FASILITAS PEMERIKSA EJAAN DAN ANALISIS STATISTIK KATA

Bobby Nazief
Fakultas Ilmu Komputer - Universitas Indonesia

1. Latar

Komputer, yang semula dirancang sebagai alat pengolah data yang umumnya bertipe numerik, telah berkembang fungsi dan cakupannya. Saat ini komputer tidak lagi semata-mata dikaitkan dengan perhitungan numerik, namun juga mencakup banyak aspek pengolahan informasi yang menyangkut data bertipe non-numerik. Mulai dari pengolahan dokumen tekstual dalam bentuk penyunting kata elektronis, sampai dengan pembuatan efek-efek visual yang menjadikan sebuah film lebih menakjubkan untuk ditonton.

Dari luasnya fungsi komputer pada saat ini, kami akan menyoroti satu penggunaannya yang mungkin merupakan penggunaan terbesar, yaitu sebagai alat bantu yang memungkinkan penyuntingan kata secara elektronis. Bersama suatu paket perangkat lunak penyunting kata, komputer menggantikan peran mesin ketik sebagai alat bantu yang memungkinkan para penulis/wartawan/penyunting menghasilkan karya-karya tulis mereka dengan cepat.

1.1. Fasilitas-fasilitas Pendukung Pengolahan Kata Elektronis

Saat ini banyak tersedia paket-paket perangkat lunak yang berfungsi sebagai penyunting kata, walaupun sayangnya hampir semua penyunting kata elektronis tersebut masih ditujukan untuk penyuntingan dokumen dalam bahasa Inggris. Akibatnya, penulis dokumen berbahasa Indonesia tidak mungkin memanfaatkannya secara maksimal. Keterbatasan ini disebabkan penulis tersebut tidak dapat menggunakan fasilitas-fasilitas pengolahan kata yang lazim tersedia dalam suatu penyunting kata

elektronis seperti misalnya: *pemeriksa ejaan, pemenggal kata, tesaurus,* atau *pemeriksa tata bahasa.*

Jika kita menulis dokumen dalam bahasa yang memiliki dukungan fasilitas-fasilitas pengolahan kata yang disebutkan di atas, maka kita akan dapat memanfaatkan fasilitas-fasilitas tersebut untuk menghasilkan dokumen dengan kualitas bahasa yang baik. Fasilitas pemeriksa ejaan akan menjamin penggunaan kata-kata yang sesuai dengan acuan yang baku. Kesalahan ejaan, yang tentunya akan dapat menimbulkan kesalahan interpretasi yang mungkin dapat menyesatkan, dapat dihindarkan.

Fasilitas pemenggal kata akan membantu proses pemenggalan kata terutama jika kita berhadapan dengan dokumen berukuran besar, atau jika kita harus menyunting banyak dokumen sedangkan waktu yang tersedia untuk melakukan pemenggalan kata amat singkat. Walaupun kita dapat memilih untuk tidak membolehkan adanya kata-kata yang terpenggal di dalam dokumen kita, untuk beberapa jenis dokumen seperti artikel-artikel di surat kabar misalnya, hal tersebut akan sangat berpengaruh pada jumlah kertas yang dibutuhkan.

Selanjutnya, bagi banyak penulis ketersediaan sebuah tesaurus elektronis yang dapat diakses secara langsung pada saat ia tengah menyunting dokumennya akan terasa membantu sekali. Pada saat menulis suatu dokumen, seringkali kita dihadapkan pada rasa "kebosanan" yang diakibatkan penggunaan satu kata yang sama berulang-ulang. Dengan adanya tesaurus yang dapat diakses tanpa meninggalkan dokumen yang tengah disunting tersebut, maka dengan segera kita dapat memilih kata pengganti yang sesuai tanpa kehilangan konteks tulisannya.

Akhirnya, untuk menghindari penggunaan tata bahasa yang tidak mengikuti aturan yang berlaku, kita dapat menggunakan fasilitas pemeriksa tata bahasa. Kesalahan jenis ini akan menyulitkan pemahaman isi dokumen yang kita tulis pada saat dibaca oleh pembaca nantinya.

1.2. Ketidaktersediaan Fasilitas-fasilitas Pendukung Pengolahan Kata Khusus Bahasa Indonesia

Seperti telah disinggung di atas, fasilitas-fasilitas pendukung pengolahan kata yang umumnya merupakan bagian dari paket-paket perangkat lunak pengolah kata elektronis yang banyak digunakan saat ini tidak ada yang dirancang khusus untuk Bahasa Indonesia. Belum adanya fasilitas-fasilitas pengolahan kata khusus untuk Bahasa Indonesia pada penyunting-penyunting kata elektronis tersebut tentunya memprihatinkan karena hal ini berarti hampir semua pengguna penyunting kata elektronis tersebut bagaikan pengendara kendaraan yang tidak dilengkapi dengan peralatan yang sesuai dengan kondisi atau aturan-aturan lalu-lintas yang berlaku pada jalan yang dilaluinya. Jika peralatan kendaraan tersebut dibuat sesuai dengan kondisi jalannya, maka si pengendara akan dapat mencapai tempat tujuannya tanpa harus melanggar rambu-rambu lalu lintas selama perjalanannya.

Pada makalah ini kami akan memusatkan perhatian pada fasilitas pemeriksa ejaan kata-kata Bahasa Indonesia. Secara spesifik kami akan menguraikan strategi perancangan "modul" perangkat lunak pemeriksa ejaan beserta beberapa hasil awal yang kami peroleh dalam mewujudkan strategi tersebut. Hasil-hasil awal ini berkaitan dengan karakteristik penggunaan kata dalam dokumen-dokumen berbahasa Indonesia.

2. Kebutuhan Pemeriksaan Ejaan Otomatis Khusus Bahasa Indonesia

Sebelum menyajikan strategi perancangan modul perangkat lunak pemeriksa ejaan, kami akan menayangkan beberapa contoh kesalahan ejaan pada beberapa naskah tulisan yang merupakan bagian dari sejumlah artikel di beberapa surat kabar nasional. Setelah itu, kami akan menguraikan suatu cara sederhana yang dapat dilakukan oleh pengguna paket pengolah kata elektronis untuk memeriksa ejaan kata-kata yang digunakan dalam dokumennya secara otomatis.

2.1. Kesalahan Ejaan: beberapa contoh

Sebagai bahan ilustrasi, pada bagian berikut akan kami tayangkan beberapa contoh kesalahan ejaan yang kami temui pada beberapa artikel yang terdapat pada tiga surat kabar nasional:

KOMPAS:

- "Kehadiran Menristek **BJ** Habibie menjadi Ketua Umum ICMI, jelas membawa sejumlah manfaat yang positif bagi organisasi yang lahir **Desember** 1990 itu, terutama **mobilidasi** dana dan akses ke **birokrasi**."
  [*Struktur Kepemimpinan ICMI Lemah, 4/1/95*]

- "Pada saat sekarang, terutama pada perusahaan yang **pada** karya, ..."
  [*Upah Minimum Regional Naik 10 Sampai 35 Persen Mulai April, 4/1/95*]

SUARA PEMBARUAN:

- "Secara tidak langsung, fasilitas pengolahan kata otomatis pada pemroses kata elektronis ini dapat **membiasaan** penulis menyajikan bahan tulisan dengan kata-kata dan tata bahasa yang benar."
  [*Pemanfaatan Teknologi Informasi Dalam Pengembangan Bahasa Indonesia, 6/12/93*]

- "..., terlebih yang duduk di legislatif agar jangan menyakiti hati rakyat dan jangan melakukan tindakan **penyelewangan** atau penyalah-gunaan wewenang."
  [*Golkar Tidak Lakukan "Litsus," 8/1/95*]

MEDIA INDONESIA:

- "Perkembangan yang kurang menguntungkan tersebut, antara lain **disebebkan** oleh kebutuhan pelat baja sebagai bahan baku, ..." [*Kontraktor Jangan Banting Harga, 9/1/95*]

- "Sebab, pada akhirnya **kulitas** pekerjaan proyek akan menurun dan cepat rusak." [*Kontraktor Jangan Banting Harga, 9/1/95*]

Kesalahan-kesalahan ejaan yang ditampilkan pada cuplikan-cuplikan artikel di atas menunjukkan bahwa proses pemeriksaan ejaan yang ada saat ini belum dilakukan secara sempurna. Menurut informasi yang kami peroleh, proses pemeriksaan ejaan memang masih dilakukan secara *manual*. Dapat dibayangkan jika jumlah dokumen yang harus diperiksa kebenaran ejaannya besar, waktu yang tersedia terbatas, serta tenaga pemeriksa pun terbatas, maka kesalahan-kesalahan ejaan seperti yang kami tampilkan pada contoh-contoh di atas memang sulit dihindarkan.

Keadaannya mungkin akan berbeda jika terdapat suatu pemeriksa ejaan otomatis. Sebagian besar kesalahan ejaan di atas dapat dihindari sehingga yang tersisa adalah kesalahan ejaan yang memang sulit dideteksi oleh pemeriksa ejaan yang canggih sekalipun. Contohnya dapat dilihat pada cuplikan artikel di harian KOMPAS yang mengandung kata **pada** yang seharusnya ditulis **padat**. Kesalahan ejaan seperti ini akan sulit dideteksi karena kata **pada** sendiri merupakan kata yang benar.

2.2. Teknik Sederhana Memeriksa Ejaan Secara Otomatis

Sekarang, mari kita lihat kemungkinan pembuatan suatu pemeriksa ejaan otomatis sederhana dengan memanfaatkan fasilitas-fasilitas yang telah tersedia pada paket-paket perangkat lunak penyunting kata yang ada. Di antara sekian banyak paket-paket penyunting kata elektronis yang ada saat ini seperti *Microsoft Word, Word Perfect, Wordstar*, dan lain-lainnya, kami memilih produk buatan *Microsoft* sebagai objek uji coba. Seperti halnya paket-paket penyunting kata lainnya, *Microsoft Word* telah dilengkapi fasilitas-fasilitas pendukung pengolahan kata yang kami

uraikan di atas untuk beberapa bahasa seperti bahasa *Inggris, Jerman, Belanda, Perancis, Denmark, Finlandia, Norwegia, Swedia, Italia, Spanyol,* dan *Portugis*. Terlihat bahwa Bahasa Indonesia belum masuk ke dalam daftar tersebut.

Pada paket *Microsoft Word*, kita dapat menambahkan kamus kata buatan kita sendiri ke dalam daftar kamus yang digunakan *Word* untuk memeriksa ejaan. Dengan menyusun kata-kata baku Bahasa Indonesia pada kamus tambahan tersebut, *Word* akan "mengerti" kata-kata Bahasa Indonesia selama kata-kata yang dijumpainya pada dokumen yang diperiksa memiliki padanannya dalam kamus tambahan yang kita buat tadi. Cara ini memungkinkan kita memeriksa ejaan kata-kata dalam dokumen-dokumen berbahasa Indonesia.

Ketersediaan pemeriksa ejaan otomatis seperti di atas akan banyak membantu pengolahan dokumen berbahasa Indonesia. Namun demikian, dari hasil pengujicobaan yang kami lakukan, ternyata besarnya kamus tambahan ini mempengaruhi kecepatan *Word* memeriksa ejaan. Akibatnya, ukuran kosa kata yang dapat disimpan di dalam kamus tersebut akan terbatas sekali jika masih ingin mempertahankan waktu pemeriksaan ejaan yang dapat diterima. Tentunya hal ini akan membatasi kosa kata yang dapat digunakan untuk menulis suatu dokumen.

Berdasarkan hasil uji coba ini, kami menyimpulkan bahwa dengan menggunakan mekanisme yang tersedia pada paket-paket penyunting kata elektronis yang ada, kita dapat membuat pemeriksa ejaan otomatis sederhana yang dapat "mengerti" Bahasa Indonesia. Kita cukup menyediakan sebuah daftar kata-kata Bahasa Indonesia baku yang sering kita gunakan untuk menulis dokumen. Daftar ini kemudian akan digunakan oleh pemeriksa ejaan yang telah tersedia pada paket tersebut sebagai kamus tambahan selain kamus standarnya. Walaupun cara ini memiliki banyak keterbatasan, namun seperti kata pepatah *"tidak ada rotan, akar pun jadi,"* maka cara ini dapat digunakan sementara fasilitas pemeriksa ejaan yang lebih baik belum tersedia.

3. Strategi Pengembangan Pemeriksa Ejaan Otomatis Bahasa Indonesia

Pada bagian ini kami akan memaparkan strategi yang menurut hemat kami merupakan salah satu pilihan dalam merancang modul perangkat lunak yang berfungsi sebagai pemeriksa ejaan kata-kata Bahasa Indonesia. Sebelum sampai ke pembahasan teknis, perlu kiranya kami utarakan suatu kriteria penting yang harus diperhatikan berkaitan dengan perancangan modul perangkat lunak tersebut.

Kriteria penting yang kami maksudkan di atas adalah modul pemeriksa ejaan yang dirancang harus dapat dihubungkan dengan paket-paket perangkat lunak penyunting kata elektronis yang ada. Hal ini penting karena jika modul tersebut dirancang sebagai bagian dari suatu paket penyunting kata yang baru, yang belum banyak dikenal orang, maka mungkin tidak banyak yang dapat memanfaatkannya karena mereka telah terbiasa menggunakan paket-paket yang ada.

Kembali ke aspek teknis perancangan modul pemeriksa ejaan itu sendiri, faktor penting yang perlu diperhatikan adalah aspek kinerja pemeriksa ejaan tersebut. Agar pemeriksa ejaan ini digunakan banyak orang maka ia harus mampu memeriksa ejaan dengan benar dan cepat. Hal ini menuntut pengetahuan yang mendalam tentang karakteristik penggunaan kata-kata Bahasa Indonesia pada suatu dokumen.

3.1. Algoritme Pemeriksaan Ejaan

Langkah-langkah (algoritme) yang harus dilakukan dalam memeriksa kebenaran ejaan suatu kata dapat dinyatakan sebagai berikut:

```
Bandingkan kata dengan kata-kata di dalam KAMUS.
Jika ditemukan, maka pencarian selesai.
Jika tidak ditemukan, maka
    Jika kata = awalan-kata* dan kata* ada di KAMUS,
          maka pencarian selesai.
    Jika kata = kata*-akhiran dan kata* ada di KAMUS,
          maka pencarian selesai.
    Jika kata = awalan-kata*-akhiran dan
          kata* ada di KAMUS, maka pencarian selesai.
Jika tidak berhasil, maka ejaan kata salah.
```

Terdapat dua komponen penting pada algoritme di atas. Pertama, proses pemeriksaan kebenaran ejaan suatu kata membutuhkan adanya *kamus* yang berisi kata-kata baku. Pada algoritme di atas, diasumsikan kamus yang dimaksud hanya berisi kata-kata dasar sehingga pemeriksaan kebenaran ejaan kata-kata turunan akan membutuhkan prosedur pemilahan kata ke dalam *kata dasar* dan bentuk-bentuk *imbuhannya*. Prosedur pemilahan ini merupakan komponen penting yang kedua, yang mencakup tiga kemungkinan pemilahan kata berimbuhan: kata berawalan, kata berakhiran, serta kata berawalan dan berakhiran.

Untuk mengimplementasikan algoritme di atas sehingga dihasilkan pemeriksa ejaan otomatis yang berkinerja tinggi (cepat dan efisien) dibutuhkan pengetahuan karakteristik penggunaan kata dalam dokumen-dokumen berbahasa Indonesia. Pengetahuan ini sangat dibutuhkan dalam menyusun baik kamus baku, maupun prosedur pemilahan kata yang kami uraikan di atas; kamus harus disusun sedemikian rupa sehingga proses pencarian kata dapat dilakukan dengan cepat, sedangkan prosedur pemilahan kata harus dirancang sedemikian rupa sehingga pemilahannya dapat dilakukan dengan cepat tanpa mengurangi kebenaran hasil pemilahan itu sendiri.

3.2. Kebutuhan Penelitian Karakteristik Dokumen Berbahasa Indonesia

Berdasarkan kebutuhan algoritme pemeriksa ejaan di atas, kami mencoba memberikan gambaran tentang penelitian-penelitian kebahasaan yang mungkin dilakukan untuk membantu pembuatan suatu pemeriksa ejaan otomatis yang berkinerja tinggi.

Dalam penyusunan kamus baku, kita perlu menentukan ukuran kamus karena dampaknya yang besar pada kecepatan pencarian kata. Di samping itu karena penataan kata-kata dalam kamus tersebut amat mempengaruhi proses pemeriksaan ejaan secara keseluruhan, maka kita perlu mengetahui karakteristik kemunculan kata dalam suatu dokumen. Sebagai contoh, jika kita mengetahui kata-kata mana yang sering digunakan, maka kita dapat menyusun kamus sehingga kata-kata yang sering digunakan tersebut

merupakan kata-kata pertama yang akan dibandingkan pada proses pencarian kata.

Selanjutnya, dalam perancangan prosedur pemilahan kata, kita membutuhkan pengetahuan tentang bentuk-bentuk kata turunan mana saja yang sering digunakan atau yang tidak mungkin ada walaupun secara morfologis benar. Kesemuanya akan membantu pengoptimalan prosedur pemilahan kata.

4. Analisis Karakteristik Penggunaan Kata

Melihat minimnya hasil-hasil penelitian yang dapat kami gunakan untuk merancang modul pemeriksa ejaan berkinerja tinggi, kami mencoba melakukan beberapa kegiatan penelitian untuk memenuhi kebutuhan yang berkaitan dengan perancangan modul tersebut. Sebagaimana telah kami uraikan pada bagian terdahulu, kami ingin mengetahui karakteristik penggunaan kata dalam dokumen-dokumen berbahasa Indonesia.

Pada penelitian ini, kami menggunakan berkas-berkas tugas akhir mahasiswa S1 bidang Ilmu Komputer di Fakultas Ilmu Komputer, Universitas Indonesia dan Jurusan Informatika, Institut Teknologi Bandung sebagai dokumen sampel.

Pertama-tama, mari kita lihat beberapa data statistik sederhana yang berkaitan dengan dokumen-dokumen sampel tersebut. Tabel di halaman berikut menunjukkan jumlah kata[1] yang digunakan, jumlah jenis katanya, dan jumlah kemunculan tiap jenis kata untuk masing-masing dokumen.

[1]Jumlah kata tidak termasuk kata-kata pada bagian *Kata Pengantar*, *Daftar Isi/Gambar/Tabel*, *Gambar/Tabel*, dan *Lampiran*.

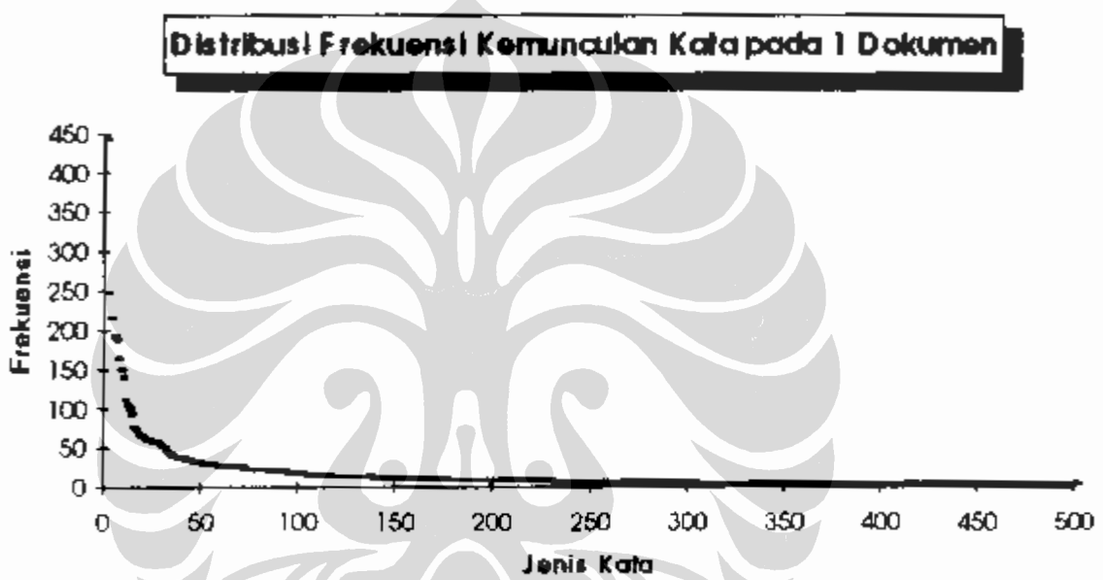
Dokumen Sumber	jumlah kata	jumlah jenis kata	rata-rata kata/jenis
dok01:	24956	2590	9,64
dok02:	17321	2104	8,23
dok03:	11241	1281	8,78
dok04:	10871	1431	7,60
dok05:	10553	1137	9,28
dok06:	9225	1446	6,38
dok07:	9129	1579	5,78
dok08:	9288	1534	6,05
dok09:	8146	1029	7,92
dok10:	6601	1201	5,50
dok11:	6591	1222	5,39
dok12:	6230	1237	5,04
dok13:	6283	1359	4,62
dok14:	5832	1037	5,62
dok15:	5807	1280	4,54
rata-rata:	9871,6	**1431,13**	**6,69**
total:	148074	9067	**16,33**

Selanjutnya, berdasarkan pengamatan data pada tabel di atas, kita dapat membuat beberapa catatan penting:

- Jumlah kata rata-rata yang dipakai pada setiap dokumen adalah sebesar 9871.6 kata. Dari kesembilan ribu lebih kata per dokumen tersebut, 1431.13 kata di antaranya merupakan kata-kata yang unik (jenis kata), selebihnya merupakan duplikasi. Jadi, setiap kata rata-rata digunakan sebanyak 6,69 kali. Pada kenyataannya, seperti yang nanti dapat dilihat pada pengamatan distribusi frekuensi kemunculan kata, kesimpulan ini agak menyesatkan karena pola sebaran yang tidak merata.
- Jika kita amati kelima belas dokumen sebagai satu kesatuan, maka ternyata kemunculan tiap kata meningkat dari 6,69 kali menjadi 16,33 kali. Angka ini, yang jauh lebih besar dari angka frekuensi kemunculan kata untuk satu dokumen, menunjukkan bahwa jumlah kata-kata yang digunakan pada lebih dari satu dokumen cukup besar.

4.1. Frekuensi Penggunaan Kata

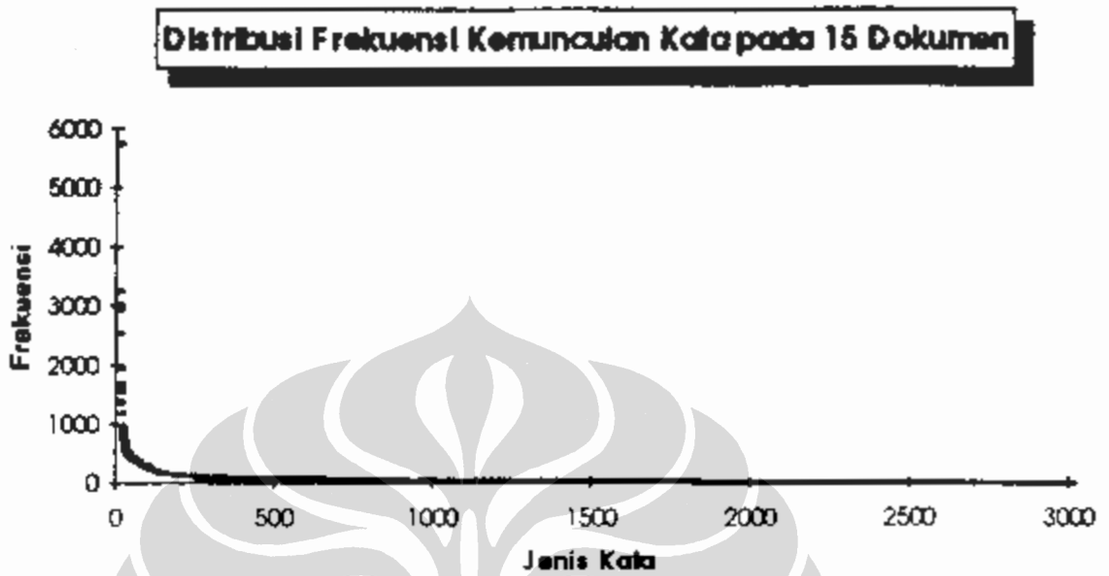
Untuk melengkapi hasil pengamatan di atas, kami mencoba mengamati karakteristik kemunculan kata lebih jauh dengan melihat distribusi frekuensi kemunculan kata sebagai fungsi dari jenis kata yang digunakan. Gambar berikut menunjukkan distribusi yang dimaksud untuk satu dokumen.



Terlihat bahwa pola sebarannya sangat tidak merata. Gambar di atas menunjukkan bahwa sebagian kecil kata digunakan amat sering, sedangkan sebagian besar lainnya digunakan amat jarang. Dengan kata lain, jumlah kata yang digunakan dalam satu dokumen ternyata didominasi oleh sekelompok kecil kata saja.

Untuk mengamati pola distribusi frekuensi kemunculan kata pada cakupan yang lebih luas, kami mengamati distribusi frekuensi kemunculan kata untuk kelima belas dokumen yang kami gunakan sebagai sampel. Ternyata polanya serupa. Bahkan, kesan kelompok kata minoritas mendominasi frekuensi kemunculan kata menjadi lebih kuat.[2]

[2]Pola seperti ini tampaknya universal sifatnya karena distribusi kemunculan kata dalam bahasa Inggris pun mirip.



No.	Jenis Kata	Frekuensi Kemunculan	No.	Jenis Kata	Frekuensi Kemunculan
1.	yang	5751	11.	data	1194
2.	dan	3252	12.	tersebut	947
3.	dengan	3022	13.	tidak	926
4.	pada	2933	14.	akan	909
5.	untuk	2524	15.	sistem	887
6.	dari	1955	16.	di	835
7.	ini	1936	17.	sebagai	754
8.	dalam	1688	18.	oleh	730
9.	dapat	1540	19.	satu	712
10.	adalah	1382	20.	dilakukan	674

Pengamatan lebih rinci menunjukkan hal-hal berikut:

- *lima puluh persen* dari jenis kata yang dipakai hanya muncul satu kali,
- perbedaan frekuensi kemunculan di antara kata-kata yang sering digunakan amat mencolok, sebagai contoh: kata peringkat pertama, *yang*, muncul sebanyak 5751 kali, sedangkan kata peringkat ke-dua puluh, *dilakukan*, muncul sebanyak 674 kali yang berarti hampir sepersepuluh frekuensi kemunculan kata peringkat pertama.

Selanjutnya, melalui analisis kualitatif kami dapat membuktikan kesimpulan kami sebelumnya bahwa beberapa kata ternyata memang sering digunakan pada lebih dari satu dokumen seperti yang dapat dilihat pada tabel berikut:

dok04:	dok06:	dok08:	dok10:	dok14:
yang	**yang**	**yang**	**yang**	*artikel*
fuzzy	**dan**	**dari**	*prosesor*	**yang**
dan	***pada***	*data*	**dan**	**untuk**
dengan	**dari**	**untuk**	***ini***	**dari**
linier	***ini***	**dan**	***pada***	*berkas*
himpunan	*cahaya*	**dengan**	**dengan**	***ini***
solusi	**untuk**	dalam	*algoritma*	tersebut
pada	**dengan**	*entity*	**untuk**	**dan**
untuk	*objek*	***pada***	**dari**	**dengan**
dari	*permukaan*	*attribute*	*proses*	*grup*

Tabel di atas menampilkan sepuluh kata yang paling sering muncul pada lima dokumen sampel. Terlihat bahwa kata-kata seperti *yang, dan, untuk, dari,* dan *dengan* muncul di kelima dokumen; beberapa kata lainnya seperti *ini* dan *pada* muncul di lebih dari dua dokumen.

Selain itu, memang terdapat kata-kata yang nampaknya "khas" bagi masing-masing dokumen seperti *artikel, berkas,* dan *grup* pada dokumen dok14 atau *prosesor, algoritma,* dan *proses* pada dokumen dok10. Bahkan, kami mendapatkan dokumen yang masih banyak menggunakan istilah-istilah asing seperti dokumen dok08: *entity* dan *attribute.*

4.2. Komposisi Kata Berimbuhan

Pada uraian terdahulu kami telah menyatakan bahwa salah satu komponen penting algoritme pemeriksa ejaan adalah prosedur pemilahan kata berimbuhan. Untuk merancang prosedur yang optimal, kita perlu mempelajari pola penggunaan kata-kata berimbuhan dalam dokumen. Sebagai langkah awal, kami mencoba mengamati distribusi kemunculan kata-kata yang diawali bentuk-bentuk awalan yang baku: *ber-, di-, ke-, me-, per-, se-,* dan *ter-.*

Berdasarkan "bank" kosa kata yang kami bangun dari kelima belas dokumen sampel yang digunakan pada penelitian ini, kami dapatkan bentuk-bentuk berimbuhan di atas mencakup *tiga puluh persen* dari seluruh kata yang terdapat di bank kosa kata tersebut. Lebih lanjut, di antara bentuk-bentuk berimbuhan tersebut, kata-kata yang berawalan *di-*, *me-*, dan *per-* merupakan kelompok terbesar, diikuti kelompok kata-kata yang berawalan *se-*, dan terakhir kelompok kata-kata berawalan *ber-*, *ke-*, dan *ter-*.

Jika diamati lebih jauh, dalam suatu kelompok kata yang diberi awalan tertentu, seperti *me-* misalnya, hanya mengandung sebagian kombinasi awalan-akhiran tertentu seperti terlihat pada tabel berikut:

Jenis kombinasi	Frekuensi kemunculan
me-	2929
me-i	2053
me-kan	3084
member-kan	136
memper	95
memper-i	22
memper-kan	135
menge-kan	40
menye-i	11
menye-kan	153

Hasil-hasil pengamatan di atas memang belum dapat menjawab seluruh kebutuhan pembuatan modul pemeriksa ejaan otomatis yang diinginkan. Masih diperlukan penelaahan lebih lanjut dengan menggunakan sampel yang lebih besar kandungan kosa katanya diikuti dengan analisis karakteristik penggunaan kata yang lebih rinci sifatnya. Walaupun demikian, hasil-hasil pengamatan awal tersebut telah membantu kami menentukan langkah-langkah lanjut pembuatan modul pemeriksa ejaan otomatis.

5. Penutup

Untuk menyempurnakan hasil-hasil penelitian yang telah diperoleh sehingga karakteristik penggunaan kata dalam dokumen-dokumen berbahasa Indonesia dapat diketahui dengan rinci, kami mengusulkan penelitian-penelitian berikut:

- Penelitian karakteristik penggunaan kata pada dokumen-dokumen berbahasa Indonesia dalam skala yang lebih besar, baik dalam cakupan jenis dokumennya maupun cakupan jenis analisis statistik katanya.
- Penelitian yang khusus mempelajari metode pemilahan kata-kata turunan, termasuk bentuk-bentuk kata yang kompleks seperti kata-kata gabungan yang berimbuhan, terutama pada teknik-teknik *heuristic* yang memungkinkan pemilahan kata dilakukan dengan cepat tanpa mengorbankan hasil pemilahannya.

6. Bahan Bacaan

Adriani, Mirna dan Bobby Nazief. Pemanfaatan Teknologi Informasi dalam Pengembangan Bahasa Indonesia." *Suara Pembaruan*. 6 Des. 1993. hal. 10.

Knuth, Donald E. *The Art of Computer Programming, Volume 3: Sorting and Searching*. Massachusetts: Addison-Wesley, 1973.

Peterson, James L. "Computer Programs for Detecting and Correcting Spelling Errors." *Communication of the ACM*, Vol. 23, No. 12, Des. 1980.

Tim Penyusun Kamus Pusat Pembinaan dan Pengembangan Bahasa. *Kamus Besar Bahasa Indonesia*. Edisi Kedua, Jakarta: Balai Pustaka, 1993.

Tim Tata Bahasa Baku Bahasa Indonesia. *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka. 1988.