Penerapan Pohon dan Himpunan dalam Klasifikasi Bahasa

Jeremia Jason Lasiman - 13514021[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]*jeremia_jason@s.itb.ac.id*

***Abstrak*—Bahasa merupakan cara komunikasi manusia yang paling beragam dan juga rumit. Ada lebih dari 6500 bahasa yang masih digunakan sampai saat ini. Dengan begitu banyaknya bahasa yang ada, tentu diperlukan penggolongan bahasa untuk mempermudah mempelajarinya. Salah satu teknik penggolongan atau lebih dikenal dengan klasifikasi adalah induksi pohon keputusan. Hasil klasifikasi bahasa ini juga beragam tergantung pendekatannya. Pendekatan genetis yang paling umum menggolongkan bahasa ke dalam relasi seperti keluarga. Pendekatan ini juga dapat direpresentasikan ke dalam dua model, yaitu model pohon dan model gelombang. Makalah ini akan membahas hubungan antara pemodelan matematis pohon dan himpunan dengan klasifikasi bahasa melalui induksi pohon keputusan.**

***Keywords*—Klasifikasi, Model Pohon, Model Gelombang, Bahasa Purba, Rumpun Bahasa.**

I. PENDAHULUAN

Bahasa adalah salah satu cara komunikasi manusia yang paling umum dan rumit. Penggunaan bahasa adalah suatu hal yang umum dalam hidup manusia, baik dalam bentuk tertulis, lisan, ataupun gerakan. Meskipun bahasa bersifat umum, hampir setiap daerah mempunyai bahasa masing-masing sehingga seringkali menjadi suatu ciri khas daerah.

Setiap bahasa memiliki keunikan sendiri. Cara pelafalan aksara, dialek, tata bahasa, serta arti kata dari tiap bahasa pasti terdapat perbedaan. Tetapi, tidak semua perbedaan sangatlah jauh sehingga tidak dapat dibandingkan. Beberapa bahasa memiliki kemiripan yang cukup tinggi terutama dari segi tata bahasa dan arti katanya. Dengan demikian, bahasa juga dapat digolongkan sesuai dengan kesamaannya.

Berbagai bahasa di dunia ini dapat digolongkan ke dalam kategori sesuai dengan rumpunnya, yang biasa dikenal dengan rumpun bahasa (*language family*). Tiap rumpun merupakan cabang dari bahasa purba (*proto language*). Bahasa purba yang secara umum diakui adalah *Proto-Indo-European, Proto-Uralic,* dan *Proto-Dravidian.* Dengan mengklasifikasi bahasa purba, maka dapat mempermudah para linguistik untuk menggolongkan bahasa-bahasa yang lain yang ada sekarang maupun yang sudah tidak ada lagi sebagai bahasa turunan dari bahasa-purba yang merupakan akar dari tiap rumpun bahasa.

Klasifikasi bahasa ini biasanya menggunakan representasi pohon (*tree mode*). Hal ini disebabkan karena hubungan antar bahasa yang ditemukan dari suatu akar

bahasa purba memiliki turunan yang banyak dan tiap anak bahasa juga mempunyai sifat rekursif yang sama dengan akarnya. Tetapi tidak semua jenis bahasa dapat direpresentasikan dalam pohon. Karena itu, para linguistik menambahkan suatu model atau teori baru bernama *wave model* (model gelombang). Model gelombang ini tidak meniadakan reperesentasi pohon, namun hanya menambahkan untuk beberapa bahasa yang sulit dikategorikan ke dalam salah satu anak bahasa saja.

II. DASAR TEORI

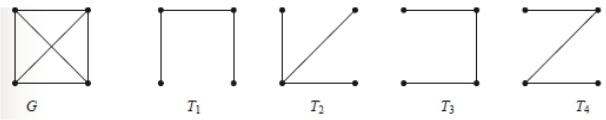
A. Pohon

Pohon adalah graf tidak-berarah terhubung yang tidak mengandung sirkuit. Istilah pohon pertama kali dikemukakan tahun 1857 oleh matematikawan Inggris, Arthur Carley.

Sebuah graf adalah pohon jika memenuhi syarat-syarat berikut [5] :

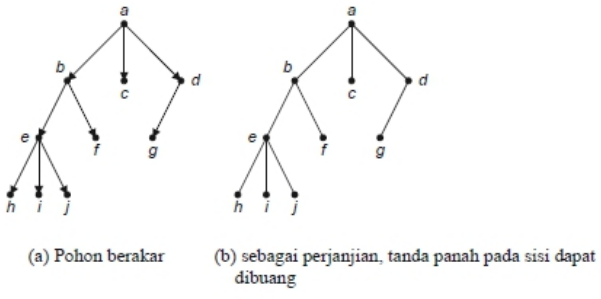
1. Setiap pasang simpul di graf terhubung dengan lintasan tunggal
2. Graf terhubung dan memiliki $m = n-1$ buah sisi
3. Graf tidak mengandung sirkuit dan memiliki $m = n-1$ buah sisi
4. Graf tidak mengandung sirkuit dan penambahan satu sisi pada graf akan membuat hanya satu sirkuit
5. Graf terhubung dan semua sisinya adalah jembatan.

Ada dua jenis pohon secara umum, yaitu pohon merentang (*spanning tree*) dan pohon berakar (*rooted tree*). Pohon merentang diperoleh dengan memutus sirkuit di dalam graf.



Gambar 2.1 Pohon merentang dari graf G

Sedangkan, pohon berakar adalah pohon yang salah satu simpulnya diperlakukan sebagai akar dan sisi-sisinya diberi arah sehingga menjadi graf berarah. Salah satu aplikasi pohon berakar adalah pohon keputusan (*decision tree*).



Gambar 2.2 Pohon berakar

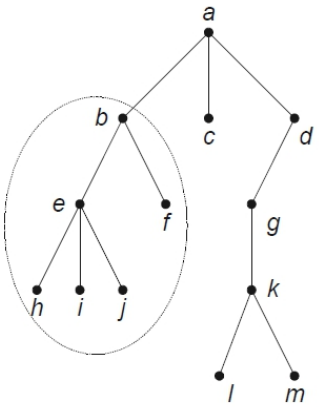
Pohon berakar memiliki beberapa terminologi sebagai berikut [5] :

1. Anak (*child*) dan Orang tua (*parent*)
   Pada gambar 2.2 *b,c,d* merupakan anak dari *a*, dan *a* adalah orang tua dari *b,c,d*

2. Lintasan (*path*)
   Pada gambar 2.2, lintasan dari *a* ke *i* adalah *a,b,e,i*.

3. Saudara kandung (*sibling*)
   f adalah saudara kandung *e* tetapi *f* bukan saudara kandung *g*

4. Upapohon (*subtree*)
   Simpul *b* dapat diambil menjadi akar dan dengan simpul-simpul dibawahnya menjadi upapohon.



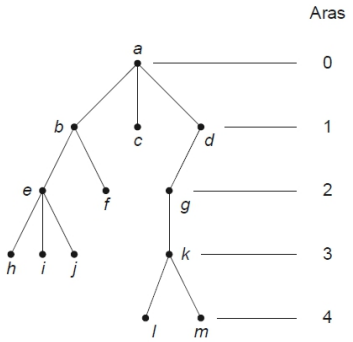
Gambar 2.3 Upapohon *b*

5. Derajat (*degree*)
   Derajat simpul adalah jumlah upapohon (atau jumlah anak) pada simpul tersebut.
   Derajat *a* adalah 3, *b* adalah 2.

6. Daun (*leaf*)
   Daun adalah simpul yang berderajat nol.

7. Simpul Dalam (*internal nodes*)
   Simpul dalam adalah simpul yang memiliki anak.

8. Aras (*level*) atau tingkat



Gambar 2.4 Aras pohon

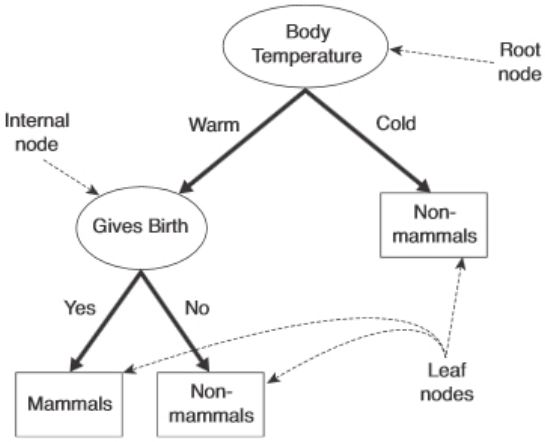
9. Tinggi (*height*) atau kedalaman (*depth*)
   Tinggi atau kedalaman merupakan aras maksimum sebuah pohon.

B. Klasifikasi dan Penggolong

Klasifikasi adalah proses untuk memasukkan objek ke dalam salah satu dari beberapa kategori yang telah ada [1]. Data masukan yang akan diklasifikasi dapat dilakukan dengan dua acara, yaitu pemodelan deskriptif (*descriptive modeling*) dan pemodelan prediktif (*predictive modeling*). Pemodelan deskriptif digunakan sebagai alat pemisah antara objek dari kelas yang berbeda, sedangkan pemodelan prediktif digunakan untuk memprediksi kelas dari catatan data baru.

Teknik klasifikasi, atau biasa disebut penggolong (*classifier*) adalah pendekatan sistematik untuk membuat model klasifikasi dari data-data yang ada. Penggolong ini digunakan sebagai pembeda dari tiap kelas yang ada. Salah satu cara yang paling umum digunakan untuk menentukan penggolong adalah dengan induksi pohon keputusan.

Pada pohon keputusan, setiap daun diberikan label kelas, sedangkan simpul yang bukan daun berisikan kondisi pemisah yang memiliki berbagai macam karakteristik [1]. Contoh induksi pohon keputusan gambar 2.4.



Sumber : [1]
Gambar 2.4 Pohon keputusan untuk klasifikasi mamalia

Salah satu algoritma untuk menentukan pohon keputusan adalah algoritma Hunt. Dalam algoritma Hunt, pohon keputusan bersifat rekursif. Misalkan D_t merupakan himpunan catatan yang berhubungan dengan simpul t dan $y = \{y_1, y_2, ..., y_c\}$ menjadi label kelas [1].

1. Jika semua catatatan di D_t termasuk dalam kelas y_t, maka t adalah simpul daun yang dilabelkan sebagai y_t [1].
2. Jika D_t berisi catatan yang termasuk lebih dari satu kelas, tes kondisi atribut digunakan untuk memisahkan catatan ke dalam himpunan yang lebih kecil. Node anak dibuat setiap keluaran dari tes kondisi dan catatan di D_t didistribusikan ke anak berdasarkan hasil keluaran. Algoritma kemudian dilakukan secara rekursif ke node anak [1].

C. Linguistik Historis

Linguistik historis adalah cabang linguistik yang mempelajari bagaimana dan mengapa perubahan bahasa terjadi sepanjang sejarah. Bidang ini juga membahas mengenai perubahan dan perbandingan terhadap bahasa lain yang serumpun serta perkembangan dialek dan sejarah kata.

Linguistik historis meliputi beberapa sub-bidang.

a. Comparative linguistics
Mempelajari kekerabatan suatu bahasa untuk mendapatkan hubungan historisnya.

b. Etymology
Mempelajari asal usul suatu kata.

c. Dialectology
Mempelajari bahasa berdasar distribusi geografis serta mencangkup percabangan dua dialek lokal dari induk bahasa yang sama.

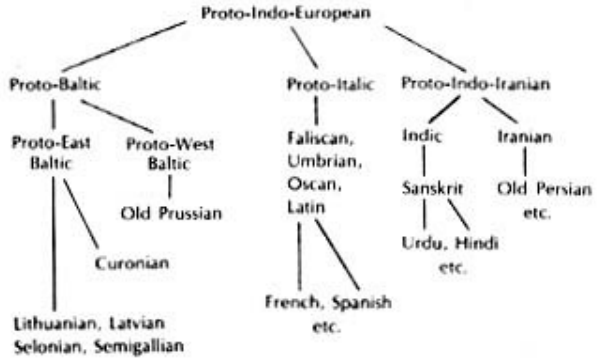
d. Phonology
Kajian bahasa yang mempelajari tentang bunyi-bunyi bahasa yang diproduksi alat ucap manusia.

e. Morphology
Morfologi mempelajari seluk-beluk kata serta pengaruh perubahan-perubahan bentuk kata terhadap golongan dan arti kata.

f. Syntax
Mempelajari prinsip dan peraturan membuat kalimat dalam bahasa.

Klasifikasi bahasa pada linguistik historis dilakukan pada cabang *comparative linguistics* yang kemudian dikelompokkan sesuai kekerabatan dan rumpun. Klasifikasi tersebut juga biasa direpresentasikan dalam dua bentuk model, yaitu *tree model* (model pohon) dan *wave model* (model gelombang).

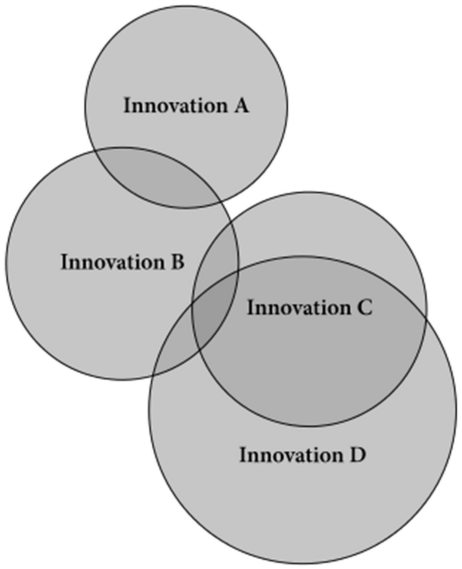
Model pohon merepresentasikan klasifikasi bahasa seperti analogi dari pohon keluarga. Konsep anak dari bahasa berarti bahasa yang terhubung mengalami perubahan sepanjang waktu dari bahasa awal sebagai orang tua menjadi suatu bahasa baru yang merupakan anak simpulnya.



Sumber : http://www.lituanus.org/1982_1/82_1_03.jpg

Gambar 2.5 Model pohon sederhana *Proto-Indo-European*

Model gelombang pertama kali dikemukakan oleh seorang linguistik Jerman, Johannes Schmidt, pada tahun 1872. Model ini menggunakan konsep yang mrirp dengan himpunan degnan representasi diagram Euler. Perbedaan dari model gelombang ini adalah jangkauan (diameter) dari tiap himpunan yang dapat semakin membesar tiap waktunya.



Sumber : http://www.qub.ac.uk/images/cramlap/logo-explan1.gif

Gambar 2.6 Representasi Model Gelombang Schmidt

III. INDUKSI POHON KEPUTUSAN UNTUK KLASIFIKASI BAHASA

Klasifikasi bahasa menggunakan cara penggolongan yang sama dengan klasifikasi pada umumnya. Dengan menggunakan pohon keputusan dan penggolong yang banyak. Selain itu, bahasa juga memiliki beberapa jenis penggolongan tergantung dengan pendekatannya. Ada empat jenis pendekatan pada klasifikasi bahasa, yaitu

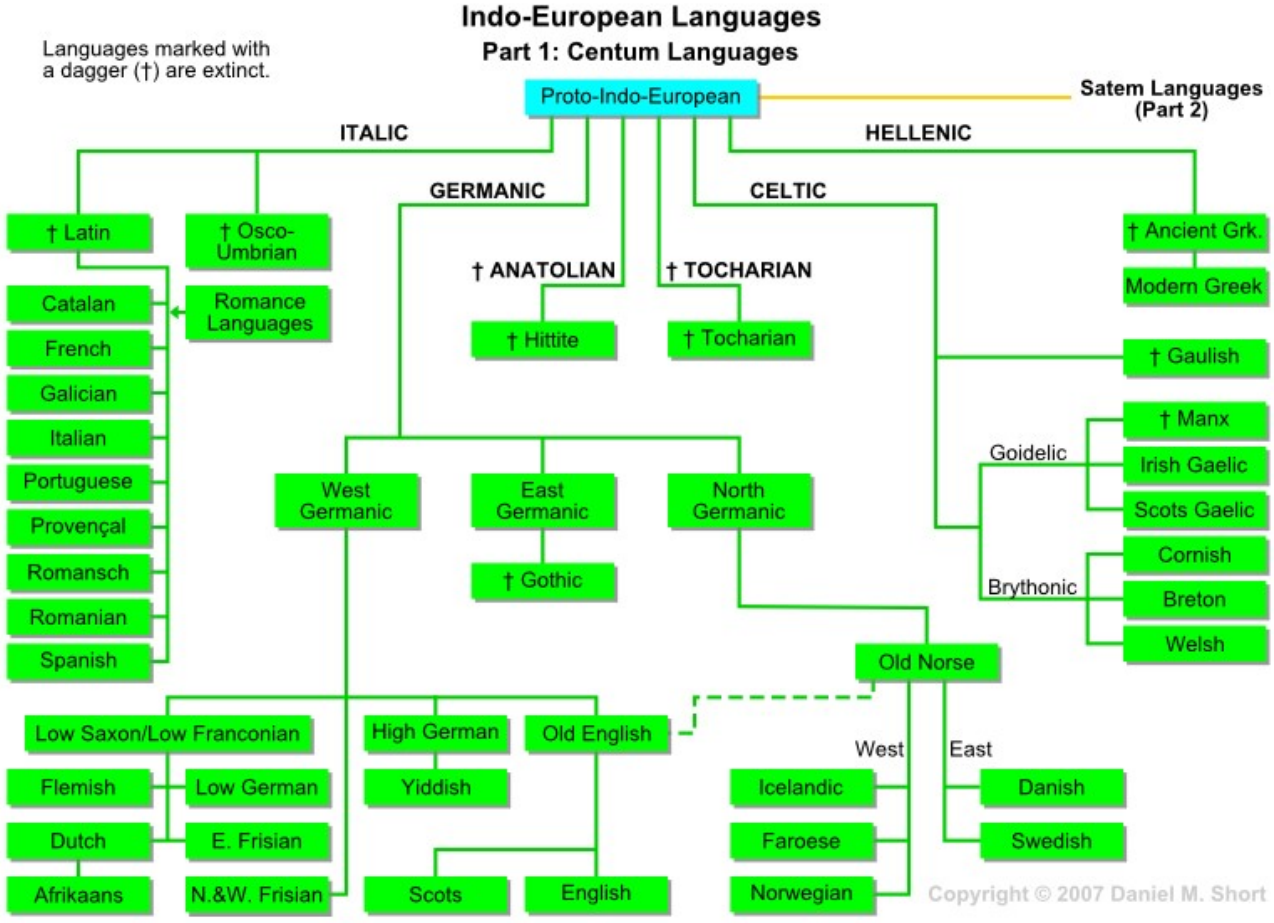
pendekatan genetis, pendekatan tipologis, pendekatan areal, dan pendekatan sosiolinguistik [2].

Klasifikasi yang paling umum digunakan adalah klasifikasi secara genetis. Klasifikasi genetis juga disebut klasifikasi geneologis, yang berarti suatu bahasa berasal dari bahasa yang lebih tua. Menurut klasifikasi genetis, suatu bahasa akan pecah menjadi dua bahasa baru atau lebih dan pecahan tersebut akan pecah lagi seperti sebelumnya. Teori ini dikemukakan oleh A.Schleicher pada tahun 1866 dan dilengkapi oleh Johannes Schmidt pada tahun 1872 dengan teori gelombang.

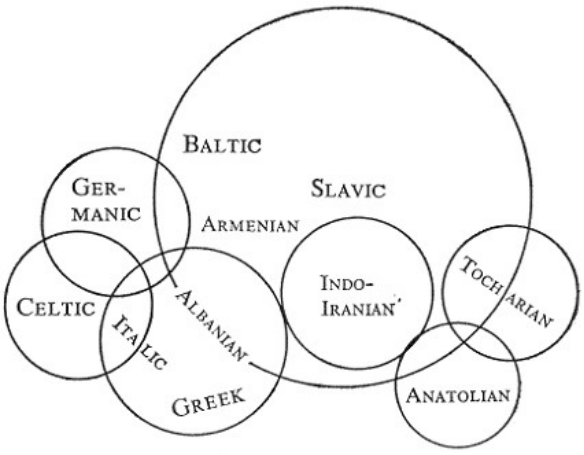
Klasifikasi genetis dilakukan berdasarkan kriteria kesamaan bunyi dan makna yang dikandugnya. Jika sebuah bahasa memiliki kesamaan dengan bahasa purba, maka bahasa itu akan digolongkan ke dalam anak dari bahasa purba. Kriteria yang paling dasar pada klasifikasi ini adalah garis keturunan atau dasar sejatah perkembangan yang sama [2].

Berdasarkan kriteria tersebut, maka pertama dilakukan adalah menggolongkan bahasa menurut urutan waktunya dan beberapa bahasa paling tua akan menjadi bahsa purba yang menjadi akar dari semua bahasa-bahasa di bawahnya. Selain menjadi akar, bahasa tersebut juga menjadi penggolong bagi bahasa lain yang kemudian dicocokan ke dalam beberapa bahasa purba. Bahasa yang memiliki kesamaan paling banyak pada suatu bahasa purba tertentu.

Hasil setelah penggolongan bahasa kemudian menjadi suatu kelas tersendri. Kelas ini biasa direpresentasikan dalam model pohon dan dilengkapi dengan model gelombang. Salah satu bahasa purba paling besar yang diketahui adalah *Proto-Indo-European*.



Sumber : http://www.danshort.com/ie/trees/iecentum1.png
Gambar 3.1 Keluarga bahasa Proto-Indo-Eurpean model pohon



Sumber:
http://www.qub.ac.uk/images/cramlap/logo-explan1.gif
Gambar 3.2 Keluarga bahasa Proto-Indo-Eurpean model gelombang

Penggunaan kedua model ini merepresentasikan dua jenis klasifikasi yang sama, tetapi menekankan hal yang berbeda. Penggunaan model pohon mempunyai penekanan faktor asal dan secara keturunan bahasa satu dengan yang lainnya, sedangkan penggunaan model gelombang memiliki penekanan pada relasi antar bahasa. Bahasa yang berada pada model pohon tidak dapat masuk ke dalam golongan lain meskipun bahasa itu mempunyai keunikan dari golongan lain juga. Setelah ditemukan model gelombang, dapat dinjelaskan hubungan satu bahasa dengan bahasa lain, tetapi untuk bahasa yang secara unik berbeda dengan yang lain tidak dapat mempunyai tempat di model gelombang. Karena itu, model pohon dan model gelombang dibutuhkan keduanya karena klasifikasi bahasa yang sangat luas adanya.

IV. KESIMPULAN

Klasifikasi adalah metode pengelompokan yang salah satu caranya adalah degan menggunakan induksi pohon keputusan berdasar pada penggolong yang ada. Penggolong dalam klasifikasi bahasa ditentukan berdasar pendekatan yang digunakan. Pendekatan yang paling sering digunakan adalah pendekatan secara genetis. Klasifikasi bahasa dapat dimodelkan secara umum sebagai model pohon dan model gelombang. Model pohon digunakan untuk merepresentasikan hubungan berubahan bahasa sepanjang sejarah, sedangakan model gelombang dapat menjelaskan hubungan kedekatan antar bahasa yang mungkin lebih dari satu rumpun bahasa.

Konsep Model pohon memanfaatkan aplikasi pohon yang dapat menunjukkan relasi orang tua dan anak. Model gelombang memakai himpunan yang dapat saling beririsan untuk menunjukkan hubungan antar bahasa yang dapat dekat ke lebih dari satu jenis saja.

V. REFERENSI

[1] Kumar, Vipin. *Introduction to Data Mining*. < http://www-users.cs.umn.edu/~kumar/dmbook/index.php> Tanggal Akses : 7 Desember 2015 pukul 21.08.

[2] Sanjaya, Rizki. Metode Pengelompokan Bahasa. < http://rizkimasbox.blogspot.co.id/2013/04/metode-pengelompokan-bahasa.html> Tanggal Akses : 8 Desember 2015 pukul 09.08.

[3] Lehmann, Winfred P. *Historical Linguistics : An Introduction*. < https://books.google.co.id/books?id=z0Ip3LUwexQC&printsec=frontcover&source=gbs_atb#v=onepage&q&f=false> Tanggal Akses : 7 Desember 2015 pukul 22.20.

[4] François, Alexandre. *The Routledge Handbook of Historical Linguistics, Chapter 6*. < http://alex.francois.free.fr/data/AlexFrancois_2014_HHL_Trees-waves-linkages_Diversification.pdf > Tanggal Akses : 7 Desember 2015 pukul 22.35.

[5] Munir, Rinaldi. *Matematika Diskrit*, Bandung : Informatika, 2003.

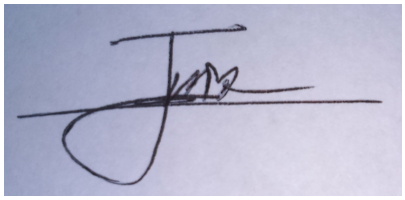
[6] Lehmann, Winfred P. *Historical Linguistics.* < https://books.google.co.id/books?id=z0Ip3LUwexQC&printsec=frontcover&source=gbs_atb#v=onepage&q&f=false> Tanggal Akses : 7 Desember 2015 pukul 23.45.

[7] Schmalstieg, William R. (1982). Special issue: The Lithuanian Language—Past and Present. *Lithuanian Quarterly Journal of Arts and Sciences, 82(1)*,01.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 8 Desember 2015



Jeremia Jason Lasiman, 13514021