MENJARING DATA DARI TEKS

Muhadjir

1. Pengantar

Penyaji makalah ini bukan ahli atau tehnolog kumputer, tetapi pemakai biasa saja, yang memiliki sedikit pengalaman penggunaan komputer untuk penelitiannya. Penyajian pengalaman di forum ini lebih bersifat memancing masukan daripada memberikan informasi.

Selama ini para peneliti bahasa tampak merasa tidak perlu benar akan bantuan teknologi komputer sebagai alat bantu untuk penelitian struktur intern bahasa. Memang penelitian struktur bahasa yang dilaksanakan dengan 500-600 contoh kalimat, masih dapat dilakukan dengan tidak terlalu memerlukan peralatan pembantu; sekalipun demikian dengan alat bantu komputer akan lebih cepat dan teliti. Tetapi, bilamana kita ingin merencanakan tata bahasa yang bersifat komprehensif, dengan polulasi yang bersifat nasional, seperti misalnya bila kita ingin menyusun tata bahasa seperti dilakukan Rondolph Quirk untuk tata-bahasa bahasa Inggris *A Comprehensive Grammar of The English Grammar* (1985), bantuan komputer mutlak diperlukan. Sama halnya kalau kita ingin menyusun kamus berdasarkan bahasa yang hidup, yang memerlukan contoh yang sebenarnya hidup dalam masyarakat bahasa, bantuan komputer mutlak diperlukan.

2. Berikut beberapa contoh penggunaan program yang saya lakukan untuk pengolahan data dari teks. Perangkat lunak yang saya pergunakan adalah **Word Perfect**. selanjutnya saya singkat WP; jadi program pengolah kata yang biasa saja. Tentu lebih lambat, kurang akurat dibanding dengan penggunaan yang dilakukan oleh para pakar komputer yang secara khusus menyusun program sesuai dengan keperluannya.

Setelah korpus ditentukan berdasarkan metode penelitian yang kita pilih, maka korpus yang berasal dari bahasa lisan maupun bahasa tulis harus diketik lebih dahulu dalam berkas komputer. Bila penelitin yang kita lakukan menyangkut satuan-satun fonemis, seperti fonomiks atau

morfofonemiks, maka harus diketik dalam huruf-huruf fonemis. Kebetulan progran WP memiliki huruf-huruf yang kita perlukan, huruf fonetis.

Kunci kegunaan program ini ialah kemampuannya untuk mengenali kembali lambang-lambang yang kita masukkan dalam berkas data. Seperti umumnya program pengolah kata. WP dapat diperintahkan untuk: mencari lambang-lambang huruf dan tanda-tanda baca, memblok bagian yang kita tunjuk, membedakannya dari bagian lain yang tidak diblok, lalu mengutip atau memindahkannya. Serta mengatur data yang sudah kita kutip sedemikian rupa sehingga memudahkan kita mengidentifikasi pola-pola yang sama untuk melakukan klasifikan. Agar perintah-perintah yang kita berikan bekerja cepat kita dapat menggunakan fasilitas yang disebut macra.

3. Contoh-contoh Penarikan Data dari Teks

Untuk tidak membuang-buang waktu, saya akan mulai saja dengan contoh yang saya lakukan dalam penelitian-penelitian saya.

3,1 Penelitian fonologi

Penelitian fonologi, lazimnya dilakukan terhadap bahasa yang belum pernah diteliti, yang belum diketahui jumlah fonem dan sistem yang terdapat di dalamnya. Langkah pertama biasanya peneliti melakukan fonemisasi, termasuk penyusunan sistem fonemnya. Kalau kita melakukan fonemisasi bahasa yang belum kita kenal lebih baik kita mengumpulkan datanya lewat rekaman data yang diperlukan baik dengan teknik penterjemahan, maupun dengan teknik memancing dan hasilnya direkam. Pengolahan data dapat dilakukan secara manual, dengan kartu-kartu, atau dengan menggunakan vasilitas yang disediakan komputer. Bila cara kedua yang dipilih, maka hasil rekaman itu lebih dahulu harus di ketik ke dalam berkas komputerm; tentu saja dengan menggunakan huruf fonetik. Jadi sama dengan apabila kita menggunakan teks sebagai korpus untuk penelitian struktur bahasa.

3.2 Mencari data Pasangan Minimal

Penelitian fonologi biasanya menyangkut identifikasi bunyi bahasa secara

Namun harus disadari bahwa pengelompokan huruf dalam komputer kita tidak sejajar dengan klasifikasi bunyi fonetis berdasarkan artikulasi. Jadi peneliti harus menyusunnya lagi menurut klasifikasi fonetis untuk memperoleh data pasangan minimal yang kita perlukan untuk menentukan apakah sebuah atau sekelompok bunyi yang berdekatan secara fonetis itu merupakan fonem atau hanya variasi fonetis dari fonem yang sama.

Dari contoh tersebut kita menemukan kata *galah*. Bila kita ingin mengetahui apakah *g* dalam bahasa yang kita teliti mempuyai perbedaan fungsional dengan kelompok bunyi yang berdekatan secara fonetis atau tidak, maka kita tinggal lagi mencari bentuk yang mengandung susunan bunyi *Konsonan + alah*. Dalam korpus kita temukan bentuk *kalah* dan *malah* yang dengan cara yang mudah dan cepat dapa kita identifikasi bila telah disorter menurut alfabet.

Data untuk penelitian ini tentunya berupa bentuk-bentuk yang kita sebut kata, Kalau misalnya kita mulai dengan menggunakan daftar kata Swadesh, di antara daftar itu akan ada bentuk kata kerja, sehingga dapat diduga ada bentuk-bentuk yang polimorfemis. Oleh sebab itu pertama-tama kita harus memisahkan kata-kata polimorfemis menjadi bentuk-bentuk monomorfemis. Jadi mau tidak mau kita sebenarnya harus lebih dahulu melakukan penelitian morfofonemiks.

3.3 Mencari bentuk kanonik

Kalau kita sudah mengetahui bentuk-bentuk monomorfemis ketika kita menghadapi teks yang cukup besar, maka kita tinggal lagi mengambil bentuk-bentuk dasar dari teks, lalu melakukan hal yang sama seperti dijelaskan. Selain itu dari daftar yang sama juga sekaligus kita mengetahui distribusi tiap fonem bahkan fonotaktiknya. Dan dengan demikian, kita juga sekaligus dapat menyusun bentuk-bentuk kanonik yang kita perlukan dalam fonologi dan morfologi.

Marilah sekarang kita coba menggunakan teks untuk mengetahui jumlah dan sistem fonem bahasa Melayu. Saya akan menggunakan teks dari hikayat lama yang cukup dikenal, yaitu *Hikayat Raja Pasai* yang kebetulan sudah ada dalam berkas komputer saya.

Untuk mencari pola bentuk monomorfemis, kita pisahkan bentuk-bentuk yang ada dalam teks menurut jumlah fonem, misalnya, dengnn menggunakan perintah agar komputer mencari bentuk-bentuk berfonem dua, tiga, empat, dan seterusnya hingga bentuk-bentuk yang sebesar-besarnya. Dari naskah Melayu itu, kita susun bentuk-bentuk seperti terpampang dalam Daftar 2.:

Daftar 2

2	3	4	5	6	7	8	9	10
di	lah	oraŋ	hamba	həndak	kubunuh	bərkenan	pərəmpuan	dimandikan
ɲa	yaŋ	maka	dalam	*alaihi*	*kəmbali*	*Samudəra*	*hulubalaŋ*	bərbunuhan
si	ada	kata	kapal	tunduk	bərmula	məmaŋgil	bərtəmpat	peŋlihatan
ia	itu	raja	*hatta*	səoraŋ	bərnama	bəraŋkat	*səmbahyaŋ*	*bersəmbuyi*
kə	pun	alam	Merah	*səgala*	pərdana	bərjalan	bərmulaan	[illegible]
wa	ini	dari	mimpi	kəluar	*Məntəri*	bərsabda	bərwasiat	bərkənalan
hɛ	dan	anak	dəŋan	bərapa	məlawan	bərdataŋ	pəkərjaan	*gəmərinciŋ*
kɘ	saŋ	cucu	əmpat	məɲala	sələlah	*dəmikian*	məɲuratmu	bərhəntian
al		naik	baraŋ	nəgeri	- bərlalu	bərlayar	məŋataan	dilepasan
		bawa	*agama*	məŋapa	*baginda*	tiadakah	hərsədara	məɲuruhan
						dipəŋgal	kəsudahan	bərpatutan
						pohonan	bəlantara	dipəlajari
						bərsədar	bərburuan	məŋərjaan

Kaidah morfologi mengatakan bahwa bentuk-bentuk minim yang berulang dengan makna yang mirip satu sama lain disebut morfem. Dari data yang lebih lengkap, kita temukan bentuk berulang: ***kə, sə, mə, bər, ku, tər, di, kah, i,* dan *an*.** Kalau bentuk-bentuk berulang itu kita pisahkan dari bagian lain dari bentuk bersangkutan, sehingga sisanya masih mempunya maknam, berarti kita sudah melakukan identifikasi morfem.

Dengan memisahkan bentuk-bentuk berulang dari bagian lainnya, maka sisanya atau bagian lain yang ditinggalkan oleh bentuk berulang itu ternyata sebagian besar berupa bentuk yang bersuku dua. Sebagian kecil terdiri lebih dari dua suku-- ditulis miring--. merupakan bentuk monomorfemis, tidak lagi dapat dipecah menjadi dua yang masing-masing bermakna.

Melihat frekuensinya ternyata bentuk-bentuk yang terdiri dari lebih dari dua suku itu jumlahnya tidak banyak. Dengan mengecualikan yang frekuensinya kecil, yakni yang terdiri dari lebih dari dua suku, kita dapat merumuskan bahwa bentuk kanonik bahasa Melayu Pasai itu adalah:

A. Bersuku satu

KV
VV
KVK
AK

B. Bersuku dua

V-KV	: ada
KV-KV	: cucu
VK-KV	: əmpu
VK-KVK	: angkat
KVK-KV	: mim-pi
KVK-KVK	: təm-pat
KV-VK	: na-ik

= (K)V(K)-(K)V(K)

Selanjutnya dari data diketahui bahwa suku pertama yang berpola KVK, konsonan kedua selalu NASAL. Bila kita menemukan struktur morfem yang mengizinkan munculnya fonem bukan nasal pada akhir suku pertama, dapat dicurigai berasal dari bahasa non Melayu.

4. **Data morfofonemik**

Contoh yang baik untuk data morfofonemik adalah untuk bahasa Melayu adalah awalan me-, yang variasinya cukup banyak. Untuk itu kita bisa minta tolong komputer untuk mengumpulkan semua bentuk *mə-* dari teks.

Daftar 3

mələlui
mələraŋkan
məmakan
məmbawa
məməluk
məminaŋ
məminta
mənamai
mənampuŋ
mənaŋkap
məɲcari
məɲcarik-carik
məndapat
məndataŋi
məndəŋar
mənəbas
mənəgahkan
mənəŋar
məniŋgalkan
məniŋgalkan
məɲjamu
məntabalkan
mərajakan

məŋadap
məɲahut
məŋapa
məɲapu
məɲəsal
məŋgərakkan
məŋhimpunkan
məŋhukumkan
məŋikut
məɲjadi
məŋkhatamkan
məŋucap
məɲuruh
məɲala
məɲalak
məŋambil
məŋambil
məŋapa
məŋajar

Dengan mengurutkan data saja --yang dengan mudah dan cepat dapat dilakukan oleh komputer-- kita sudah dapat menjelaskan adanya jumlah alomorf awalan {me}, yaitu *mə-*, *məŋ-*, *məŋ-*, *məɲ-* dan *məŋ-*. Sayangnya komputer hanya dapat menyajikan data untuk analisis penataan unsur-unsur, dan tidak dapat merumuskan proses menurut model Item & Process. Komputer tidak dapat menjelaskan beda antara *məɲala* yang berbentuk dasar *ɲala* dengan *məɲəsal* yang berbentuk dasari *səsal*. Untuk analisis itu diperlukan pertolongan pemunculan bentuk-bentuk *səsal* dan *ɲala* dari bentuk turunan lain. Misalnya kita ketahui ada bentuk-bentuk *ɲala api*, dan *səsal kəmudian*.

5. **Analisis morfologi**

Bahasa Indonesia hanya memiliki sejumlah morfem pembentuk kata yang terbatas, sehingga seringkali sebuah ujud morfem memiliki makna morfologis yang lebih dari satu. Atau sebaliknya sermngklali beberapa makna yang sama dinyatakan dengan dua tiga morfem yang berbeda. Untuk itu kita perlu data yang banyak dalam konteks kalimat. Karena makna-makna itu hanya menjadi jelas bila sudah dalam konteks. Untungnya komputer dapat membantu kita mengumpulkan data dengan amat mudah, cepat, dan teliti. Untuk analisis makna awalan *me* misalnya kita dapat menyuruh komputer mengutip dari teks bentuk awalan itu dalam konteks seperti berikut:

Daftar 4

liau telah	menghadiahkan	naskah yang berisi "Wejangan S
ang pernah	menjadi	Residen Yogyakarta di antara tahun 1
991:78). #	Menurut	catatan Chambert-Loir, lima belas n
las naskah	Melayu	koleksi Raffles yang sekarang tersimp
ty, London,	merupakan	hadiah Kyai Suradimenggala, Bupati
la seconde	*mention*	*a: nagari) Demak nagari Bagor warso*
mbert-Loir	mencatat	sebanyak lima belas naskah lain di
lam tulisan	mengenai	Hikayat Raja-raja Pasai, Russell Jo
sell Jones	mengutarakan	bahwa Kyai Suradimanggala telah
penerjemah	Melayu	dan Jawa di kantor "Penterjemah Jawa"
Jawa, dia	membantu	Raffles dalam penyelidikan ilmiahnya
elanjutnya	memperkirakan	bahwa sejumlah naskah telah di
ones tidak	melihat	bukti bahwa naskah-naskah itu telah
ebuah buku	mengenai	tata bahasa Melayu yang disusun ole
dly dengan	memanfaatkan	naskah-naskah Melayu (1736). #
anyak pula	memanfaatkan	naskah sebagai bahan untuk bela
jar bahasa	Melayu.	Naskah-naskah yang dahulu telah dihi
embaga itu	menghentikan	aktivitasnya. Naskah-naskah yan
ah Inggris	melalui'	pegawainya. Ada juga yang mengumpulk
ahnya telah	menerima	tugas dari Betawi untuk mencari nas
tab-kitab	Melayu")	ke berbagai tempat , yaitu ke Riau,
t Abdullah	menyebutkan	pula betapa giatnya Raffles dalam
anya untuk	memperoleh	atau menyalin naskah. Diceritakan
orang yang	menjual	kitab dan hikayat Melayu kepada Raff
n, Raffles	mempekerjakan	empat sampai lima juru tulisny[1]

Sayangnya komputer hanya mengenal lambang-lambang fisik yang disimpan dalam berkasnya, sehingga apa yang kita perintahkan dilakukan seperti apa adanya. Karena ciri linguistis awalan *me* hanya dapat dikenali dengan susunan huruf *m* dan *e* saja, ditambah spasi yang mendahuluinya, maka komputer menandai dan mengutip bentuk apa saja yang mulai dengan *spasi* **ditambah** ***me*** , dan dengan itu maka semua bentuk *me* termasuk yang bukan kata kerja dikutipnya pula. seperti *Melayu* pada contoh tersebut. Atau juga mengutp *mention* karena dalam teks terdapat kata Perancis itu. Olek karena kenyataan-kenyataan itu maka hasil kutipan itu mau tidak mau harus kita seleksi lagi sehingga bisa dijadikan data linguistis sesuai dengan tujuan penelitian sang peneliti.

Sekalipun ada pekerjaan tambahan itu, yakni menyeleksi kutipan mana yang mengandung unsur linguistis yang kita maksud, namun cara atau

[1] Cara ini sangat baik utnuk data yang diperlukan dalam penyusunan kamus

yang mengandung unsur linguistis yang kita maksud, namun cara atau bantuan komputer ini tetap jauh lebih memudahkan sang [eneliti, apa lagi dibandingkan dengan sustem pengartuan yang konvensional itu.

Data tersenut dikutip dengan mmisahkan bentuk *me* dari lainnya agar kita mudah memeriksanya untuk mencari maknanya, atau distribusi sintaksisnya. kita bisa juga mengutip seluruh kalimat, misalnya:

> Dain sendiri **menjelaskan** bahwa kodikologi ialah ilmu ***mengenai*** naskah-naskah dan bukan ilmu yang ***mempelajari*** apa yang tertulis di dalam naskah.

Dan untuk menandainya bentuk yang kita maksud untuk dianalisis kita cetak tebal atau miring seperti contoh di atas. Atau bisa juga kita pisahkan dengan perintah indent sesudah bentuk yang kita fokuskan:

> Dain sendiri menjelaskan **bahwa kodikologi ialah ilmu *mengenai* naskah-naskah dan bukan ilmu yang *mempelajari* apa yang tertulis di dalam naskah.**

Komputer dalam hal ini Program WP mampu disuruh menandai apa saja yang ada dalam berkas komputer dan kemudian kita suruh mengutip bagian yang kita minta itu untuk mengutipnya satu kata, satu baris, satu alenia, bahkan satu halaman. Jadi tergantung peneliti seberapa banyak yang akan diperlulan dalam penelitiannya. Dengan demikian ciri fisik kategori kata yang berupa imbuhan seperti *me, ber, pe, ter*, dan sebagainya dapat kita tunjuk, lalu kita kutip dalam kalimat atau dalam satuan lain yang mengandung ciri morfologis itu.

6. Data Penelitian sintaksis

Ciri linguistis tentu saja bukan hanya bentuk fisik fonem atau huruf dalam teks saja. Ada ciri lain yang dapat kita manfaatkan. Bahasa umumnya memiliki tiga sintaksis: morfem, urutan atau lingkungan, dan intonasi. Alat berupa morfem (segmental) merupakan gejala fisik yang dapat dikenali komputer, sehingga kita dapat mengumpulkan data dengan cara seperti dijelaskan pada contoh pengumpulan data morfologi yang telah disebut

Ciri fisik sintaksis tersebut adalah apa yang kita sebut kata sambung seperti *dan, bahwa, lalu*, dan sebagainya, yang dapat kita pakai untuk analisis kalimat kompleks atau kalimat majemuk. Ciri fisik sintaksis itu dapat digunakan untuk memperoleh data, dari teks, dengan cara yang sama seperti kita lakukan untuk morfologi. Misalnya bila kita ingin menganalisis klausa relatif, maka kita kumpulkan semua kalimat yang mengandung *yang, dimana, tempat*, dan kata lain yang kita duga mempunyai fungsi sintaksis yang sama.

Daftar 5

masa lampau	yang	sampai kepada kita sebagai warisan
berupa naskah	yang	bermacam-macam bentuk dan ragamnya,yang
Indonesia dan	yang	ditulis dalam berbagai bahasa daerah dan
Naskah	yang	ditulis itu berancka ragam isinya, antara

Dari data semacam itu kita dapat mengetahui kategori klausa, frasa, atau kata yang mendahului dan yang mengikuti klausa *yang*, bagaimana prilaku sintaksisnya, atau fungsi sintaksisnya, dan sebagainya.

Selain cara mengutip seperti terpampang, kita bisa juga mengutipnya dalam konteks kalimat lengkap, seperti dapat kita lakukan dalam pengumpulan data morfologi. (Lihat kembali contoh pada bagian 4 di atas).

Alat sintaksis kedua, urutan atau lingkungan sintaktis, ialah ciri yang dapat diperoleh dengan melihat unsur mana mendahului unsur lainnya. Misalnya apa saja yang dapat mendahului nomina, termasuk nomina yang tidak memiliki ciri fisik morfologis sebagai penanda fisiknya. Jadi untuk menentukan unsur tertentu termasuk nomina atau bukan misalnya dapat dilakukan melalui pemeriksaan urutannya dengan verba. Unsur yang muncul sesudah verba transitif umumnya adalah nomina atau frasa nomina. Dengan demikian kita dapat menjaring nomina dengan mengutip kalimat yang mengandung verba, misalnya dengan menutip bentuk verba berawlan *me, me- -kan/i*. Unsur sesudah verba berawalan *me* itu semuanya nomina atau frase nomina.

Selain itu nomina bisa juga dilihat dari unsur lain -- yang sering disebut sebagai kata tugas-- yang munculnya hanya mendahului atau mengikuti

nomina untuk membentuk susunan yang disebut frasa nominal. Kata tugas *di, dari,* dan *ke* misalnya hanya muncul di depan nomina. Dengan kata lain nomina dapat ditandai dengan kata tugas yang--oleh karena selalu mendahului nomina--disebut preposisi atau kata depan.

Intonasi juga ciri fisik yang dapat dikenali komputer, tetapi harus dengan program khusus. Program itu dibahas pada makalah yang ditulis oleh Myrna Laksman. Jadi tidak dibahas di sini.

Demikianlah sekedar contoh penggunaan Program Pengolah Kata untuk dimanfaatkan sebagai pengumpul data dari teks.

Sumber Acuan

Quirk, Rudolph, et.al
1985 *A Comprehensive Grammarf English Language*, Londom, New York: Longman

Word Perfect
1993 Version 6.0, U.S.A: Word Perfect Corporation