Perbandingan distribusi frekuensi kata bahasa Indonesia di Kompas, Wikipedia, Twitter, dan Kaskus

Ivan Lanin
ivan@ardwort.com

Jim Geovedi
jim@ardwort.com

Wicak Soegijoko
wicak@ardwort.com

Ardwort
Jakarta, Indonesia

Abstrak. Dalam bahasa yang sama, kata yang paling sering digunakan, jumlah huruf per kata, serta berbagai statistik lain yang terkait dengan distribusi frekuensi sangat bergantung kepada ragam yang digunakan. Makalah ini menelaah perbandingan distribusi frekuensi kata antara empat ragam bahasa Indonesia yang populer di internet, yaitu Kompas (media massa), Wikipedia bahasa Indonesia (ensiklopedia), Twitter (mikroblog), dan Kaskus (forum). Kajian dilakukan dengan menggunakan korpus yang diambil dari data yang tersedia secara publik di internet serta diproses dengan menggunakan bahasa pemrograman Python serta beberapa pustaka pemrograman yang bersumber terbuka. Hasil kajian menunjukkan adanya perbedaan distribusi yang cukup tajam di antara keempat ragam bahasa Indonesia ini. Kompas banyak menggunakan kata *akan* karena sifat beritanya; Wikipedia banyak menggunakan kata *adalah* karena sifat deskriptifnya; Twitter banyak menggunakan kata *aku* karena sifat subjektifnya; Kaskus banyak menggunakan kata *gan* yang merupakan kata khas komunitas ini. Kajian ini juga memberikan beberapa hal yang harus diperhatikan dalam kajian serupa seperti penyiapan dan pembersihan data korpus dan leksikon. Kajian ini diharapkan dapat memberikan dasar penelitian lebih lanjut dalam bidang distribusi frekuensi dan analisis korpus bahasa Indonesia.

Kata kunci: distribusi frekuensi, korpus, NLP

Pendahuluan

Frekuensi penggunaan kata dalam sebuah tulisan maupun percakapan sangat memengaruhi waktu tanggap penutur. Semakin sering suatu kata digunakan, semakin cepat kata pula tersebut dipahami (Whaley 1978; Ford, Marslen-Wilson, and Davis 2003). Frekuensi penggunaan sebuah kata juga sering menjadi variabel pembeda antara ragam lisan dan ragam tulis (Tryk 1968). Di sisi lain, pemilihan kata dapat dianggap sebagai representasi pengetahuan penutur. Pengalaman penutur terhadap penggunaan sebuah kata harus dipertimbangkan pada semua modalitas yang berhubungan dengan kata tersebut (Gernsbacher 1984). Hal yang sama berlaku untuk makna dari kata yang dimaksud. Keakraban penutur dengan kata tidak hanya diperoleh dari menghadapi dan merasakannya, namun juga dari operasi semantik yang terlibat dalam pengolahan, khususnya dalam pemahaman, yang kemudian mengharuskan penutur memahami representasi bentuk kata dan maknanya. Keakraban dengan kedua representasi bentuk kata dan maknanya dapat secara bersamaan meningkat setiap kali kata tersebut diproses. Makalah ini menyajikan peringkat 20 kata yang paling sering digunakan dalam beberapa empat ragam bahasa Indonesia, yaitu jurnalistik, ensiklopedia, mikroblog, dan forum daring. Analisis terhadap data tersebut menggambarkan keakraban penutur terhadap kata-kata tertentu berdasarkan sifat ragam bahasa yang dipilih.

Metodologi

Korpus yang digunakan untuk keperluan penelitian ini diperoleh dari 4 situs daring berbahasa Indonesia yang populer di internet, yaitu Kompas (media massa), Wikipedia bahasa Indonesia (ensiklopedia, selanjutnya

disingkat Wikipedia), Twitter (mikroblog), dan Kaskus (forum). Keempat korpus ini dipilih sebagai bahan penelitian karena mewakili ragam bahasa yang berbeda. Informasi detail data korpus dipaparkan pada Tabel 1.

Data mentah diperoleh melalui teknik scraping untuk situs Kaskus dan Kompas, ekstraksi salinan data XML dari Wikipedia, dan *API query* pada layanan Twitter. Data mentah tersebut selanjutnya dibersihkan dari kode HTML, nama pengguna (pada korpus Twitter dan Kaskus), templat atau pola acu (pada korpus Wikipedia), alamat URL (Uniform Resource Locator) internet, serta sebagian besar tanda baca, kecuali tanda hubung ("-") pada kata ulang. Data korpus diproses dengan menggunakan bahasa pemrograman Python serta pustaka NLTK, NumPy, dan SciPy.

Korpus Data	Informasi Pengambilan Data	Jumlah Kata Unik	Jumlah Kata Keseluruhan
Kompas	Diambil pada bulan Januari tahun 2013 untuk artikel berita berbahasa Indonesia daring tahun 2012.	343.532	32.724.503
Wikipedia	Diambil dari salinan "idwiki" bulan Januari tahun 2013.	936.288	43.545.242
Twitter	Diambil pada bulan Januari tahun 2013 untuk percakapan bulan Oktober-Desember 2012 oleh pengguna Twitter yang berlokasi di Indonesia.	798.078	34.769.573
Kaskus	Diambil pada bulan Januari tahun 2013 dan dari 1,000 threads terakhir sub-forum "The Lounge".	761.795	109.292.156

Tabel 1: Informasi Pengambilan Data Korpus

Frekuensi kemunculan kata dihitung dengan menggunakan rumus dari Ramisch (2012) yang menyatakan bahwa frekuensi (*fréquence*, f) merupakan hasil pembagian jumlah kemunculan suatu kata atau token (*nombre d'occurrences*, $C_w(\cdot)$) dengan jumlah total kata atau token (n). Sebagai contoh, berikut rumus penghitungan kemunculan kata *yang* dalam korpus data Wikipedia:

$$f = \frac{C_w(\cdot)}{N} = \frac{C_{wikipedia}(yang)}{N} = \frac{975156}{43545242} = 0.02239408842876565$$

Namun hasil perhitungan ini tidak dapat langsung digunakan sebagai model linguistik karena banyaknya kata yang jarang digunakan (*low-rank words*) memiliki frekuensi sama. Statistik suatu korpus data sangat bergantung pada besar korpus, jenis bahasa, dan ragam yang digunakan. Walaupun demikian, kaitan logaritmik antara jumlah kata (token) dan jumlah jenis kata (token) biasanya akan mengikuti Hukum Zipf (Zipf 1936; Baayen 2001). Hukum Zipf menyatakan bahwa pada korpus data bahasa alami, peringkat (r) dari sebuah kata adalah berdasarkan frekuensinya (f) yang diformulakan menjadi $f \propto \frac{1}{r}$ dan formula ini ekuivalen dengan $f = \frac{k}{r}$ untuk k sebagai konstanta faktor.

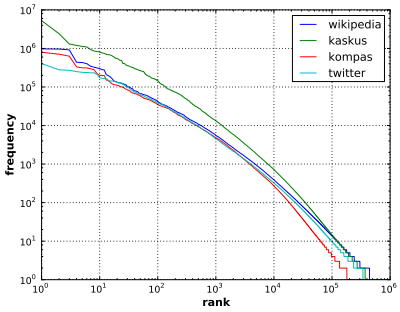
Hasil dan Analisis

Jumlah keseluruhan kata dari keempat korpus adalah sebanyak 220.331.474 kata dengan distribusi jumlah kata unik dan jumlah keseluruhan kata per korpus disajikan pada Tabel 1. Karena keterbatasan tempat, tabel penghitungan yang ditampilkan dalam makalah ini dibatasi hanya 20 kata teratas dari masing-masing korpus (Tabel 2). Data contoh yang lebih besar tersedia daring di repositori Github.

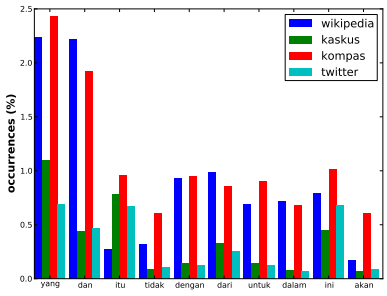
No	Kompas	%	Wikipedia	%	Twitter	%	Kaskus	%
1	yang	2,429	yang	2,239	di	1,162	gan	4,808
2	di	2,168	dan	2,214	yg	0,803	ane	2,202
3	dan	1,923	di	2,107	ya	0,778	di	1,194
4	ini	1,017	pada	1,007	aku	0,719	yang	1,097
5	itu	0,958	dari	0,987	yang	0,690	yg	1,034
6	dengan	0,953	dengan	0,927	ini	0,682	ya	0,998
7	untuk	0,907	ini	0,791	itu	0,670	ada	0,854
8	dari	0,858	adalah	0,749	ada	0,669	itu	0,786
9	dalam	0,679	dalam	0,714	d	0,613	tuh	0,758
10	akan	0,610	untuk	0,689	aja	0,498	aja	0,739
11	pada	0,609	kategori	0,649	ga	0,481	bisa	0,701
12	tidak	0,604	tahun	0,633	dan	0,470	juga	0,680
13	juga	0,463	sebagai	0,476	gak	0,469	kalo	0,642
14	ke	0,449	oleh	0,457	i	0,435	keren	0,626
15	tersebut	0,410	indonesia	0,426	mau	0,412	ga	0,624
16	ada	0,378	ke	0,390	ke	0,410	banget	0,599
17	bisa	0,359	the	0,349	udah	0,410	nya	0,567
18	saat	0,352	ia	0,322	lagi	0,405	wah	0,532
19	jakarta	0,344	tidak	0,318	kalo	0,389	nih	0,508
20	tahun	0,337	menjadi	0,303	the	0,379	jadi	0,502

Tabel 2: Peringkat dan persentase kemunculan kata

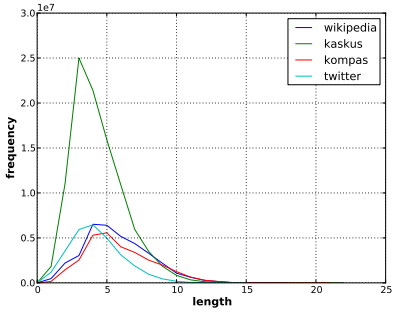
Gambar-gambar di bawah ini memberikan ilustrasi dari analisis yang akan diuraikan selanjutnya.



Gambar 1: Perbandingan distribusi frekuensi kemunculan kata



Gambar 2: Perbandingan peringkat frekuensi kemunculan kata



Gambar 3: Perbandingan distribusi jumlah huruf per kata

Kata yang tersering digunakan

Beberapa penelitian terdahulu (Kirby 2001; Roberts, Onnis, and Chater 2005) mengusulkan model linguistik yang dapat menjelaskan fenomena ketidakteraturan dalam bahasa. Fenomena itu dapat diketahui lewat distribusi Zipf (Gambar 1) dan korelasi frekuensi kata dengan konstruksi dan ketidakteraturan morfologi atau sintaksis dari kata yang dimaksud. Hal tersebut dapat menjelaskan bahwa kata yang sering muncul di sejumlah korpus adalah kata yang tidak teratur bentuknya. Berikut akan disoroti beberapa hal menarik seputar 20 kata yang tersering digunakan dalam tiap korpus.

Semua kata teratas di Kompas merupakan ragam baku dengan satu kata nama tempat. Kata *akan* (#10) banyak dipakai karena ragam jurnalistik yang memberitakan hal-hal di masa depan. Kata *Jakarta* (#19) muncul untuk menjelaskan tempat, sedangkan kata *tahun* (#20) muncul untuk menjelaskan waktu.

Korpus Wikipedia banyak mengandung kata *adalah* (#8) karena sifat deskriptif ensiklopedia. Kata bahasa Inggris, *the* (#17), berasal dari nama yang dicantumkan dalam artikel. Kata *kategori* (#11) berasal dari kategorisasi artikel yang muncul hampir di semua artikel.

Korpus Twitter banyak mengandung kata *aku* (#4) dan *mau* (#15) karena sifat subjektif media sosial. Pemendekan kata seperti *yg* (#2), *d* (#9), dan *ga* (#11) banyak dipakai karena pembatasan jumlah huruf pada media ini. Ragam cakapan *aja* (#10), *gak* (#13), *udah* (#17), dan *kalo* (#19) muncul sesuai dengan sifat percakapan pada Twitter. Kata *i* (#14) merupakan cerminan campur kode (*code-mixing*) atau alih kode (*code-switching*) pada Twitter.

Dua kata teratas pada korpus Kaskus adalah *gan* (#1) dan *ane* (#2) yang merupakan pronomina persona khas komunitas ini. Pemendekan seperti *yg* (#5) dan *ga* (#15), serta ragam cakapan seperti *kalo* (#13) dan *banget* (#16) juga muncul dalam korpus ini. Partikel *tuh* (#9), *wah* (#18), dan *nih* (#19), serta kata *keren* (#14) menggambarkan sifat ekspresif media ini. Pronomina *-nya* yang merupakan bentuk terikat yang ditulis serangkai dengan kata sebelumnya muncul sebagai kata terpisah *nya* (#17).

Perbandingan peringkat frekuensi kemunculan kata

Sebagai perbandingan, kata yang tersering digunakan pada empat korpus dibandingkan dengan hasil penelitian Tala (2003) yang menggunakan korpus kepala berita Kompas pada tahun 2001. Dari perbandingan pada Gambar 2, terlihat bahwa korpus Kompas dan Wikipedia memberikan hasil yang cukup serupa dengan penelitian Tala. Namun, korpus Twitter dan Kaskus memberikan hasil yang sangat berbeda. Hal ini menunjukkan adanya perbedaan kata yang tersering digunakan dalam dua kelompok ragam bahasa yang berbeda, yaitu ragam formal (Kompas dan Wikipedia) dan ragam informal (Twitter dan Kaskus).

Dari 20 kata yang tersering digunakan menurut penelitian Tala, ada dua kata (*yang* dan *di*) yang juga masuk peringkat 20 besar dalam empat korpus dalam penelitian ini. Lima kata lain (*ada*, *dan*, *ini*, *itu*, dan *ke*) masuk dalam peringkat 20 besar dalam tiga dari empat korpus. Hal ini menunjukkan bahwa meskipun ada perbedaan antarragam bahasa, ada beberapa kata yang kerap digunakan bersama dalam semua ragam bahasa Indonesia.

Jumlah huruf per kata

Analisis jumlah huruf per kata yang diilustrasikan dalam Gambar 3 sejalan dengan pendapat Zipf (1932) bahwa kata pendek (jumlah huruf sedikit) lebih sering digunakan daripada kata panjang (jumlah huruf banyak). Hal ini disebabkan oleh dorongan penutur untuk meminimalkan waktu dan upaya yang diperlukan dalam penuturan atau penulisan kata (Piantadosi, Tily, and Gibson 2011).

Sebaran frekuensi kemunculan kata menurut jumlah huruf per kata pada keempat korpus menunjukkan pola distribusi Gauss yang condong ke kiri ke arah kata pendek. Wikipedia dan Twitter sama-sama paling sering menggunakan kata dengan empat huruf, sedangkan Kompas paling sering menggunakan kata dengan lima huruf. Kaskus paling sering menggunakan kata dengan tiga huruf dengan kurva distribusi Gauss yang cukup curam.

Hal ini menunjukkan kecenderungan yang cukup tinggi dari komunitas Kaskus untuk meminimalkan waktu dan upaya dalam penulisan kata.

Kesimpulan

Kata yang tersering dipakai di Kompas, Wikipedia, Twitter, dan Kaskus berbeda. Perbedaan ini timbul karena perbedaan ragam bahasa dan sifat khas masing-masing. Walaupun demikian, frekuensi kemunculan kata di keempat korpus ini secara umum tetap mengikuti hukum Zipf. Di samping itu, hasil analisis sebaran jumlah huruf per kata pada keempat korpus ini juga sesuai dengan hukum Zipf yang menyatakan kecenderungan penutur untuk menggunakan kata dengan jumlah huruf sedikit demi meminimalkan upaya untuk berkomunikasi.

Metode pemrosesan data korpus yang telah dikembangkan di sini akan digunakan untuk melakukan penelitian-penelitian lain yang lebih aplikatif dan langsung dapat dimanfaatkan, seperti kamus kata bahasa Indonesia sederhana berdasarkan kata yang paling sering muncul dalam berbagai korpus, kamus variasi ejaan (atau salah eja) kata, serta penerjemahan otomatis antarragam bahasa Indonesia (misalnya dari SMS ke ragam formal). Penyempurnaan yang perlu dilakukan dalam metode ini antara lain adalah identifikasi nama diri (*named-entity*) dan *stopwords*.

Referensi

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers.

Ford, Michael A., William D. Marslen-Wilson, and Matthew H. Davis. 2003. Morphology and frequency: Contrasting methodologies. In *Morphological structure in language processing*, ed. R. Harald Baayen and Robert Schreuder, 89–124. Berlin: Mouton de Gruyter.

Gernsbacher, Morton Ann. 1984. Resolving 20 Years of Inconsistent Interactions Between Lexical Familiarity and Orthography, Concreteness, and Polysemy. *Journal of Experimental Psychology: General* 113, no. 2: 256–81.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*: 102–10.

Piantadosi, Steven T., Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 9 (March): 3526–29.

Ramisch, Carlos. 2012. A generic and open framework for multiword expressions treatment: from acquisition to applications. Grenoble: Université de Grenoble.

Roberts, Matthew, Luca Onnis, and Nick Chater. 2005. Acquisition and Evolution of quasi-regular languages: Two puzzles for the price of one. In *Language Origins: Perspectives on Evolution*, ed. Maggie Tallerman. Oxford Linguistics.

Tala, Fadillah Z. 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Amsterdam: Universiteit van Amsterdam.

Tryk, H. Edward. 1968. Subjective scaling of word frequency. *Journal Verbal Learning and Verbal Behavior* 81, no. 2: 170–77.

Whaley, C. P. 1978. Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior* 17, no. 2: 143–54.

Zipf, George Kingsley. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press.

---. 1936. *The Psychobiology of Language*. Routledge.