PEMBENTUKAN DERET PENDEKATAN UNTUK BAHASA INDONESIA MENGGUNAKAN METODE SHANNON

Muhammad Jazilun Niam
Achmad Hidayatno
Rizal Isnanto

ABSTRACT

At the beginning of existing of information theory, Shannon has shown a visual representation of how a series of processes approaches a language in letter and word level, in this case is English. Then the question arises, whether the method can be used for other languages in the world such as Bahasa Indonesia. Therefore a research is required to be done to make a series of approximation to Bahasa Indonesia with a method which has been used by Shannon in forming the series of approximation to English.

A previous research by Shannon has produced a series of approximation to English. The same method was also used in this research. Research done by these following steps. The first step is data collection process of letter probabilities, digram probabilities, trigram probabilities, word probabilities, and digram word probabilities of Bahasa Indonesia by using samples of 2 books and 2 articles. Opportunities for character data, digram and trigram done with 3 variations of one paragraph, half page and one page of the article. Then the next stage is designing the program that consists of 6 stages, zero, first, second, and third order letter approximation level, and first and second order word approximation level, then at last step is analyzing the results obtained.

Based on the result of the series of approximation to Bahasa Indonesia, at each series length, the emergence of Indonesian words increases as the existing level increase.The emergence of Indonesian words also increases at every increase of series length except for zeroth order approximation. In the simulation of series length influence ,the emergence of a minimum value of occurrence the Indonesian word is 0% and the maximum value of occurrence the Indonesian word is 44.70899471%. Effect of using different database, occurrence of the word is the difference in value of occurrence the Indonesian word and generated Indonesian words itself, generated Indonesian word suitable with the used source of database. Effect of variation in database source of one paragraph, a half page and one page is the Indonesian word emergence increases as the existing source database increase from a single source of one paragraph, a half page and one page, this applies to both articles.

Keywords: Shannon Method, Series Approximation, Bahasa Indonesia.

I. PENDAHULUAN

1.1 Latar Belakang

Perkembangan ilmu pengetahuan dan teknologi semakin pesat, terutama teknologi informasi dan komputer. Perkembangan teknologi informasi dan teknologi komputer yang sangat pesat ini, terutama teknologi komputer memicu perkembangan di berbagai bidang yang salah satu di antaranya adalah teknologi bahasa manusia (TBM) yang termasuk didalamnya pembentukan deret pendekatan untuk sebuah bahasa alami yang sering juga disebut sebagai pemodelan bahasa.

Pada awal munculnya teori informasi, Shannon telah memperlihatkan dalam makalahnya yang berjudul *A Mathematical Theory of Communication* mengenai gambaran visual tentang bagaimana rangkaian proses pendekatan suatu bahasa dalam hal ini adalah Bahasa Inggris. Kemudian muncul pertanyaan, apakah metode tersebut digunakan untuk bahasa-bahasa lain di dunia, dengan dasar tersebut penulis mengambil penelitian tentang pendekatan untuk Bahasa Indonesia dengan metode yang Shannon gunakan.

Proses pendekatan untuk Bahasa Indonesia ini dilakukan dalam perangkat lunak pemrograman Matlab. Dari hasil yang diperoleh dari program pendekatan untuk Bahasa Indonesia tersebut akan diketahui bagaimana hubungan antara deret pendekatan untuk Bahasa Indonesia dengan metode yang Shannon gunakan pada pembentukan deret pendekatan untuk Bahasa Inggris.

1.2 Tujuan

Tujuan dari Tugas Akhir ini adalah membuat deret pendekatan untuk Bahasa Indonesia dengan pemrograman Matlab dan mengetahui hubungan antara deret pendekatan untuk Bahasa Indonesia dengan metode yang Shannon gunakan pada pembentukan deret pendekatan untuk Bahasa Inggris.

Achmad Hidayatno, Rizal Isnanto (achmad@elektro.ft.undip.ac.id, rizal_isnanto@yahoo.com), adalah dosen di Jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro (Undip) Semarang Jl. Prof. Soedarto, S.H. Tembalang Semarang 50275.
Muhammad Jazilun Niam adalah mahasiswa di Jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro (Undip) Semarang Jl. Prof. Soedarto, S.H. Tembalang Semarang 50275.

1.3 Batasan Masalah

Pembatasan masalah pada penulisan tugas akhir ini sebagai berikut :

1. Penelitian menggunakan metode yang telah ada sebelumnya yaitu Metode Shannon.
2. Penelitian menggunakan 27 karakter Bahasa Indonesia yaitu 26 abjad dan Spasi.
3. Penelitian menggunakan tabel frekuensi karakter, tabel digram, tabel trigram, tabel kata, dan tabel digram kata yang telah disusun dengan program terpisah sebelumnya.
4. Buku yang digunakan dalam pembuatan tabel frekuensi karakter, tabel digram, tabel trigram, tabel frekuensi kata, dan tabel digram kata adalah gabungan 2 buku yaitu Pendidikan Kewarganegaraan untuk SMP/MTs kelas IX oleh Sugiyono dan Sosiologi untuk SMA/MA kelas X oleh Vina Dwi Laning yang penulis dapat dari pusat pembukuan Departemen Pendidikan Nasional.
5. Laporan ini tidak membahas program penyusun tabel frekuensi karakter, tabel digram, tabel trigram, tabel kata, dan tabel digram kata.

II. LANDASAN TEORI

2.1 Teori Probabilitas

Teori probabilitas mempelajari rata-rata gejala massa yang terjadi secara berurutan atau bersama-sama, seperti pancaran elektron, hubungan telepon, deteksi radar, pengendalian kualitas, kegagalan sistem, permainan untung-untungan, mekanika statistik, gangguan, laju kelahiran dan kematian, dan teori antrian. Pada bidang-bidang tersebut atau lainnya telah diamati bahwa suatu rata-rata akan mendekati suatu nilai konstan jika jumlah observasi bertambah besar dan nilai-nilai ini tetap sama bila dihitung pada sembarang barisan bagian yang ditentukan sebelum eksperimen dilakukan.

Tujuan teori probabilitas adalah menggambarkan dan menaksir rata-rata tersebut dalam bentuk probabilitas peristiwa.

2.2 Probabilitas Bersyarat

Probabilitas bersyarat kejadian A dengan diketahui kejadian B, ditulis P(A|B) didefinisikan sebagai

$$P(A|B) = \frac{\mathrm{P(A \cap B)}}{P(B)} \quad \text{..... (1)}$$

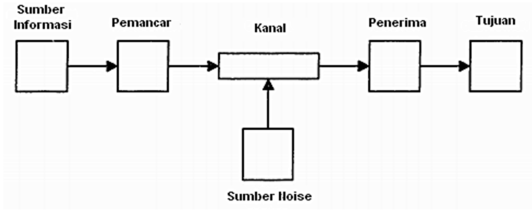
Dimisalkan bahwa P(B) tidak nol.

P(A|B) secara sederhana memperlihatkan kenyataan bahwa probabilitas kejadian A tergantung pada kejadian B. Jika A dan B adalah kejadian mutual eksklusif, $A \cap B = \emptyset$, dan P(A|B) = 0.

2.3 Teori Informasi

Asal-usul dari teori informasi dapat dilihat kembali ke publikasi makalah dari Claude E. Shannon dalam *Bell System Technical Journal* pada tahun 1948 yang berjudul *A Mathematical Theory of Communication*. Dalam hal makna informasi sehari-hari, makalah Shannon berkaitan dengan membawa informasi dalam bentuk simbol dan tidak dalam bentuk informasi itu sendiri.

Komunikasi menurut Shannon digambarkan sebagai aliran pesan sepanjang serangkaian tahap, yang diwakili bagan dengan diagram blok pada gambar 1.



Gambar 1. Skema klasik Diagram Shannon dari sistem komunikasi secara umum.

Blok pertama adalah sumber, di mana pesan tersebut berasal. Blok kedua adalah pemancar, yang mengubah atau menyandikan pesan asli menjadi bentuk yang sesuai untuk pengiriman. Pesan disandikan kemudian dikirim melalui saluran komunikasi. Selama perjalanan melalui saluran, pesan dapat terpengaruh oleh kesalahan. Dengan kata lain, saluran terkendala oleh kebisingan (*noise*). Kebisingan tersebut berada di mana-mana, dalam udara karena badai magnetik dapat mengganggu sinyal, dalam peralatan elektronik karena arus dapat merusak data, dalam suatu serat optic karena kerugian menurunkan energi cahaya yang ditransmisikan. *Noise* ini tidak mungkin untuk dihilangkan dari kanal atau saluran. Pesan yang disandikan akan keluar dari saluran dan mencapai penerima di mana penerima melakukan operasi kebalikan dari pemancar, yaitu menerjemahkan pesan itu dan memberikannya kepada blok akhir yaitu tujuan.

2.4 Deret Pendekatan untuk Bahasa Inggris

Dalam teori informasi, proses stokastik digunakan sebagai model untuk menghasilkan pesan. Proses stokastik ini adalah mesin matematika yang berjalan tanpa henti menumpahkan pesan menurut aturan probabilitas. Dan aturan-aturan tersebut dapat didefinisikan. Dimulai dengan aturan-aturan sederhana kemudian diperkenalkan lebih dan lebih aturan, lebih khusus bagaimana pesan dihasilkan. Shannon memberikan beberapa contoh bagaimana aturan dapat mengubah suatu proses stokastik yang menghasilkan teks bahasa Inggris dan untuk memberikan gambaran visual tentang bagaimana rangkaian proses pendekatan bahasa, deret khas dalam pendekatan untuk Bahasa Inggris telah dibangun dan diberikan di bawah ini.

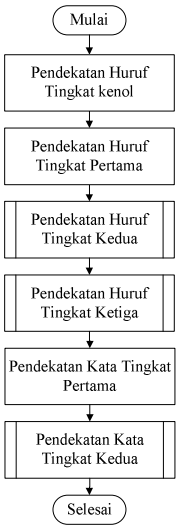
Dalam semua kasus diasumsikan 27 simbol yaitu 26 huruf abjad dan spasi.

1. Pendekatan karakter tingkat kenol (simbol saling bebas dan masing-masing karakter mempunyai probabilitas kemunculan yang sama).
   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
2. Pendekatan karakter tingkat pertama (simbol saling bebas namun dengan probabilitas frekuensi karakter dalam teks Bahasa Inggris).
   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
3. Pendekatan karakter tingkat kedua (menggunakan probabilitas struktur digram sebagaimana dalam Bahasa Inggris).
   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE
4. Pendekatan karakter tingkat ketiga (menggunakan probabilitas struktur dtrigram sebagaimana dalam Bahasa Inggris).
   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE
5. Pendekatan kata tingkat pertama. Daripada melanjutkan struktur tetragram,…, n-gram lebih mudah dan lebih baik untuk melompat pada titik ini ke unit kata. Berikut kata-kata yang dipilih secara independen tetapi dengan frekuensi masing-masing yang sesuai.
   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
6. Pendekatan kata tingkat kedua. Menggunakan probabilitas transisi kata tetapi tidak ada struktur lebih lanjut yang disertakan.
   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

III. PERANCANGAN SIMULASI

3.1 Gambaran Umum

Perancangan simulasi pembentukan deret pendekatan untuk Bahasa Indonesia dimulai dengan pembuatan diagram alir secara umum seperti pada Gambar 3.1. Masing-masing prosedur dapat dijabarkan lagi melalui diagram alir yang lebih rinci dalam pembahasan selanjutnya.



Gambar 2. Diagram alir program utama.

3.2 Pendekatan Huruf Tingkat Kenol

Pada pembentukan deret pendekatan huruf tingkat kenol digunakan 27 karakter dengan masing-masing karakter saling bebas dan mempunyai peluang kemunculan karakter yang sama, dalam hal ini adalah 1/27 atau 0,037037.

3.3 Pendekatan Huruf Tingkat Pertama

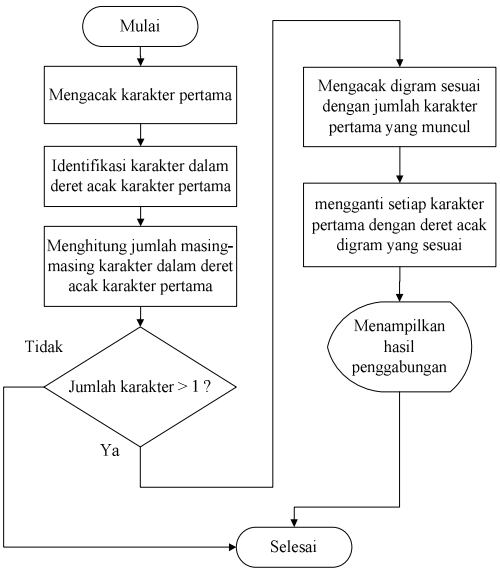
Pada pembentukan deret pendekatan huruf tingkat pertama digunakan 27 karakter dengan masing-masing karakter saling bebas, tetapi mempunyai peluang kemunculan karakter yang berbeda sesuai dengan Tabel 1.

Tabel 1. Peluang kemunculan karakter untuk Bahasa Indonesia

Karakter	Peluang	Karakter	Peluang
A	0,1769780	P	0,0273651
Spasi	0,1271980	B	0,0223276
N	0,0900539	O	0,0188681
E	0,0717472	H	0,0177528
I	0,0678818	Y	0,0148547
R	0,0451218	J	0,0061414
S	0,0433427	C	0,0043016
T	0,0431454	W	0,0030916
U	0,0407746	F	0,0025074
K	0,0395683	V	0,0010545
M	0,0363288	Z	0,0000720
D	0,0359722	Q	0,0000417
G	0,0331348	X	0,0000341
L	0,0303391		

3.4 Pendekatan Huruf Tingkat Kedua

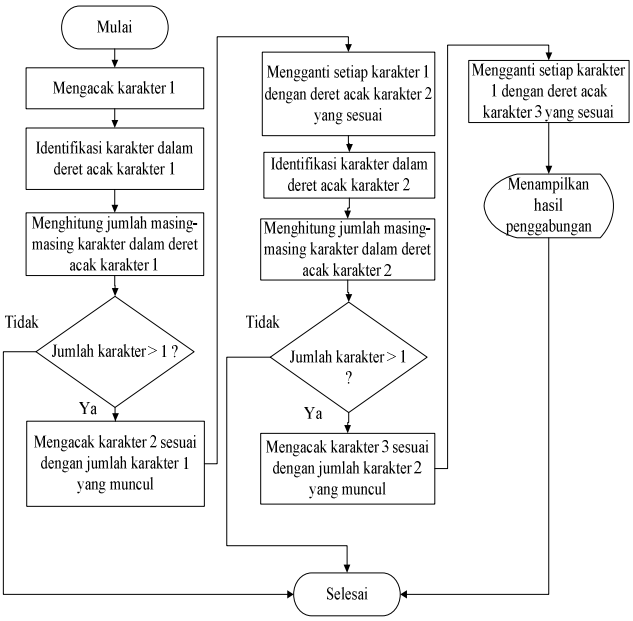
Pada pembentukan deret pendekatan huruf tingkat kedua digunakan pasangan huruf (digram). Diagram alir pembentukan deret pendekatan huruf tingkat kedua adalah seperti gambar 3.



Gambar 3. Diagram alir pembentukan deret pendekatan huruf tingkat kedua.

3.5 Pendekatan Huruf Tingkat Ketiga

Pada pembentukan deret pendekatan huruf tingkat ketiga digunakan pasangan 3 huruf (trigram). Diagram alir pembentukan deret pendekatan huruf tingkat kedua adalah seperti gambar 4.



Gambar 4. Diagram alir pembentukan deret pendekatan huruf tingkat kedua.

3.6 Pendekatan Kata Tingkat Pertama

Pada pembentukan deret pendekatan kata tingkat pertama proses yang digunakan sama dengan proses pada pembentukan deret pendekatan huruf tingkat pertama, perbedaannya hanya pada data yang digunakan. Pada pembentukan deret pendekatan kata tingkat pertama ini digunakan kata dalam Bahasa Indonesia yaitu sebanyak 3.774 kata dengan peluang masing-masing.

3.7 Pendekatan Kata Tingkat Kedua

Pada pembentukan deret pendekatan kata tingkat kedua, proses yang digunakan sama dengan proses pada pembentukan deret pendekatan huruf tingkat kedua, perbedaannya hanya pada data yang digunakan. Pada pembentukan deret pendekatan kata tingkat kedua ini kita menggunakan pasangan 2 kata dalam Bahasa Indonesia yaitu sebanyak 20.654 pasangan 2 kata dengan peluang masing-masing.

IV. HASIL SIMULASI DAN PEMBAHASAN

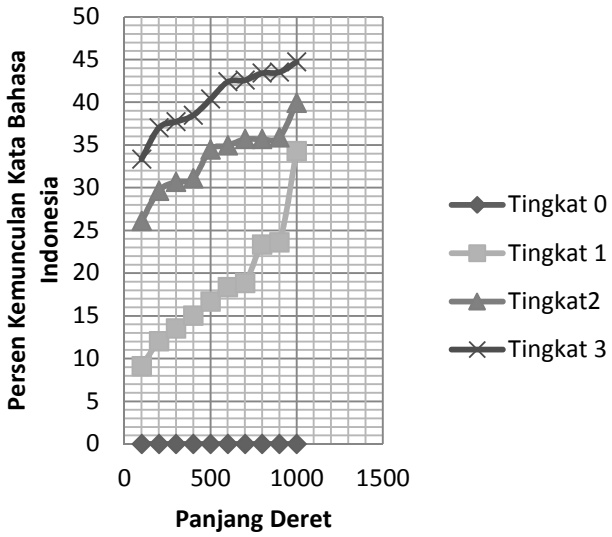
Hasil simulasi untuk database dari buku acuan dibuat dalam variasi panjang deret, yaitu panjang deret 100, 200, 300, 400, 500, 600, 700, 800, 900, dan panjang deret 1000. Dari hasil ini dilakukan analisis terhadap pengaruh panjang deret dengan hasil deret yang diperoleh. Selain itu dilakukan pula pengujian terhadap simulasi dengan database lain yaitu artikel kesehatan dan artikel teknologi untuk mengetahui pengaruh perbedaan database dengan sumber satu alinea, setengah halaman dan satu halaman dari artikel tersebut.

4.1 Pengaruh Panjang Deret dengan Hasil Deret Pendekatan untuk Bahasa Indonesia yang Diperoleh

Pengujian dilakukan dengan variasi panjang deret 100, 200, 300, 400, 500, 600, 700, 800, 900, dan panjang deret 1000. Pengujian ini dilakukan untuk mengetahui pengaruh panjang deret dengan hasil deret pendekatan untuk Bahasa Indonesia.

Dari data, diperoleh bahwa pada setiap panjang deret, kemunculan kata Bahasa Indonesia meningkat seiring peningkatan tingkat pendekatannya. Contohnya pada panjang deret 100, pada pendekatan huruf tingkat kenol kemunculan kata Bahasa Indonesia adalah 0 %, pada pendekatan huruf tingkat pertama adalah 9,090909091 %, pendekatan huruf tingkat kedua adalah 26,08695652 %, dan pendekatan huruf tingkat ketiga adalah 33,3333333 %. Selain itu terlihat juga peningkatan kemunculan kata Bahasa Indonesia disetiap kenaikan panjang deret kecuali untuk pendekatan tingkat kenol. Contohnya pada pendekatan huruf tingkat ketiga, untuk panjang deret 100 kemunculan kata Bahasa Indonesia adalah 33,33333333 %, untuk panjang deret 200 adalah

36,98630137 %, untuk panjang deret 300 adalah 37,71929825 %, untuk panjang deret 400 adalah 38,46153846 %, untuk panjang deret 500 adalah 40,36144578 %, untuk panjang deret 600 adalah 42,38095238 %, untuk panjang deret 700 adalah 42,5855513 %, untuk panjang deret 800 adalah 43,39622642 %, untuk panjang deret 900 adalah 43,51585014 %, dan untuk panjang deret 1000 adalah 44,70899471 %.



Gambar 5. Perbandingan panjang deret dengan kata Bahasa Indonesia yang muncul.

Dari gambar 5 terlihat bahwa pada setiap panjang deret, kemunculan kata Bahasa Indonesia meningkat seiring peningkatan tingkat pendekatannya dan terlihat juga peningkatan kemunculan kata Bahasa Indonesia disetiap kenaikan panjang deret kecuali untuk pendekatan tingkat kenol.

4.2 Pengaruh Perbedaan Database dengan Hasil Deret Pendekatan untuk Bahasa Indonesia yang Diperoleh

Pengujian ini dilakukan untuk mengetahi pengaruh perbedaan database yang digunakan terhadap deret pendekatan untuk Bahasa Indonesia yang diperoleh. Pengujian ini menggunakan database dari 2 artikel yaitu artikel kesehatan berjudul “Serba-serbi Stroke” dan artikel teknologi dan internet yang berjudul “Facebook, Bermanfaat Atau Berbahaya?”. Database ini dibuat dari kedua artikel tersebut dengan masing-masing 3 variasi yaitu, database dengan sumber 1 alinea dari artikel, database dengan sumber setengah halaman dari artikel dan database dengan sumber satu halaman dari artikel.

Dari hasil diperoleh bahwa terdapat nilai kemunculan kata Bahasa Indonesia yang berbeda, hal ini juga disebabkan perbedaan dalam tabel peluang karakter, digram dan trigram yang dihasilkan dari kedua sumber database tersebut.

Pengaruh penggunaan database yang berbeda adalah perbedaan nilai kemuncuan kata Bahasa Indonesia pada deret pendekatan yang dihasilkan dan kata Bahasa Indonesia yang dihasilkan itu sendiri, kata Bahasa Indonesia yang dihasilkan sesuai dengan sumber database yang digunakan.

Pengaruh variasi dari sumber database satu alenia, setengah halaman dan satu halaman adalah nilai kemunculan kata Bahasa Indonesia semakin meningkat seiring peningkatan sumber database yaitu meningkat dari sumber satu alenia, setengah halaman dan satu halaman, hal ini berlaku untuk kedua artikel. Contohnya adalah pada pendekatan tingkat ketiga, sumber database dari artikel kesehatan dengan panjang deret 100 meningkat dari 40,47619048 % untuk sumber database satu alenia, 41,86046512 % untuk sumber database setengah halaman dah 44,68085106 % untuk sumber database satu halaman.

V. PENUTUP

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari hasil pengujian dan pembahasan adalah sebagai berikut :

1. Pada setiap panjang deret, kemunculan kata Bahasa Indonesia meningkat seiring peningkatan tingkat pendekatannya dan terlihat juga peningkatan kemunculan kata Bahasa Indonesia di setiap kenaikan panjang deret kecuali untuk pendekatan tingkat kenol.
2. Pada simulasi pengaruh panjang deret dari buku acuan, nilai minimal kemunculan kata Bahasa Indonesia dalam deret pendekatan untuk Bahasa Indonesia adalah 0 % yaitu pada pendekatan huruf tingkat kenol. Sedangkan nilai maksimal kemunculan kata Bahasa Indonesia dalam deret pendekatan untuk Bahasa Indonesia adalah 44,70899471 % yaitu pada pendekatan huruf tingkat ketiga dengan panjang deret 1000.
3. Pengaruh penggunaan database yang berbeda adalah perbedaan nilai kemunculan kata Bahasa Indonesia pada deret pendekatan yang dihasilkan dan kata Bahasa Indonesia yang dihasilkan itu sendiri, kata Bahasa Indonesia yang dihasilkan sesuai dengan sumber database yang digunakan.
4. Pengaruh variasi dari sumber database satu alenia, setengah halaman dan satu halaman

adalah nilai kemunculan kata Bahasa Indonesia semakin meningkat seiring peningkatan sumber database yaitu meningkat dari sumber satu alenia, setengah halaman dan satu halaman, hal ini berlaku untuk kedua artikel.

5.2 Saran

Adapun saran yang dapat diberikan sehubungan dengan pelaksanaan penelitian ini adalah :

1. Pencarian kata Bahasa Indonesia dalam deret pendekatan untuk Bahasa Indonesia yang acak akan lebih mudah jika telah ada suatu program yang dapat mendeteksi kata Bahasa Indonesia secara otomatis sehingga pencarian tidak lagi dilakukan secara manual.
2. Perlu dikembangkan suatu progran yang dapat menyusun kata-kata acak menjadi kalimat yang baik dan benar.

DAFTAR PUSTAKA

[1] Abramson, N., "Information Theory and Coding", McGraw Hill,Inc., New York, 1963.

[2] Guizzo, E.M., *The Essential Message: Claude Shannon and the Making of Information Theory*, Master Degree, Massachusetts Institute of Technology, Massachusetts, 2003.

[3] Papoulis, A., "Probabilitas, Variabel Random dan Proses Stokastik", Gadjah Mada University Press, Yogyakarta, 1992.

[4] Peebles Jr, P.Z., "Probability, Random Variables, and Random Signal Principles 3th Edition", McGraw Hill International, New York, 1993.

[5] Shannon, C.E., "A Mathematical Theory of Communication", *The Bell System Technical Journal*, 27, pp. 349-423, 623-656, 1948.

[6] ---, *Bigram*, http://en.wikipedia.org/, Oktober 2009.

[7] ---, *Statistical Distributions of English Text*, http://data-compression.com/, Oktober 2009.

[8] ---, *Tigram*, http://en.wikipedia.org/, Oktober 2009.

BIODATA PENULIS

Muhammad Jazilun Niam (L2F005553). Lahir di Magelang, 18 Desember 1987. Menempuh pendidikan dasar di MI Al-Huda Sampangan, MTs N Kaliangkrik, dan SMA N 3 Magelang. Pada tahun 2005 melanjutkan studi strata satu di Teknik Elektro Universitas Diponegoro Semarang, konsentrasi Elektronika Telekomunikasi.

Menyetujui,
Dosen Pembimbing I,

Achmad Hidayatno, S.T., M.T.
NIP. 196912211995121001

Dosen Pembimbing II,

Rizal Isnanto, S.T.,M.M, M.T.
NIP. 197007272000121001