

RESENSI BUKU

Judul : **Quantitative Corpus Linguistics with R: A Practical Introduction (2nd Edition)**
Penulis : Stefan Th. Gries
ISBN : 9781138816275 (hardback)
Penerbit : Routledge Taylor & Francis (2017)
Tebal : 274 hal.

Ferdinan Okki Kurniawan
Universitas Katolik Indonesia Atma Jaya
Ferdinan.Kurniawan@atmajaya.ac.id

Dalam beberapa dekade terakhir, penggunaan korpus untuk kajian linguistik semakin marak. Ada beberapa hal yang menyebabkan korpus makin banyak digunakan. Salah satu di antaranya adalah semakin majunya teknologi yang memungkinkan para ahli bahasa untuk mengkaji korpus teks maupun tuturan dalam skala besar. Sifat korpus yang memiliki data yang bersifat natural membuat korpus semakin diminati. Jika dibandingkan dengan penelitian yang berbasis eksperimen, di mana peneliti memiliki peran kontrol yang besar dalam pengumpulan data, penggunaan korpus memungkinkan peneliti bahasa mendapatkan hasil penelitian yang lebih alamiah.

Namun demikian, bukanlah berarti penggunaan korpus tanpa tantangan. Salah satu yang dihadapi oleh para ahli bahasa adalah pengolahan data korpus secara statistik. Bagaimana data yang diambil dari korpus dapat memberikan hasil yang valid dan informatif secara statistik juga bukan suatu pertanyaan yang mudah dijawab. Di samping itu, data korpus yang cenderung besar dan tidak seimbang merupakan tantangan tersendiri yang dihadapi para peneliti bahasa. Banyak pelatihan dan buku yang menawarkan pengantar ilmu statistik yang biasanya menggunakan data yang berasal dari cabang ilmu lain seperti ekonomi atau sosial. Hal ini mungkin dapat membantu namun juga cukup menyulitkan para ahli bahasa ketika harus menerapkan metode analisa statistik dari cabang ilmu lain ke dalam data bahasa. Belum banyak buku statistik yang mengkaji secara khusus data bahasa terutama data korpus.

Di tengah kekurangan tersebut, para ahli bahasa yang menekuni kajian korpus boleh bersenang hati karena Stefan Th. Gries, seorang profesor linguistik dari University of California, Santa Barbara, di tahun 2017 menyusun buku yang berjudul *Quantitative Corpus Linguistics for Linguistics with R: A Practical Introduction*. Analisa statistik dalam buku ini menggunakan data yang diambil dari korpus bahasa. Hal ini memudahkan ahli bahasa untuk memahami dan menerapkan analisa statistik yang digunakan di buku ini pada data korpus yang mereka gunakan.

Meskipun ada kata-kata *Corpus Linguistics*, buku ini tidak menawarkan pembahasan tentang bagaimana mengumpulkan korpora bahasa, menganotasi tuturan, strategi melakukan *tagging* dan *parsing*, teori korpus, dan aspek-aspek korpus lain yang biasa dibicarakan dalam buku-buku pengantar korpus linguistik lainnya. Buku ini disusun secara khusus untuk melakukan analisa statistik pada korpus bahasa dengan menggunakan perangkat lunak R (R Core Team, 2017). Tidak seperti perangkat lunak statistik lain yang cenderung berbayar, perangkat lunak R merupakan *open-source* yang dapat diunduh dan digunakan secara gratis.

Meskipun pembaca memerlukan kemampuan bahasa pemrograman untuk mengoperasikan perangkat lunak *R*, Gries menyusun buku ini dengan gaya bahasa yang praktis, informatif, dan dapat dimengerti bahkan oleh pembaca yang memiliki kemampuan statistik dan bahasa pemrograman yang terbatas. Kata *Practical* di judul buku ini merupakan kata kunci dalam kesuksesan pembaca mempelajari materi di dalamnya. Buku ini bukanlah buku yang bisa dibaca di taman, ruang baca, pesawat, atau perpustakaan tanpa mempraktikkan meterinya langsung di perangkat lunak *R* pada layar komputer.

Buku ini dibuka dengan pembahasan definisi korpus. Pembaca diajak untuk mendapatkan gambaran tentang jenis-jenis korpus dan juga tentang empat metode utama dalam korpus bahasa, yaitu daftar frekuensi, dispersi, kolokasi, dan konkordansi. Meskipun buku ini bukan merupakan pengantar perangkat lunak *R*, namun pengenalan dan dasar-dasar penggunaan perangkat lunak *R* juga dijelaskan. Penjelasannya meliputi cara mengunduh dan meng-*install* *R* untuk beberapa macam sistem operasi. Pembaca dipandu untuk dapat menggunakan *RStudio* maupun *R console* yang merupakan panel untuk menulis *R script*. Di bab ini, penulis juga menjelaskan tentang cara-cara dasar pengoperasian *R*, di antaranya cara membaca *data structure*, *functions*, *arguments*, *vectors*, dan *factors*. Yang juga penting dalam buku ini bagi ahli bahasa adalah cara menyusun *data frame* yang akan sangat berpengaruh dalam analisa statistik korpus mereka. Gries juga menyampaikan kesalahan-kesalahan umum yang dilakukan para ahli bahasa ketika menyusun *data frame* yang diambil dari korpus. Sebagai pelengkap buku ini, Gries menyediakan tautan pelengkap yang salah satu isinya berupa *input files* di mana pembaca dapat melihat contoh-contoh *data frame* untuk analisis statistik korpus.

Buku ini tidak hanya menawarkan hal-hal yang melulu teknis. Gries juga menjelaskan hal-hal yang bersifat fundamental seperti jenis-jenis variabel dan penyusunan hipotesis. Ia mendiskusikan peran variabel bebas dan variabel terikat dalam penelitian korpus. Selain itu, nilai informasi yang dimiliki oleh variabel, yaitu kategoris dan numerik, juga dijelaskan. Sebagai contoh, variabel *stress* (fonologi): *stressed vs. unstressed*, variabel *animacy* (semantik): *animacy vs. non-animacy* memiliki nilai variabel kategoris. Untuk variabel yang bersifat numeris, Gries mengambil contoh dari nilai frekuensi dalam skala Hertz (fonetik), frekuensi kata dan jumlah klausa dalam korpus (sintaks).

Cara melakukan tes statistik dengan menggunakan perangkat lunak di *R* juga dicontohkan di buku ini. Yang terpenting dalam bagian ini adalah ketika Gries menyampaikan bahwa tidak semua variabel dalam kajian korpus memiliki sifat yang sama. Oleh sebab itu, tes statistik yang digunakan harus sesuai dengan skenario yang dipakai dalam kajian korpus tersebut. Contohnya adalah jika sebuah kajian korpus tidak memiliki variabel bebas dan hanya memiliki variabel terikat yang bersifat kategoris, maka tes statistik yang digunakan adalah *chi-squared test for goodness of fit*. Lain lagi jika sebuah kajian korpus memiliki sebuah variabel bebas dan variabel terikat yang dimiliki bersifat kategoris. Dalam hal ini tes statistik yang digunakan adalah *chi-squared test for independence*.

Walaupun ditulis dengan baik, buku ini tidaklah disusun tanpa kekurangan. Salah satunya adalah contoh file *data frame* dalam bentuk csv yang disediakan di tautan pelengkap bukanlah suatu bentuk yang memadai ketika pembaca buku ini hendak menggunakannya sebagai model penyusunan *data frame* di file csv. File csv yang disediakan oleh Gries sepertinya merupakan file yang dihasilkan dari perangkat lunak *R*. Cara ini menghasilkan variabel-variabel yang tersusun sekaligus dalam satu kolom pada file csv dan bukanlah *data frame* yang disusun secara mentah

dari awal penelitian. Hal ini menyulitkan pembaca ketika hendak menggunakan *input files* sebagai contoh pembuatan *data frame* penelitian mereka.

Kekurangan lain adalah tidak didiskusikannya tes statistik untuk kajian korpus yang memiliki dua atau lebih variabel bebas. Variabel bebas ini sangat umum dijumpai dalam kajian sosiolinguistik atau cabang ilmu linguistik lain yang melibatkan faktor non-linguistik, seperti latar belakang pendidikan penutur, jenis kelamin, umur, dan sebagainya. Contohnya: penutur dengan latar belakang pendidikan tinggi dan berjenis kelamin wanita (dua variabel bebas) memproduksi varian linguistik A (variabel terikat) ketimbang varian linguistik B (variabel terikat) atau sebaliknya. Kajian yang melibatkan lebih dari satu faktor non-linguistik dalam penggunaan korpus belum terjawab di buku ini.

Terlepas dari kekurangannya, materi di dalam buku ini sangat layak untuk dipelajari, terutama oleh para ahli bahasa yang menggunakan korpus dalam kajiannya. Kehadirannya sangatlah membantu para ahli bahasa, terutama yang belum terbiasa betul dengan teknik statistik dan bahasa pemrograman.

Referensi

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.