ISSN 2477-0040
e-ISSN 2460-7900

Jurnal Teknologi Terpadu
Vol. 1, No. 2, Desember 2015

# PENYUSUNAN KORPUS BERITA TERBUKA BERBAHASA INDONESIA

**Ahmad Rio Adriansyah**
STT Terpadu Nurul Fikri, arasy@nurulfikri.ac.id , ahmad.rio.adriansyah@gmail.com

## Abstrak

Korpus dalam bahasa Indonesia dibutuhkan untuk menunjang penelitian dalam bahasa atau sistem temu kembali informasi. Sebelumnya, untuk membuat sebuah korpus dibutuhkan waktu yang lama dan biaya yang mahal. Tapi sejak internet mulai populer dan laman web semakin banyak, informasi yang menggunakan bahasa tertentu jadi lebih mudah diakses sehingga penyusunan korpus pun menjadi lebih cepat dan murah. Dalam penelitian pengolahan bahasa alami juga dibutuhkan korpus yang sama untuk membandingkan dua buah metode yang berbeda. Tapi sayangnya korpus berbahasa Indonesia yang terbuka masih minim. Ada yang menyediakan tetapi hanya bisa diakses melalui website tersebut saja. Karena pertimbangan kecepatan jaringan dan kecepatan proses, terkadang dibutuhkan korpus yang bisa diakses lokal. Penelitian ini menyediakan korpus khusus yang diambil dari beberapa laman web berita, metode pengambilan, beserta statistiknya. Korpus yang dihasilkan dari metode ini dapat digunakan secara terbuka oleh peneliti lain untuk diolah secara lokal.

**Kata Kunci:** Korpus, Sistem Temu Kembali Informasi, Bahasa Indonesia, Open Data

## 1. PENDAHULUAN

Istilah korpus (*corpus*) digunakan untuk menjelaskan sekumpulan dokumen, baik berbentuk tulisan atau lisan, yang disimpan dan diproses di dalam komputer untuk tujuan penyelidikan linguistik. (Renouf, A. 1988).

Penggunaan komputer untuk mengolah data linguistik sudah dimulai sejak tahun 1961. Yaitu pada saat W. Nelson Francis dan Henry Kucera mulai menggunakan komputer untuk menyusun Brown Corpus di Universitas Brown, AS. Penyusunan korpus tersebut bertujuan untuk meneliti penggunaan kata-kata tertentu dalam bahasa Inggris dengan cara mengumpulkan dokumen-dokumen berbahasa Inggris.

Brown Corpus terdiri dari 500 sampel teks yang masing masing memiliki sekitar 2000 kata. Teks tersebut dikompilasi dari 15 kategori yang berbeda agar dapat digunakan sebagai referensi standard yang baik. Jika dibandingkan dengan korpus lain dewasa ini, Brown Corpus termasuk berukuran kecil (hanya sekitar 1 juta kata), dan tidak mutakhir (datanya dikumpulkan dari teks tahun 1960an), tapi masih digunakan.

Pada era yang sama, Randolph Quirk juga melakukan pengumpulan data bahasa untuk penelitian tata bahasa Inggris secara empiris. Saat itu datanya belum terkomputerisasi, hingga pertengahan tahun 1980an, Quirk dan Greenbaum melakukan komputerisasi terhadap data bahasa tersebut yang lebih dikenal dengan nama International Corpus of English (ICE).

Pada tahun 1980an pula, Universitas Birmingham dan Collins mulai merancang dan memulai kerjasama pengelolaan korpus yang dikenal dengan korpus Birmingham.

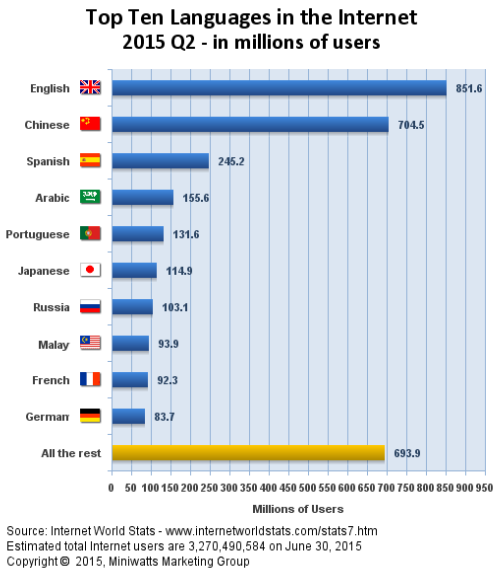
Sejak saat itu pengembangan korpus berkembang pesat. Pengembangan tersebut tak terbatas hanya pada korpus berbahasa Inggris, tetapi juga bahasa lain. Korpus yang umum digunakan tercantum pada tabel 1.

Tabel 1.
Daftar Korpus

| No | Nama Korpus | | Bahasa | Alamat | #Kata |
|---|---|---|---|---|---|
| 1 | Brown Corpus | | Inggris | https://archive.org/details/BrownCorpus | 1 juta |
| 2 | British National Corpus | | Inggris British | http://www.natcorp.ox.ac.uk/ | 100 juta |
| 3 | Corpus of Contemporary American English | COCA | Inggris Amerika | corpus.byu.edu/coca/ | 520 juta |
| 4 | International Corpus of English | ICE | Inggris (variasi sesuai daerah) | http://ice-corpora.net/ice/index.htm | |
| 5 | SEAlang | | Bahasa nasional di Asia Tenggara dan beberapa bahasa daerahnya (termasuk Indonesia) | http://sealang.net/library/ | |
| 6 | Korpus PoS Tag Bahasa Indonesia | | Indonesia | http://bahasa.cs.ui.ac.id/postag/corpus | 256 ribu |

Yang paling populer adalah korpus teks berbahasa Inggris karena pengguna bahasa Inggris di dunia, baik lisan maupun tulisan masuk ke dalam 10 besar bahasa yang paling banyak digunakan. Tapi jika kita memandang bahasa Indonesia, potensi bahasa Indonesia juga besar karena penduduknya di atas 200 juta orang.

Diagram 1. Sepuluh Bahasa yang Paling Banyak Digunakan di Internet (dalam juta pengguna).
Sumber : http://www.internetworldstats.com/stats7.htm




Sementara itu, upaya untuk pengembangan korpus berbahasa Indonesia sebagai sarana penelitian dalam bidang bahasa (linguistik), pengolahan bahasa alami, temu balik informasi, dll masih terbatas. Manurung dkk dari Universitas Indonesia, Faisal dan Rohadi dari Politeknik Negeri Malang, juga membuat korpus berbahasa Indonesia. Beberapa korpus yang terbuka dapat diakses hanya melalui web lembaga yang bersangkutan, seperti Universitas Indonesia (http://bahasa.cs.ui.ac.id/postag/corpus), atau SEAlang (http://sealang.net/library/).

Untuk pengembangan algoritma atau metode baru, objek yang digunakan sebagai penguji dalam perbandingan diharapkan dapat berupa objek yang sama, agar perbandingan jadi *apple to apple*. Karena itu data yang bersifat terbuka sangat penting.

Selain itu, kebutuhan untuk proses yang cepat terkadang membutuhkan data yang disimpan secara lokal, tidak melalui jaringan. Hal ini juga berguna bagi peneliti yang berada di daerah yang tidak terhubung ke dalam jaringan internet setiap saat.

Seperti yang disampaikan pula oleh Rahutomo dan Erfan [2], bahwa sistem penunjang SKTI yang bersifat terbuka belum tersedia, termasuk di dalamnya koleksi dokumen (korpus) yang bisa diakses offline. Karena itu penulis menyusun sebuah korpus khusus berbahasa Indonesia yang datanya bersifat terbuka.

## 2. METODE

Sebuah laman korpus yang bernama Sketch Engine (http://sketchengine.co.uk/) membuat korpus menggunakan bantuan search engine melalui tahapan-tahapan berikut :
1. mengumpulkan daftar 'seed word' dari beberapa ratus kata berfrekwensi menengah dalam suatu bahasa.
2. ulangi beberapa kali (hingga korpusnya berukuran cukup besar) :
a. pilih 3 kata untuk membuat sebuah query
b. kirimkan query tersebut ke search engine populer (Google, Yahoo, Bing) yang mengembalikan halaman 'hasil pencarian'.
c. buka halaman pada hasil pencarian tersebut, lalu disimpan.
3. Bersihkan teks dari navigasi bar, iklan, dan skrip lain yang muncul berulang.
4. Hapus duplikat
5. Tokenisasi, lematisasi, dan beri tag POS jika memungkinkan.
6. Masukkan ke dalam perangkat untuk corpus query

Penelitian yang dilakukan oleh Kilgarriff dkk [1] tersebut menghasilkan korpus umum yang berukuran besar. Korpus tersebut akan terdiri dari beragam kategori dan mungkin tercampur dengan bahasa lain, tergantung terhadap respon dari search engine yang bersangkutan.

Berbeda dengan penelitian tersebut, pada penelitian ini, yang akan dibuat adalah korpus khusus. Yaitu korpus yang sumbernya diambil dari web berita. Dokumen korpus ini disusun dari artikel berita yang bisa didapat dari web yang bisa diakses oleh umum seperti detik.com dan kompas.com. Korpus dari website berita dipilih karena informasi yang terkandung di dalamnya. Meskipun banyak redundansi (perulangan) kalimat, kalimat yang dihasilkan oleh website berita cukup banyak dan beragam. Bahasa yang digunakan dalam penyampaian berita biasanya adalah bahasa formal atau semi formal.

Alur penyusunan korpus ini adalah dengan menggerayang (crawling) web. Ada 2 metode yang digunakan.

Metode 1. Berdasarkan Indeks Berita
Input awal :
alamat web utama (**web**),
alamat web indeks (**indeks**),
tanggal awal (**ta**) ,
dan tanggal akhir (**tb**)

Proses :
1. untuk setiap tanggal **day** dari **ta** s.d. **tb**,
a. buka halaman **indeks** berita tanggal **day** tersebut
b. catat semua link pada halaman tersebut, masukkan ke dalam **daftarlink** tanpa duplikasi
2. untuk setiap **link** dalam **daftarlink**,
a. periksa apakah **link** masih mengandung **web**
b. jika tidak, drop **link** tersebut, lanjutkan **link** selanjutnya dari poin 2.a.
c. jika iya, buka laman **link**
d. simpan
3. bersihkan halaman tersebut dari script, tag, dan informasi lain yang tidak berkaitan dengan isi berita.
4. simpan informasi berupa judul, penulis, isi berita ke dalam file teks.

Alur di atas mengecualikan halaman-halaman khusus seperti login, juga halaman yang sepenuhnya berupa gambar atau video. Link yang tidak dapat diakses karena lamannya sudah menghilang atau error karena salah alamat juga tidak dimasukkan.

Menyimpan informasi halaman pada poin 2.d. memang memakan memori yang banyak. Bagi sebagian orang dianggap tidak efektif. Tapi informasi yang tersimpan tersebut dapat digunakan untuk mengecek ulang informasi pada website tanpa membebani kembali website tersebut.

Jika halaman indeks tidak ditemukan, maka pengambilan laman-laman dari web dilakukan dengan cara kedua. Dengan crawling menggunakan metode pencarian breadth first search (BFS) menggunakan halaman utama web tersebut sebagai seed.

Metode 2. BFS
Input awal :
halaman utama web sebagai **seed**

Proses :
1. buka **seed**
2. catat semua link yang terkandung dalam halaman tersebut ke dalam **daftarlink**
3. untuk semua **link** dalam **daftarlink** :
a. periksa apakah **link** termasuk ke dalam kategori laman berita berupa teks
b. jika tidak, maka **link** tersebut dicatat ke dalam daftar pengecualian, dan dilanjutkan ke **link** selanjutnya mulai dari 3.a.
c. jika iya, buka laman **link**
d. catat semua **link** pada halaman tersebut, tambahkan ke **daftarlink** jika belum pernah diakses
e. simpan
4. bersihkan halaman tersebut dari script, tag, dan informasi lain yang tidak berkaitan dengan isi berita.
5. simpan informasi berupa judul, penulis, isi berita ke dalam file teks.

Implementasi dari kedua algoritma tersebut sudah mempertimbangkan beban server. Dalam pengambilan setiap lamannya diberikan jeda (*delay*).

Yang membedakan dengan [1], selain konten yang lebih terarah, hasil dari pengumpulan informasi tersebut tidak disimpan dalam perangkat untuk corpus query tertentu, melainkan dalam bentuk file. File ini dapat disalin oleh peneliti yang membutuhkan, diproses dalam bentuk yang sesuai kebutuhan.

Selain dipisahkan berdasarkan websitenya, dokumen juga dipisahkan berdasarkan kanal informasi dan waktunya (dalam tahun). Pembagian berdasarkan kanal ditujukan untuk mempermudah pengguna yang membutuhkan informasi terkait bidang tertentu seperti teknologi, atau kesehatan. Disusun per tahun agar dapat digunakan juga sebagai korpus historis.

## 3. URAIAN PENELITIAN

Algoritma pada bagian sebelumnya diimplementasikan untuk dua buah website berita populer, yaitu Detik.com dan Kompas.com dalam rentang waktu dari 1 Januari 2014 hingga 31 Desember 2014. Kedua website tersebut memiliki indeks sehingga metode pertama yang digunakan.

Korpus tersebut dibagi jadi beberapa file untuk mempermudah pengolahan. Tiap file dialokasikan untuk rentang waktu satu tahun dari sebuah website, kanal tertentu. Pembagian per kanal juga dilakukan untuk membedakan kemunculan istilah tertentu dalam bidang yang sama. Dari situ pengguna bisa mencari dan menuliskan istilah yang terkait suatu bidang (*glossary*).

Selain dari kanal berita (*news*), diambil pula dari kanal lain. Yang ditampilkan dalam paper ini hanya 4 kanal, yaitu berita, kesehatan, ekonomi, dan teknologi. Tiap file diberi kode sesuai dengan website, kanal, dan tahunnya.

Tabel 2.
Daftar Nama File, Sumber, dan Ukuran Korpus

| No | Nama File | Sumber | Ukuran |
|---|---|---|---|
| 1 | KH2014 | health.kompas.com | 7,1 MB |
| 2 | KE2014 | bisniskeuangan.kompas.com | 19,7 MB |
| 3 | KN2014 | kompas.com | 140.2 MB |
| 4 | KT2014 | tekno.kompas.com | 8,1 MB |
| 5 | DH2014 | health.detik.com | 19,0 MB |
| 6 | DF2014 | finance.detik.com | 43,8 MB |
| 7 | DN2014 | news.detik.com | 136,7 MB |
| 8 | DI2014 | inet.detik.com | 14,4 MB |

Dari hasil implementasi tersebut, didapat 71.436 artikel dari kanal berita website news.detik.com dan 83.545 artikel dari kanal berita kompas.com. Data statistik dari kumpulan artikel tersebut ditampilkan dalam tabel 3.

Tabel 3.
Data Statistik Artikel dari Beberapa Website tahun 2014

| No | Nama File | # artikel | # Kata | # Kata Unik |
|---|---|---|---|---|
| 1 | KH2014 | 2.828 | 980.201 | 41.281 |
| 2 | KE2014 | 10.350 | 2.742.234 | 66.046 |
| 3 | KN2014 | 71.436 | 19.453.245 | 214.319 |
| 4 | KT2014 | 3.886 | 1.132.712 | 43.219 |
| 5 | DH2014 | 11.075 | 2.694.257 | 73.239 |
| 6 | DF2014 | 23.085 | 6.336.933 | 95.960 |
| 7 | DN2014 | 83.545 | 19.452.949 | 226.757 |
| 8 | DI2014 | 7.733 | 2.034.535 | 63.126 |

Dari file-file tersebut, ditampilkan data 20 kata (*token*) dengan frekwensi terbanyak dari file KN2014 dan DN2014 pada tabel 4 dan tabel 5.

Tabel 4.
Kata-kata Dengan Kemunculan Terbanyak pada Korpus KN2014

| KN 2014 | | | |
|---|---|---|---|
| no | token | frekwensi | % |
| 1 | yang | 430770 | 2.21% |
| 2 | di | 387187 | 1.99% |
| 3 | dan | 336346 | 1.73% |
| 4 | itu | 205655 | 1.06% |
| 5 | ini | 150615 | 0.77% |
| 6 | untuk | 148754 | 0.76% |
| 7 | dengan | 143970 | 0.74% |
| 8 | dari | 141009 | 0.72% |
| 9 | tidak | 122015 | 0.63% |
| 10 | 2014 | 117070 | 0.60% |
| 11 | akan | 109593 | 0.56% |
| 12 | dalam | 105107 | 0.54% |
| 13 | ke | 89784 | 0.46% |
| 14 | tersebut | 89640 | 0.46% |
| 15 | pada | 89299 | 0.46% |
| 16 | kata | 85737 | 0.44% |
| 17 | juga | 83930 | 0.43% |
| 18 | Jakarta | 82616 | 0.42% |
| 19 | ada | 82340 | 0.42% |
| 20 | com | 80397 | 0.41% |

Tabel 5.
Kata-kata Dengan Kemunculan Terbanyak pada Korpus DN2014

| DN 2014 | | | |
|---|---|---|---|
| no | token | frekwensi | % |
| 1 | yang | 437589 | 2.25% |
| 2 | di | 420480 | 2.16% |
| 3 | dan | 320349 | 1.65% |
| 4 | ini | 189671 | 0.98% |
| 5 | itu | 184558 | 0.95% |
| 6 | dengan | 144423 | 0.74% |
| 7 | dari | 138155 | 0.71% |
| 8 | untuk | 133954 | 0.69% |
| 9 | Jakarta | 133677 | 0.69% |
| 10 | tidak | 112307 | 0.58% |
| 11 | ke | 109196 | 0.56% |
| 12 | 2014 | 107040 | 0.55% |
| 13 | ada | 100822 | 0.52% |
| 14 | dalam | 100330 | 0.52% |
| 15 | akan | 98572 | 0.51% |
| 16 | juga | 92045 | 0.47% |
| 17 | tersebut | 80562 | 0.41% |
| 18 | sudah | 71588 | 0.37% |
| 19 | pada | 68822 | 0.35% |
| 20 | kata | 67908 | 0.35% |

## 4. KESIMPULAN RINGKASAN

Penelitian ini menghasilkan sekumpulan dokumen yang bersifat terbuka untuk digunakan oleh komunitas yang meneliti pemrosesan bahasa natural atau bahasa Indonesia.

Korpus yang diambil dengan metode di atas belum diberikan tag untuk PoS (*Part-of-Speech*). Untuk tagging, dapat digunakan tagger otomatis seperti yang disampaikan oleh [3] atau secara manual.

Yang dicantumkan dalam jurnal ini adalah sebagian kecil data yang sudah berhasil diambil. Ke depannya, korpus ini akan dikembangkan ke website lain dan dengan rentang waktu yang lebih lebar.

Data tersebut dapat diakses secara umum melalui email penulis.

## 5. DAFTAR PUSTAKA

[1] A. Kilgarriff, S. Reddy, J. Pomikalek, Avines PVS, "A Corpus Factory for Many Languages". 2010. *https://www.sketchengine.co.uk/*

[2] F. Rahutomo, R. Erfan, "Pengembangan Piranti Penelitian Sistem Temu Kembali Informasi Bahasa Indonesia". *Seminar Nasional Sistem Informasi Indonesia (SESINDO). Nov 2015*, pp. 313-319, 2015.

[3] F. Rashel, A. Luthfi, A. Dinakaramani, R. Manurung, "Building an Indonesian Rule-Based Part-of-Speech Tagger". In International Conference on Asian Language Processing (IALP 2014). *http://bahasa.cs.ui.ac.id/postag/tagger*

[4] R. Manurung, B. Distiawan, D.D. Putra, "Developing an Online Indonesian Corpora Repository". *PACLIC 24 Proceeding*. 2010