam *hapaxes*. Jadual 4 menunjukkan variasi kata jauh dalam pemakaiannya KWIC.

JADUAL 4. KWIC variasi kata *jauh*

Indeks	Fail	Kiri	Nod	Kanan
1	TAJUL MULUK.txt	*rakaat memohonkan berkat pada allah ta'ala*	*menjauhkan*	*sekalian syaitan dan meluputkan daripada marabahaya dan*
1	MSS1521.txt	*sentosa dianugerahi allah ta'ala adapun belangan itu*	*dijauhkan*	*allah ta'ala maka yang adalah utama*
1	MSS741.txt	*adakan terbit dari dalam negeri pergi pergian*	*jauh*	*alamat akan beroleh laba dan kesukaan di*
1	MSS1521.txt	*tolong daripada allah dan rasullah minta beri*	*jauhkan*	*bala dari dunia dan akhirat supaya selamat*
1	MSS1521.txt	*pedas jika kediaman keluarganya sekalian alamat akan*	*berjauhan.*	

Selain itu, terdapat juga variasi kata kelahi iaitu berkelahi, kelahian dan perkelahian-perkelahian. Jadual 5 menunjukkan variasi kata dalam KWIC. Nilai ko kejadian adalah 0.68.

JADUAL 5. KWIC variasi kata *kelahi*

Indeks	Fail	Kiri	Nod	Kanan
1	MSS741.txt	*Itu dan segala orang besar-besar akan jadi*	*berkelahi*	*Hujung tahun padahnya. Bab jikalau pada tahun*
1	TAJUL MULUK.txt	*Tiada beroleh arta akan beroleh penyakitan dan*	*kelahian*	*Padahnya dan kedua pintu ini tiada beroleh*
1	MSS1849.txt	*Rumah nescaya empunya rumah itu terlalu jahat*	*Perkelahian-perkelahian*	*Lagi kehilangan dan pada bulan ramadhan mendirikan*

Akhir *hapaxes* yang mengalami pembentukan imbuhan bagi variasi kata adalah kata akar *bayang* dan *likur*. Kata akar juga berkedudukan sekali bersama kata yang mengalami penambahan. Manakala, Jadual 6 merupakan variasi kata *likur* dalam KWIC. Imbuhan yang terlibat iaitu, imbuhan akhiran –nya dan -an.

JADUAL 6. KWIC variasi kata *likur*

Indeks	Fail	Kiri	Nod	Kanan
1	MSS741.txt	*atau mana-mana buat sampai sekelian pada dua*	*likur*	*haribulan jamadilakhir pada hari ahad hendaknya dari*
1	TAJUL MULUK.txt	*basah itu luasnya sehasta jari manusia dan*	*likuran*	*panjang tiang seri serambi basah itu seukuran*

Dapatan *hapaxes* dalam KPMR boleh diteliti dengan melihat fenomena kesilapan ejaan. Fenomena kesilapan ejaan boleh dilihat dalam pelbagai variasi kata sama ada morfologi atau berlaku penggabungan kata disebabkan teknikal. Terdapat 21% kesilapan ejaan yang dikenal pasti. Antara contoh kata yang mengalami masalah kesilapan ejaan segi morfologi iaitu, *biiangannya, sekal, jalali, sapai, paitlah, putak, bulang, dariada, rantangnya, zuhri, lat, mujrab, berganjat, sunah, itlilah, kedaksin, karku, prain, helan, pipin,*

ternai, angsatu, jaknya, hayam, rubuh, dakvva, iagi, jurian, hukama, iain, rajona, iilik, berung, quluh, berai, iyyaka, berawali, inni, bares, maplung, sayogiahlah, maaratul, tua-tiia, sanat, perabun, rifatuka, tingga, mehajat, petas, dan pelbagai lagi bentuk kesilapan ejaan kata. Kesilapan ejaan secara menyeluruh dapat dilihat daripada segi kata dasar, pengguguran morfem dan huruf. Selain itu, terdapat juga fenomena kata dari segi jarak antara kata yang mengalami penggabungan (tanpa jarak) sebanyak 1% iaitu, *tiadajadi, isteridan, berpusupusu, dudukjika, mengejutngejut, pangkatkafir, kesurutsurut, banyakbanyak, makajangan, diperbuatnasi, katalalu, barangsiapa, berbantahbantah, tuatua, dilepaslepas, sedangjuga* dan *adalahpenuh*. Bentuk kesilapan ejaan ini sering berlaku akibat daripada kecuaian pembersihan data, kesilapan transliterasi dan perbezaan variasi semasa sehingga berlaku pengguguran huruf atau penggabungan kata (Pierrehumbert & Granell, 2019).

Terdapat 78% perbendaharaan kata atau *neologisme* yang terdapat dalam kelompok *hapaxes*. Perkataan ini terdiri daripada variasi kata yang muncul sekali dalam KPMR tanpa pengulangan atau kekerapan. Antaranya contoh kata *pelapit, timah, keluang, kencang, berpelaba, rambung, siputkan, hastakan, darjatnya, miang, sang, lobangnya, keramat, syarahnya, haru, belitlah, rembang, kebaktian, perkelahian-perkelahian, garung, terajah, tambak, cupak, likur, muluk, kepiatuan, galang, cengal, laraknya, uting, beliang, perabun, kepapara, pemeleh, setapang, tantan, pinta, tebuk, bertempok, warta, tembikar, sejengkal, sedepa, berhuma, kapit, buruj, mustari, jintan, sekanak, ditohot, penelur, manjung, sepilok, setar, masyirik, berbueh,bueh, candak, berseteru-seteru* dan lain-lain variasi kata.

Fungsi neologisme menunjukkan bentuk kata yang digunakan sama ada baharu atau jarang digunakan (Fontaine, 2017). Dapatan menunjukkan kata kandungan kata nama dan kata kerja banyak digunakan pada aras ini. Kata nama seperti nama tumbuhan, objek, dan haiwan juga muncul pada aras ini. Antara kata yang menunjukkan nama-nama tumbuhan seperti *rembang, galang, uting, cengal, jintan* dan *setapang*. Kata kerja dengan berlakunya pembentukan dan perubahan kata menunjukkan kata ini masih digunakan. Misalnya kata *hastakan, belitlah, perkelahian-perkelahian, bertempok, ditohot, berhuma,* dan *berbueh-bueh.* Di samping itu, kata fungsi juga berlaku dengan melibatkan kata bilangan seperti se- menunjukkan satu bagi subjek bersifat ukuran. Sebagai contoh, *sejengkal, sedepa, sekanak,* dan *sepilok.* Peranan *hapaxes* sangat penting bagi representasi kekayaan sumber kata yang terkandung dalam KPMR dapat ditonjolkan. Kata unik ini merupakan bentuk baharu yang jarang digunakan sama ada literal mahupun non-literal (Zaini, Sarudin, Muhammad, & Abu Bakar, 2020).

KESIMPULAN

Berdasarkan kajian ini, dapat diteliti bahawa produktiviti kata hadir berdasarkan aras penilaian. Aras penilaian kekerapan ini merepresentasikan empat aras kata iaitu kata tinggi, kata sederhana, kata rendah dan *hapaxes*. Representasi ini menunjukkan kemunculan indeks kata bagi setiap aras penilaian kekerapan. Hal ini membolehkan teks klasik Melayu ini diperhalusi. Kemunculan kata pada aras tinggi, sederhana dan rendah menunjukkan keupayaan perbendaharaan kata yang kerap digunakan. Manakala, keunikan kajian ini fokus kepada dapatan *hapaxes* yang menunjukkan ragam atau plot kata dengan kemunculan sekali. Hasil daripada penelitian dapatan tersebut menunjukkan *hapaxes* muncul pada aras 50% hingga 80% iaitu 54% berbanding non-*hapaxes*. Hal ini membuktikan bahawa separuh daripada kata teks klasik Melayu ini terdiri daripada *hapaxes* dan bersifat unik. Terdapat beberapa faktor yang menyumbang kepada *hapaxes* iaitu, variasi kata, kesilapan ejaan dan

neologisme. Variasi kata menzahirkan dapatan kata yang mengalami pembentukan kata berdasarkan kata akar kepada kata terbitkan dianggap sebagai *hapaxes*. Hal ini memberikan impak kepada kata yang mempunyai berlainan fungsi makna dan pemakaiannya. Oleh hal yang demikian, kajian ini merupakan penelitian baharu dalam kajian statistik korpus yang memfokuskan kepada penilaian aras produktiviti kata. Signifikan ini telah menunjukkan bahawa kupasan teks klasik perlu diperhalusi dengan kemunculan indeks dan nilai kekerapan kata. Perisian korpus linguistik iaitu #LancsBox sebagai *machine-readable* data korpus. Hal ini juga menjawab kepada konstruk mental sarjana Melayu feudal yang kaya dengan variasi kata secara statistik kata. Kekayaan variasi kata ini mencerminkan teknik bahasa dalam penulisan dan pemikiran sarjana.

PENGHARGAAN

Kajian ini telah dijalankan di bawah Skim Geran Penyelidikan Fundamental (FRGS / 1/2018 / WAB04 / UPSI / 02/5) yang disediakan oleh Kementerian Pendidikan Malaysia.

RUJUKAN

Ahmad, M., Lukman, A. M., & Yusof, N. M. (2016). Manuskrip Kehakiman Islam : Analisis Keupayaan dan Kearifan Melayu Islamic Judiciary Manuscript : Analysis on Malay Intellectual and Ability. *Jurnal Sultan Alauddin Sulaiman Shah*, *3*(2), 251–267.

Amer Hudhaifah, H. (2017). Manuskrip Melayu : Isu Kontemporari dan Lontaran Idea. *International Journal of West Asian Studies*, *9*(1), 25–38. https://doi.org/10.22583/ijwas.2017.09.01.03

Boytsov, L. (2017). *A Simple Derivation of the Heap's Law from the Generalized Zipf's Law*. 1–5. Retrieved from http://arxiv.org/abs/1711.03066

Bradac, J. J., Davies, R. A., Courtright, J. A., Desmond, R. J., & Murdock, J. I. (1977). Richness of Vocabulary: An Attributional Analysis. *Psychological Reports*, *41*(3_suppl), 1131–1134. https://doi.org/10.2466/pr0.1977.41.3f.1131

Brezina, V. (2018). Statistics in Corpus Linguistics. In *Cambridge University Press* (First). https://doi.org/10.1017/9781316410899

Crawford, W. J., & Csomay, E. (2016). *Doing Corpus Linguistics* (First Edit). New York: Routledge.

Davis, V. (2019). Types, Tokens, and Hapaxes: A New Heap's Law Abstract. *Glottotheory*, *9*(2), 113–129. https://doi.org/10.1515/glot-2018-0014

Delgado, U., Ángela, S. Y., & García, H. (2007). Zipf's and Heap's law. *Acta Colombiana De Psicología*, *10*(2), 9–17. Retrieved from http://editorial.ucatolica.edu.co/ojsucatolica/revistas_ucatolica/index.php/acta-colombiana-psicologia/article/view/207

Ding Choo Ming. (2016). *Manuskrip Melayu: Sumber Maklumat Peribumi Melayu* (Four). Bangi: Universiti Kebangsaan Malaysia Press.

Domazakis, N. (2010). *Septuagintal Hapax Legomena and Neologisms in 2 Maccabees, 4-7*. Retrieved from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=2441258&fileOId=2441314

Faisal, A. H. F. @ A., & Wahidah, F. N. (2012). Perubatan Melayu Tradisional: Kitab Tibb Pontianak. *Journal of Al-Tamaddun*, *7*(1), 149–162. https://doi.org/10.22452/jat.vol7no1.10

Filzah, I., Mat, N. F. C., Wan, I. W. N., & Firdaus, H. M. (2018). Kearifan tempatan dalam kosa kata

persenjataan melalui kajian manuskrip MS31: satu pengenalan. *AIJLLS*, *2*(5), 115–135.

Fontaine, L. (2017). The early semantics of the neologism BREXIT: a lexicogrammatical approach. *Functional Linguistics*, *4*(1). https://doi.org/10.1186/s40554-017-0040-x

Ghani, H. A. (2015). *MSS2999 Kitab tib pandangan dan tafsiran perubatan moden terhadap manuskrip perubatan Melayu* (First). Kuala Lumpur: Institut Penyelidikan Perhutanan Malaysia.

Hassan, A. (2016). *Bahasa dan Pemikiran Melayu* (First). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Hernández-Campoy, J. M., & Conde-Silvestre, J. C. (2012). *The Handbook of Historical Sociolinguistics* (First). West Sussex, United Kingdom: Blackwell Publishing Ltd.

Lardilleux, A., & Lepage, Y. (2007). The contribution of the notion of hapax legomena to word alignment. *English*, (February), 458–462.

Lardilleux, A., & Lepage, Y. (2009). Hapax legomena: Their contribution in number and efficiency to word alignment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5603 LNAI*, 440–450. https://doi.org/10.1007/978-3-642-04235-5_38

Lü, L., Zhang, Z. K., & Zhou, T. (2013). Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. *Scientific Reports*, *3*, 1–7. https://doi.org/10.1038/srep01082

Michael P, O. (1998). Statistics for corpus linguistics edinburgh textbook in empirical linguisic. In *Edinburgh University Press*. Edinburgh: Edinburgh University Press.

Pierrehumbert, J., & Granell, R. (2019). *On Hapax Legomena and Morphological Productivity*. 125–130. https://doi.org/10.18653/v1/w18-5814

Riswadi, A., & Mustaffa, A. (2017). Manuskrip al-Qur'an di Alam Melayu: Kajian Terhadap Manuskrip al-Qur'an Terengganu. *Journal of Usuluddin*, *45*(2), 19–54. https://doi.org/10.22452/usuluddin.vol45no2.2

Safwan, R., & Zubir, I. (2018). Gambaran masyarakat Melayu tradisional dalam Syair Putera Mahkota. *Jurnal Melayu*, *17*(2), 225–240.

Smith, J. A., & Kelly, C. (2002). Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Language Resources and Evaluation*, *36*(4), 411–430.

Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. *English for Specific Purposes*, *57*, 1–12. https://doi.org/10.1016/j.esp.2019.08.002

Štichauer, P. (2016). Verb-noun compounds in Italian from the 16th century onwards: an increasing exploitation of an available word-formation pattern. *Morphology*, *26*(2), 109–131. https://doi.org/10.1007/s11525-015-9274-z

Stiles, P. (2019). Beowulf 33a and Hapax Legomena. *Neophilologus*, *104*(2), 255–261. https://doi.org/10.1007/s11061-019-09621-w

Sukawai, E., & Omar, N. (2020). Corpus Development for Malay Sentiment Analysis Using Semi Supervised Approach. *Asia-Pacific Journal of Information Technology and Multimedia*, *09*(01), 94–109. https://doi.org/10.17576/apjitm-2020-0901-08

Tiun, S., Abdullah, R., Kong, T. E., & Muhammad, S. K. (2013). *Korpus pertuturan sintaksis-prosodi bahasa melayu*. *2*(1), 1–12.

Wan Mohd Dasuki, W. H., & Radziah, M. S. (2014). Manuskrip Ilmu Bedil Sebagai Sumber

Etnosejarah Teknologi Senjata Api Melayu (Malay Manuscripts on Firearms as an Ethnohistorical Source of Malay Firearms Technology). *KEMANUSIAAN*, *21*(1), 53–71.

Weeber, M., Baayen, R. H., & Vos, R. (2000). Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics*, *26*(3). https://doi.org/10.1162/089120100561719

Yakob, M. A. (2018). *Keintelektualan Melayu dalam Manuskrip dan Kitab Jawi* (First). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Zaini, M. F., Sarudin, A., Muhammad, M. M., & Abu Bakar, S. S. (2020). Representasi Leksikal Ukuran sebagai Metafora Linguistik berdasarkan Teks Klasik Melayu (Representatives of Lexical Ukuran as Linguistics Metaphors Based on Malay Classic Text). *GEMA Online® Journal of Language Studies*, *20*(2), 168–187. https://doi.org/10.17576/gema-2020-2002-10

Zaytseva, V., Miralpeix, I., & Pérez-Vidal, C. (2019). Because words matter: Investigating vocabulary development across contexts and modalities. *Language Teaching Research*. https://doi.org/10.1177/1362168819852976

Muhamad Fadzllah Zaini
Mazura Mastura Muhammad
Anida Sarudin
Siti Saniah Abu Bakar
Zulkifli Osman
Fakulti Bahasa dan Komunikasi,
Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak.
muhamadfadzllahhjzaini@gmail.com, mazura@fbk.upsi.edu.my, anida@fbk.upsi.edu.my, saniah@fbk.upsi.edu.my, zulkifli@fbk.upsi.edu.my