TRANSFORMASI STATISTIK SENARAI KEKERAPAN KATA DALAM KAJIAN BERASASKAN KORPUS: MANIFESTO PILIHAN RAYA 2008

Imran Ho Abdullah

Abstrak

Artikel ini meneliti transformasi statistik maklumat kekerapan kata dalam kajian berasaskan korpus. Berdasarkan satu kajian kes terhadap manifesto pilihan raya 2008, penelitian perbandingan kekerapan, iaitu kekerapan mentah, kekerapan relatif, serta visualisasi data kekerapan antara manifesto BN, PKR, PAS dan DAP dilakukan. Seterusnya, penggunaan kekerapan untuk menjana kata kunci dengan menggunakan statistik ujian khi kuasa dua dan *log-likelihood.* Transformasi statistik data kekerapan kata menggunakan kaedah multivariat – analisis penghubungan juga dipelopori untuk melihat keberkesanan pelbagai kaedah statistik ini. Dapatan dan implikasi kajian menunjukkan transformasi statistik terhadap data kekerapan serta jenis statistik yang digunakan boleh mempengaruhi dapatan dan membawa kepada kesimpulan yang berbeza tentang hubung kait linguistik antara teks dalam sesuatu korpus.

Abstract

This article examines the statistical transformation of wordlist frequency in corpus based studies. Based on a case study of the 2008 General Election manifestos of four parties BN, PKR, PAS and DAP, comparisons are made between the manifestos based on raw frequencies, relative frequencies, and different techniques of frequencies visualisation. The article also compares the generation

of keyword (based on frequency data) using different statistical test such as chi-square and log-likelihood, and the transformation of the frequency data using a multi variate correspondence analysis to uncover the (linguistics) relationship between the different manifesto in question. The results reveal that using different statistical procedures can lead to different conclusion with regards to the linguistics relationships between the texts.

PENGENALAN

Senarai kekerapan kata merupakan sejenis data yang mudah diperoleh dan dijana dalam paradigma kajian linguistik dan bahasa berasaskan korpus. Kekerapan kata dijadikan sebagai asas untuk kajian linguistik seperti kajian variasi bahasa dalam sosiolinguistik dengan perbandingan korpus berdasarkan pengguna contohnya gender, kawasan, dialek dan umur. Juga, sebagai asas kajian genre, perbandingan korpus berdasarkan jenis genre seperti bahasa tulis dengan bahasa tutur; bahasa sains dengan bahasa sastera dan sebagainya. Perbandingan kekerapan kata juga dilakukan antara korpus yang menjadi fokus sesuatu kajian dengan korpus rujukan yang lazimnya lebih besar dari segi saiz dan yang diandaikan mewakili sesuatu bahasa secara tuntas.

Data kekerapan kata juga sering menjadi asas analisis statistik inferens. Penggunaan statistik dalam paradigma kajian linguistik korpus merupakan satu prinsip penting. Sesuatu aspek linguistik atau bahasa dikaji secara empirikal dengan menganalisis data tabii menggunakan korpus. Penggunaan statistik juga pada pendapat Leech (1992) sesuai dengan metodologi kajian berasaskan korpus yang menepati kaedah saintifik kerana data itu boleh diuji untuk kebolehpalsuan (*falsifiability*), keringkasan (*simplicity),* kesempurnaan (completeness), keutuhan lagi objektif.

Artikel ini akan memperlihatkan bagaimana maklumat kekerapan kata, khususnya kekerapan mentah dan kekerapan relatif boleh dimanfaatkan. Seterusnya, bagaimana maklumat ini ditransformasikan melalui beberapa kaedah statistik seperti ujian khi kuasa dua dan *log-likelihood* untuk menghasilkan analisis kata kunci; dan analisis multivariat, khususnya analisis penghubungan bagi membantu dalam penghuraian linguistik. Dari segi linguistik korpus, kajian linguistik atau bahasa itu mestilah berdasarkan satu himpunan teks "benar" atau gunaan yang disebut sebagai korpus. Kaedah-kaedah yang disebutkan akan diimplementasikan dalam

satu kajian kes membandingkan manifesto beberapa parti politik dalam Pilihan Raya Umum Ke-12 2008.

STATISTIK DALAM LINGUISTIK KORPUS

Senarai Kekerapan Kata

Senarai kekerapan kata merupakan data yang sering dikaitkan dengan linguistik korpus. Senarai kekerapan kata dijana berdasarkan pengiraan bilangan kata (*tokens*) untuk setiap jenis kata yang berbeza (*type*) yang berlaku dalam sesuatu teks.

Kegunaan awal senarai kekerapan kata adalah dalam bidang pendidikan, khususnya pendidikan bahasa. Menurut Nation dan Waring (1997), seawal tahun 1953 Michael West telah mengusahakan satu senarai 2000 patah perkataan yang paling kerap dalam bahasa Inggeris untuk dijadikan panduan perbendaharaan kata kepada guru dan perancang kurikulum. Senarai itu menjadi pewajaran tentang pemilihan kata dan frasa untuk setiap tahap sukatan pelajaran. Seterusnya, senarai kekerapan itu juga digunakan sebagai panduan untuk menentukan dan mengredkan tahap kesukaran bahan bacaan dan juga untuk menulis bahan bacaan yang dipermudah (*simplified readers*). Senarai kekerapan kata juga diperlukan untuk pengiraan nisbah jenis kata dengan bilangan kata (*type token ratio*). Nisbah ini menjadi pengukuran tahap kesukaran dan juga kekayaan leksikal yang ada dalam sesuatu teks atau korpus. Dalam leksikografi dan bidang perkamusan, kekerapan kata menjadi asas bagi menentukan tertib penyusunan maklumat tentang sesuatu perkataan dan pencirian statusnya sebagai dialek, varieti atau slanga mengikut keberlakuan dalam sesuatu korpus.

Dalam kajian linguistik korpus, korpus yang dianggap mewakili sesuatu bahasa, seperti korpus LOB yang mewakili bahasa Inggeris UK dan korpus Brown yang mewakili bahasa Inggeris Amerika, dianggap mampu memberikan senarai kekerapan kata dalam bahasa tersebut. Kata digolongkan kepada tiga kategori utama yakni perkataan yang mempunyai kekerapan rendah, kekerapan menengah atau kekerapan tinggi dalam bahasa tersebut. Konsep kekerapan juga digunakan untuk memprofil korpus kajian tertentu dan khusus seperti memprofil kepengarangan Shakespeare (Ilsemann, 2008) atau Munsyi Abdullah (Norhafizah, 2008). Dalam hal ini, senarai kekerapan kata, khususnya kata kandungan dan *hapax legomena*, mungkin tidak mencerminkan keseluruhan bahasa dan

unik kepada korpus kajian itu sahaja. Begitu juga, bukan keseluruhan senarai kekerapan kata digunakan tetapi kekerapan kata terpilih sahaja yang menjadi tumpuan kajian.

Dari sudut kajian linguistik pula, senarai kekerapan kata telah banyak diterapkan dalam kajian variasi bahasa. Dalam kajian ini, perbandingan senarai kekerapan kata antara korpus dilakukan untuk mengenal pasti perkataan yang berlaku lebih kerap secara signifikan dalam satu korpus berbanding dengan satu korpus yang lain, atau dengan satu korpus rujukan. Contohnya, Hofland dan Johansson (1982) dan Johansson and Hofland (1989) telah mengkaji perbezaan kepelbagaian bahasa Inggeris British dan Amerika Syarikat dengan membandingkan senarai kekerapan kata korpus Brown dan korpus LOB. Tujuannya untuk mengenal pasti perkataan yang lebih lazim dalam bahasa Inggeris British berbanding bahasa Inggeris Amerika Syarikat. Dalam kajian mereka, kaedah statistik yang digunakan ialah koefisien pembezaan Yule serta ujian khi kuasa dua. Ujian khi kuasa dua merupakan kaedah penganalisisan data kekerapan yang mudah, melibatkan jadual 1-dimensi yang mengandungi pemboleh ubah bahasa seperti kata dan sintaksis. Seterusnya, dilakukan tabulasi silang untuk menguji sama ada taburan antara kekerapan yang diperhatikan berbeza secara signifikan berbanding taburan yang dijangkakan. Taburan dijangkakan mungkin merupakan satu taburan rawak atau taburan berdasarkan satu korpus piawaian atau korpus rujukan.

Pendekatan yang sama digunakan oleh Leech dan Fallon (1992) untuk mengkaji perkataan apakah yang lebih kerap secara signifikan dalam bahasa Inggeris British berbanding bahasa Inggeris Amerika yang dapat menunjukkan perbezaan budaya. Imran Ho Abdullah dan Azhar (2004) juga membandingkan korpus teknologi maklumat bagi tujuan pembinaan istilah dengan mengkaji dan membandingkan senarai kekerapan korpus itu dengan korpus cerpen menggunakan indeks koefisien keterwakilan. Kekerapan kata juga digunakan dalam kajian sosiolinguistik berasaskan linguistik korpus pertuturan bagi tujuan mengenal pasti perbezaan antara variabel sosial pengguna bahasa. Contohnya, Rayson, Leech, and Hodges (1997) telah menggunakan data daripada subkorpus pertuturan British National Corpus (BNC) untuk kajian sosiolinguistik dengan mengenal pasti perbezaan kekerapan kata antara penutur lelaki dan perempuan; antara penutur berbeza umur serta antara penutur berbeza latar belakang dengan melakukan ujian khi kuasa dua terhadap senarai kekerapan yang dijana. Begitu juga, Imran Ho Abdullah (1996) telah membandingkan kekerapan preposisi *by* penulisan penutur natif dengan penulisan penutur bukan natif

dengan menggunakan ujian khi kuasa dua. Rayson and Garside (2000) pula menggunakan statistik *log-likelihood* untuk membandingkan korpus komunikasi kawalan udara dengan korpus rujukan. Imran Ho Abdullah dan Charlie Laman (1997) menggunakan statistik multivariat –analisis penghubungan untuk melihat hubung kait 30 kata paling kerap dalam korpus rujukan bahasa Inggeris (Brown) Amerika, (LOB) British dengan korpus surat khabar New Zealand dan Malaysia.

Kebelakangan ini, senarai kekerapan mentah juga menjadi asas kepada analisis visual yang dikenali sebagai awan kata (*word cloud*). Kandungan perkataan sesuatu teks atau korpus dapat digambarkan melalui perisian seperti Tagcrowd atau Wordle. Dalam suatu awan kata, perkataan yang berlaku dengan lebih kerap dan juga ketumpatan kata kunci (*keyword density*) dalam korpus akan diberikan keutamaan melalui saiz fon, susun atur serta skema warna. Konsep ketumpatan kata kunci berasaskan peratusan kali sesuatu kata kunci atau frasa berlaku dalam korpus berbanding jumlah keseluruhan kata dalam korpus tersebut. Kaedah ketumpatan kata kunci juga digunakan dalam kebanyakan enjin carian internet untuk menentukan sama ada sesuatu laman web adalah relevan untuk sesuatu kata carian. Kaedah analisis kata awan telah digunakan Kirkpatrick (2009) untuk meneliti ucapan sulung Obama yang kemudian dibandingkan dengan presiden terdahulu seperti Bush, Clinton, Reagan, dan Lincoln.

Jadual 1 Rumusan beberapa kajian menggunakan senarai kekerapan kata dan jenis statistik yang digunakan.

Pengkaji	**Kajian**	**Statistik**
Hofland dan Johansson (1982) Johansson and Hofland (1989)	Perbandingan senarai kekerapan kata korpus Brown dan korpus LOB untuk mengenal pasti perbezaan varieti bahasa Inggeris British dan bahasa Inggeris Amerika Syarikat.	Koefisien pembezaan Yule serta ujian khi kuasa dua
Leech dan Fallon (1992)	Perbandingan antara kekerapan kata bahasa Inggeris British dengan bahasa Inggeris Amerika yang dapat menunjukkan perbezaan budaya.	Koefisien pembezaan Yule serta ujian khi kuasa dua
Imran Ho Abdullah dan Azhar (2004)	Perbandingan korpus teknologi maklumat dengan korpus cerpen bagi tujuan pembinaan istilah dengan mengkaji senarai kekerapan kata.	Indeks koefisien keterwakilan

Pengkaji	Kajian	Statistik
Rayson, Leech, dan Hodges (1997)	Perbandingan senarai kekerapan kata korpus BNC untuk variabel sosiolinguistik bagi mengenal pasti perbezaan kekerapan kata antara penutur lelaki dan perempuan; antara penutur berbeza umur serta antara penutur berbeza latar belakang.	Ujian khi kuasa dua
Rayson dan Garside (2000)	Perbandingan senarai kekerapan kata korpus komunikasi kawalan trafik udara dengan korpus rujukan.	*Log-likelihood*
Imran Ho Abdullah (1996)	Perbandingan korpus penulis natif dengan bukan natif dan kekerapan preposisi *by*.	Ujian khi kuasa dua
Imran Ho Abdullah dan Charlie Laman (1997)	Perbandingan senarai 30 kata paling kerap dalam korpus British, Amerika, New Zealand dan Malaysia.	Analisis penghubungan
Imran Ho Abdullah dan Azhar Jaludin (2004)	Perbandingan korpus teknologi maklumat dan cerpen.	Indeks koefisien
Hundt and Smith (2009)	Perbandingan tensa *present perfect* dalam bahasa Inggeris British dan Amerika.	Ujian khi kuasa dua
Kirkpatrick (2009)	Perbandingan ucapan sulung Obama, Bush, Clinton, Reagan dan Lincoln.	Awan kata

Kata kunci

Kata kunci ditakrifkan sebagai satu kata yang berlaku pada kekerapan yang lebih tinggi (berlaku lebih kerap) dalam satu korpus X berbanding dengan satu korpus rujukan Y. Analisis kata kunci pula merupakan analisis perbandingan data berdasarkan dua senarai kekerapan kata.

Penjanaan kata kunci mudah dilakukan dengan menggunakan perisian seperti MonoConc (mengikut prosedur *Corpus Comparison*) atau WordSmith (mengikut prosedur *Keyword*). Kaedah penjanaan kata kunci membandingkan kekerapan setiap perkataan dalam senarai kata korpus kajian dengan kekerapan dalam korpus bandingan. Untuk mengira keutamaan (*keyness*) setiap perkataan, kekerapan serta bilangan perkataan dalam korpus kajian diambil kira bersama dengan kekerapan serta bilangan perkataan dalam korpus bandingan. Seterusnya, data ini ditabulasisilangkan untuk tujuan analisis ujian statistik. Lazimnya, ujian statistik yang digunakan ialah ujian khi kuasa dua kesignifikanan dengan pembetulan Yates untuk jadual 2 X 2. Selain itu, ujian *log-likelihood* Ted Dunning juga boleh digunakan. Ujian ini dianggap dapat memberi anggaran keutamaan yang lebih jitu terutamanya untuk korpus yang lebih besar berbanding korpus rujukan.

Meskipun senarai kata kunci menggunakan senarai kekerapan kata sebagai asasnya, perkataan yang tinggi kekerapannya tidak semestinya tersenarai sebagai kata kunci. Ini kerana sekiranya kata yang tinggi kekerapan dalam korpus kajian juga tinggi kekerapannya dalam korpus bandingan, maka kata itu mungkin tiada perbezaan dan bukan kata kunci yang membezakan dua korpus berkenaan. Contohnya, jika kata X berlaku 9 peratus dalam senarai kata korpus kajian and 10 peratus dalam senarai korpus yang dibandingkan, maka kata X mungkin tidak muncul dalam senarai kata kunci, meskipun kekerapan kata itu sangat tinggi. Kata yang tersenarai sebagai kata kunci ialah perkataan yang luar biasa kekerapannya, sama ada luar biasa tinggi atau luar biasa rendah, berbanding dengan kekerapan dalam korpus bandingan.

Kata kunci amat berguna untuk memprofilkan sesuatu teks atau genre terutama untuk tujuan perbandingan. Kaedah kata kunci membolehkan analisis makroskopik (keseluruhan teks) untuk membantu dalam tafsiran pada tahap mikroskopik (yang memberi fokus pada satu aspek atau unsur linguistik tertentu). Senarai kata kunci dapat memberi petunjuk unsur linguistik (atau kata) yang mana yang perlu diteliti dengan lebih mendalam dan serasi dengan pendekatan kajian linguistik korpus "berpandukan data".

Analisis Penghubungan

Analisis penghubungan merupakan satu kaedah penghuraian deskriptif dan bersifat penerokaan yang mula dibangunkan oleh ahli matematik Perancis,

Jean-Paul Benzérci, sekitar 1960-an. Kaedah ini mampu menganalisis data dalam jadual dan pelbagai hala yang mengandungi perhubungan antara baris dan lajur jadual tersebut. Hasil analisis dapat memberikan maklumat yang sama seperti yang dijana kaedah analisis faktor, dan membolehkan penganalisis meneroka struktur variabel dalam jadual. Jadual yang paling lazim digunakan ialah jadual tabulasi silang kekerapan dua hala. Dalam kaedah statistik analisis penghubungan, kekerapan dalam jadual ialah kekerapan relatif, iaitu jumlah keseluruhan kekerapan relatif adalah bersamaan 1.0. Hasil daripada analisis dapat mengenal pasti kekuatan hubung kait antara unsur dalam baris serta hubung kait antara unsur dalam lajur. Analisis penghubungan juga mampu menghasilkan interpretasi data kekerapan secara visual apabila setiap unsur dipetakan sebagai satu titik dalam satu graf dua dimensi atau tiga dimensi.

Memandangkan analisis penghubungan berupa satu kaedah penerokaan yang bertujuan untuk menghasilkan satu representasi visual dalam bentuk graf, berdasarkan maklumat yang terkandung dalam jadual kekerapan, tiada ujian untuk tahap kesignifikanan secara statistik dibuat terhadap hasil analisis penghubungan. Ini sesuai dengan falsafah analisis penghubungan yang menekankan pembangunan/penjanaan satu model yang menepati data (*models that fit the data*) dan bukan untuk tujuan menguji sesuatu hipotesis.

KAJIAN KES

Untuk memperlihatkan penggunaan senarai kekerapan kata serta mengkaji perbezaan antara kaedah statistik berasaskan data senarai kekerapan, satu kajian kes telah dilakukan menggunakan subkorpus manifesto Pilihan Raya Umum Ke-12 (PRU 12) 2008 untuk empat parti, Barisan Nasional (BN), Parti Keadilan Rakyat (PKR), Parti Islam SeMalaysia (PAS) dan Parti Tindakan Demokrasi (DAP). Dokumen manifesto ini membentuk korpus kajian dan telah dimuat turun daripada laman web[1]. Manifesto BN *Selamat Aman Makmur* merupakan satu fail PDF bersaiz 5334 Kb (3045 patah perkataan) yang mengandungi 24 muka surat bercetak (termasuk gambar dan ilustrasi). Muka surat pertama dan terakhir merupakan kulit dokumen. Fail PDF ini telah diubah kepada format teks biasa bersaiz 24 Kb. Manifesto PKR- *Harapan Baru Untuk Malaysia* dimuat turun dalam

1 Manifesto BN-www.bn2008.org.my/downloads/manifesto/bm_manifesto2008.pdf; Manifesto PKR - http://www.keadilanrakyat.org/index.php/content/view/544/; Manifesto PAS - http://pru12.pas.org.my/manifesto/ManifesFinalDetail.pdf; Manifesto DAP - http://www.esnips.com/doc/900a21f5-3b46-4106-a697-241469d3d6be/Manifesto-DAP

bentuk MHTML bersaiz 286 Kb dan selepas diubah format kepada teks biasa dokumennya bersaiz 43 Kb (5767 patah perkataan). Manifesto PAS *Kerajaan Beramanah, Adil dan Bersih Negara Berkebajikan* dimuat turun dalam bentuk PDF bersaiz 28 Kb dengan 3302 patah perkataan. Selepas diubah format kepada teks biasa dokumennya bersaiz 28 Kb. Manifesto DAP *Malaysia Boleh Lebih Baik Lagi* merupakan dokumen satu muka surat dimuat turun dalam bentuk JPEG bersaiz 505Kb. Cetakan JPEG ini telah diubah kepada format PDF dan seterusnya ke format teks biasa bersaiz 7 Kb dengan 459 patah perkataan. Statistik asas korpus dibentangkan dalam Jadual 2.

Jadual 2 Statistik asas korpus.

	BN	**PKR**	**PAS**	**DAP**	**Keseluruhan**
Bil. Perkataan (*tokens*)	3045	5767	3302	852	12 966
Bil. Perkataan Berbeza (*types*)	957	1393	1211	459	2555
Nisbah kata berbeza	31.43	24.15	36.67	53.87	19.71

Seterusnya, senarai kekerapan kata telah dijana menggunakan WordSmith 5.0 untuk setiap korpus. Senarai kekerapan kata ini menjadi asas pelbagai analisis kekerapan menggunakan kaedah statistik yang berbeza. Tujuan ini dilakukan adalah untuk melihat perbezaan dapatan yang mungkin dicapai dengan menggunakan kaedah statistik yang berbeza tetapi berdasarkan data kekerapan yang sama.

Analisis Kekerapan Mentah dan Kekerapan Relatif

Senarai kekerapan kata untuk manifesto Barisan Nasional mengandungi 957 kata berbeza, manifesto PKR mengandungi 1392 kata berbeza dan manifesto PAS mengandungi 1210 kata berbeza, sementara manifesto DAP mengandungi 459 kata berbeza. Jadual 3 menunjukkan 30 kata yang paling kerap dalam setiap senarai.

Jadual 3 Kata paling kerap dalam manifesto BN, PKR, PAS, dan DAP.

	BN		PKR		PAS		DAP	
Kedudukan	**Kata**	**F**	**Kata**	**F**	**Kata**	**F**	**Kata**	**F**
1	dan	126	yang	267	dan	188	dan	38
2	yang	54	dan	511	yang	97	yang	33
3	di	52	untuk	140	dengan	67	untuk	29
4	untuk	42	Malaysia	103	rakyat	46	kita	19
5	**negara**	40	akan	89	untuk	33	Malaysia	10
6	bagi	39	dengan	85	bagi	31	rakyat	8
7	kepada	38	kita	76	**negara**	30	di	7
8	dalam	38	di	75	dalam	29	kerajaan	7
9	meningkatkan	37	rakyat	74	di	27	lebih	7
10	ke	25	ekonomi	59	serta	25	telah	7
11	lebih	24	tidak	56	kepada	23	peluang	6
12	dari	24	**negara**	54	ini	22	sediakan	6
13	Malaysia	21	dalam	48	PAS	20	DAP	5
14	bilion	18	ini	43	tidak	19	dengan	5
15	polis	17	mereka	42	kerajaan	18	jenayah	5
16	melalui	17	keadilan	41	seperti	17	kadar	5
17	usaha	16	lebih	39	pelbagai	16	kebebasan	5
18	tahun	16	semua	39	manifesto	15	rasuah	5
19	serta	15	kepada	38	teras	15	semua	5
20	nasional	15	dasar	33	akta	14	ubah	5
21	dengan	15	memastikan	32	amalan	14	wujudkan	5
22	sekolah	14	undang	32	atau	14	anak	4
23	awam	14	demi	30	maka	14	baik	4
24	aman	13	menjamin	30	undang	14	bebas	4
25	pembangunan	13	juga	28	kos	13	dasar	4
26	juta	13	harga	23	media	13	ekonomi	4
27	Islam	13	polis	23	pilihan raya	13	kawasan	4
28	bandar	13	ke	22	rasuah	13	pendidikan	4
29	makmur	13	kerajaan	22	dasar	12	pilihan raya	4
30	selamat	13	sama	22	pendidikan	12	saksama	4

Analisis visual awan kata (*word cloud*) yang dijana berdasarkan kekerapan kata juga dilakukan menggunakan perisian Tagcrowd (http://tagcrowd.com/) untuk mendapatkan gambaran tentang kandungan setiap korpus berdasarkan visualisasi kekerapan kata (lihat Rajah 1). Berdasarkan awan kata, kata yang paling kerap dapat dikenal pasti berdasarkan saiz

fon. Secara sepintas lalu, kata paling kerap seperti *dan, yang, di, untuk* dikongsi oleh semua teks. Perkataan kandungan yang mencagun dengan ketara dalam manifesto BN ialah *negara, Malaysia* dan *meningkatkan*, sementara dalam manifesto PKR - *ekonomi, Malaysia* dan *rakyat*. Dalam manifesto PAS perkataan *rakyat* dan *negara* membentuk awan kata yang agak besar. Manakala dalam manifesto DAP, perkataan yang membentuk awan besar adalah seperti *jenayah*, *kebebasan*, *rakyat*, *rasuah*, *kerajaan*, dan *Malaysia.*

Awan Kata Manifesto BN

Awan Kata Manifesto PKR

Awan Kata Manifesto PAS

Awan Kata Manifesto DAP

Rajah 1 Awan kata manifesto BN, PKR, PAS, dan DAP.

Beberapa perkara harus diberikan perhatian dalam analisis kekerapan kata. Pertama, perbandingan berdasarkan kekerapan mentah sahaja adalah

terbatas kerana saiz korpus yang berbeza. Manifesto PKR hampir dua kali lebih panjang daripada manifesto BN dan PAS, serta lebih sepuluh kali saiz manifesto DAP. Bilangan kata manifesto BN melebihi 3045 patah perkataan berbanding dengan manifesto PKR 5767 dan PAS 3302. Satu kaedah statistik yang lebih wajar yang membolehkan perbandingan kekerapan dilakukan adalah dengan normalisasi kekerapan untuk mendapatkan kekerapan relatif mengikut saiz korpus. Ini dilakukan dengan menukarkan kekerapan mentah kepada nilai peratusan atau nilai per seribu patah perkataan. Sebagai contoh, perkataan *negara* berlaku sebanyak 54 kali dalam manifesto PKR dan 40 kali dan 30 kali masing-masing dalam manifesto BN dan PAS. Adalah tidak benar untuk menyatakan bahawa manifesto PKR mempunyai penggunaan perkataan *negara* yang lebih tinggi. Ini kerana perbandingan kekerapan mentah sahaja tidak mengambil kira saiz korpus (manifesto PKR) yang lebih panjang. Sebaliknya, berdasarkan kekerapan relatif perkataan *negara* dalam manifesto PKR ialah 0.93 peratus atau 93 bagi setiap 1000 patah perkataan, sementara dalam manifesto BN kekerapan relatif perkataan *negara* ialah 1.31 peratus atau 131 bagi setiap 1000 patah perkataan, dan dalam manifesto PAS pula, perkataan *negara* berlaku pada kadar 0.91 peratus atau 91 bagi setiap 1000 patah perkataan. Maka sebenarnya, perkataan *negara* berlaku dengan dengan nilai kekerapan relatif yang lebih tinggi dalam manifesto BN berbanding PKR dan PAS.

Selain kekerapan (berapa kali sesuatu kata muncul dalam teks), senarai kekerapan kata juga mengandungi maklumat *ranking* atau kedudukan dalam senarai bagi setiap perkataan mengikut tertib dari paling kerap kepada yang kurang kerap. Dari sudut kajian perbendaharaan kata linguistik, kata paling kerap yang lazim terdiri daripada leksis gramatikal seperti kata penghubung, kata pancangan, kata bantu, kata sendi (contohnya *yang, dan, untuk, di*) akan muncul dalam senarai dan menduduki kedudukan teratas – tidak kira jenis teks. Namun demikian, untuk kajian genre, kedudukan perkataan boleh dibandingkan dengan kedudukan kata dalam korpus rujukan. Dalam hal ini, perkataan *kita* mendapat kedudukan tinggi dalam senarai kekerapan kata manifesto DAP (ke-4) dan PKR (ke-7). Persoalannya, adakah perbezaan penggunaan kata ganti nama ini satu perkara yang luar biasa? Untuk menjawab persoalan ini, kedudukan kata ini dalam korpus rujukan boleh menjadi panduan. Berdasarkan Jadual 4, adalah jelas bahawa kedudukan *kita* dalam korpus rujukan berada dalam senarai 20 kata paling kerap pada kedudukan ke-17. Oleh itu, kedudukan 4

dan 7 merupakan satu perkara yang agak luar biasa (senarai 10 kata paling kerap). Begitu juga kedudukan *kita* dalam senarai kekerapan manifesto BN dan PAS juga luar biasa kerana begitu corot sekali. Pemerhatian yang sama juga didokong oleh data kekerapan relatif. Norma dalam korpus rujukan ialah 0.42 berbanding dengan kekerapan relatif yang sangat tinggi bagi manifesto PKR (1.32) dan DAP (2.33), serta kekerapan relatif yang sangat rendah bagi manifesto BN (0.30) dan PAS (0.09).

Jadual 4 Perbandingan kedudukan perkataan *kita.*

	Kedudukan	**Kekerapan**	**Kekerapan Relatif**
BN	49	9	0.30
PKR	7	76	1.32
PAS	234	3	0.09
DAP	4	19	2.33
Rujukan	17	21 888	0.42

Namun demikian, penganalisis korpus perlu berhati-hati kerana perbandingan kedudukan juga mungkin tidak memadai dan memberikan analisis yang kurang tepat. Sebagai contoh, perkataan *negara* menduduki kedudukan ke-5 dalam senarai BN. Sebaliknya, perkataan yang sama hanya menduduki kedudukan ke-12 dalam senarai PKR. Apabila kita membandingkan kekerapan relatif kata itu dalam kedua-dua teks, penggunaan dalam manifesto BN sebanyak 1.31 peratus sememangnya lebih tinggi daripada PKR sebanyak 0.94 peratus. Hakikatnya, perbezaan ini sebenarnya adalah tidak signifikan (nilai *log-likelihood* pada 0.57 pada 1 d.f), tetapi seseorang penganalisis mungkin membuat kesimpulan yang kurang tepat berdasarkan kedudukan (relatif) dalam senarai.

Dalam hal ini, persoalan seterusnya sama ada sesuatu perbezaan dalam senarai kekerapan (sama ada kekerapan mentah, kekerapan relatif mahupun kedudukan) adalah signifikan atau tidak memerlukan transformasi angka-angka kekerapan itu dengan penggunaan statistik inferens atau *second order statistics*. Kaedah statistik yang dapat menjawab persoalan sama ada kekerapan yang diperhatikan mempunyai perbezaan yang signifikan adalah seperti ujian khi kuasa dua atau *log-likelihood.*

Analisis Kata kunci – Perbandingan Ujian Khi Kuasa Dua dan *Log-Likelihood*

Seperti dinyatakan dalam bahagian **Kata Kunci** halaman 221, senarai dan analisis kata kunci dijana dengan melakukan perbandingan kekerapan setiap kata dalam senarai kekerapan kata sesuatu korpus kajian dengan senarai kekerapan kata korpus bandingan, atau korpus rujukan. Senarai kata kunci hanya memaparkan kata yang mempunyai perbezaan kekerapan yang signifikan. Dalam hal ini, kelazimannya statistik yang digunakan untuk menjana kata kunci ialah ujian khi kuasa dua atau *log-likelihood.* Untuk tujuan perbincangan hanya perbandingan antara subkorpus BN dan PKR dilakukan.

Dengan melakukan analisis kata kunci bagi membandingkan senarai kekerapan kata PKR dengan BN, menggunakan log-linear, pada tahap keyakinan 99 peratus (atau $p < 0.01$), sebanyak 68 perkataan dikesan sebagai lebih representasi (*overused*) dan kurang representasi secara signifikan antara senarai kekerapan PKR berbanding senarai kekerapan BN. Pada tahap keyakinan 99.9 peratus ($p < 0.001$) dengan nilai kritikal 10.89, senarai kata yang berbeza secara signifikan antara manifesto PKR dan BN mengurang kepada 28 kata. Jalur terakhir pada Jadual 5 menunjukkan sama ada kata itu terlebih representasi atau lebih kerap ditemui secara signifikan dalam manifesto PKR berbanding manifesto BN (+), sementara (-) menunjukkan kata itu terkurang representasi dalam manifesto PKR berbanding manifesto BN atau lebih kerap ditemui dalam manifesto BN berbanding PKR.

Apa yang menarik ialah penggunaan statistik yang berbeza boleh menghasilkan dapatan yang berbeza. Sekiranya ujian khi kuasa dua digunakan untuk menjana kata kunci pada tahap keyakinan yang sama yakni, 99.9 peratus ($p < 0.001$), hanya 17 kata didapati berbeza secara signifikan antara senarai PKR dan BN.

Penggunaan log-linear menghasilkan lebih banyak kata kunci pada tahap keyakinan yang sama berbanding dengan ujian khi kuasa dua. Sebelas kata kunci yang dikatakan berbeza secara signifikan menggunakan log-linear yang tidak terdapat dalam senarai kata kunci ujian khi kuasa dua ialah *sebarang* (+), *menjamin* (+), *demi* (+), *mana* (+), *juga* (+), *tetapi* (+), *untuk* (+), *ekonomi* (+), *mestilah* (+), *mengurangkan* (-), *ini* (+). Daripada senarai ini, sepuluh kata kunci yang terlebih representasi dalam manifesto PKR (ditandakan dengan (+)) dan hanya satu perkataan, iaitu *mengurangkan* yang terlebih representasi dalam manifesto BN (ditandakan dengan (-)).

Jadual 5 Kata berbeza (signifikan) antara manifesto BN dan PKR (log-linear).

N	Kata kunci	Frek.	%	RC. Frek.	RC. %	Nilai Keutamaan	P	
1	bagi	3	0.052	39	1.2808	64.12	0.00000000	-
2	yang	267	4.6298	54	1.7734	51.98	0.00000000	+
3	tidak	56	0.971	0		47.67	0.00000000	+
4	akan	89	1.5433	6	0.197	43.82	0.00000000	+
5	meningkatkan	11	0.1907	37	1.2151	36.50	0.00000000	-
6	mereka	42	0.7283	0		35.72	0.00000000	+
7	dari	5	0.0867	24	0.7882	28.68	0.00000008	-
8	keadilan	41	0.7109	1	0.0328	27.53	0.00000015	+
9	rakyat	74	1.2832	8	0.2627	27.53	0.00000015	+
10	kita	76	1.3178	9	0.2956	26.35	0.00000028	+
11	dasar	33	0.5722	1	0.0328	21.14	0.00000426	+
12	usaha	3	0.052	16	0.5255	20.02	0.00000767	-
13	dengan	85	1.4739	15	0.4926	19.60	0.00000952	+
14	Malaysia	103	1.786	21	0.6897	19.40	0.00001058	+
15	sebarang	22	0.3815	0		18.68	0.00001543	+
16	undang	32	0.5549	2	0.0657	16.22	0.00005642	+
17	menjamin	30	0.5202	2	0.0657	14.77	0.00012165	+
18	demi	30	0.5202	2	0.0657	14.77	0.00012165	+
19	Islam	3	0.052	13	0.4269	14.76	0.00012227	-
20	mana	17	0.2948	0		14.43	0.00014530	+
21	juga	28	0.4855	2	0.0657	13.33	0.00026094	+
22	bandar	4	0.0694	13	0.4269	12.49	0.00040805	-
23	tetapi	14	0.2428	0		11.88	0.00056657	+
24	untuk	140	2.4276	42	1.3793	11.56	0.00067456	+
25	ekonomi	59	1.0231	12	0.3941	11.10	0.00086498	+
26	mestilah	13	0.2254	0		11.03	0.00089496	+
27	mengurangkan	4	0.0694	12	0.3941	10.92	0.00095092	-
28	ini	43	0.7456	7	0.2299	10.89	0.00096461	+

Kedua-dua kaedah statistik ini juga menghasilkan senarai kata kunci yang sedikit berbeza dari segi urutan atau tertib nilai keutamaan (*keyness*). Senarai kata kunci berdasarkan *log-likelihood*, perkataan *tidak* merupakan kata utama yang ke-3 berbanding dengan senarai kata kunci berdasarkan khi kuasa dua (kedudukan ke-5). Sebaliknya, perkataan *rakyat* serta *Islam* menduduki tempat yang ke-9 dan ke-19 dalam senarai kata kunci berdasarkan *log-likelihood* tetapi menduduki tempat ke-7 dan ke-15 dalam senarai kata kunci berdasarkan khi kuasa dua.

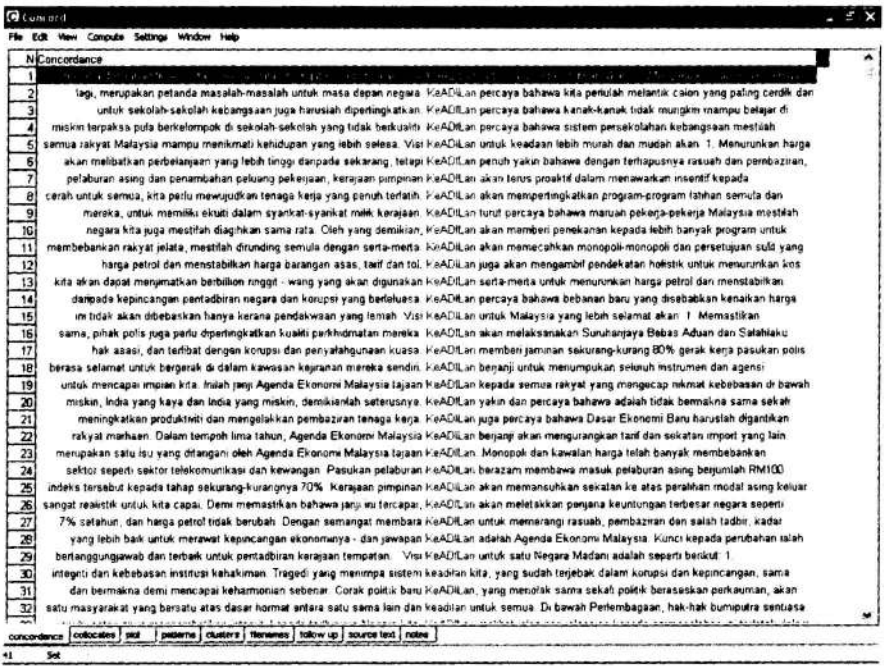
Jadual 6 Kata berbeza (signifikan) antara manifesto BN dan PKR (ujian khi kuasa dua).

N	Kata kunci	Frek.	%	RC. Frek.	RC. %	Nilai Keutamaan	P	
1	bagi	3	0.052	39	1.2808	60.867195	4.12499E-14	+
2	yang	267	4.6298	54	1.7734	45.511368	5.24456E-13	-
3	meningkatkan	11	0.1907	37	1.2151	36.731529	3.9253E-11	+
4	akan	89	1.5433	6	0.197	32.614544	8.31237E-09	-
5	tidak	56	0.971	0		28.239429	1.04274E-07	-
6	dari	5	0.0867	24	0.7882	27.794641	1.31975E-07	+
7	rakyat	74	1.2832	8	0.2627	21.415705	3.69438E-06	-
8	mereka	42	0.7283	0		20.773504	5.1664E-06	-
9	kita	76	1.3178	9	0.2956	20.743412	5.24827E-06	-
10	usaha	3	0.052	16	0.5255	18.618217	1.59658E-05	+
11	keadilan	41	0.7109	1	0.0328	17.914436	2.31033E-05	-
12	Malaysia	103	1.786	21	0.6897	16.484468	4.90473E-05	-
13	dengan	85	1.4739	15	0.4926	16.240269	5.57925E-05	-
14	dasar	33	0.5722	1	0.0328	13.71367	0.000212896	-
15	Islam	3	0.052	13	0.4269	13.45529	0.000244313	+
16	bandar	4	0.0694	13	0.4269	11.440763	0.0007185	+
17	undang	32	0.5549	2	0.0657	11.168065	0.000832171	-

Namun begitu, banyak juga persamaan antara senarai kata kunci kedua-dua kaedah statistik ini. Misalnya, kesemua 17 perkataan dalam senarai kata kunci berdasarkan khi kuasa dua juga ditemui dalam senarai kata kunci berdasarkan *log-likelihood*. Meskipun terdapat perbezaan pada urutan keutamaan, urutan bagi perkataan paling utama adalah sama.

Dalam kajian berpandukan korpus, berdasarkan senarai kata kunci ini, analisis kualitatif dapat dilakukan dengan bantuan konkordans. Berdasarkan senarai kata kunci, perkataan *keadilan*, *rakyat*, *akan* dan *tidak* yang terlebih representasi (secara signifikan) dalam manifesto PKR serta perkataan *bagi* yang terkurang representasi mungkin menjadi sasaran untuk kajian kualitatif dalam mengkaji perbezaan antara manifesto PKR dan BN.

Hakikatnya, perkataan *keadilan* tersenarai sebagai satu kata kunci yang terlebih representasi dalam manifesto PKR berbanding BN tidaklah begitu memeranjatkan. Ini kerana perkataan itu sebahagian daripada nama parti. Dengan merujuk kepada konkordans bagi *keadilan*, (Rajah 2), daripada 41 keberlakuan kata ini, 37 kali merujuk kepada nama parti

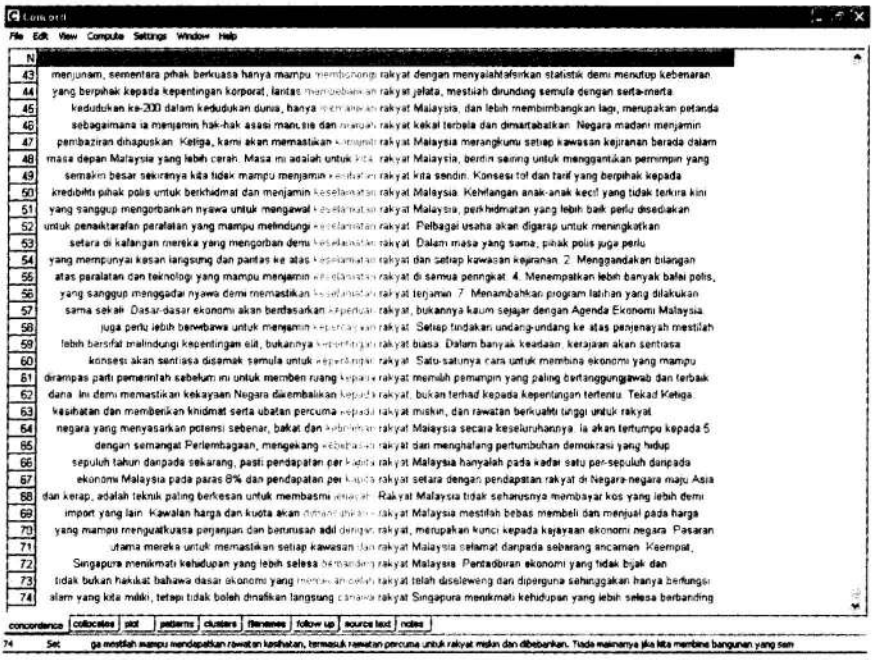


Rajah 2 Paparan konkordans *keadilan* dalam manifesto PKR.

yang mengandungi kata keadilan. Nama parti dibezakan daripada konsep keadilan dalam teks dengan pemaparan ortografik keADILan, dan hanya empat kali kata itu digunakan merujuk kepada konsep keadilan.

Satu lagi kata kunci yang terlebih representasi dalam manifesto PKR ialah perkataan *rakyat* dengan nilai keutamaan 27.53. Perkataan ini berlaku 74 kali dalam manifesto PKR dan hanya lapan kali dalam manifesto BN. Dari segi kekerapan relatif pula, perkataan *rakyat* hampir lima kali lebih kerap dalam manifesto PKR (1.28 peratus) berbanding dalam manifesto BN (0.26 peratus). Persoalannya, adakah perbezaan ini juga disebabkan perkataan rakyat itu sebahagian daripada nama *Parti Keadilan Rakyat* atau sememangnya manifesto itu banyak membangkitkan dan merujuk secara langsung kepada *rakyat*? Berdasarkan pemeriksaan konkordans *rakyat* dalam manifesto PKR (Rajah 2), ternyata kekerapan penggunaan perkataan *rakyat* bukanlah kerana nama parti tetapi merujuk kepada *rakyat* sebagai sebahagian daripada strategi wacana manifesto PKR. Setiap pernyataan dalam manifesto PKR menyasarkan *rakyat* sebagai penerima (*recipient*) sesuatu tindakan. Ini amat sesuai dalam bahasa manifesto:

1. Setiap *rakyat* Malaysia juga mestilah mampu mendapatkan rawatan kesihatan, termasuk rawatan percuma untuk *rakyat* miskin dan dibebankan.
2. Menjamin maruah semua *rakyat* Malaysia dengan melaksanakan gaji minimum pada kadar RM1500 sejajar dengan kenaikan kos sara hidup.
3. Penyalahgunaan pihak berkuasa polis demi tujuan politik sempit yang berleluasa, meluasnya perpecahan dalaman dan sikap menerima korupsi, semuanya merupakan duri dalam daging kepada kredibiliti pihak polis untuk berkhidmat dan menjamin keselamatan *rakyat* Malaysia.



Rajah 3 Paparan konkordans *rakyat* dalam manifesto PKR.

Satu lagi kata kunci yang terlebih representasi dalam manifesto PKR ialah kata bantu *akan* (nilai keutamaan 43.82). Penggunaan kata bantu modus *akan* berkait dengan cadangan, saranan dan janji. Rayson (2004) yang mengkaji manifesto Liberal Democrats berbanding Labour di Britain, mendapati penggunaan kata bantu modus seperti *would* dan *will* adalah lebih tinggi bagi parti yang lazimnya menjangka peluang mereka membentuk kerajaan adalah tipis, oleh itu, lebih banyak menyebut tentang

rancangan masa hadapan berbanding dengan parti yang sedang memerintah atau berpeluang cerah untuk membentuk kerajaan yang akan menghuraikan perancangan secara lebih konkrit dan jelas tanpa perlu menggunakan kata bantu atau modus. Hipotesis Rayson (2004) ini ternyata ada kebenaran memandangkan kekerapan relatif PKR menggunakan kata bantu ini lima kali lebih tinggi (1.54 peratus) berbanding BN (0.20 peratus).

Kata kunci yang ketiga paling signifikan ialah kata nafi *tidak*. *Tidak* digunakan hanya dalam manifesto PKR (56 kali dengan kekerapan relatif 0.97 peratus) dan tiada keberlakuan dalam manifesto BN. Horn (1989:202) berpendapat bahawa fungsi utama ayat negatif dan ayat nafi adalah untuk memperbetul dan menyanggah *(contradict)*. Namun, dalam penggunaan kata nafi yang tinggi mungkin kerana PKR secara sengaja mahu menggunakan ayat nafi bagi menidakkan usaha parti yang memerintah (1–2) atau memburukkan parti pemerintah (3):

1. Pentadbiran ekonomi yang *tidak* bijak dan kucar-kacir, terutamanya dalam tempoh empat tahun yang lepas, telah membawa prestasi ekonomi Malaysia berada di tahap yang terendah setakat ini.
2. Tetapi memperbaiki infrastruktur sahaja *tidak* akan mencukupi, selagi kandungan pembelajaran itu sendiri *tidak* diolah semula.
3. Satu ketika dahulu kita berada sama taraf dengan beberapa buah negara luar dalam pertumbuhan ekonomi, perlaksanaan undang-undang dan juga mutu pendidikan, tetapi kini dunia memandang rendah keupayaan Malaysia dengan ekonomi yang *tidak* kompetitif, pencapaian universiti yang bertambah merosot, dan sistem mahkamah yang *tidak* ada integriti dan maruah sama sekali.

Banyak juga penggunaan kata nafi yang berfungsi untuk memburukkan parti pemerintah dari segi kekangan terhadap rakyat:

1. Beban yang amat berat lagi *tidak* adil kini terpaksa dipikul oleh rakyat jelata sejak beberapa tahun yang lepas.
2. Kini, kita *tidak* boleh membeli lebih daripada satu paket gula di pasaraya, ataupun membeli kereta pada harga yang dibayar pengguna di negara-negara lain.

Penggunaan kata nafi *tidak* yang begitu banyak juga mencerminkan serta mencorakkan prosodi semantik mesej manifesto PKR yang mengajak pembaca dan pengundi menolak kerajaan sedia ada.

Penggunaan perkataan *bagi* yang begitu kerap dalam manifesto BN, kekerapan relatif 1.28 peratus dalam manifesto BN berbanding 0.05 peratus

dalam manifesto PKR, juga menuntut satu penjelasan. Hipotesis awal pengkaji ialah penggunaan perkataan *bagi* yang kerap dalam manifesto BN adalah untuk menjelaskan tujuan sesuatu dilakukan (1–2), atau mengenal pasti kumpulan sasar sesuatu (3–4):

1. Melaksanakan usaha berterusan *bagi* memupuk kerjasama, keamanan dan kestabilan ASEAN.
2. Merangka garis panduan *bagi* memastikan penyediaan tanah simpanan untuk tempat beribadat dalam kawasan yang baru dibangunkan.
3. Meningkatkan liputan bekalan air *bagi* kawasan luar bandar di Sabah dan Sarawak kepada 70 peratus menjelang tahun 2010.
4. Memperuntukkan RM3 bilion untuk biasiswa dan keperluan-keperluan *bagi* golongan pelajar dari keluarga berpendapatan rendah.

Begitu juga kata kunci lain yang terlebih representasi dalam manifesto BN seperti *meningkatkan, dari, usaha, Islam* dapat dijelaskan berdasarkan struktur manifesto BN yang mendahului setiap tajuk dengan penjelasan pencapaian usaha semasa, sebagai parti pemerintah, dan apa yang mahu ditingkatkan lagi.

Satu kekangan analisis kata kunci dengan kaedah statistik khi kuasa dua dan juga *log-likelihood* ialah perbandingan terhad kepada dua senarai kekerapan kata sahaja. Ini kerana statistik perbandingan untuk analisis kata kunci berdasarkan jadual kontingensi 2 x 2. Untuk tujuan kajian ini, perbandingan antara senarai kata kekerapan BN dan PKR dijadikan contoh perbincangan dan bagaimana analisis kata kunci menggunakan kaedah statistik khi kuasa dua dan *log-likelihood* berbeza, serta manfaat analisis kata kunci sebagai satu kaedah linguistik korpus berpandukan data yang dapat memacu persoalan kajian lanjutan.

Untuk melihat perhubungan atau kesepadanan secara serentak bagi beberapa korpus, kaedah statistik multivariat diperlukan. Dalam hal ini, kaedah statistik yang diterokai dalam artikel ini ialah analisis kesepadanan atau analisis penghubungan.

Analisis Penghubungan

Untuk tujuan perbincangan, perbandingan antara manifesto analisis penghubungan BN, PKR, PAS dan DAP akan dikemukakan. Dalam kajian ini, 49 kata paling kerap serta kekerapannya dalam manifesto BN, PKR, PAS, DAP membentuk jadual dua hala, Jadual 7. Prosedur analisis penghubungan dijana menggunakan SPSS.

Analisis penghubungan ialah kaedah pemvisualan secara statistik dan geometrik yang memperlihatkan hubungan antara tahap dengan memaparkan baris dan lajur dalam jadual kontingensi dua hala sebagai posisi titik baris dan lajur konsisten dengan hubungan dalam jadual. Matlamat akhir adalah untuk mendapatkan satu pandangan global data bagi membolehkan interpretasi dilakukan. Dalam kajian kes ini, himpunan pertama terdiri daripada manifesto parti dan himpunan kedua terdiri daripada perkataan paling kerap. Dalam jadual kontingensi dua hala, hubungan yang diperhatikan antara dua variabel disimpulkan oleh kekerapan sel dan statistik inferens. Ini dilakukan untuk melihat sama ada terdapat hubung kait antara unsur dalam satu variabel dengan unsur dalam variabel yang lain.

Jadual 7 Kekerapan kata paling kerap dalam manifesto BN, PKR, dan PAS.

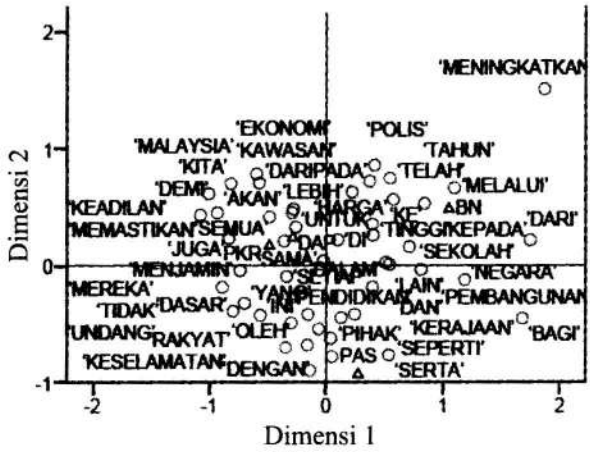
NAMA	PAS	BN	PKR	DAP
LABEL	Numerik	Numerik	Numerik	Numerik
AKAN	7	6	89	1
BAGI	31	39	3	3
DALAM	29	38	48	3
DAN	188	126	244	38
DARI	11	24	5	0
DARIPADA	3	6	21	0
DASAR	12	1	33	4
DEMI	1	2	30	0
DENGAN	67	15	85	5
DI	27	52	75	7
EKONOMI	3	12	59	4
HARGA	8	11	23	3
INI	22	7	43	2
JUGA	4	2	28	0
KAWASAN	3	11	14	4
KE	9	25	22	1
KEADILAN	3	1	41	2
KEPADA	23	38	38	3
KERAJAAN	18	10	22	7
KESELAMATAN	10	3	15	1
KITA	3	9	76	19

Samb. Jadual 7

NAMA LABEL	PAS Numerik	BN Numerik	PKR Numerik	DAP Numerik
LAIN	8	8	13	0
LEBIH	8	24	39	7
MALAYSIA	3	21	103	10
MELALUI	5	17	10	2
MEMASTIKAN	5	9	32	1
MENINGKATKAN	1	37	11	0
MENJAMIN	7	2	30	0
MEREKA	11	0	42	2
NEGARA	30	40	54	1
OLEH	9	4	15	0
PEMBANGUNAN	9	13	7	0
PENDIDIKAN	12	5	20	4
PIHAK	11	5	16	0
POLIS	3	17	23	2
RAKYAT	46	8	74	8
SAMA	5	5	22	1
SEKOLAH	10	14	13	0
SEMUA	6	8	39	5
SEPERTI	17	6	19	3
SERTA	25	15	20	2
SETIAP	7	7	19	1
TAHUN	4	16	18	3
TELAH	5	13	11	7
TIDAK	19	0	56	2
TINGGI	6	10	14	2
UNDANG	14	2	32	4
UNTUK	33	42	140	29
YANG	97	54	267	33

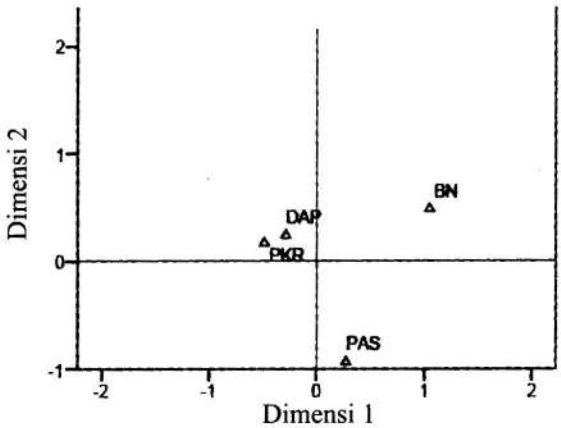
Analisis penghubungan data dalam Jadual 7 menghasilkan paparan grafik dalam Rajah 5. Terdapat dua jenis titik yang terhasil. Titik jenis pertama merupakan titik lajur. Dalam hal ini, empat titik lajur (Δ) setiap

satu mewakili BN, PKR, PAS dan DAP terhasil. Titik jenis kedua merupakan titik baris (o), 49 kesemuanya yang mewakili setiap perkataan dalam senarai kata paling kerap.



Rajah 5 Paparan graf analisis penghubungan (baris dan lajur).

Jarak antara setiap titik baris ialah ukuran kesamaan atau kesepadanan antara profil kekerapan baris. Misalnya, titik perkataan *meningkatkan* adalah jauh dari titik perkataan *rakyat* kerana kedua-dua perkataan mempunyai profil kekerapan yang berbeza. Sebaliknya, perkataan *dasar* dan *rakyat* berdekatan kerana profilnya serupa atau sepadan. Begitu juga, jarak antara titik lajur ditafsirkan dengan cara yang sama. Untuk tujuan pemaparan yang lebih jelas, Rajah 6 memaparkan hanya kedudukan titik lajur.



Rajah 6 Paparan graf analisis penghubungan (lajur sahaja).

Berdasarkan Rajah 6, empat titik lajur – satu bagi setiap parti dipetakan dalam rajah dua dimensi. Terdapat tiga kedudukan titik yang jelas. Manifesto PKR dan DAP berada dalam satu kuadran sementara manifesto PAS dan BN berada dalam dua lagi kuadran yang berlainan dengan jarak yang agak jauh antara satu dengan yang lain. Ini bermaksud manifesto PKR dan DAP mempunyai kesamaan. Jarak antara mereka pada dimensi 2 hanyalah 0.07 dan pada dimensi 1 pula 0.20. Manakala, terdapat perbezaan signifikan antara manifesto PKR/DAP dengan PAS dan juga dengan BN. Pada masa yang sama, PKR 0.17, DAP 0.24 dan BN 0.49 berbeza dengan PAS -0.93 pada dimensi 2 (menegak). Sebaliknya, pada dimensi 1 (mendatar) BN dan PAS terletak pada kuadran positif manakala PKR dan DAP berada pada kuadran negatif. Sungguhpun demikian, dari segi dimensi ini, jarak dari paksi asal bagi PKR -0.49, DAP -0.29 dan PAS 0.27 adalah lebih dekat antara mereka berbanding dengan BN 1.05.

Jadual 8 Sumbangan lajur kepada dimensi.

Lajur	Kelompok	Skor dalam Dimensi		Inersia	Sumbangan Titik ke Inersia Dimensi		Sumbangan Dimensi ke Inersia Titik		
		1	2		1	2	1	2	Total
PAS	0.22	0.27	-0.93	0.05	0.04	0.74	0.11	0.89	1.00
BN	0.20	1.05	0.49	0.09	0.61	0.19	0.87	0.13	1.00
PKR	0.52	-0.49	0.17	0.05	0.34	0.06	0.88	0.08	0.96
DAP	0.06	-0.29	0.24	0.03	0.01	0.01	0.06	0.03	0.09
Jumlah	1.00			0.23	1.00	1.00			

Perkataan yang menyumbang kepada perbezaan yang diperhatikan dalam dimensi 1 adalah seperti dalam Jadual 9. Perkataan dalam Jadual 9 adalah lebih dekat kepada PKR dan DAP mempunyai nilai negatif. Manakala, perkataan yang lebih dekat kepada BN bernilai positif.

Jadual 9 Perkataan yang menyumbang kepada Dimensi 1.

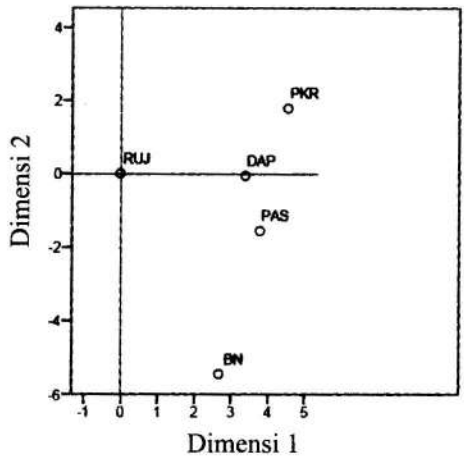
Lajur	Kelompok	Skor dalam Dimensi	
		1	2
KEADILAN	0.01	-1.08	0.44
DEMI	0.01	-1.01	0.62
AKAN	0.02	-0.93	0.45
MEREKA	0.01	-0.89	-0.18
JUGA	0.01	-0.83	0.24
KITA	0.03	-0.82	0.70
TIDAK	0.02	-0.80	-0.39
MENJAMIN	0.01	-0.74	-0.04
DASAR	0.01	-0.70	-0.32
MALAYSIA	0.03	-0.60	0.79
MENINGKATKAN	0.01	1.87	1.51
DARI	0.01	1.75	0.22
BAGI	0.02	1.68	-0.45
PEMBANGUNAN	0.01	1.19	-0.12
MELALUI	0.01	1.10	0.67
KE	0.01	0.84	0.53
SEKOLAH	0.01	0.81	-0.03
KEPADA	0.02	0.71	0.16

Persamaan antara BN, PKR dan DAP berbanding PAS dengan merujuk kepada posisi pada paksi kedua (dimensi 1) (paksi X) disumbang oleh perkataan dalam Jadual 10. Interpretasi persamaan antara BN dan PKR berbanding PAS pada paksi ini adalah kerana manifesto kedua-dua parti ini mempunyai perkataan skor positif pada dimensi 2, yang didapati lebih kerap secara signifikan berbanding Pas seperti dalam Jadual 10. Perkataan yang mempunyai skor negatif pada dimensi 2 pula adalah lebih hampir kepada PAS berbanding BN dan PKR.

Jadual 10 Perkataan yang menyumbang kepada Dimensi 2.

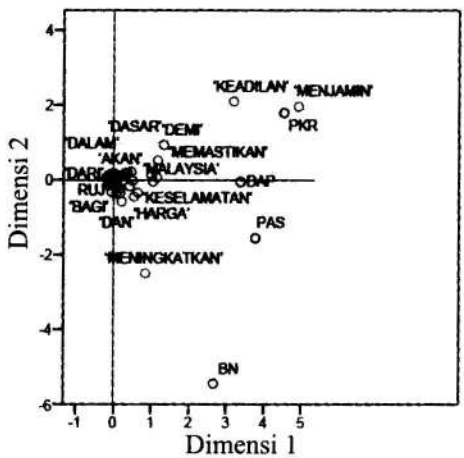
Lajur	Kelompok	Skor dalam Dimensi	
		1	2
MENINGKATKAN	0.01	1.87	1.51
POLIS	0.01	0.41	0.86
MALAYSIA	0.03	-0.60	0.79
TAHUN	0.01	0.55	0.75
KAWASAN	0.01	0.37	0.72
EKONOMI	0.02	-0.57	0.71
KITA	0.03	-0.82	0.70
MELALUI	0.01	1.10	0.67
LEBIH	0.02	0.22	0.63
DEMI	0.01	-1.01	0.62
DENGAN	0.04	-0.14	-0.90
SEPERTI	0.01	0.05	-0.78
SERTA	0.01	0.54	-0.77
RAKYAT	0.03	-0.35	-0.70
KESELAMATAN	0.01	-0.16	-0.68

Untuk tujuan kajian kes, senarai kekerapan kata perbandingan manifesto ketiga-tiga parti juga telah dilakukan analisis penghubungan dengan korpus rujukan (5 juta patah perkataan). Hasilnya adalah seperti dalam Rajah 7.



Rajah 7 Paparan graf analisis penghubungan manifesto dan korpus rujukan (lajur sahaja).

Apa yang menarik hasil analisis penghubungan ialah manifesto bersifat sebagai satu genre tersendiri. Ketiga-tiga titik lajur membentuk satu garis yang terpisah dari titik korpus rujukan. Seterusnya persoalan, apakah perkataan yang menyumbang kepada genre manifesto? Jawapannya, perkataan yang berkedudukan jauh dari titik korpus rujukan termasuklah perkataan seperti kata kerja seperti *menjamin, meningkatkan, memastikan* serta kata nama seperti *Malaysia, keselamatan, harga, kerajaan, kawasan* dan *ekonomi.*



Rajah 8 Paparan graf analisis penghubungan manifesto dan korpus rujukan (baris dan lajur).

RUMUSAN

Artikel ini telah membincangkan bagaimana maklumat kekerapan kata dalam korpus dapat dimanfaatkan. Kekerapan kata merupakan data lazim dalam kajian linguistik korpus. Data kekerapan ini boleh digunakan dalam banyak cara, bergantung pada kaedah dan transformasi statistik kekerapan kata.

Berdasarkan kajian kes menggunakan manifesto Pilihan Raya Umum Ke-12 sebagai korpus kajian, kebergantungan hanya pada kekerapan mentah mungkin menghasilkan interpretasi yang tidak tepat kerana keterbatasan kekerapan mentah dan juga kedudukan kata paling kerap. Keterbatasan ini boleh diatasi dengan penggunaan normalisasi data kekerapan mentah atau kekerapan relatif. Kaedah visualisasi melalui analisis awan kata juga boleh dihasilkan melalui data kekerapan. Seterusnya, perbandingan antara dua senarai kekerapan kata telah dilakukan. Transformasi kekerapan mentah kepada nilai log-linear atau khi kuasa dua bagi menghasilkan senarai kata kunci. Senarai kata kunci memberikan kepastian tentang perkataan yang secara signifikan berbeza antara dua korpus. Dalam hal ini, analisis kuantitatif perlu ditokok dengan analisis kualitatif dengan meneliti konkordans bagi perkataan yang berbeza agar aspek linguistik tentang perbezaan itu dapat dihuraikan. Transformasi kekerapan untuk kajian korelasi atau kajian hubung kait pula memerlukan analisis statistik multivariat. Dalam kajian ini, analisis penghubungan telah digunakan. Kaedah ini dapat mengambarkan perbezaan serta sumbangan setiap unsur yang dikaji melalui perbezaan yang wujud antara korpus. Perbandingan dengan korpus rujukan yang besar, iaitu 5 juta patah perkataan, jelas menunjukkan bahawa berdasarkan senarai kata paling kerap manifesto pilihan raya mempunyai ciri-ciri tersendiri atau membentuk satu genre.

Yang jelas, apapun analisis kuantitatif dalam kajian linguistik korpus dilakukan bukan bermaksud analisis kualitatif boleh diabaikan, terutamanya dalam pemerhatian kata kunci. Namun demikian, penerokaan data kekerapan melalui kaedah statistik dalam paradigma kajian linguistik dan bahasa berpandukan korpus dapat memberi petunjuk dan pewajaran kepada persoalan lanjutan yang menuntut penyelesaian dalam pemantapan dan pembinaan ilmu linguistik.

Rujukan

Gries, S.T., . "Some Proposals towards more Rigorous Corpus Linguistics". *Zeitschrift für Anglistik und Amerikanistik*, 54:2, 191–202, 2006.

Gries, S.T., "Dispersion and Adjusted Frequencies in Corpora". *International Journal of Corpus Linguistics*, 13:4, 403–37, 2008.

Gries, S.T., 2009. Useful statistics for corpus linguistics. http://www.linguistics.ucsb.edu/faculty/ stgries/research/UsefulStatsForCorpLing.pdf (dicapai 10 September 2009).

Hidalgo-Downing, L., 2000. *Negation, Text Worlds, and Discourse: The Pragmatics of Fiction*. (Advances in Discourse Processes, V. 66.) Norwood: Ablex

Hofland, K. & Johansson, S., 1982. Word frequencies in British and American English. Bergen. The Norwegian Computing Centre for Humanities.

Horn, Laurence R., 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.

Ilsemann, H., 2008. More Statistical Observations on Speech Lengths in Shakespeare's Plays Literary and Linguistic Computing Advance Access published online on September 29, 2008. Literary and Linguistic Computing, doi:10.1093/llc/fqn011.

Imran Ho Abdullah, 1996. By ESL writers vs. by native writers: A Corpus Analysis of Native and Non-native Speakers' Written English. *Deep South* v.2 n.3.

Imran Ho Abdullah & C. Laman, 1997.Comparing word frequencies across corpora: a Correspondence Analysis of varieties of English. Zymurgix. The 4th New Zealand Postgraduate Conference. Refereed Proceedings. 85–90.

Imran Ho Abdullah & Azhar Jaludin, "Perbendaharaan Kata dalam Bidang Komputer dan Teknologi Maklumat: Satu Kajian Korpus" dlm *Jurnal Bahasa Jendela Alam,* Jilid 3, 270–88, 2004.

Jeong, H., "Discourse analysis of public debates using corpus linguistics methodologies" dlm. *Journal of Computers*, Jilid 3 No. 8. 58–68, 2008.

Johansson, S. and Hofland, K., 1989. *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus*. Oxford: Oxford University Press.

Kirkpatrick, M., 2009. Word Cloud Analysis of Obama's Inaugural Speech Compared to Bush, Clinton, Reagan, Lincoln's. http://www.readwriteweb.com/archives/ tag_clouds_of_ obamas_inaugural_speech_compared_to_ bushs.php) (Capaian 6 Ogos 2009).

Lee, BL., 1996. *Correspondence analysis. www.uv.es/prodat/ViSta/vista-frames/ pdf/chap11 .pdf.*

Leech, G.N., 1992. Corpus linguistics and theories of linguistic performance. In: J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm*, 4–8 August 1991 (hlm. 105–22). Berlin: Mouton de Gruyter.

Leech, G & Fallon, R., "Computer corpora – what do they tell us about culture?" *ICAME Journal,* 16, 29–50, 1992.

Leech, G.N., P. Rayson and A. Wilson. 2001. *Word Frequencies in Written and Spoken English Based on the British National Corpus.* London: Longman.

Norhafizah Mohamed Husin, 2008. *"Analisis Kolokasi Leksikal: Citra Melayu dalam Hikayat Abdullah."* Tesis Sarjana. Universiti Kebangsaan Malaysia.

Nation, P & R. Waring, 1997. Vocabulary size, text coverage and word lists. In Schmitt Norbert and Michael McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (hlm. 6–20). Cambridge: Cambridge University press.

Rayson, P., "From key words to semantic domains" dlm. *International Journal of Corpus Linguistics*, Vol 13:4:519–49, 2004.

Rayson, P., Leech, G., and Hodges, M., Social Differentiation In The Use Of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus dlm. *International Journal of Corpus Linguistics*, Jilid 2, nombor 1, hlm. 133–52, 1997.

Rayson P, & Garside R., 2000. Comparing corpora using frequency profiling. In: *Proceedings of the workshop on Comparing Corpora*, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), hlm. 1–6.

Sampson, G.R., 1987. "Probabilistic Methods of Analysis" dlm. R.G. Garside, G. Leech and G.R. Sampson (eds) *The Computational Analysis of English: A Corpus-Based Approach.* London: Longman.

Sampson, G.R., 2001. *Empirical Linguistics.* London: Continuum.

Thorndike, E.L. and I. Lorge, 1944. The Teacher's Word Book of 30,000 Words. Teachers College, Columbia University.

Thorndike, E.L., 1924. "The vocabularies of school pupils" dlm. J. Carelton Bell (ed.) *Contributions to Education.* New York: World Book Co.

West, M., 1953. *A General Service List of English Words.* London: Longman, Green & Co.