KAJIAN AWAL DENSITAS LEKSIKAL UNTUK PENGEMBANGAN PELABELAN OTOMATIS KELAS KATA BAHASA JAWA

Totok Suhardijanto, Renny Pradina Kusumawardani
Departemen Linguistik FIB Universitas Indonesia, Departemen Sistem Informasi FTEIC
Institut Teknologi Sepuluh Nopember
totok.suhardijanto@ui.ac.id; renny.pradina@gmail.com, renny@is.its.ac.id

ABSTRAK

Identifikasi fitur linguistik pada jenis teks sangat krusial bagi banyak proyek pemerolehan bahasa alami, salah satunya pelabelan kelas kata (part-of-speech tagging). Informasi tersebut bermanfaat bagi penyusunan algoritma untuk meningkatkan akurasi dan kecepatan pelabelan. Densitas leksikal mengukur kompleksitas kebahasaan dalam sebuah teks tertulis atau lisan melalui komposisi kata gramatikal atau kata kontennya (Halliday 1985). Menurut Halliday, bahasa lisan dicirikan dengan struktur kalimat kompleks dengan densitas leksikal rendah (lebih banyak klausa, tetapi lebih sedikit rasio kata konten per klausa), sementara bahasa tertulis dengan struktur kalimat sederhana tetapi dengan densitas leksikal tinggi. Kajian ini bertujuan untuk melihat pengaruh densitas dan diversitas leksikal sebuah teks terhadap pelabelan otomatis kelas kata, khususnya bahasa Jawa. Data dalam kajian ini diambil dari Korpus Bahasa Jawa Universitas Indonesia yang terdiri atas berbagai jenis teks, seperti teks naratif, prosedural, ekspositoris, deskriptif, dan hortatoris (Larson 1984). Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif dengan memanfaatkan metode statistika untuk mengetahui perbedaan persebaran antar-jenis teks. Hasil yang diperoleh menunjukkan bahwa distribusi kata tertentu sangat bergantung pada jenis teks tertentu. Temuan tersebut mempunyai implikasi terhadap pelabelan kelas kata yang akan dikembangkan.
Kata Kunci: *densitas leksikal, distribusi leksikal, pelabelan kelas kata, linguistik korpus*

PENDAHULUAN

Dalam kajian teks, dikenal istilah kepadatan leksikal atau densitas leksikal (lexical density) dan keragaman leksikal atau diversitas leksikal (lexical diversity) (Halliday 1985, Johansson 2008). Kedua istilah ini kadang-kadang ada yang mencampuradukkannya. Sejatinya, keduanya merujuk pada fenomena yang berbeda. Johansson (2008) mengatakan bahwa diversitas leksikal merupakan ukuran seberapa beragam kata yang digunakan di dalam sebuah teks, sedangkan densitas leksikal merupakan ukuran proporsi satuan leksikal (misalnya nomina, verba, adjektiva, dan beberapa adverb di dalam sebuah teks.

Menurut Stubb (1986), densitas leksikal mengukur kompleksitas linguistis pada teks tertulis atau lisan berdasarkan komposisi kata fungsi (satuan gramatikal) dan kata kontennya (satuan leksikal, leksem). Densitas leksikal berpengaruh terhadap keterbacaan sebuah teks dan kemudahan teks yang berpengaruh terhadap kemungkinan seorang pendengar atau pembaca dapat memahami sebuah komunikasi atau tidak (To et al, 2013, Kieran 1995). Kepadatan leksikal mungkin juga berpengaruh terhadap daya ingat dan retensi sebuah kalimat dan pesannya (Perfetti 1969).

Sementara itu, diversitas leksikal sering dipadankan dengan kekayaan leksikal (lexical richness) (misalnya Daller, van Hout & Treffers-Daller 2003). Meskipun demikian, baik Read (2000) maupun Malvern et al. (2004) menyampaikan bahwa terdapat perbedaan di antara keragaman leksikal dan kekayaan leksikal dalam hal diversitas leksikal hanya merupakan salah satu bagian dari fitur multidimensional dari kekayaan leksikal. Dengan demikian, diversitas leksikal lebih spesifik daripada kekayaan leksikal.

Kajian ini bertujuan untuk melihat pengaruh densitas dan diversitas leksikal sebuah teks terhadap pelabelan otomatis kelas kata (part-of-speech tagging), khususnya dalam bahasa Jawa.

Dalam banyak kasus pemrosesan bahasa alami, sering kali ditemukan pengaruh ranah topik, genre, densitas dan diversitas leksikal terhadap tingkat presisi (precision) dan pemanggilan (recall).

METODOLOGI

Dalam kajian ini digunakan pendekatan gabungan kuantitatif-kualitatif dengan desain penelitian sequential explanatory. Jadi, dalam penelitian ini, data kuantitatif dikumpulkan dan dianalisis terlebih dahulu, baru kemudian data kuantitatif dikumpulkan dan dianalisis untuk menjelaskan temuan pada tahap analisis kuantitatif (Ivankova 2006). Metode dengan pendekatan gabungan ini lazim juga disebut metode gabungan.

Data penelitian ini diambil dari Korpus Bahasa Jawa yang disusun oleh Program Studi Daerah untuk Sastra Jawa FIB UI. Korpus tersebut terdiri atas 6 juta token. Namun, tidak semua bagian dari korpus tersebut digunakan karena belum semua kata di dalam korpus tersebut telah dilabeli kelas kata. Hanya ada sekitar 3954 kalimat yang telah mengandung informasi kelas kata. Dari jumlah kalimat tersebut, diperoleh 8915 tipe atau kata unik.

ANALISIS

Distribusi Kalimat dalam Korpus Sampel

Tulisan ini merupakan eksplorasi awal terhadap bagaimana jenis teks/genre mempengaruhi pemrosesan teks berbahasa Jawa. Dari 83 file yang dipilih dari korpus, terdapat 3954 kalimat terlabeli kelas kata bahasa Jawa yang berasal dari enam jenis teks: akademik (academic), fiksi (fiction), majalah (magazine), surat kabar (newspaper), bahan rujukan (reference), dan buku teks (textbook). Distribusi file dan kalimat tersebut dapat dilihat pada Tabel 1.

Diversitas Leksikal

Terdapat 51354 token dan 8915 tipe dalam Korpus Sampel Bahasa Jawa yang terlabeli POS (*part-of-speech*). Tabel 2 menampilkan metrik untuk tiap-tiap jenis teks/genre, yaitu sebagai berikut:

1. # of documents : jumlah dokumen pada jenis teks tertentu
2. # of types : jumlah tipe pada jenis teks tertentu
3. # of tokens : jumlah tipe pada jenis teks tertentu
4. Lexical diversity : jumlah tipe/jumlah token
5. Ave. TF : rata-rata frekuensi kemunculan tipe
6. Stdev. TF : deviasi standar frekuensi kemunculan tipe
7. Coeff. Var. TF : koefisien variasi frekuensi tipe (rata-rata dibagi deviasi standar)
8. Ave. DF : rata-rata jumlah dokumen tempat tipe muncul
9. Stdev. DF : deviasi standar jumlah dokumen tempat tipe muncul
10. Coeff. Var. TF : koefisien variasi frekuensi tipe (rata-rata dibagi deviasi standar)

Tabel 1. Distribusi file dan kalimat berdasarkan tipe teks/ genre.

Genre	Jumlah dokumen	% dokumen	Jumlah kalimat	% kalimat
Academic	22	26.506	760	19.221
Fiction	38	45.783	2222	56.196
Magazine	5	6.024	299	7.562
Newspaper	7	8.434	239	6.045
Reference	8	9.639	263	6.651
Textbook	3	3.614	171	4.325
Total	83	100	3954	100

Tabel 2. Data metriks terkait tipe untuk tiap jenis teks/genre.

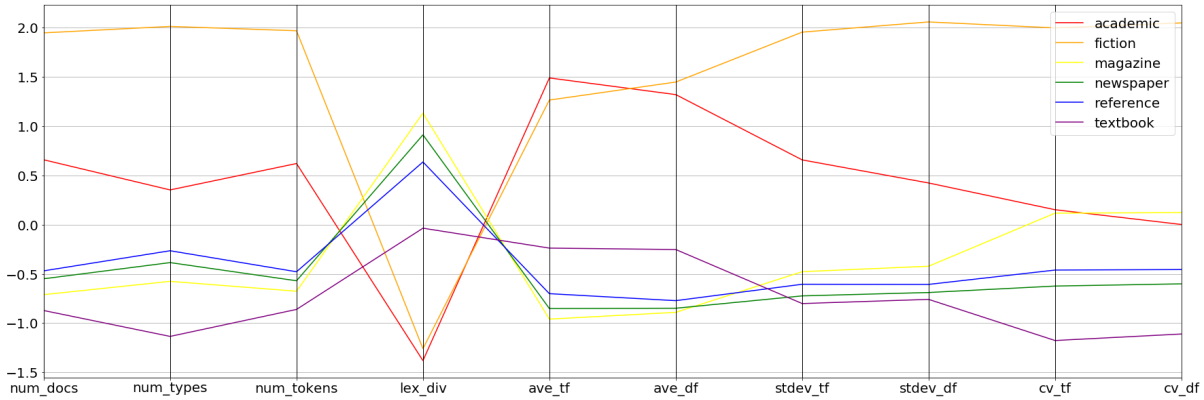
Genre	# Docs	# Types	# Tokens	Lex. Div.	Ave. TF	Stdev. TF	Coeff. Var. TF	Ave. DF	Stdev. DF	Coeff. Var. DF
academic	22	2779	13421	0.207	4.829	22.530	4.666	4.244	16.823	3.964
fiction	38	5220	23996	0.218	4.597	36.181	7.871	4.363	31.834	7.296
magazine	5	1414	3251	0.435	2.299	10.587	4.605	2.178	9.063	4.161
newspaper	7	1694	4083	0.415	2.410	8.001	3.320	2.217	6.614	2.983
reference	8	1871	4801	0.390	2.566	9.244	3.602	2.289	7.369	3.219
textbook	3	592	1802	0.329	3.044	7.183	2.360	2.774	5.974	2.154

Perhatikan bahwa terdapat urutan besaran jumlah yang berbeda secara signifikan pada kolom # of tokens dan lex div atau densitas leksikal. Oleh karena itu, untuk mendapatkan visualisasi informatif kami menormalisasi nilai tiap kolom dengan menggunakan distribusi z menggunakan Persamaan 1.

$$z_x = \frac{x - mean(values_{in_{column}})}{sdev(values_{in_{column}})}$$

Persamaan 1. Formula penghitungan skor Z untuk masing-masing parameter karakterisasi korpus.

Hasil visualisasi nilai metriks terstandar dengan koordinat parallel dapat dilihat pada Gambar 1 berikut ini. Secara umum, tampak bahwa ukuran jenis teks dalam korpus berpengaruh kuat terhadap nilai metriks lainnya.



Gambar 1. Grafik koordinat parallel dari metriks ternormalisasi berdasarkan genre.

Pada gambar di atas, tampak bahwa jenis teks majalah, surat kabar, dan bahan rujukan mempunyai kemiripan yang tinggi. Hal ini berbeda dengan jenis teks yang lainnya sehingga tampak ada kecenderungan bahwa jenis teks sangat mempengaruhi diversifikasi leksikal.

Densitas Leksikal

Penghitungan densitas leksikal di sini dilakukan dengan menghitung persentase tiap kelas kata pada tiap jenis teksnya. Hasil penghitungan tersebut dapat dilihat pada Tabel 4. Namun, sebelumnya perlu disampaikan jenis label kelas kata yang digunakan dalam penelitian ini (lihat Tabel 3).

Tabel 3. Label POS (Part of Speech) yang digunakan dalam penelitian ini.

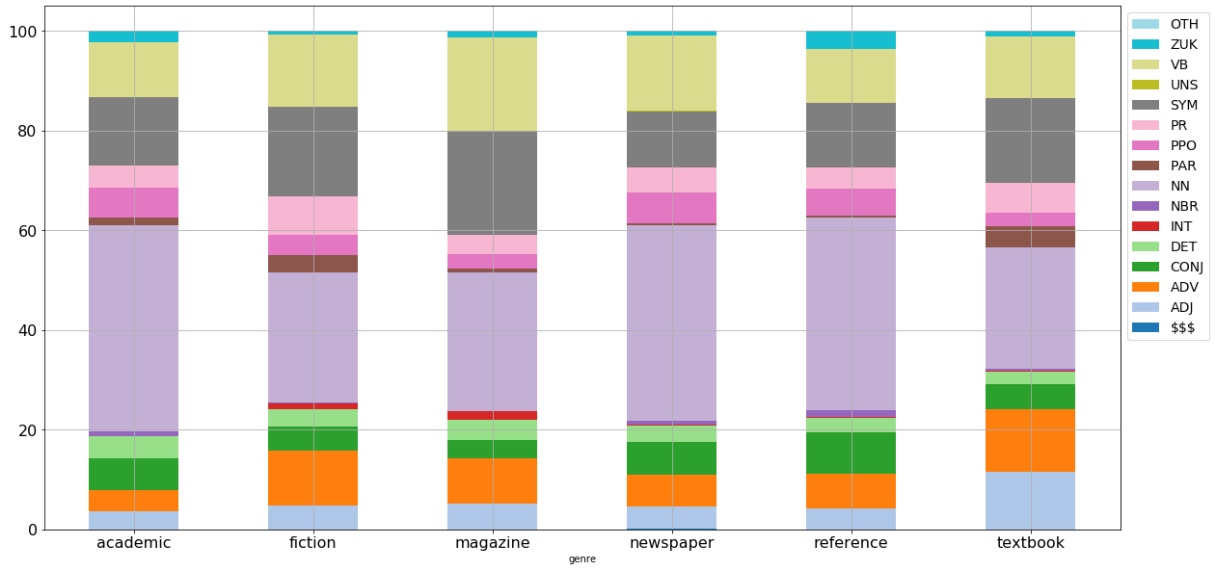
Label	*Part-of-Speech*	Kelas Kata
NN	Nouns	Nomina
PR	Pronouns	Pronomina
VB	Verbs	Verba
ADJ	Adjectives	Adjektiva
ADV	Adverbs	Adverbia
CONJ	Conjunctions	Konjungsi
PPO	Prepositions	Preposisi
INT	Interjections	Interjeksi
DET	Determiners	Determina
PAR	Particles	Partikel
NBR	Numbers	Bilangan
UNS	Unit symbols	Lambang satuan
$$$	Currency	Mata uang
SYM	Character symbols	Lambang karakter
ZUK	Foreign terms	Istilah asing
PUN	Punctuation	Tanda baca
QUOT	Quotation marks	Tanda petik
OTH	Others	Lain-lain (selain kelas di atas)

Tabel 2. Persentase kemunculan label terkait jumlah total token pada tiap-tiap genre dalam subkorpus bahasa Java berlabel POS tertentu.

Genre	ADJ	ADV	CONJ	DET	INT	NBR	NN	PAR	PPO	PR	VB
academic	3.63	4.19	6.5	4.37	0.01	0.95	41.41	1.41	6.06	4.45	11.01
fiction	4.73	11.05	4.89	3.37	1.16	0.28	26.15	3.29	4.18	7.65	14.57
magazine	5.11	9.07	3.84	3.97	1.69	0.06	27.84	0.8	2.86	3.81	18.76
newspaper	4.58	6.34	6.49	3.31	0.24	0.81	39.16	0.39	6.25	4.95	14.99
reference	4.21	7.04	8.29	2.85	0.1	1.44	38.55	0.54	5.33	4.23	10.81
textbook	11.54	12.49	5.16	2.5	0.06	0.5	24.36	4.22	2.61	5.99	12.43

Dari persentasenya, tampak bahwa distribusi nomina dan verba sangat signifikan dibandingkan kelas kata lain. Namun, yang cukup menarik adalah distribusi adjektiva dan adverbia mengalami peningkatan drastik dibandingkan distribusinya pada jenis teks lain. Hal ini bisa saja terjadi karena faktor keterbatasan korpus berjenis teks buku teks.

Gambar 2. Diagram batang vertikal persentase kemunculan label terkait dengan jumlah total token pada tiap genre.



Gambar 2 menunjukkan bagaimana persentase kemunculan label POS atau kelas kata dengan menggunakan diagram batang vertikal. Pada gambar ini, tampak lebih jelas lagi secara visual bahwa nomina (abu-abu) dan verba (kuning kehijauan) cukup mendominasi distribusi pada semua genre. Dari sini juga tampak bahwa pemakaian adverbia (jingga) sangat signifikan perbedaannya pada teks berjenis fiksi dan majalah, misalnya jika dibandingkan dengan teks akademik yang jarang penggunaan adverbianya.

KESIMPULAN

Dapat disimpulkan secara ringkas, bahwa pada studi awal ini terlihat bahwa perbedaan jenis teks/genre akan sangat mempengaruhi kinerja pelabelan POS secara otomatis pada data teks berbahasa Jawa. Terkait dengan diversitas leksikal atau keragaman leksikal, kinerja pelabelan POS otomatis itu sangat bergantung pada perbedaan distribusi tipe dan token yang sangat bervariasi pada jenis-jenis teks tertentu. Sementara itu, terkait dengan densitas leksikal, tampaknya hal tersebut tidak terlalu banyak mempengaruhi kinerja pelabelan POS secara otomatis karena sebaran kelas kata cenderung mempunyai kemiripan, kecuali pada kelas adverbia yang distribusinya cukup mencolok perbedaannya pada jenis teks fiksi dan majalah.

DAFTAR PUSTAKA:

Michael Halliday (1985). *Spoken and Written Language*. Deakin University. pp. 61–64. ISBN 978-0-7300-0309-0.

Michael Stubbs (1986). "Lexical density: A technique and some findings". In Malcolm Coulthard (ed.). *Talking about Text*. University of Birmingham: English Language Research. pp. 27–42

Ure, J (1971). Lexical density and register differentiation. In G. Perren and J.L.M. Trim (eds), *Applications of Linguistics*, London: Cambridge University Press. 443-452.

V To; S Fan; DP Thomas (2013). "Lexical density and Readability: A case study of English Textbooks". *The International Journal of Language, Society and Culture*. 37 (7): 61–71.

O'Loughlin, Kieran (1995). "Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test". *Language Testing*. SAGE Publications. 12 (2): 217–237. doi:10.1177/026553229501200205. S2CID 145638000.

Perfetti, Charles A. (1969). "Lexical density and phrase structure depth as variables in sentence retention". Journal of Verbal Learning and Verbal Behavior. Elsevier BV. 8 (6): 719–724.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. Working papers/Lund University, Department of Linguistics and Phonetics, 53, 61-79.

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field methods*, 18(1), 3-20.

Biodata 1:

a. Nama Lengkap : Totok Suhardijanto
b. Universitas: Universitas Indonesia
c. Alamat Surel: totok.suhardijanto@ui.ac.id
d. Pendidikan Terakhir: S3
e. Minat Penelitian: corpus linguistics, computational linguistics

Biodata 2:

a. Nama Lengkap : Renny Pradina Kusumawardani
b. Universitas: Institut Teknologi 10 Nopember Surabaya
c. Alamat Surel: renny@is.its.ac.id, renny.pradina@gmail.com
d. Pendidikan Terakhir: S2
e. Minat Penelitian: natural language processing, computational linguistics, machine learning, deep learning