BAB I
PENDAHULUAN

I.1 Latar Belakang

Bahasa merupakan alat yang digunakan untuk berkomunikasi. Tidak hanya untuk melakukan komunikasi antara manusia dengan manusia yang lainnya, namun dalam hal ini bahasa juga menjembatani komunikasi antara manusia dengan komputer. Bahasa yang digunakan manusia untuk berkomunikasi dengan komputer dikenal dengan bahasa pemrograman. Untuk mengolah bahasa dari manusia dan *computer* maka diperlukan sistem *Natural Languange Processing* (NLP).

Natural Language Processing (NLP) adalah salah satu bidang ilmu komputer, kecerdasan buatan, dan bahasa (linguistik) yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia, seperti bahasa Indonesia atau bahasa Lainnya. Tujuan utama dari studi NLP adalah membuat mesin yang mampu mengerti dan memahami makna bahasa manusia lalu memberikan respon yang sesuai.

Ada beberapa metode yang dilakukan untuk membuat mesin penerjemah yang mampu menerjemah bahasa Indonesia ke bahasa Sunda. Salah satu cara untuk meningkatkan akurasi mesin penerjemah stastik (MPS) adalah dengan menambah kuantitas korpus monolingual. Oleh karena itu, untuk menyelesaikan permasalahan yang ada maka diperlukan penelitian tentang peningkatan akurasi mesin penerjemah statistik tersebut dengan menambah kuantitas korpus monolingual.

I.2 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah dalam penelitian ini adalah seberapa besar penambahan kuantitas korpus monolingual agar dapat meningkatkan akurasi Mesin Penerjemah Statistik bahasa Indonesia – bahasa Sunda.

1.3. Tujuan Penelitian

Tujuan penelitian ini adalah meningkatkan akurasi terjemahan pada Mesin Penerjemah Statistik untuk penerjemahan bahasa Sunda ke bahasa Indonesia dengan cara menambah kuantitas korpus monolingual. Adapun manfaat dari penelitian ini adalah mengetahui seberapa besar penambahan kuantitas korpus monolingual agar dapat meningkatkan akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Sunda.

I.1 Pembatasan Masalah

Untuk menghindari meluasnya permasalahan yang ada, maka batasan penelitian ini adalah sebagai berikut.

1. Penerjemahan yang dilakukan adalah penerjemahan satu arah dari bahasa Indonesia ke bahasa Sunda.
2. Bahasa Sunda yang digunakan adalah bahasa Sunda yang terdapat dalam cerita rakyat yang digunakan sebagai korpus teks.
3. Korpus teks paralel yang digunakan berasal dari cerita rakyat yang menggunakan bahasa Sunda dan bahasa Indonesia.
4. Sistem yang dibangun tidak menggunakan fitur linguistik.

I.2 Sistematika Penulisan

Sistematika dalam penulisan pada tugas akhir ini dibagi dalam 5 (lima) bab yang terdiri dari Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Metodologi Penelitian, Bab IV Hasil dan Analisis serta Bab V Penutup.

Bab I Pendahuluan adalah bab yang berisi mengenai latar belakang, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan.

Bab II Tinjauan Pustaka berisi mengenai definisi penerjemahan, proses penerjemahan, korpus, bahasa Sunda, Mesin Penerjemah Statistik, Moses, model bahasa, *translation model, decoder, automatic evaluation*, *case folding*,

tokenizing. Dalam bab II juga terdapat kajian terkait yang berisi hasil-hasil penelitian yang telah dilakukan oleh peneliti terdahulu.

Bab III Metodologi Penelitian merupakan bab yang berisi tentang data dan perangkat penelitian, metode penelitian, diagram alir penelitian dan arsitektur system, pengumpulan data, pembuatan korpus teks paralel, membangun mesin penerjemah statistik, implementasi mesin penerjemah statistik, pengujian hasil terjemahan mesin translasi oleh BLEU, penambahan kuantitas korpus monolingual, implementasi mesin penerjemah statistik setelah dilakukan penambahan kuantitas korpus monolingual, pengujian ulang hasil terjemahan mesin translasi oleh BLEU, pengujian ahli bahasa, analisis hasil pengujian dan penarikan kesimpulan.

Bab IV Hasil dan Analisis adalah bab yang berisi data hasil penelitian, pengamatan, observasi yang telah dirancang pada bab III. Setiap hasil yang disajikan akan dilakukan analisis hasil pengujian untuk mengarah kepada suatu kesimpulan.

Bab V Penutup adalah bab terakhir dalam sistematika penulisan tugas akhir yang berisi kesimpulan dari penelitian yang telah dilakukan dan saran atau rekomendasi untuk perbaikan, pengembangan atau kelengkapan penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

II.1 Kajian Terkait

Hansel Tanuwijaya dan Hisar Maruli Manurung (2009), melakukan penelitian dengan menggunakan aturan restrukturisasi teks pada Bahasa Inggris sesuai dengan aturan struktur Bahasa Indonesia yang berupa *word reordering*, *phrase reordering*, atau keduanya. Terdapat dua korpus paralel yang digunakan, yaitu korpus paralel *bible* dan korpus paralel novel yang masing-masing korpus paralel berjumlah 12000 kalimat. Aturan-aturan yang dibuat terdiri dari 7 buah aturan *word reordering*, 7 buah *aturan phrase reordering*, dan 2 buah aturan gabungan *phrase reordering* dan *word reordering*. Hasil eksperimen menunjukkan peningkatan akurasi dan kualitas yang efektif diperoleh dengan *word reordering* yang dapat meningkatkan kualitas dan nilai akurasi mesin penerjemah statistik pada BLEU sebesar 1.3896% dan 0.6218% pada NIST.

Rahmat Izwan Heroza (2014) melakukan penelitian dampak kata yang dapat memiliki beberapa kelas kata terhadap hasil mesin penerjemah berbasis statistik dalam menerjemahkan dokumen berbahasa Arab ke dalam Bahasa Indonesia. Korpus paralel yang digunakan dalam penelitian ini berjumlah 6226 kalimat yang bersumber dari ayat-ayat Al-Qur'an. Penelitian ini mengungkap bahwa kalimat yang memiliki kata yang memiliki beberapa makna diterjemahkan tidak sesuai dengan kelas katanya. Penelitian ini akhirnya mengusulkan solusi untuk mengurangi kesalahan pemilihan makna tersebut dengan cara menambahkan informasi mengenai kedudukan suatu kata dalam kalimat pada dokumen sumber sebelum proses training dilakukan (PoS tag). Hasilnya adalah tingkat efektivitas sebesar 20% atau sebanyak 5 dari 25 kalimat yang mengandung kata yang dapat memiliki beberapa kelas kata bisa diterjemahkan dengan baik menggunakan mesin penerjemah berbasis statistik yang telah mengimplementasi solusi yang diusulkan dalam penelitian ini.

Steffi Melinda (2012), melakukan penelitian yang berjudul "Mesin Penerjemah Statistik Mesin Penerjemah Statistik Dua Arah untuk Bahas

Indonesia dan Bahasa Mandarin Menggunakan Dictionary-Based Lookup dan Model Penerjemahan Berfaktor". Penelitian ini mengembangkan mesin penerjemah statistik untuk penerjemahan bahasa Indonesia dan Mandarin secara dua arah. Penelitian dilakukan dengan memanfaatkan korpus paralel kitab suci untuk kedua bahasa. Mesin penerjemah statistik dikembangkan menggunakan model penerjemahan tanpa faktor dengan atau tanpa postprocessing berupa dictionary based lookup serta dengan faktor berupa part-of-speech. Hasil penerjemahan kemudian dibandingkan dengan hasil penerjemahan sistem dictionary-based lookup paling sederhana dan Google Translate. Penerjemahan kualitas terbaik diberikan sistem mesin penerjemah statistik sederhana yang mencapai nilai akurasi NIST 5.7096 dan BLEU 0.2475 untuk penerjemahan dari bahasa Mandarin ke Indonesia dan sistem mesin penerjemah statistik dengan post-processing berupa dictionary-based lookup dengan NIST score 4.6713 dan BLEU score 0.2022 untuk penerjemahan dari bahasa Indonesia ke bahasa Mandarin.

Herry Sujaini dkk (2012), melakukan penelitian menggunakan penandaan *part of speech* yang dimasukkan sebagai fitur linguistik dalam model penerjemah faktor (MPF) menggunakan sistem MPS Moses dan menggunakan BLEU sebagai alat evaluasi. Dari hasil penelitian, penggunaan *part of speech* memiliki dampak terhadap meningkatnya kualitas terjemahan untuk bahasa Inggris-Indonesia, hal tersebut terlihat dari hasil eksperimen bahwa dengan menambahkan fitur *part of speech*, akurasi mesin penerjemah meningkat sebesar 2%.

Danny Indrayana (2016) melakukan penelitian peningkatan akurasi mesin penerjemah statistik bahasa Indonesia ke bahasa Melayu Pontianak. Pengujian akan dilakukan dengan cara membandingkan nilai akurasi mesin penerjemah statistik sebelum penambahan *part of speech* dan sesudah penambahan *part of speech*. Dari hasil penelitiannya didapatkan persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia ke bahasa Melayu Pontianak yang dicapai dengan korpus uji berasal dari dalam korpus sebesar 0.6% pada pengujian otomatis oleh BLEU dan korpus uji dari luar korpus sebesar 24.57%.

Hasbi Hardiansyah (2016) melakukan penelitian yang berjudul "*Tuning For Quality* Untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia – Bahasa Dayak Kanayatn". Penelitian ini dilakukan untuk mengetahui apakah dengan melakukan *Tuning* dapat meningkatkan akurasi mesin penerjemah statistik bahasa Indonesia ke bahasa Dayak Kanayatn. Berdasarkan hasil pengujian, proses *tuning* dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia-bahasa Dayak Kanayatn. Sebelum dilakukan proses *tuning*, skor BLEU pada korpus uji 3667 sebesar 89,47 dan setelah dilakukan proses *tuning* didapat skor BLEU sebesar 92,19. Terdapat peningkatan nilai BLEU sebesar 3,04% dilihat dari perbandingan sebelum dan sesudah dilakukan proses *tuning*.

Robby Darwis melakukan penelitian untuk meningkatkan akurasi mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda dengan menambah kuantitas korpus monolingual. Sumber korpus paralel yang dibuat berasal dari teks berbahasa Sunda yang disajikan dalam dua bahasa yaitu bahasa Sunda dan bahasa Indonesia. Implementasi dilakukan dengan cara melakukan *training* untuk memperoleh model bahasa dan model *translasi* dan proses selanjutnya yaitu *decoding* dan *decoder*. Untuk pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi. Pengujian dilakukan dengan penilaiaan (*scoring*) dengan perbandingan hasil terjemahan menggunakan BLEU dan pengujian oleh ahli bahasa. Analisis hasil pengujiannya dengan membandingkan nilai akurasi dari beberapa mesin penerjemah statistic dan berdasarkan hasil pengujian didapatkan seberapa besar penambahan korpus monolingual untuk meningkatkan akurasi mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.

Tabel 2.1 Perbandingan Penelitian

No.	Penulis	Judul	Metode	Keterangan
1	Hansel Tanuwijay	Penerjemahan Dokumen	*Word Reorderin*	Penelitian melakukan

	a dan Hisar Maruli Manurung (2009)	Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan *Word Reordering* dan *Phrase Reordering*	*g* dan *Phrase Reorderin g*	penerjemahan bahasa Inggris-bahasa Indonesia Peningkatan diperoleh dengan *word reordering* dibandingkan *phrase reordering* sebesar 1.3896% pada nilai BLEU dan 0.6218% pada nilai NIST
2	Rahmat Izwan Heroza (2014)	Dampak Kelas Kata Bahasa Arab Terhadap Hasil Mesin Penerjemah Berbasis Statistik	*Pos Tag*	Menerjemahkan bahasa Arab-bahasa Indonesia dengan menambahkan informasi mengenai kedudukan suatu kata dalam kalimat sebelum proses *training* dilakukan (*PoS tag).* Penelitian meningkatkan efektifitas sebesar 20% atau sebanyak 5 dari 25 kalimat
3	Steffi Melinda (2012)	Mesin Penerjemah Statistik Mesin Penerjemah Statistik Dua Arah untuk Bahasa Indonesia dan Bahasa Mandarin Menggunakan *Dictionary-Based*	*Dictionary-Based Lookup* dan *Part of Speech*	Penelitian melakukan terjemahan dua arah bahasa Indonesia-bahasa Mandarin Hasil akurasi NIST score 5.7096 dan BLEU score 0.2475 untuk penerjemahan bahasa Mandarin-Indonesia dan NIST score 4.6713 dan

		Lookup dan Model Penerjemahan Berfaktor		BLEU score 0.2022 untuk penerjemahan bahasa Indonesia-bahasa Mandarin dengan *Dictionary-Based Lookup*.
4	Herry Sujaini dkk (2012)	Pengaruh *Part-Of-Speech* Pada Mesin Penerjemah Bahasa Inggris-Indonesia Berbasis *Factored Translation Model*	*Part of speech* dengan *Factored Translation Model*	Penelitian dilakukan pada Mesin Penerjemah Bahasa Inggris-Indonesia Hasil peningkatan akurasi sebesar 2% untuk korpus 15.000 kalimat
5	Danny Indrayana (2016)	Meningkatkan Akurasi Mesin Penerjemah Bahasa Indonesia ke Bahasa Melayu Pontianak dengan *part of speech*	*Part of speech*	Korpus paralel bersumber dari dokumen bahasa Melayu Pontianak dan bahasa Indonesia. Pengujian akan dilakukan dengan cara membandingkan nilai akurasi mesin penerjemah statistik sebelum penambahan *part of speech* dan sesudah penambahan *part of*

				speech.
6	Muhammad Hasbiansyah (2016)	*Tuning For Quality* Untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia – Bahasa Dayak Kanayatn	*Tuning For Quality*	Mesin penerjemah statistik dapat diimplementasikan untuk menerjemahkan bahasa Indonesia-bahasa Dayak Kanayatn. Berdasarkan hasil pengujian, proses *tuning* dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia-bahasa Dayak Kanayatn. Sebelum dilakukan proses *tuning*, skor BLEU pada korpus uji 3667 sebesar 89,47 dan setelah dilakukan proses *tuning* didapat skor BLEU sebesar 92,19. Terdapat peningkatan nilai BLEU sebesar 3,04% dilihat dari perbandingan sebelum dan sesudah dilakukan proses *tuning*.

II.2 Definisi Penerjemahan

Ada beberapa definisi dari berbagai sumber mengenai penerjemahan. Penerjemahan berasal dari Bahasa Arab Tarjammah yang berarti mengalihbahasakan suatu bahasa ke bahasa lain. Di dalam Kamus Besar Bahasa

Indonesia (KBBI) edisi ketiga terjemah/ menerjemahakan merupakan menyalin / memindahakan suatu bahasa ke bahasa lain atau mengalihbahasakan.

Selain itu, penerjemahan menurut Hoed (23:2006) adalah kegiatan mengalihkan secara tertulis pesan dari teks suatu bahasa (misalnya bahasa Inggris) ke dalam tekas bahasa lain (misalnya bahasa Indonesia). Memang bukan suatu hal yang mudah untuk menerjemahkan suatu teks. Menyampaikan pesan merupakan kegiatan menerjemahkan yang paling utama wajib dilakuakan.

Larson menuliskan bahwa pada dasarnya penerjemahan ialah suatu perubahan bentuk dari suatu bahasa. Perubahan ini dapat berupa frasa, klausa, kalimat, paragraf dan sebagainya. dalam kaitan lisan maupun tulisan. Ini dilihat dari struktur luarnya saja. Artinya, selain membawa pesan, kegiatan menerjemahkan juga merupakan kegiatan mengubah bentuk bahasa dengan tujuan hasil terjemahan dapat dipahamai sebagai teks yang dapat dinikmati pembaca dan bahkan teks dirasa tidak seperti teks hasil terjemahan.

Jadi, penerjemahan itu proses mengalih bahasa atau mengalih eja secara tulisan suatu bahasa ke bahasa lain tanpa mengubah pesan yang ingin disampaikan. Walaupun terjadi perubahan bentuk (frasa, klausa, kalimat dan paragraf). Seperti yang ditulis Nida dan Taber (12:1974) penerjemahan harus bertujuan untuk menyampaikan pesan. Tetapi penyampaian pesan ini akan mengalami penyesuaian bentuk leksikal dan gramatikal.

2.3. Proses Penerjemahan

Proses penerjemahan adalah suatu model yang dimaksudkan untuk meneerangkan proses berfikir (internal) yang dilakukan seorang penerjemah saat melakukan penerjemahan. secara sederhana, proses penerjemahan terdiri dari dua tahap yaitu :

1. Analisis teks asli dan pemahaman makna atau pesan teks asli.
2. Pengungkapan kembali makna atau pesan tersebut ke dalam bahasa sasaran.
3. Dengan kata-kata atau kalimat yang difahami oleh pembaca bahasa sasaran.

Sedang Menurut Ibnu Burdah dalam proses menerjemah itu membutuhkan tiga tahapan yakni :

1. Penyelaman pesan naskah sumber yang hendak diterjemah.
2. penuangan pesan naskah sumber ke dalam bahasa sasaran
3. Proses Editing

Menurut E. Sadtono, proses penerjemahan terdiri dari empat tahap, yaitu :

1. Analisis

   Pada tahap ini penerjemah melakukan analisis struktur lahiriyah bahasa sumber. Tujuanya adalah untuk menemukan hubungan tata bahasa dan maksud suatu perkataan/kombinasi perkataan/frase. Dalam tahap ini ada tiga langkah yang perlu diperhatikan yaitu :

   1. menentukan hubungan yang mengandung arti antara perkataan-perkataan dan gabungan perkataan.
   2. menetukan maksud acuan perkataan atau kombinasi perkataan-perkataan atau idiom
   3. menentukan makna konotasi, yaitu reaksi pemakai bahasa itu terhadap suatu perkataan atau gabungan perkataan, baik positif maupun negatif.

   Dengan melakukan analisi bahasa sumber, seorang penerjemah akan bisa memahami maksud, arti dan konteks bahasa tersebut yang mempermudah penerjemah untuk bisa memahami teks secara keseluruhan.

2. Transfer

   Setelah melakukan proses analisa teks sumber, maka dalam kondisi kedua ini penerjemah melakukan olah bahasa di dalam otaknya guna mentransfer apa yang ada di dalam bahasa sumber tadi kedalam bahasa sasaran. Dengan mencoba memahami teks tersebut dari sudut bahasa sasaran. Dalam aktivitas menerjemah tahapan ini pasti dilalui karena proses memahami teks bahasa sumber terjadi pada saat ini. Proses tranfer ini terjadi dipikiran seorang penerjemah.

3. Restrukturisasi

   Atau pada tahapan Ibnu Burdah adalah. Penuangan pesan ke bahasa sasaran. Pada tahap ini pemahaman akan makana atau pesan bahasa sasaran

itu distrukturkan kembali atau ditulis kembali namun ke dalam bahasa sasaran. Langkah inilah yang merupakan kegiatan menerjemahkan yang sesungguhnya. Penerjemah memilih padanan kata yang sesuai dalam bahasa penerima, agar pesan penulis dapat tersampaikan. Terkadang penerjemah mengikuti struktur bahasa sumber jika tidak ditemukan kejanggalan dalam menerjemahkanya kedalam bahasa sasaran, namun jika struktur bahasa sumber tersebut dirasa tidak sesuai maka penerjemah bisa merubahnya dengan catatan pesan atau maknanya tidak berubah.

4. Revisi atau Penghalusan hasil terjemah.

Apabila proses restrukturisasi sudah selesai maka selanjutnya adalah menguji atau mengevalusai teks terjemahan. Tahap ini sama dengan tahapan ke tiga yang diberikan Ibnu Burdah yakni proses Editing. Upaya mengedit kembali adalah usaha mengolah terjemahanaya agar hasilnya (dalam bahasa sasaran) menjadi cukup lugas. Proses ke empat ini penulis anggap sangat penting, karena proses editing adalah proses dimana kita menelaah/ membaca kembali dan memahami teks terjemahan yang kita hasilkan untuk menghasilkan penerjemahan yang benar-benar baik dan mudah difahami. Apabila terjemahan dinilai sudah memiliki ketepatan yang tinggi terhadap pesan bahasa sumber dengan menggunakan bahasa sasaran, maka proses ini dikatakan sudah cukup.

2.4. Korpus

Korpus didefinisikan sebagai koleksi atau sekumpulan contoh teks tulis atau lisan dalam bentuk data yang dapat dibaca dengan menggunakan seperangkat mesin dan dapat diberi catatan berupa berbagai bentuk informasi linguistic (McEnery dkk, 2006:5). Lebih lanjut, McEnery merangkum kriteria korpus, yang telah menjadi kesepakatan banyak ahli, yakni:

1. dapat dibaca dengan menggunakan seperangkat mesin,
2. berupa teks otentik,
3. digunakan sebagai sampel,
4. mewakili bahasa atau variasi bahasa tertentu.

Korpus dapat diklasifikasikan ke dalam beberapa jenis tergantung tujuan penggunaannya. Hunston (2002: 14-16) membaginya ke dalam delapan jenis yakni korpus khusus (specialised corpus), korpus umum (general corpus), korpus komparatif (comparable corpus), korpus paralel (parallel corpus), korpus pemelajar (learner corpus), korpus pedagogis (pedagogic corpus), korpus historis atau diakronis (historical or diachronic corpus), dan korpus monitor (monitor corpus). Berdasarkan jenis korpus tersebut, untuk penelitian ini penulis akan focus pada korpus paralel.

Korpus paralel adalah dua atau lebih korpus dalam bahasa yang berbeda. Masing-masing korpus memuat teks yang telah diterjemahkan dari satu bahasa ke bahasa lain, misalnya sebuah teks yang memuat kumpulan peraturan bagi Negara-negara Uni Eropa yang telah diterjemahkan ke dalam semua bahasa anggota Uni Eropa. Korpus ini dapat digunakan penerjemah untuk melihat adanya persamaan ekspresi di masing-masing bahasa atau melihat perbedaan yang ada diantara dua bahasa.

2.5 Bahasa Sunda

Indonesia merupakan negara kesatuan yang terdiri dari beragam suku, budaya, dan bahasa. Selain bahasa Indonesia sebagai bahasa nasional, bahasa daerah merupakan khasanah kekayaan yang sangat penting untuk dijaga dan dilestarikan agar terhindar dari jamahan asing yang mampu menghapus jejak budaya kita.

Salah satu suku terbesar di Indonesia adalah Suku Sunda. Suku Sunda memiliki bahasa daerahnya sendiri yang disebut bahasa Sunda. Secara geografis, suku Sunda terletak di Pulau Jawa bagian barat. Dengan kata lain, Suku Sunda terletak di Jawa Barat.

Dalam bahasa Sunda, ada yang dikenal dengan nama bahasa Sunda Kuna. Bahasa Sunda Kuna biasanya tertulis pada benda-benda peninggalan sejarah, seperti tulisan-tulisan di batu yang disebut prasasti maupun naskah-naskah yang ditulis pada daun lontar.

Bahasa Sunda merupakan bahasa yang unik dengan tingkatan-tingkatan berbahasa, atau lebih dikenal dengan istilah undak-usuk yang nyaris tidak dimiliki oleh bahasa lain.

2.6 Mesin Penerjemah (Machine Translation)

Mesin Penerjemah adalah perangkat lunak otomatis yang cerdas yang mampu menerjemahkan sumber data dalam jumlah besar ke berbagai bahasa. Menurut Safaba, salah satu pengembang mesin terjemahan, terdapat tiga jenis mesin penerjemah yang dikenal pada saat ini, yaitu:

1. *Rule-Based Machine Translation*

   Rule-Based Machine Translation menggunakan kumpulan aturan yang sangat banyak, dikembangkan secara manual dari waktu ke waktu oleh ahli bahasa dalam memetakan struktur dari bahasa sumber ke bahasa target. Faktor ahli bahasa dalam sistem *rule-based* membantu memberikan terjemahan otomatis cukup baik dengan hasil yang diprediksi. Dikarenakan tenaga ahli bahasa yang terbatas, sistem rule-based membutuhkan biaya yang tinggi, memakan waktu dalam mengimplementasikan dan melakukan pemeliharaan dan sistem ini memiliki potensi menghasilkan ambiguitas dan degradasi terjemahan dari waktu ke waktu.

2. *Statistical Machine Translation*

   Statistical Machine Translation menggunakan algoritma computer untuk menghasilkan terjemahan yang terbaik secara statistik dari sekian banyak permutasi. Model statistik terdiri dari kata dan frasa yang dihasilkan dari korpus paralel. Daya tarik dari sistem statistik adalah tingkat otomatisasi dalam membangun sistem baru, tidak memakan banyak waktu, dan biaya yang rendah dibutuhkan untuk membangun dan mengoperasikan model statistik ini.

3. *Hybrid Machine Translation*

   Hybrid machine Translation adalah perpaduan antara Rule-Based Machine Translation dan Statistical machine translation. Ada beberapa

teknik *hybrid* MT, antara lain: keluaran *rule based* MT, kemudian hasilnya diatur lagi berdasar statistical atau hasil penerjemahan dari SMT kemudian diatur ulang tata bahasanya berdasar aturan yang baku. Hasil terjemahannya dari SMT yang kemudian diatur ulang tata bahasanya.

2.7 Mesin Penerjemah Statistik (MPS)

Mesin Penerjemah Statistik merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel (Hadi, 2014). Sebagai contoh, proses penerjemahan dari kalimat bahasa Indonesia ke bahasa Sunda dapat diinterpretasikan dengan P(i|s). P(i|s) merupakan ditribusi probabilitas kalimat i (bahasa Indonesia) terhadap kalimat s (bahasa Sunda). Nilai P(i|s) dapat dihtung dengan menggunkaan teorema Bayes, yaitu:

$$P(i|s) = \frac{P(s \vee i).P(i)}{P(s)}$$

(2.1)

Keterangan:

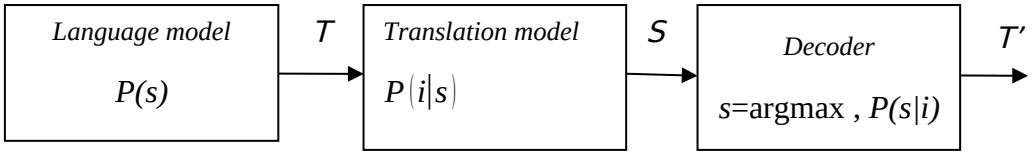
P(*i*) : Kalimat bahasa *i* (bahasa Sunda)

P(*s*) : Kalimat bahasa *j* (bahasa Indonesia)

P(*i*/*s*) : Kalimat bahasa *i* sebagai kalimat bahasa *j*

P(*s*/*i*) : Kalimat bahasa *j* sebagai kalimat bahasa *i*

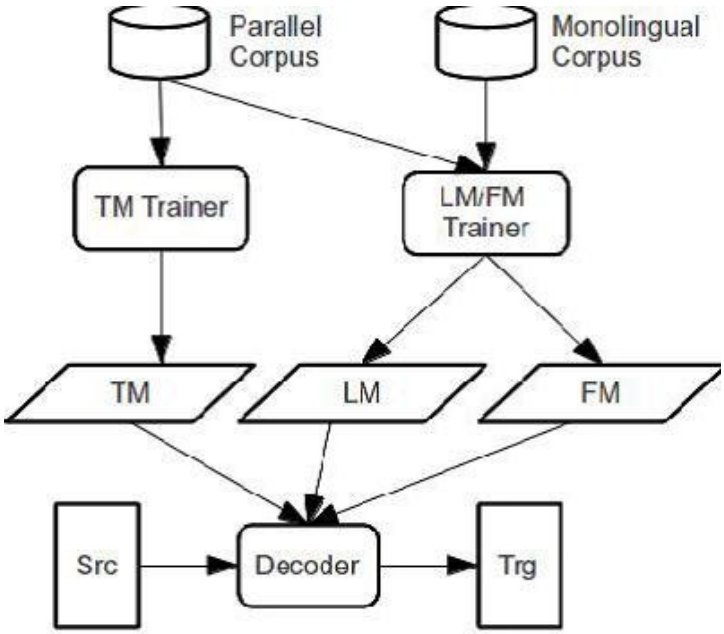
Terdapat 3 komponen yang terlibat dalam proses penerjemahan dari satu bahasa ke bahasa lain pada mesin penerjemah statistik, yaitu *language model, translation model, dan decoder.*



Gambar 2.1 Komponen Mesin Penerjemah Statistik [Manning dan Schutze, 2000:486]

Penjelasan singkat dari Gambar 2.1 yaitu, *language model* menghasilkan kalimat bahasa *s* (bahasa Sunda), *translation model* mengirimkan kalimat bahasa *s* sebagai kalimat bahasa *s* (bahasa Sunda). *Decoder* mencari kalimat bahasa *s* yang paling mungkin yang telah menimbulkan kalimat *s.*

Secara umum, arsitektur mesin penerjemah statistik Moses ditunjukkan pada Gambar 2.2.



Gambar 2.2 Arsitektur Mesin Penerjemah Statistik Moses [Sujaini dan Negara, 2015]

Sumber data utama yang dipergunakan adalah *parallel corpus* dan *monolingual corpus*. Proses *training* terhadap *parallel corpus* menggunakan GIZA++ menghasilkan *translation model* (TM). Proses *training* terhadap bahasa target pada *parallel corpus* ditambah dengan *monolingual corpus* bahasa target

menggunakan SRILM menghasilkan *language model* (LM), sedangkan *PoS model* (PoS-M) dihasilkan dari bahasa target pada *parallel corpus* yang setiap katanya sudah ditandai dengan PoS. *TM, LM dan PoS-M* digunakan untuk menghasilkan *decoder* Moses. Selanjutnya Moses digunakan sebagai mesin penerjemah untuk menghasilkan bahasa target dari *input* kalimat dalam bahasa sumber.

2.8 Moses

Moses adalah salah satu Mesin Penerjemah Statistik yang memungkinkan untuk menerjemahkan secara otomatis setiap pasangan bahasa. Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa sasaran. Saat melakukan penerjemahkan bahasa, Moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa sasaran. Moses dirilis di bawah lisensi LGPL (*Lesser General Public License*) dan tersedia sebagai kode sumber dan *binary* untuk *Windows* dan *Linux*. Perkembangannya didukung oleh proyek *EuroMatrix*, dengan pendanaan oleh European Commission (Koehn, 2007).

2.9 Model Bahasa (Language Model)

Model bahasa digunakan pada aplikasi *Natural Language Processing* seperti *speech recognition, part of speech tagging* dan *syntactic parsing*. Dalam *language model* statistik, bagian-bagian yang merupakan elemen kunci adalah probabilitas dari rangkaian-rangkaian kata yang dituliskan sebagai $P(w_1,w_2,\ldots,w_n)$ atau $P(w_{1,n})$. *Language model* menetapkan probabilitas $P(w_{1,n})$ ke serangkaian *n* kata dengan *means* sebuah distribusi probabilitas. Rangkaian-rangkaian tersebut bisa berupa frase-frase atau kalimat-kalimat dan probabilitasnya dapat diperkirakan dari korpus dokumen-dokumen yang besar. Salah satu contoh pendekatan *language model* adalah *n-gram model.* Model bahasa n-gram merupakan jenis probalilistik *language model* untuk memprediksi *item*

berikutnya dalam urutan tersebut dalam bentuk (n-1). Dalam model n-gram, probabilitas $P(w_1,\ldots,w_m)$ mengamati kalimat w $_{1,}$ ..., w $_m$ diperkirakan sebagai :

$$P(w_1,w_2,\ldots,w_n)=\prod_{i=1}^{m} P\left(w_n \vee w_{1,\ldots} w_{n-1}\right) \approx \prod_{i=1}^{m} P\left(¿\, w_n \vee w ¿¿ n-(n-1),\ldots,w_{n-1}\right) ¿¿$$

Pada persamaan diatas diasumsikan bahwa probabilitas dari mengamati kata w_n ke-n dalam sejarah konteks *n*-1 kata-kata terdahulu, dapat diaproksimasikan oleh probabilitas mengamatinya dalam sejarah konteks *n*-1 kata-kata terdahulu yang diperpendek.

Probabilitas bersyarat dapat dihitung dari jumlah frekuensi n-gram :

$$P(w_n \vee w_{n-(n-1)},\ldots,w_{n-1})=\frac{jumlah\left(w_{n-(n-1)},w_{n-1},\ldots,w_i\right)}{jumlah\left(w_{n-(n-1)},\ldots w_{n-1}\right)}$$

Berikut merupakan contoh model bahasa n-gram,yaitu :

$$Unigram \quad : P(w_i) = \frac{count\left(w_i\right)}{total\, words\, observed}$$

$$Bigram \quad : P(w_i) = \frac{count\left(w_{i-1}w_i\right)}{count\left(w_{i-1}\right)}$$

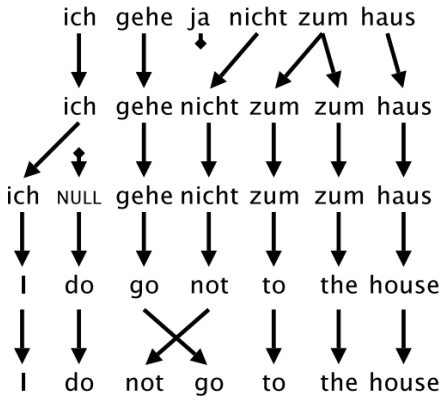
$$Trigram \quad : P(w_i) = \frac{count\left(w_{i-1}w_{i-2}w_i\right)}{count\left(w ¿¿ i-1 w_{i-2}\right) ¿}$$

Salah-satu keunggulan menggunakan *n-gram* dan bukan suatu kata utuh secara keseluruhan adalah bahwa *n-gram* tidak akan terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen.

2.10 Model Translasi (Translation Model)

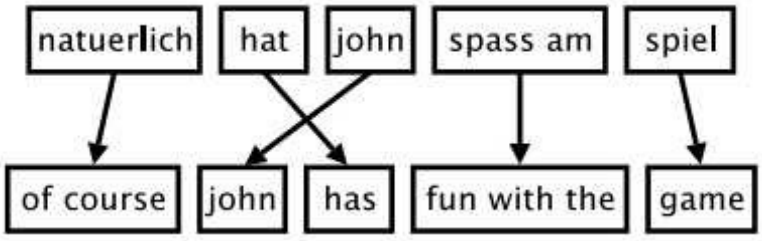
Translation model digunakan untuk memasangkan teks input dalam bahasa sumber dengan teks output dalam bahasa sasaran. Terdapat dua model penerjemahan dalam mesin penerjemah statistik, yaitu word-based translation

model (model translasi berbasis kata) dan phrase-based translation model (model translasi berbasis frase) (Tanuwijaya dan Manurung, 2009).



Gambar 2.3 Contoh penerjemahan dengan word-based translation model [Koehn, 2009]

Proses yang terjadi pada word-based translation model adalah penerjemahan kata, reordering, duplication, dan insertion (Koehn, 2009).



Gambar 2.4 Contoh penerjemahan dengan phrase-based translation model [Koehn, 2009]

Gambar 2.4 menunjukkan ilustrasi phrase-based translation model. Proses penerjemahan dipecah menjadi beberapa bagian yaitu dengan membagi kalimat bahasa sumber menjadi barisan frase, menerjemahkan setiap frase ke bahasa sasaran, dan reordering (Koehn, 2009). Penggunaan phrase-based translation model saat ini lebih diuntungkan dibanding word-based translation model karena pada phrase-based translation model semakin banyak data yang dikumpulkan maka semakin banyak pula frase yang dapat dipelajari dan phrase-

based translation model dapat menangani lebih banyak terjemahan sehingga dapat meningkatkan kualitas terjemahan disbanding word based translation model.

2.11 Decoder

Fungsi *decoder* adalah untuk mencari teks dalam bahasa sasaran yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan *language model*. Perhitungan $\hat{T}$ (hasil terjemahan) dapat dituliskan sebagai berikut:

$$\hat{T} = arg_T \, max \, P(T|S) = arg_T \, max \frac{P(S|T).P(T)}{P(S)} = arg_T \, max \, P(S|T).P(T)$$

Keterangan:

$\hat{T}$: Hasil terjemahan

$arg_T max$: Nilai maksimal

$P(T|S)$: Probabilitas bahasa target terhadap bahasa sumber

$P(S|T)$: Probabilitas bahasa sumber terhadap bahasa target

$P(S)$: Probabilitas bahasa sumber

$P(T)$: Probabilitas bahasa target

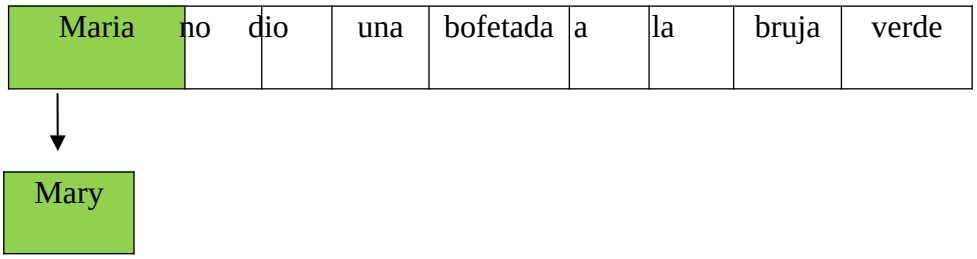
Fungsi $arg_T max$ mencari T (bahasa sasaran) yang dapat memberikan nilai probabilitas terbesar yang diperoleh. Proses *decoding* menggunakan algoritma *beam search. Beam search* adalah algoritma pencarian heuristik yang merupakan optimasi dari pencarian *best-first search* yang mengurangi kebutuhan memorinya. Terdapat dua konsep penting dalam algoritma *beam search* yang digunakan, yaitu konsep pemangkasan (*pruning*) dan estimasi *future cost*. Berikut ini adalah contoh proses penerjemahan pada *phrase-based statistical machine translation* (Koehn, 2004) :

a. Memilih kata-kata asing dalam bahasa sumber

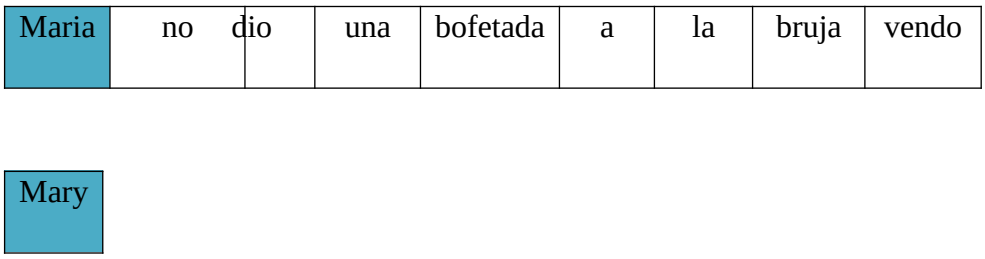
Maria	no	dio	una	bofetada	a	la	bruja	verde

b. Mencari terjemahan frase bahasa sasaran

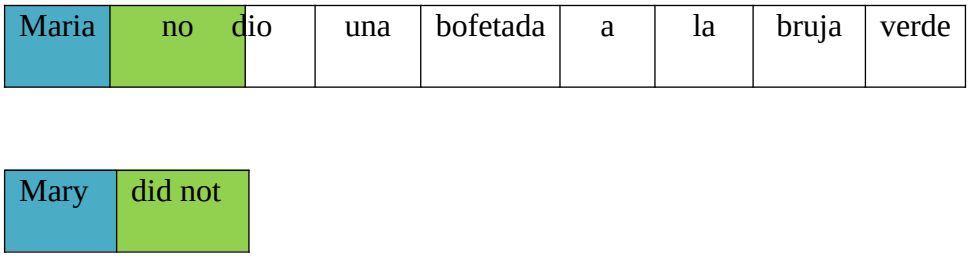
c. Menambahkan frasa bahasa sumber pada akhir terjemahan parsial



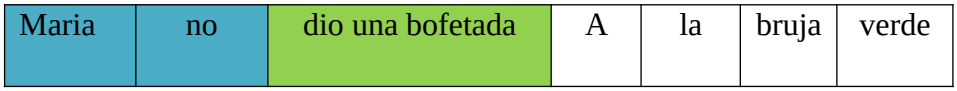
d. Menandai bahasa sumber yang telah diterjemahkan



e. Penerjemahan dari satu- ke- banyak (*one-to-many*)



f. Penerjemahan dari banyak- ke- satu (*many-to-one*)



Mary	did not	slap

g. *Reordering* (penyusunan kembali)

Maria	no	dio una bofetada	A	la	bruja	verde

Mary	did not	slap	the	green

Keterangan :

- Blok berwarna hijau merupakan proses pencarian frasa
- Blok berwarna biru merupakan frasa sudah diterjemahkan

2.12 Evaluasi Otomatis

Suatu mesin penerjemah statistik membutuhkan suatu sistem evaluasi otomatis untuk menentukan kualitas terjemahan. Kualitas mesin penerjemah statistik secara umum dinilai dari hasil terjemahan yang dihasilkan. Penilaian dapat dilakukan secara manual dan otomatis. Penilaian secara manual merupakan cara penilaian yang terbaik karena memberikan nilai akurasi yang lebih tinggi. Namun penilaian secara manual memiliki kekurangan yaitu membutuhkan sumber daya manusia (ahli bahasa) yang banyak dan tentunya membutuhkan waktu yang lama.

Sistem evaluasi otomatis yang populer saat ini adalah BLEU (*Bilingual Evaluation Understudy*). BLEU adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah hasil terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. Ide utama dibalik ini adalah " semakin dekat terjemahan sebuah mesin dengan terjemahan manusia makan akan semakin baik" (Papineni, 2002). BLEU mengukur *modified* n-*gram precission score* antara hasil terjemahan otomatis dengan tejemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty*.

Untuk menghitung nilai *precission score* dapat dilakukan dengan menghitung jumlah kata pada hasil terjemahan (unigrams) yang sesuai dengan

rujukan dan dibagi dengan total jumlah kata (unigrams) yang ada pada hasil terjemahan. Namun sayangnya mesin penerjemah statistik dapat menghasilkan kata-kata dengan berlebihan, sehingga menghasilkan terjemahan yang mustahil tetapi memiliki *precission* yang tinggi. Dalam penelitian yang dilakukan oleh Papieni, dihasilkan beberapa perubahan yang dikenal dengan metode *modified* n-*gram precision.* Untuk menghitungnya, pertama kali hitung berapa kali jumlah maksimal dari kata yang muncul dalam terjemahan rujukan tunggal. Selanjutnya gabungkan jumlah total dari setiap kalimat terjemahan dengan jumlah maksimal rujukan (Papineni, 2002).

Tabel 2.2 merupakan contoh dengan penghitungan *modified* unigram *precision* :

Tabel 2.2 Contoh Perhitungan *Modified Unigram Precission* [Tanuwijaya dan Manurung, 2009]

Hasil Terjemahan	Rujukan
the the the the the the the	*The cat is on the mat*

Nilai precision unigram adalah 7/7. Sedangkan nilai *modified* unigram 2/7. Jumlah maksimum kata "*the*" pada rujukan adalah 2 dan jumlah unigram pada hasil terjemahan adalah 7 (Tanuwijaya, 2009).

Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score*. Semakin tinggi nilai BLEU,maka semakin akurat dengan rujukan. Nilai dari BLEU berada pada rentang 0 sampai 1. Suatu terjemahan akan mencapai nilai 1 jika terjemahan tersebut identik dengan terjemahan rujukan. Oleh karena itu, meskipun dengan penerjemahan oleh manusia tidak mungkin akan menghasilkan nilai 1. Sangat penting untuk diketahui bahwa semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasikan nilai BLEU yang tinggi, panjang kalimat hasil terjemahan harus mendekati panjang dari kalimat referensi dan kalimat hasil terjemahan harus memiliki kata dan urutan yang sama dengan kalimat referensi. Rumus BLEU sebagai berikut (Tanuwijaya, 2009) :

$$BP_{BLEU} = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}$$

$$P_n = \frac{\sum_{C \epsilon\ corpus\ n-gram \epsilon C} \sum count_{clip^{(n-gram)}}}{\sum_{C \epsilon\ corpus\ n-gram \epsilon C} \sum count_{(n-gram)}}$$

$$\text{BLEU} = BP_{BLEU} \cdot e^{\sum_{n-1}^{N} w_n \log p_n}$$

Keterangan :

BP = *brevity penalty*

c = jumlah kata dari hasil terjemahan otomatis

r = jumlah kata rujukan

P_n = *modified precission score*

w_n = 1/N (standar nilai N untuk BLEU adalah 4)

p_n = jumlah n-gram hasil terjemahan yang sesuai dengan rujukan dibagi jumlah n-gram hasil terjemahan.

Berikut merupakan contoh pasangan hasil kalimat terjemahan dan rujukan.

Tabel 2.3 Contoh Pasangan Kalimat Sumber dan Rujukan

Baris	Sumber	Rujukan
1	Ery suka makan nasi goreng kambing	Ery suka makan nasi goreng dan minum teh
2	Ery suka makan soto	Ery suka makan bakso

Maka nilai BLEU untuk kalimat pertama dari contoh di atas` (“Ery suka makan nasi goreng kambing”) adalah 0.6912316215478422. Adapun proses perhitungannya yaitu :

- Nilai c = 6 dan r = 8
- c ≤ r , maka nila brevity penalty = $e^{\left(1-\frac{r}{c}\right)}$

$$= e^{\left(1-\frac{8}{6}\right)}$$

$$= e^{-2/6}$$

$$= 0.77880783$$

- Nilai *modified* 1-gram *precision* atau p_1 adalah 5/6

Hasil terjemahan :

Ery	Suka	makan	nasi	goreng	kambing

Rujukan :

Ery	suka	makan	nasi	goreng	dan	minum	teh

- Nilai *modified* 2-gram *precision* atau p_2 adalah 4/5

Hasil terjemahan :

Ery	suka	makan	nasi	goreng	kambing

Rujukan :

Ery	suka	makan	nasi	Goreng	dan	minum	teh

- Nilai *modified* 3-gram *precision* atau p_3 adalah 3/4

Hasil terjemahan :

Ery	suka	makan	nasi	goreng	kambing

Rujukan :

Ery	suka	makan	nasi	goreng	dan	minum	teh

o Nilai *modified* 4-gram *precision* atau p_3 adalah 2/3

Hasil terjemahan :

Ery	suka	makan	nasi	goreng	kambing

Rujukan :

Ery	suka	makan	nasi	goreng	dan	minum	teh

Jadi dapat diperoleh nilai BLEU sebagai berikut:

BLEU = 0.77880783. $e^{\sum_{n-1}^{4} \frac{1}{4} \log p_n}$

= 0.77880783. $e^{\{\frac{1}{4}¿¿}$

= 0.77880783 . $e^{\{-0,119280336799156¿¿}$

= 0.6912316215478422

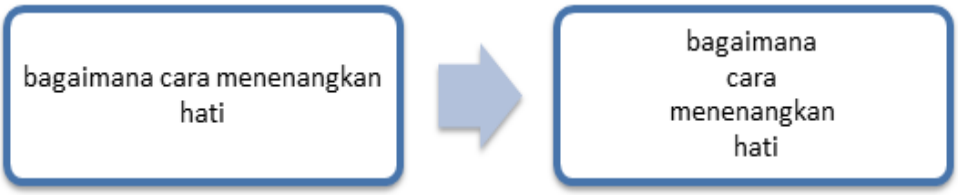
2.13 Case Folding

Case folding merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (pembatas) (Triawati, 2009). Gambar 2.6 merupakan contoh penggunaan *case folding*.

Gambar 2.5 Contoh case folding [Triawati, 2009]

2.14 Tokenizing

Tahap *tokenizing*/*parsing* adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya (Triawati, 2009). Selain itu, spasi digunakan untuk memisahkan antar kata tersebut.



Gambar 2.6 Contoh tokenizing/parsing [Triawati, 2009]

BAB III
METODOLOGI PENELITIAN

3.1 Data dan Perangkat Penelitian

3.1.1 Data Penelitian

Data penelitian yang digunakan berupa buku cerita rakyat dari Bandung. Buku cerita tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Indonesia dan bahasa Sunda.

3.1.2 Perangkat Penelitian

Perangkat penelitian yang digunakan dalam penelitian ini terdiri dari perangkat keras dan perangkat lunak. Perangkat tersebut diantaranya.

1. Perangkat Keras

   Laptop Acer Aspire 4755G dengan spesifikasi sebagai berikut.

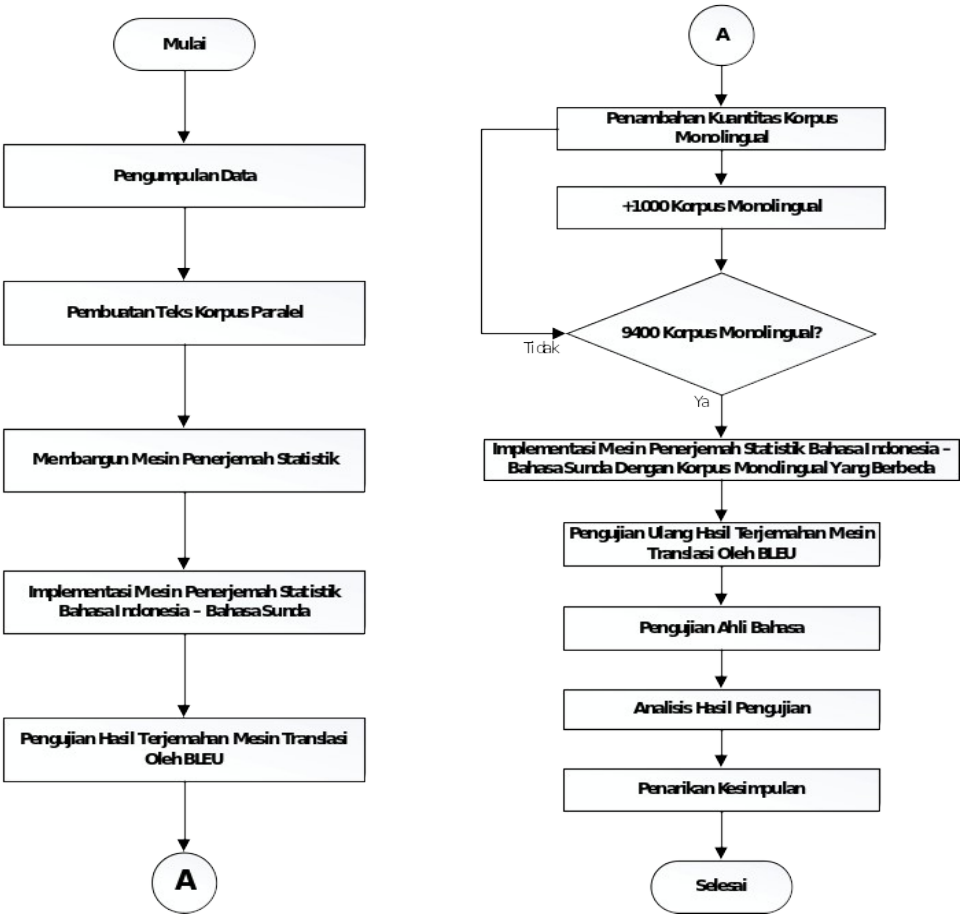
   a. *Processor* Intel Core i5-2430M 2.40 GHz
   b. *Graphic Processor* NVIDIA GeForce GT-540M
   c. RAM 2 GB
   d. *Hard Drive* 1 TB

2. Perangkat Lunak

   a. Sistem Operasi Linux Ubuntu 14.04 LTS 64 Bit
   b. SRILM untuk pemodelan bahasa
   c. Giza++ untuk pemodelan translasi
   d. Moses untuk *decoding*
   e. Sublime Text 3 untuk teks editor
   f. Cool Retro Term untuk menjalankan *command-line*

3.2 Metodologi Penelitian

Metodologi penelitian yang dilakukan dijelaskan pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

a. Pengumpulan Data

Pengumpulan data dilakukan dengan cara mencari buku dari beberapa toko buku yang terdapat di daerah Bandung berupa cerita rakyat berbahasa Sunda.

b. Pembuatan Korpus Teks Paralel

Korpus teks paralel dibuat dari terjemahan cerita rakyat Sunda, yang menghasilkan dua buah dokumen teks yaitu teks dalam bahasa Sunda dan teks dalam bahasa Indonesia.

c. Membangun Mesin Penerjemah Statistik

Membangun Mesin Penerjemah Statistik dilakukan dengan cara melakukan instalasi perangkat lunak. Lihat lampiran A.

d. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Sunda.

   Implementasi dilakukan dengan cara melakukan pemodelan bahasa, pemodelan translasi dan *decoding*.

e. Pengujian Hasil Terjemahan Mesin Translasi Oleh BLEU

   Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi. Pengujian dilakukan dengan cara pengujian secara otomatis menggunakan BLEU dengan menggunakan metode *K-Fold Cross-Validation*.

f. Penambahan Kuantitas Korpus Monolingual

   Penambahan kuantitas korpus monolingual dilakukan dengan cara menambahkan korpus monolingual bahasa Sunda yang berbeda

g. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Sunda Setelah Penambahan Kuantitas Korpus Monolingual

   Implementasi dilakukan dengan cara melakukan pemodelan bahasa, pemodelan translasi dan *decoding* setelah melalui tahap penambahan kuantitas korpus monolingual.

h. Pengujian Ulang Hasil Terjemahan Mesin Translasi Oleh BLEU

   Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi setelah menambah kuantitas korpus monolingual. Pengujian dilakukan dengan cara pengujian secara otomatis menggunakan BLEU dengan menggunakan metode *K-Fold Cross-Validation*..

i. Pengujian Ahli Bahasa

   Pengujian oleh ahli bahasa dilakukan apabila telah terdapat peningkatan pada pengujian ulang. Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi setelah ditambahkan kuantitas korpus monolingual dan sebelum ditambahkan kuantitas korpus monolingual.

j. Analisis Hasil Pengujian

   Analisis hasil pengujian dilakukan untuk mengetahui karakteristik mesin penerjemah statistik dan mengidentifikasi apakah sudah sesuai dengan kebutuhan serta membandingkan nilai akurasi mesin penerjemah statistik setelah menambah kuantitas korpus monolingual.

k. Penarikan Kesimpulan

Kesimpulan dirumuskan berdasarkan pengujian yang telah dilakukan apakah sistem yang dirancang mampu memberikan solusi berdasarkan permasalahan yang ada.

3.3 Pengumpulan Data

Penulis menggunakan metode wawancara dan observasi untuk memperoleh data yang diperlukan dalam penelitian. Langkah-langkah yang dilakukan adalah sebagai berikut.

1. Wawancara

   Diperlukan ahli bahasa sebagai narasumber dalam pengujian hasil translasi mesin penerjemah dalam penelitian ini. Penulis melakukan wawancara kepada orang yang akan menjadi ahli bahasa dengan cara berdiskusi sejauh mana mereka mengerti mengenai bahasa Sunda dan kesediaan mereka untuk menjadi ahli bahasa dalam penelitian ini.

2. Observasi

   Setelah mendapatkan data berupa orang-orang yang menjadi narasumber, maka penulis selanjutnya melakukan observasi dengan cara mencari buku cerita berbahasa Sunda yang nantinya akan dijadikan korpus dalam melakukan penelitian. Didapatkan sebuah kendala yakni tidak ada buku cerita yang menggunakan bahasa Sunda yang terdapat di Pontianak. Adapun solusi yang diambil dalam mendapatkan buku cerita berbahasa Sunda tersebut adalah dengan meminta bantuan kepada keluarga penulis di Bandung untuk mencari buku cerita berbahasa Sunda yang nantinya akan diterjemahkan dalam bahasa Indonesia dengan bantuan ahli bahasa.

3.4 Pembuatan Korpus Teks Paralel

Korpus merupakan kumpulan dari beberapa teks sebagai sumber penelitian bahasa dan sastra. Penelitian yang akan dilakukan terdapat dua buah korpus paralel yang digunakan yaitu korpus bahasa Sunda dan bahasa Indonesia. Adapun korpus paralel pada penelitian ini berupa cerita-cerita rakyat dari daerah

Bandung. Sebelum menggunakan kedua korpus paralel tersebut dalam penelitian, terlebih dahulu melakukan penerjemahan secara manual karena data yang diperoleh menggunakan Bahasa Sunda yang kemudian dilakukan pengetikan ulang untuk menjadikannya berupa file teks, karena data yang didapat dalam bentuk buku sehingga tidak mempunyai *softcopy*-nya.

Korpus paralel yang digunakan disimpan dengan nama yang sama, tetapi berbeda format berkasnya. Korpus paralel yang dibuat dalam bentuk *file* teks disimpan dengan format .sd untuk bahasa Sunda dan .id untuk korpus bahasa Indonesia. Contoh korpus paralel yang dibuat dalam bentuk *file* teks dapat dilihat pada Gambar 3.2.

Gambar 3.2 Contoh korpus paralel bahasa Indonesia – bahasa Sunda

3.5 Mesin Penerjemah Statistik

Membangun mesin penerjemah statistik dilakukan dengan cara melakukan instalasi perangkat lunak. Adapun perangkat lunak yang akan di-*install* untuk membangun mesin penerjemah statistik adalah sebagai berikut.

a. Mosesdecoder untuk membangun mesin penerjemah.
b. Boost dan GIZA++ untuk *compile* Moses.
c. SRILM untuk pemodelan bahasa.
d. *Mteval-11b.pl* untuk pengujian akurasi.

Seluruh perangkat lunak yang digunakan dalam membangun mesin penerjemah statistik merupakan perangkat lunak yang bersifat *open source* dan di-

install pada sistem operasi Ubuntu. Seluruh langkah-langkah instalasi perangkat lunak dapat dilihat pada bagian lampiran A.

3.6 Implementasi Mesin Penerjemah Statistik Bahasa Indonesia - Sunda

Setelah membangun mesin penerjemah statistik, maka langkah selanjutnya adalah membangun mesin penerjemah statistik bahasa Indonesia ke Bahasa Sunda. Dikarenakan dalam menguji mesin penerjemah statistik bahasa Indonesia ke Bahasa Sunda ini menggunakan metode *K-Fold Cross-Validation*, maka akan dibuat 5 buah mesin penerjemah statistik bahasa Indonesia ke Bahasa Sunda dengan cara korpus paralel yaitu 3000 kalimat dibagi menjadi 5 bagian dengan nama file Fold A, Fold B, Fold C, Fold D, Fold E. Dan setiap korpus berisi 600 Kalimat.

Tabel 3.1 Data Pembagian Kalimat Uji Bahasa Indonesia-Sunda

Nama	Jumlah Kalimat Uji
Fold A	1-600
Fold B	601-1200
Fold C	1201-1800
Fold D	1801-2400
Fold E	2401-3000

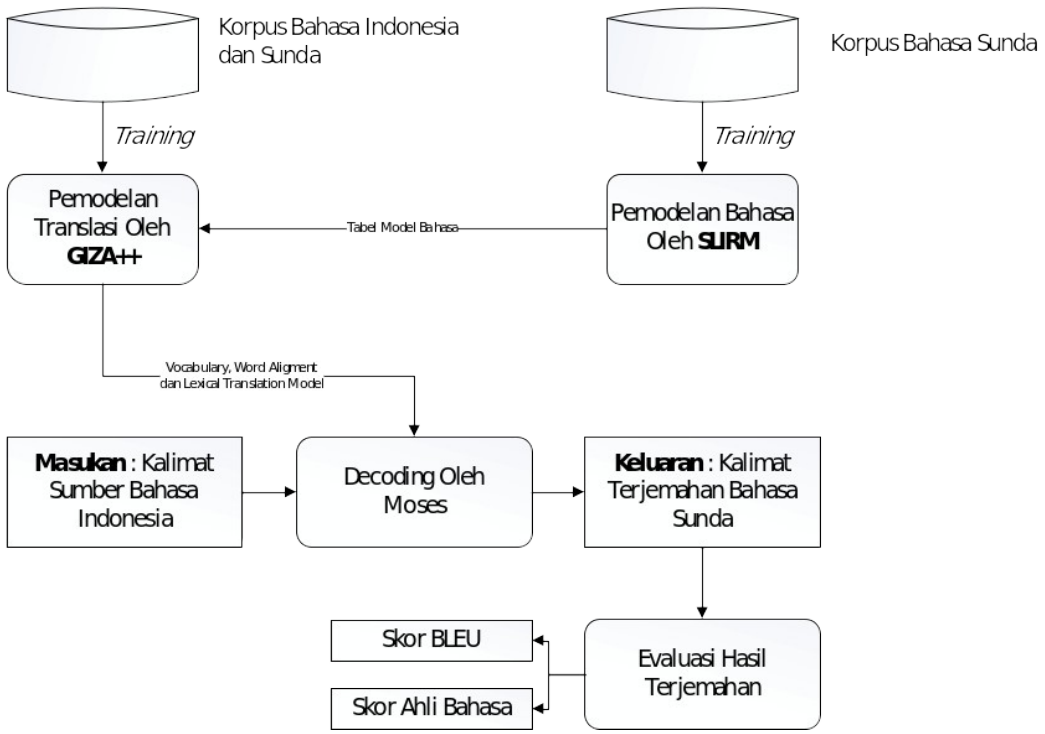
Pada tabel 3.1 terdapat jumlah kalimat uji yang telah dibagi dari korpus awal menjadi 5 bagian. korpus tersebut yang akan digunakan untuk membangun 5 buah mesin penerjemah statistik.

Tabel 3.2 Training Mesin Penerjemah Statistik Bahasa Indonesia-Sunda

Mesin	Training Korpus	Korpus Uji
M-A	Fold B+C+D+E	Fold A
M-B	Fold A+C+D+E	Fold B
M-C	Fold A+B+D+E	Fold C
M-D	Fold A+B+C+E	Fold D
M-E	Fold A+B+C+D	Fold E

Pada tabel 3.2 dijelaskan cara membangun mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.

Arsitektur sistem pada penelitian ini terdiri dari beberapa proses yaitu pemodelan bahasa, pemodelan translasi, *decoding* dan evaluasi hasil terjemahan. Arsitektur sistem mesin penerjemah statistik ditunjukkan pada Gambar 3.3.



Gambar 3.3 Arsitektur sistem Mesin Penerjemah Statistik bahasa Indonesia - Sunda [Modifikasi: Hadi, 2014]

Korpus paralel dalam *file* teks terdiri dari dua buah korpus yaitu korpus bahasa Indonesia dan korpus bahasa Sunda. Gambar 3.3 merupakan arsitektur sistem mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda yang terdiri dari beberapa tahapan, yaitu pemodelan bahasa, pemodelan translasi, *decoding*, dan proses evaluasi hasil terjemahan yang mana proses evaluasi hasil terjemahan akan dijelaskan pada poin selanjutnya. Berikut penjelasan dari tahapan pemodelan bahasa, pemodelan translasi dan proses *decoding* tersebut.

3.6.1. Pemodelan Bahasa Oleh SRILM

Pemodelan bahasa oleh SRILM (*Standarford Research Institute Language Modelling*) dilakukan pada bahasa target dan menghasilkan tabel model bahasa dengan n-gram data. Model bahasa n-gram memiliki nilai probabilitas dalam bahasa target. Proses pemodelan bahasa oleh SRILM dapat dilihat pada Gambar 3.4.

Gambar 3.4 Proses pemodelan bahasa dengan bahasa Sunda sebagai bahasa target [Modifikasi: Hadi, 2014]

Contoh tabel model bahasa yang dihasilkan oleh SRLIM dapat dilihat pada Gambar 3.5.

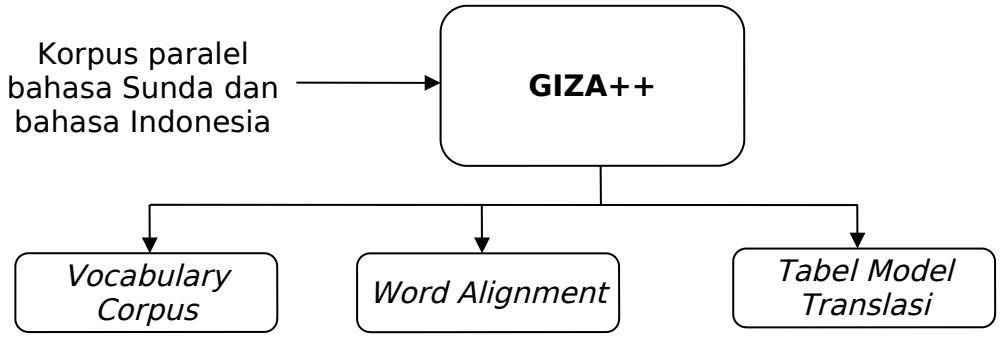
```
\data\
ngram 1=3619
ngram 2=13264
ngram 3=1019

\1-grams:
-2.81221    abdi  -0.2256695
-3.475367   acan  -0.08614869
-4.016886   aeyuna      -0.08614868
……

\2-grams
-2.210994   abdi gaduh  -0.1470668
-0.811752   albeum dibuka     -0.1470667
```

```
-1.840778   ari ayeuna  -0.06591805



\3-grams
-0.4813389  tos ampir sasasih
-0.5218764  ka angin peuting
-0.577717   mawa cai kopi
```

Gambar 3.5 Contoh tabel model bahasa dengan bahasa Sunda sebagai bahasa target

3.6.2. Pemodelan Translasi Oleh Giza++

Proses pemodelan translasi oleh Giza++ menghasilkan *vocabulary corpus*, *word alignment* dan *lexical model table*. Proses translasi oleh Giza++ dapat dilihat pada Gambar 3.6.



Gambar 3.6 Proses pemodelan translasi [Modifikasi: Hadi, 2014]

3.6.2.1. *Vocabulary Corpus*

Dokumen *vocabulary corpus* berupa dokumen yang berisi setiap kata pada masing-masing korpus dimana setiap kata-kata tersebut memiliki *uniq id* yang diikuti oleh kata (token) dan frekuensi kemunculannya. Dokumen *vocabulary corpus* yang akan dihasilkan terdiri dari *vocabulary corpus* bahasa Indonesia dan *vocabulary corpus* bahasa Sunda. Contoh dokumen *vocabulary corpus* yang dihasilkan ditunjukkan pada Gambar 3.7 dan 3.8.

```
1       UNK         0
2       yang         520
3       dia         113
4       ada         147
5       sekarang     42
```

Gambar 3.7 Contoh dokumen *vocabulary corpus* bahasa Indonesia

```
1       UNK       0
2       nu          477
3       manehna 115
4       aya       93
5       ayeuna     86
```

Gambar 3.8 Contoh dokumen *vocabulary corpus* bahasa Sunda

Angka 1 sampai 5 pada dokumen *vocabulary corpus* merupakan *uniq id* untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan.

3.6.2.2. *Word Alignment*

Proses *word alignment* akan menghasilkan dokumen korpus *alignment* yaitu kalimat bahasa target dipetakan pada kalimat bahasa sumber. Contoh Dokumen *alignment* yang dihasilkan dapat dilihat pada Gambar 3.9.

```
# Sentence pair (30) source length 8 target length 7 alignment score : 0.00118982

sini sama awit , dasar nenek-nenek pemalas

NULL ({ }) ka ({ }) dieuh ({ 1 }) ku ({ 2 }) awit ({ 3 }) , ({ 4 }) dasar ({ 5 }) nini-nini ({ 6 }) pangedulan ({ 7 })
```

Gambar 3.9 Contoh dokumen *alignment* bahasa Indonesia - bahasa Sunda

Dokumen *alignment* terdapat tiga baris kalimat. Baris pertama berisi letak kalimat sumber dalam korpus, panjang kalimat sumber, panjang kalimat sumber dan nilai *alignment.* Baris kedua merupakan bahasa sumber dan baris ketiga merupakan *alignment* kalimat bahasa target terhadap kalimat bahasa sumber.

Gambar 3.9 merupakan dokumen *alignment* dengan bahasa Indonesia sebagai bahasa sumber dan bahasa Sunda sebagai bahasa target. Gambar tersebut

terdapat kata “pangedulan” pada bahasa target yang di-*align* dengan kata yaitu kata ke-7 pada kalimat bahasa sumber yaitu “pemalas”.

3.6.2.3. Tabel Model Translasi

Proses pemodelan translasi oleh Giza++ akan menghasilkan tabel translasi yang terdiri dari tabel kata yang berisi kosakata dari bahasa sumber yang memiliki makna pada bahasa sasaran ataupun sebaliknya (leksikal). Setiap kosakata yang dihasilkan memiliki nilai probabilitas. Contoh tabel model translasi yang dihasilkan ditunjukkan pada Gambar 3.10.

```
atas luhur 0.8333333
tertawa seuri 0.9090909
kaya beunghar 1.0000000
tidur sare 1.0000000
bukan lain 0.9574468
lama lami 0.5000000
```

Gambar 3.10 Contoh tabel translasi

3.6.3. *Decoding* Oleh Moses

Moses adalah *decoder* untuk menterjemahkan bahasa sumber ke bahasa target berdasarkan model bahasa dan model translasi yang telah dilakukan sebelumnya. Proses *decoding* oleh *decoder* Moses dapat dilihat pada Gambar 3.11.

Gambar 3.11 Proses *decoding* oleh Moses pada terjemahan bahasa Indonesia ke bahasa Sunda [Modifikasi: Hadi, 2014]

Gambar 3.11 menunjukkan *decoder* moses akan menerjemahkan kalimat masukan berupa kalimat sumber (Bahasa Indonesia) dari korpus paralel. Selanjutnya, kalimat masukan tersebut akan diproses oleh *decoder* moses dan akan menghasilkan kalimat keluaran berupa kalimat hasil terjemahan ke dalam bahasa target. Proses *decoding* dilakukan sebagai berikut. (Ramdhani, 2007).

1. Cari terjemahan frasa bahasa target untuk setiap frasa dari bahasa sumber pada tabel translasi frasa.
2. Cari probabilitas maksimum untuk frasa bahasa target pada tabel model translasi.

awit ({ 1 }) yakin ({ 2 }) inu ({ 3 }) belum ({ 4 }) tidur ({ 5 }) di ({ 6 }) kamar ({ 7 }) samping ({ 8 }) dia ({ 9 })
awit ({ 1 }) yakin ({ 2 }) inu ({ 3 }) can ({ 4 }) sare ({ 5 }) di ({ 6 }) kamar ({ 7 }) gigireun ({ 8 }) manehna ({ 9 })

Proses *decoding* dijelaskan sebagai berikut.

f1	f2	f3	f4	f5	f6	f7	f8	f9

a. mencari frasa dimulai dari kata f1,
b. dapat diasumsikan ada 3 frasa, yaitu F1, F1-F2, F1-F2-F3,
c. dalam tabel model translasi, “awit” diterjemahkan sebagai frasa “awit”, maka F1 diterjemahkan dengan frasa E1,
d. menandai frasa bahasa sumber yang sudah diterjemahkan,

F1	f2	f3	f4	f5	f6	f7	f8	f9

E1

e. dilanjutkan mencari frasa setelah frasa f2,
f. pada tabel model translasi, kata “yakin” diterjemahkan sebagai frasa “yakin” dan memiliki probabilitas maksimum, maka frasa F2 diterjemahkan sebagai E2 yaitu frasa “yakin” ,

F1	F2	f3	f4	f5	f6	f7	f8	f9

E1	E2

g. pada tabel model bahasa frasa “inu” diterjemahkan sebagai frasa “inu”,
h. maka frasa E3 adalah “inu”,

F1	F2	F3	f4	f5	f6	f7	f8	f9

E1	E2	E3

i. dilanjutkan mencari frasa F4,

j. frasa F4 “belum” ditranslasikan sebagai frasa “can”, maka frasa E4 sebagai “can”,

F1	F2	F3	F4	f5	f6	f7	f8	f9

E1	E2	E3	E4

k. proses decoding berlanjut dengan cara yang sama hingga ditemukan frasa F9.

Keterangan :

1. Simbol f1, f2, f3...fn merupakan setiap kata untuk kalimat bahasa sumber
2. Simbol Fn merupakan frasa dari kalimat bahasa sumber
3. Simbol En merupakan frasa dari kalimat bahasa target
4. Blok berwarna biru merupakan frasa bahasa sumber yang sudah diterjemahkan ke frasa bahasa target

3.7 Pengujian Hasil Terjemahan Mesin Translasi Oleh BLEU

Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi dan sebagai nilai awal untuk dibandingkan dengan nilai akurasi setelah dilakukan penambahan kuantitas korpus monolingual (Bahasa Sunda). Pengujian dilakukan cara pengujian secara otomatis menggunakan BLEU dengan metode K-Fold Cross-Validation. Pengujian oleh ahli bahasa pada tahap ini belum dapat dilakukan karena pada penilaian ini belum memiliki pembanding sehingga apabila ahli bahasa dilibatkan, maka akan terjadi kebingungan dalam menilai hasil translasi.

Proses evaluasi otomatis pada penelitian menggunakan sistem evaluasi otomatis BLEU (*Bilingual Evaluation Understudy*). BLEU mengukur *modified* n-

gram *precision score* antara hasil terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang disebut *brevity penalty*. Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score*. Rumus untuk menghitung nilai BLEU, *brevity penalty* dan *modified precision score* dapat dilihat pada Persamaan 2.5, Persamaan 2.6, dan Persamaan 2.7 pada bab sebelumnya.

3.8 Penambahan Kuantitas Korpus Monolingual (Bahasa Sunda)

Setelah dilakukan pengujian untuk mendapatkan skor awal, maka dilakukan inti dari penelitian ini, yaitu penambahan kuantitas korpus monolingual.

Tabel 3.3 Data Korpus Monolingual Bahasa Sunda

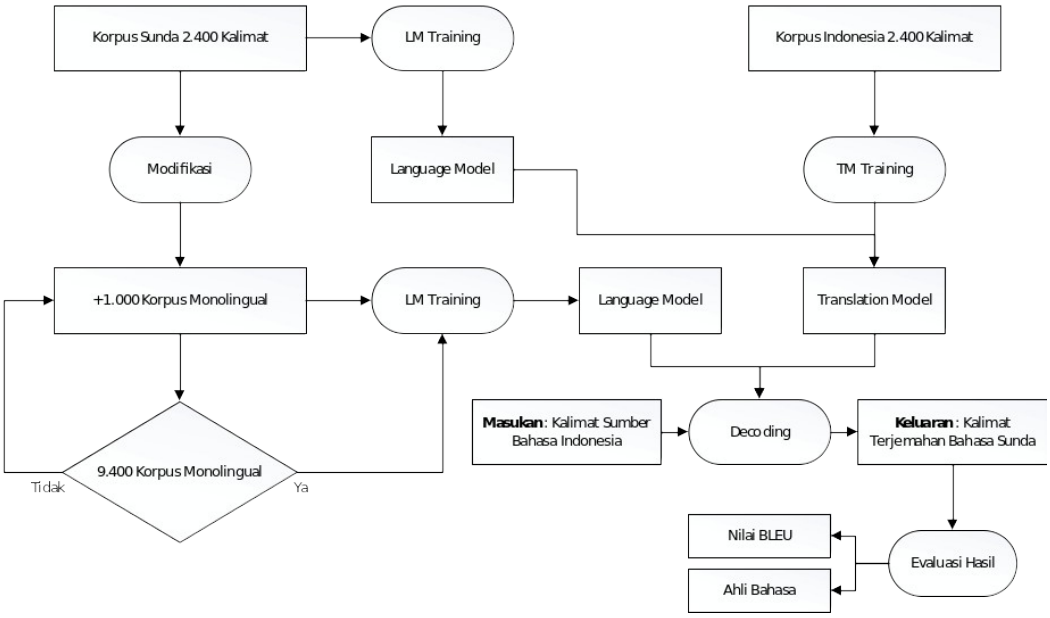
Kopus	Jumlah Korpus
Korpus Paralel (CP)	2400 Kalimat (Paralel)
CM1	3400 Kalimat (Paralel + 1000 Korpus Monolingual)
CM2	4400 Kalimat (Paralel + 2000 Korpus Monolingual)
CM3	5400 Kalimat (Paralel + 3000 Korpus Monolingual)
CM4	6400 Kalimat (Paralel + 4000 Korpus Monolingual)
CM5	7400 Kalimat (Paralel + 5000 Korpus Monolingual)
CM6	8400 Kalimat (Paralel + 6000 Korpus Monolingual)
CM7	9400 Kalimat (Paralel + 7000 Korpus Monolingual)

Tabel 3.3 merupakan data jumlah korpus monolingual yang akan ditambahkan ke dalam mesin penerjemah bahasa Indonesia ke bahasa Sunda.

Sebelum melakukan implementasi mesin penerjemah statistik, terlebih dahulu korpus teks paralel yang telah di buat dilakukan proses *cleaning*, tokenisasi dan *case folding*.

3.9 Implementasi Mesin Penerjemah Statistik Bahasa Indonesia - Sunda Setelah Penambahan Kuantitas Korpus Monolingual

Pada arsitektur sistem penambahan kuantitas korpus monolingual, akan dilakukan *preprocessing* pada korpus monolingual bahasa Sunda sebagai bahasa target. Setelah korpus monolingual dilakukan proses *cleaning*, tokenisasi dan *case folding* maka selanjutnya korpus monolingual tersebut dilakukan proses LM *Training* dan menghasilkan model bahasa.



Gambar 3.12 Diagram sistem penambahan kuantitas korpus monolingual

Gambar 3.12 merupakan sistem mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda dengan menambah kuantitas korpus monolingual. Korpus monolingual yang awalnya 2400 kalimat ditambahkan 1000 kalimat yang kemudian dilakukan proses LM Training dan menghasilkan pemodelan bahasa. Untuk proses LM Training bisa di lihat di pembahasan sub bab 3.6.1.

Adapun langkah-langkah dalam melakukan pemodelan bahasa, pemodelan translasi, decoding, dan proses evaluasi hasil terjemahan sama seperti pembahasan sub bab 3.6.

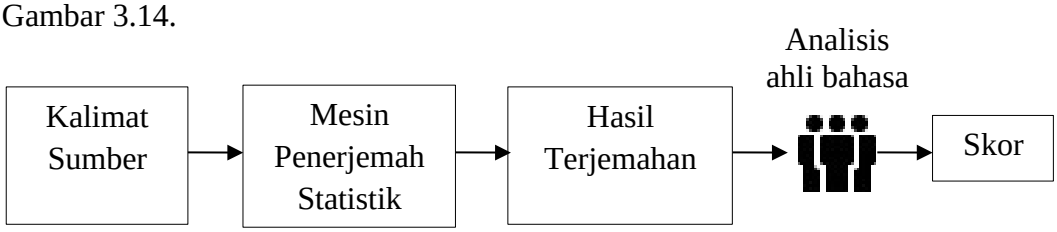
3.10 Pengujian Ulang Hasil Terjemahan Mesin Translasi Setelah Penambahan Kuantitas Korpus Monolingual

Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi setelah dilakukan penambahan kuantitas korpus monolingual. Pengujian dilakukan dengan cara pengujian secara otomatis menggunakan BLEU dengan menggunakan metode *K-Fold Cross-Validation* sesuai dengan Tabel 3.1.

Proses evaluasi otomatis pada penelitian menggunakan sistem evaluasi otomatis BLEU (*Bilingual Evaluation Understudy*). BLEU mengukur *modified* n-gram *precision score* antara hasil terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang disebut *brevity penalty*. Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score.*

3.11 Pengujian Ahli Bahasa

Pengujian Ahli bahasa dilakukan apabila telah terdapat peningkatan pada saat pengujian ulang hasil terjemahan mesin translasi oleh BLEU dikarenakan mengingat keterbatasan waktu ahli bahasa dalam melakukan penilaian. Penilaian yang dilakukan oleh ahli bahasa adalah menilai beberapa kalimat yang dihasilkan oleh mesin penerjemah sebelum dan setelah mengalami penambahan korpus monolingual (Bahasa Sunda) serta memberikan kalimat referensi menurut ahli bahasa. Proses evaluasi yang dilakukan oleh ahli bahasa Sunda dapat dilihat pada Gambar 3.14.



Gambar 3.13 Proses evaluasi ahli bahasa [Hadi, 2014]

3.12 Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan untuk mengetahui persentase peningkatan mesin penerjemah statistik sebelum dan sesudah dilakuakan penambahan kuantitas korpus monolingual serta untuk mengetahui seberapa besar

korpus monolingual yang diperlukan untuk meningkatkan akurasi mesin penerjemah statistik.

Nilai akurasi yang dibandingkan dibagi menjadi dua bagian. Pertama, nilai akurasi yang didapatkan dari mesin penerjemah sebelum melewati tahap penambahan korpus monolingual dan setelahnya. Kedua, nilai akurasi yang didapatkan dari ahli bahasa sebelum melewati tahap penambahan korpus monolingual dan setelahnya.

3.13 Penarikan Kesimpulan

Penarikan kesimpulan dilakukan dengan cara melihat hasil analisis yang telah dilakukan. Apakah dengan menambah korpus monolingual mampu memberikan peningkatan akurasi terhadap skor pada hasil terjemahan pada pengujian otomatis maupun pengujian oleh ahli bahasa dan sebera besar pengaruh penambahan kuantitas korpus monolingual terhadap peningkatan akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Sunda.

BAB IV
HASIL DAN ANALISIS

4.1. Hasil Penelitian

Berikut adalah hasil dari penelitian yang telah dibuat pada bab sebelumnya. Hasil dari penelitian ini terdiri dari pengumpulan data, pembuatan korpus teks paralel, membangun mesin penerjemah statistik, dan implementasi mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.

4.1.1. Hasil Pengumpulan Data

Berdasarkan hasil pengumpulan data yang dilakukan, data yang diperoleh selanjutnya disesuaikan dengan kebutuhan penelitian sehingga menghasilkan analisis data. Adapun analisis yang diperoleh yakni.

a. Hasil dari wawancara didapatkan ahli bahasa yaitu Bella Yuda. Penulis memilih beliau karena merupakan keturunan asli Sunda, dan memahami mengenai bahasa Sunda, baik penulisan, pengucapan dan arti dalam bahasa Indonesia. Beliau juga bersedia dilibatkan dalam penelitian ini.
b. Hasil dari pencarian dokumen berupa buku berbahasa Sunda dari buku Sanggeus Halimun Peuray yang akan diterjemahkan dalam bahasa Indonesia. Adapun judul-judul bukunya adalah sebagai berikut.
    1. Sanggeus Halimun Peuray, Karya Aam Amilia, Penerbit SundaBlog
    2. Arulin di Pilemburan, Karya Eusina, Penerbit SundaBlog
    3. Dalem Boncel, Karya Ki Umbara, Penerbit SundaBlog
    4. Ngawayan Teu Direbaban, Karya Ensa Wiarna, Penerbit SundaBlog
    5. Carita Parahiyangan, Karya Titilar Karuhun, Penerbit Jajasan Kebudajaan Nusalarang

4.1.2. Pembuatan Korpus Teks Paralel

Buku Sanggeus Halimun Peuray yang berbahasa Sunda berupa bahasa sehari-hari yang digunakan dalam berbahasa Sunda akan diterjemahkan dalam bahasa Indonesia, selanjutnya dibuat menjadi korpus teks paralel sebanyak 3000 korpus, kemudian dilakukan penggabungan korpus-korpus menjadi satu file yang

disimpan dengan format .sd untuk korpus bahasa Sunda dan .id untuk korpus Bahasa Indonesia. Gambar 4.1 merupakan hasil dari korpus teks paralel yang telah dibuat.

Gambar 4.1 Korpus paralel bahasa Indonesia – bahasa Sunda

4.1.3. Membangun Mesin Penerjemah Statistik

Sebagai Langkah dalam membangun mesin penerjemah statistik, dilakukan proses instalasi perangkat lunak. Adapun perangkat lunak yang akan diinstal adalah sebagai berikut.

a. Mosesdecoder untuk membangun mesin penerjemah,
b. GIZA++ untuk pemodelan translasi,
c. SRILM untuk pemodelan bahasa.

Seluruh langkah-langkah instalasi perangkat lunak diatas dapat dilihat pada lampiran A.

4.1.4. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia - Sunda

Setelah membangun mesin penerjemah statistik, maka langkah selanjutnya adalah membangun mesin penerjemah statistik bahasa Indonesia - Bahasa Sunda. Sebelum melakukan implementasi mesin penerjemah statistik, terlebih dahulu korpus teks paralel yang telah di buat dilakukan proses *cleaning* dan tokenisasi.

```
1. ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/
   moses/M-A/corpus/corpus id sd
   ~/moses/M-A/corpus/corpus.clean 1 80

2.  perl ~/moses/M-A/clean.plx
   ~/moses/M-A/corpus/corpus.clean.id
   ~/moses/M-A/corpus/corpus.clean1.id

3. perl ~/moses/M-A/clean.plx
   ~/moses/M-A/corpus/corpus.clean.sd
   ~/moses/M-B/corpus/corpus.clean1.sd

 4. ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   ~/moses/M-A/corpus/corpus.clean1.id >
   ~/moses/M-A/corpus/corpus.lowercased.id
```

Gambar 4.2 Kode program c*leaning*, tokenisasi, dan *case folding* untuk M-A

```
1. ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/
   moses/M-B/corpus/corpus id sd
   ~/moses/M-B/corpus/corpus.clean 1 80

2.  perl ~/moses/M-B/clean.plx
   ~/moses/M-B/corpus/corpus.clean.id
   ~/moses/M-B/corpus/corpus.clean1.id

3. perl ~/moses/M-B/clean.plx
   ~/moses/M-B/corpus/corpus.clean.sd
   ~/moses/M-B/corpus/corpus.clean1.sd

4. ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   ~/moses/M-B/corpus/corpus.clean1.id >
   ~/moses/M-B/corpus/corpus.lowercased.id
```

Gambar 4.3 Kode program c*leaning*, tokenisasi, dan *case folding* untuk M-B

```
1. ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/
   moses/M-C/corpus/corpus id sd
   ~/moses/M-C/corpus/corpus.clean 1 80

2.  perl ~/moses/M-C/clean.plx
   ~/moses/M-C/corpus/corpus.clean.id
   ~/moses/M-C/corpus/corpus.clean1.id

3. perl ~/moses/M-C/clean.plx
   ~/moses/M-C/corpus/corpus.clean.sd
   ~/moses/M-C/corpus/corpus.clean1.sd

4. ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   ~/moses/M-C/corpus/corpus.clean1.id >
   ~/moses/M-C/corpus/corpus.lowercased.id
```

Gambar 4.4 Kode program c*leaning*, tokenisasi, dan *case folding* untuk M-C

```
1. ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/
   moses/M-D/corpus/corpus id sd
   ~/moses/M-D/corpus/corpus.clean 1 80

2.  perl ~/moses/M-D/clean.plx
   ~/moses/M-D/corpus/corpus.clean.id
   ~/moses/M-D/corpus/corpus.clean1.id

3. perl ~/moses/M-D/clean.plx
   ~/moses/M-D/corpus/corpus.clean.sd
   ~/moses/M-D/corpus/corpus.clean1.sd

4. ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   ~/moses/M-D/corpus/corpus.clean1.id >
   ~/moses/M-D/corpus/corpus.lowercased.id
```

Gambar 4.5 Kode program c*leaning*, tokenisasi, dan *case folding* untuk M-D

```
1. ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/
   moses/M-E/corpus/corpus id sd
   ~/moses/M-E/corpus/corpus.clean 1 80

2.  perl ~/moses/M-E/clean.plx
   ~/moses/M-E/corpus/corpus.clean.id
   ~/moses/M-E/corpus/corpus.clean1.id

3. perl ~/moses/M-E/clean.plx
   ~/moses/M-E/corpus/corpus.clean.sd
   ~/moses/M-E/corpus/corpus.clean1.sd

4. ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   ~/moses/M-E/corpus/corpus.clean1.id >
   ~/moses/M-E/corpus/corpus.lowercased.id
```

Gambar 4.6 Kode program c*leaning*, tokenisasi, dan *case folding* untuk M-E

Kode program pada Gambar 4.2, 4.3, 4.4, 4.5 dan 4.6 merupakan kode program *cleaning, tokenisasi* dan *case folding* untuk mesin M-A, M-B, M-C, M-D, M-E.

Baris pertama digunakan untuk memenggal kalimat yang memiliki kata lebih dari 80, baris kedua sampai kelima merupakan perintah untuk menghapus tanda baca titik di akhir kalimat dan menyisipkan spasi atara kata dan tanda baca. Baris keenam dan ketujuh digunakan untuk mengubah huruf kapital yang terdapat dalam korpus menjadi huruf kecil (*case folding)*. Berikut merupakan tabel perbandingan sebelum dan setelah dilakukan prosses *cleaning* dan tokenisasi.

Tabel 4.1 Hasil Proses *Cleaning* dan *Tokenisasi*

Sebelum *Cleaning* dan Tokenisasi	**Setelah *Cleaning* dan Tokenisasi**
Tapi itu kapan, takut Mila tidak percaya.	Tapi itu kapan , takut Mila tidak percaya

Pada Tabel 4.1 sebelum dilakukan proses *cleaning* dan tokenisasi tanda baca menyatu dengan kata lain, sedangkan setelah *cleaning* dan tokenisasi tanda baca terpisah antara kata dan tanda baca dan tanda baca titik di ujung kalimat menjadi hilang.

Tabel 4.2 Hasil Proses *Case Folding*

Sebelum *Case Folding*	**Setelah *Case Folding***
Tapi itu kapan , takut Mila tidak percaya	tapi itu kapan , takut mila tidak percaya

Pada Tabel 4.2 sebelum dilakukan proses *case folding* masih terdapat huruf kapital, sedangkan setelah dilakukan proses *case folding* tidak terdapat lagi huruf kapitalnya.

4.1.4.1. Implementasi SRILM Untuk Pemodelan Bahasa

Model bahasa digunakan sebagai sumber pengetahuan berbasis teks dengan nilai-nilai probabilistik. Penelitian ini menggunakan n-gram sebagai *language model.* Model bahasa dibangun dengan *tools* SRILM. Adapun cara instalasi SRILM terdapat pada lampiran A. Gambar 4.7 merupakan perintah untuk membangun model bahasa M-A, M-B, M-C, M-D dan M-E.

```
1. ~/moses/srilm/bin/i686/ngram-count -order 3 -interpolate -
   kndiscount -unk -text
   ~/moses/M-A/corpus/corpus.lowercased.sd -lm
   ~/moses/M-A/lm/sdA.lm
2. ~/moses/srilm/bin/i686/ngram-count -order 3 -interpolate -
   kndiscount -unk -text
   ~/moses/M-B/corpus/corpus.lowercased.sd -lm
   ~/moses/M-B/lm/sdB.lm
3. ~/moses/srilm/bin/i686/ngram-count -order 3 -interpolate -
   kndiscount -unk -text
   ~/moses/M-C/corpus/corpus.lowercased.sd -lm
   ~/moses/M-C/lm/sdC.lm
4. ~/moses/srilm/bin/i686/ngram-count -order 3 -interpolate -
   kndiscount -unk -text
   ~/moses/M-D/corpus/corpus.lowercased.sd -lm
   ~/moses/M-D/lm/sdD.lm
```

Gambar 4.7 Kode program membangun model bahasa

Setelah melakukan perintah untuk membangun model bahasa, maka akan dihasilkan *output* dengan format file *.lm. Gambar 4.8 merupakan tabel model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.

```
\data\
ngram 1=3781
ngram 2=13681
ngram 3=976

\1-grams:
-3.07971    abdi  -0.1877418
-4.030138   abis  -0.08501079
-4.030138   abong -0.08501079
-4.030138   abot  -0.08501078
-----------------------------------
--
\2-grams
-2.05707    abdi gaduh  -0.1528544
-0.9287719  abdi mah    -0.1014395
-0.4887881  aing teh    -0.03498349
-0.8097574  albeum dibuka     -
0.1528544
-1.220661   angin peuting     -
0.09498782
-----------------------------------
--
\3-grams
-0.59763    asa aya nu
-0.417091   ira aya nu
-0.5078039  salawasna aya dina
-0.466377   lantaran ayana mila
-1.017561   nu ayeuna keur
```

Gambar 4.8 Tabel model bahasa dengan bahasa Sunda sebagai bahasa target

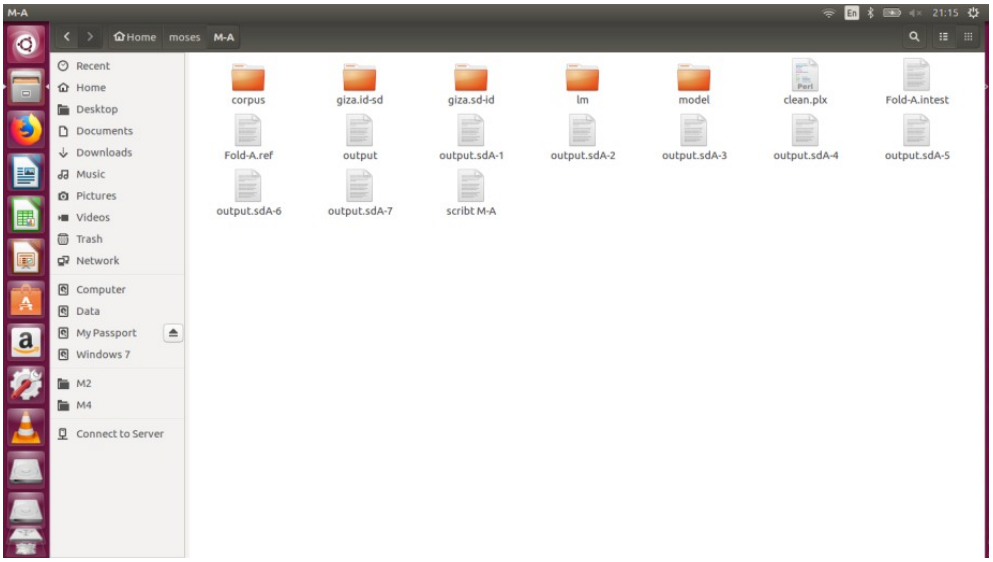
4.1.4.2. Implementasi Giza++ Untuk Pemodelan Translasi

Model translasi digunakan untuk memasangkan teks *input* dalam bahasa sumber dengan teks *output* dalam bahasa target. Model translasi dibangun dengan tools Giza++. Gambar 4.9 merupakan perintah untuk membangun model translasi.

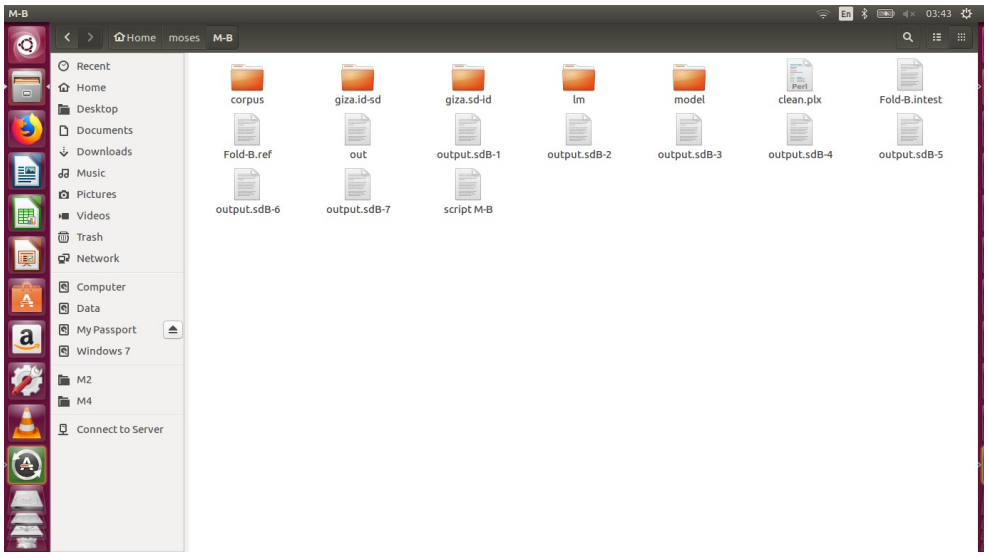
```
1.  ~/moses/mosesdecoder/scripts/training/train-model.perl -
    root-dir . --corpus ~/moses/M-A/corpus/corpus.lowercased --
    f id --e sd --lm 0:3:/home/acer/moses/M-A/lm/sdA.lm:0
2.  ~/moses/mosesdecoder/scripts/training/train-model.perl -
    root-dir . --corpus ~/moses/M-B/corpus/corpus.lowercased --
    f id --e sd --lm 0:3:/home/acer/moses/M-B/lm/sdB.lm:0
3.  ~/moses/mosesdecoder/scripts/training/train-model.perl -
    root-dir . --corpus ~/moses/M-C/corpus/corpus.lowercased --
    f id --e sd --lm 0:3:/home/acer/moses/M-C/lm/sdC.lm:0
4.  ~/moses/mosesdecoder/scripts/training/train-model.perl -
    root-dir . --corpus ~/moses/M-D/corpus/corpus.lowercased --
5.  ~/moses/mosesdecoder/scripts/training/train-model.perl -
    root-dir . --corpus ~/moses/M-E/corpus/corpus.lowercased --
    f id --e sd --lm 0:3:/home/acer/moses/M-E/lm/sdE.lm:0
```

Gambar 4.9 Kode program membangun model translasi

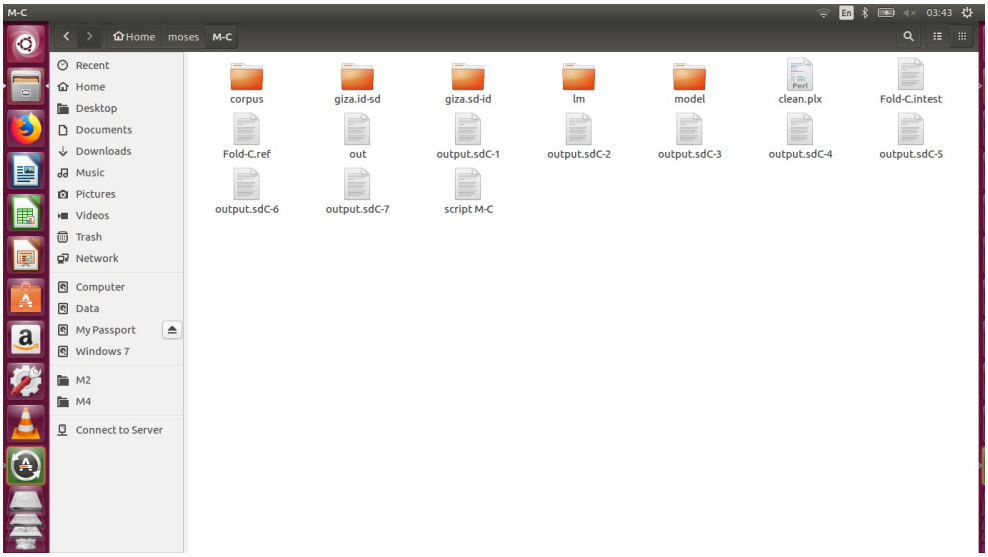
Proses pemodelan translasi oleh Giza++ menghasilkan dokumen *vocabulary corpus* dan *word alignment*. Dokumen-dokumen tersebut terdapat dalam folder “train” yang didalamnya terdapat 4 *file* yaitu “*corpus*, giza.id-sd, giza.sd-id dan model”.



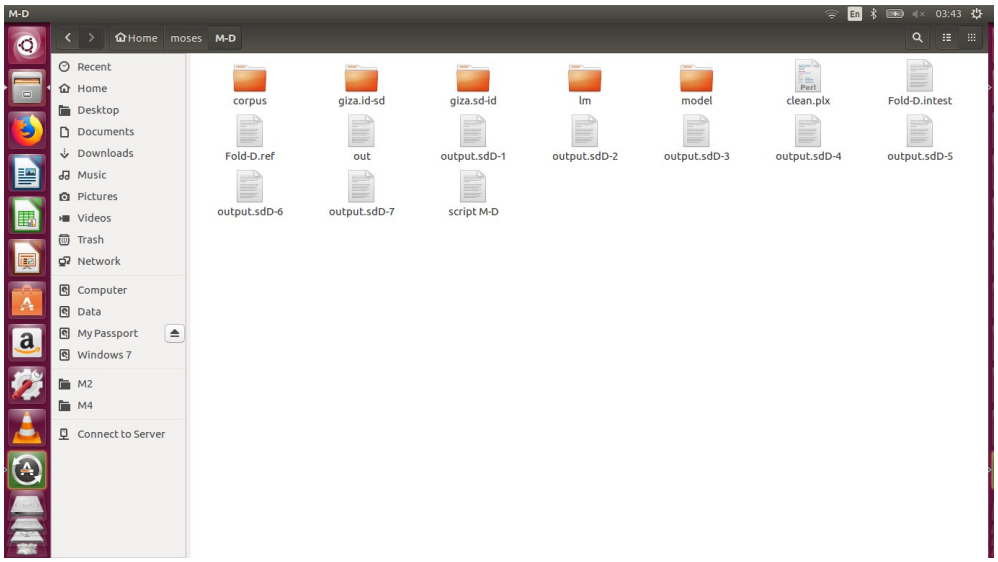
Gambar 4.10 *File* model translasi M-A



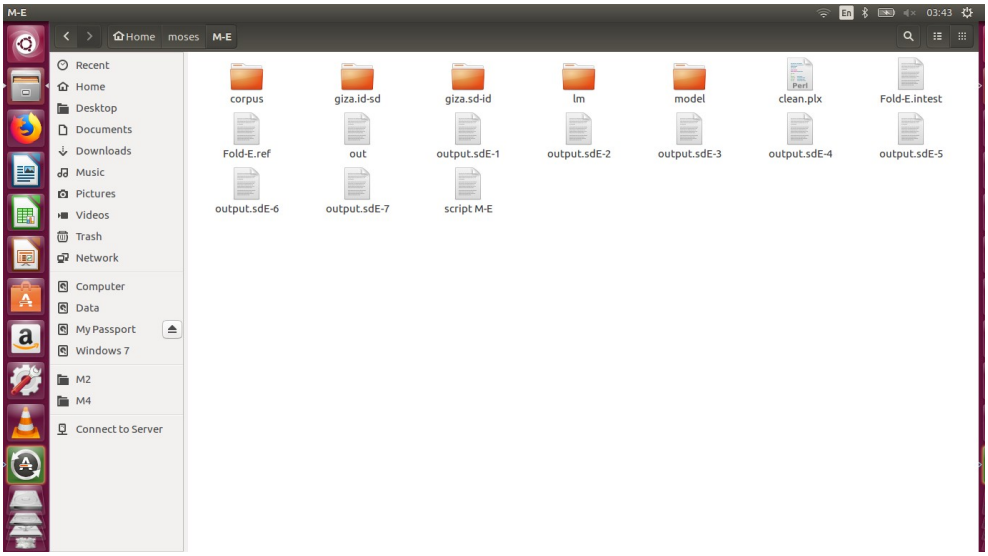
Gambar 4.11 *File* model translasi M-B



Gambar 4.12 *File* model translasi M-C



Gambar 4.13 *File* model translasi M-D



Gambar 4.14 *File* model translasi M-E

Gambar 4.10, 4.11, 4.12, 4.13, 4.14 merupakan tampilan file model translasi M-A, M-B, M-C, M-D dan M-E.

4.1.4.2.1. *Vocabulary Corpus*

Dokumen *vocabulary corpus* berisi setiap kata pada masing-masing korpus dimana setiap kata-kata tersebut memiliki *uniq id* yang diikuti oleh kata (token) dan frekuensi kemunculannya.Gambar 4.15 dan 4.16 merupakan *Vocabulary Corpus* yang terdapat pada mesin penerjemah bahasa Indonesia ke bahasa Sunda.

```
1     UNK   0
2     tuh   654
3     yang  520
4     tidak 509
5     inu   493
6     awit  306
7     ke    271
8     di    227
9     sama  216
10    mau   211
```

Gambar 4.15 Dokumen *vocabulary corpus* bahasa Indonesia

```
1     UNK    0
2     teh   717
3     inu   493
4     nu    491
5     ka    356
6     awit  306
7     mah   298
8     ku    254
9     mila  167
10    cek   160
```

Gambar 4.16 Dokumen *vocabulary corpus* bahasa Sunda

Angka 1 sampai 10 pada dokumen *vocabulary corpus* merupakan *uniq id* untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan. *Vocabulary corpus* yang dihasilkan mesin penerjemah bahasa Indonesia ke bahasa Sunda terdiri dari 2742 token untuk korpus bahasa Indonesia dan 3617 token untuk bahasa Sunda.

4.1.4.2.2. *Word Alignment*

Proses *word alignment* akan menghasilkan dokumen korpus *alignment* yaitu kalimat bahasa target dipetakan pada kalimat bahasa sumber. Gambar 4.17 merupakan dokumen *Word Alignment* yang terdapat pada mesin penerjemah Bahasa Indonesia ke Sunda.

```
# Sentence pair (55) source length 9 target length 9 alignment
score : 2.99844e-05


teh lila imut , tuluy nyekel leungeun inu pageuh
```

Gambar 4.17 Dokumen *alignment* bahasa Indonesia ke bahasa Sunda

Dokumen *alignment* Bahasa Indonesia ke Sunda terdapat tiga baris kalimat. Baris pertama berisi letak kalimat target (55) dalam korpus, panjang kalimat sumber (9), panjang kalimat target (9) dan skor *alignment* 2.99844e-05. Baris kedua merupakan bahasa sumber dan baris ketiga merupakan *alignment*

kalimat bahasa target terhadap kalimat bahasa sumber. Kata “lila” ({ 2 }) memiliki makna bahwa kata “lila” pada kalimat bahasa target, di-*align* ke kata keenam pada kalimat bahasa sumber yaitu “lama”.

4.1.5. Pengujian Hasil Terjemahan Mesin Translasi Oleh BLEU Sebelum Ditambahkan *Korpus Monolingual*

Pengujian hasil translasi dilakukan dengan cara pengujian otomatis dari mesin penerjemah. Pengujian otomatis dari mesin penerjemah menghasilkan keluaran berupa nilai akurasi yang dihasilkan oleh BLEU *(Billingual Evaluation Understudy)*. Hasil pengujian ini nantinya akan menjadi parameter untuk membandingkannya dengan hasil pengujian setelah dilakukan proses penambahan corpus monolingual.

Langkah pada pengujian otomatis, korpus yang akan diuji terlebih dahulu melalui langkah translasi otomatis yang akan memberikan *output* berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin. Pengujian mesin menggunakan metode *K-Fold Cross-Validation* yang sudah dijelaskan pada bab sebelumnya. Gambar 4.18 merupakan perintah membuat *output* dalam bahasa target.

```
1. ~/moses/mosesdecoder/moses-cmd/src/moses -f
   model/moses.ini < Fold-A.intest > output.sdA
2. ~/moses/mosesdecoder/moses-cmd/src/moses -f
   model/moses.ini < Fold-B.intest > output.sdB
3. ~/moses/mosesdecoder/moses-cmd/src/moses -f
   model/moses.ini < Fold-C.intest > output.sdC
4. ~/moses/mosesdecoder/moses-cmd/src/moses -f
   model/moses.ini < Fold-D.intest > output.sdD
5. ~/moses/mosesdecoder/moses-cmd/src/moses -f
   model/moses.ini < Fold-E.intest > output.sdE
```

Gambar 4.18 Kode program membuat *output*

Setelah membuat *output* berupa hasil translasi otomatis dari mesin penerjemah, langkah selanjutnya adalah mendapatkan nilai BLEU dari *output* dengan cara membandingkan *output* tersebut dengan korpus bahasa target yang telah dibuat sebelumnya. Gambar 4.11 merupakan perintah untuk menghitung skor.

```
1. ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl Fold-
   A.ref < output.sdA
2. ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl Fold-
   B.ref < output.sdB
3. ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl Fold-
   C.ref < output.sdC
4. ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl Fold-
   D.ref < output.sdD
5. ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl Fold-
   E.ref < output.sdE
```

Gambar 4.19 Kode program menghitung skor *output*

Perintah yang terdapat pada Gambar 4.19 akan menghasilkan *output* berupa skor BLEU dari mesin M-A, M-B, M-C, M-D dan M-E.

```
acer@acer-Aspire-4755::~/moses/M-A$
~/mosesdecoder/scripts/generic/multi-bleu.perl Fold-A.ref <
output.sdA
BLEU = 25.40, 55.3/33.1/20.2/12.9 (BP=0.966, ratio=0.966,
hyp_len=4823, ref_len=5001)
acer@acer-Aspire-4755::~/moses/M-B$
~/mosesdecoder/scripts/generic/multi-bleu.perl Fold-B.ref <
output.sdB
BLEU = 27.02, 56.7/35.1/21.3/13.2 (BP=0.986, ratio=0.986,
hyp_len=4873, ref_len=4941)
acer@acer-Aspire-4755::~/moses/M-C$
~/mosesdecoder/scripts/generic/multi-bleu.perl Fold-C.ref <
output.sdC
BLEU = 30.34, 57.8/37.8/24.7/15.8 (BP=0.998, ratio=0.998,
hyp_len=4400, ref_len=4409)
acer@acer-Aspire-4755::~/moses/M-D$
~/mosesdecoder/scripts/generic/multi-bleu.perl Fold-D.ref <
output.sdD
BLEU = 27.93, 56.9/35.3/22.4/14.3 (BP=0.986, ratio=0.986,
hyp_len=4611, ref_len=4674)
acer@acer-Aspire-4755::~/moses/M-E$
~/mosesdecoder/scripts/generic/multi-bleu.perl Fold-E.ref <
output.sdE
BLEU = 30.57, 57.9/37.5/24.6/16.7 (BP=0.996, ratio=0.996,
hyp_len=4623, ref_len=4643)
```

Gambar 4.20 Tampilan skor BLEU

Berdasarkan Gambar 4.20 diperoleh nilai BLEU dari M-A adalah sebesar 25.40%, M-B sebesar 27.02%, M-C sebesar 30.34%, M-D sebesar 27.93% dan M-E sebesar 30.57%.

Tabel 4.3 Nilai BLUE Sebelum Ditambahkan Korpus Monolingual

Mesin	Korpus Paralel	Korpus Uji	Nilai BLEU
M-A	Fold B+C+D+E	Fold A	25.40%
M-B	Fold A+C+D+E	Fold B	27.02%
M-C	Fold A+B+D+E	Fold C	30.34%
M-D	Fold A+B+C+E	Fold D	27.93%
M-E	Fold A+B+C+D	Fold E	30.57%

Berdasarkan hasil pengujian mesin penerjemah pada Tabel 4.3 diperoleh nilai skor BLEU pada mesin penerjemah bahasa Indonesia ke bahasa Sunda sebelum dilakukan penambahan kuantitas korpus monolingual adalah sebesar 28.25%.

4.1.6. Penambahan Kuantitas Korpus Monolingual Pada Bahasa Sunda

Setelah mendapatkan nilai awal dari korpus uji, maka langkah selanjutnya adalah melakukan proses penambahan kuantitas korpus monolingual pada bahasa Sunda. Proses penambahan kuantitas korpus bahasa Sunda seperti yang telah dijelasakan pada bab sebelumnya. Proses penambahan kuantitas korpus dilakukan penulis dengan menyiapkan 7 file korpus bahasa Sunda yang setiap masing-masing korpus berisi 1000 kalimat bahasa Sunda yang nantinya akan ditambahkan ke dalam korpus awal. Gambar 4.21 merupakan tampilan dari 7 file korpus monolingual.

Gambar 4.21 Tampilan File Korpus Monolingual

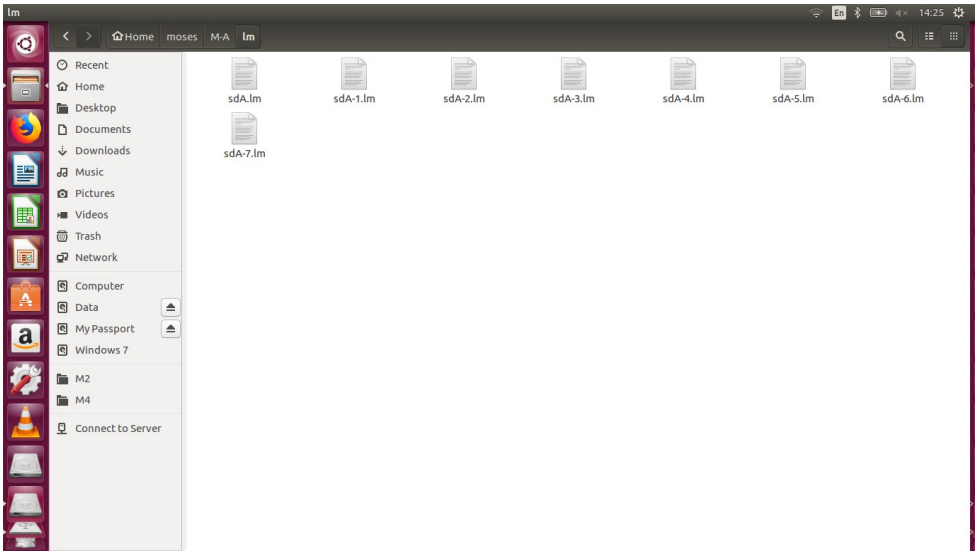
Gambar 4.21 adalah folder yang dibuat terpisah untuk mendapatkan file model bahasa dari bahasa target yaitu Sunda yang nantinya akan dipindahkan ke dalam mesin M-A, M-B, M-C, M-D dan M-E.

4.1.7. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia - Sunda Dengan Menambah Kuantitas Korpus Monolingual

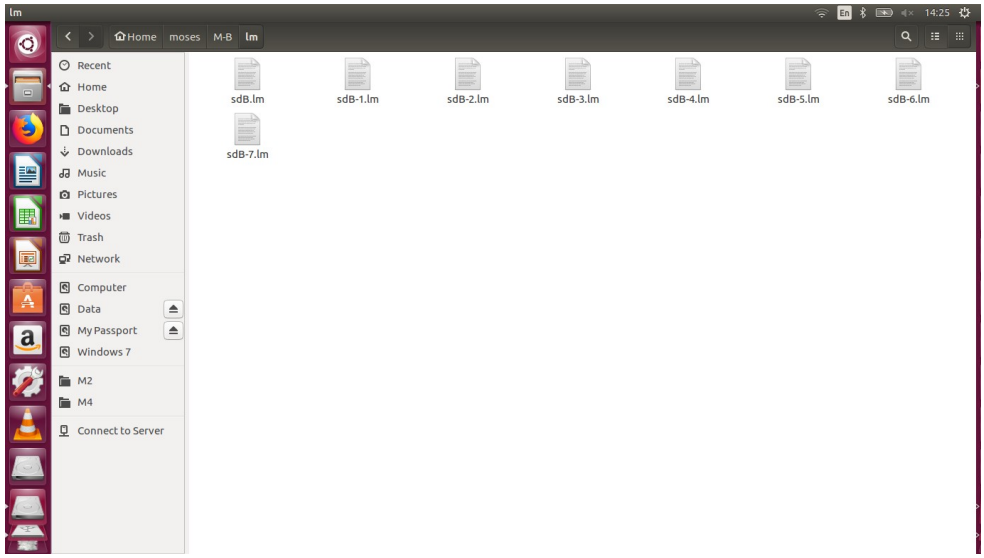
Setelah korpus monolingual telah dilakukan proses *cleaning*, tokenisasi dan *case folding*, maka selanjutnya penulis membuat model bahasa dari korpus monolingual bahasa Sunda tersebut. Kode program tersebut dapat diambil dari kode program pada Gambar 4.2, 4.3, 4.4, 4.5 dan 4.6, hanya diganti sesuai folder dimana tempat disimpan korpus monolingual bahasa Sunda.

Setelah melakukan perintah untuk membangun model bahasa, maka akan dihasilkan *output* dengan format file *.lm, model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.

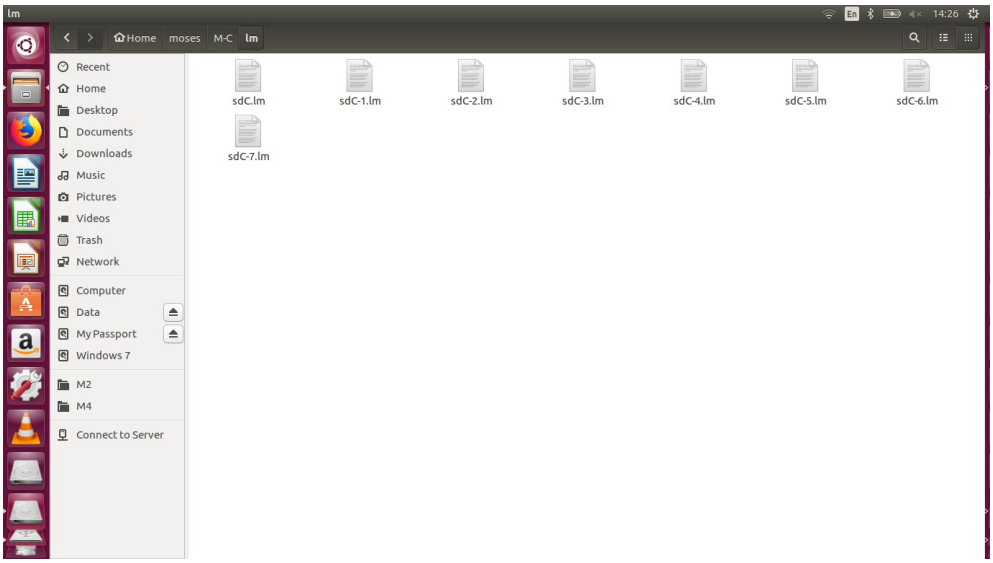
Penulis tidak lagi membangun mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda. Model bahasa yang telah didapat kemudian di pindahkan ke dalam folder lm pada mesin penerjemah M-A, M-B, M-C, M-D dan M-E.



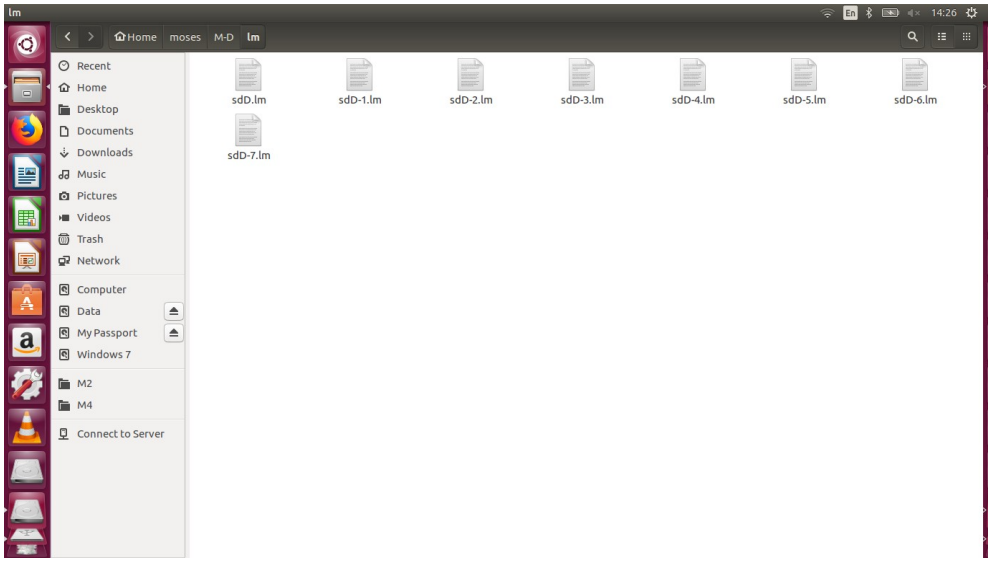
Gambar 4.22 Tampilan Model Bahasa M-A



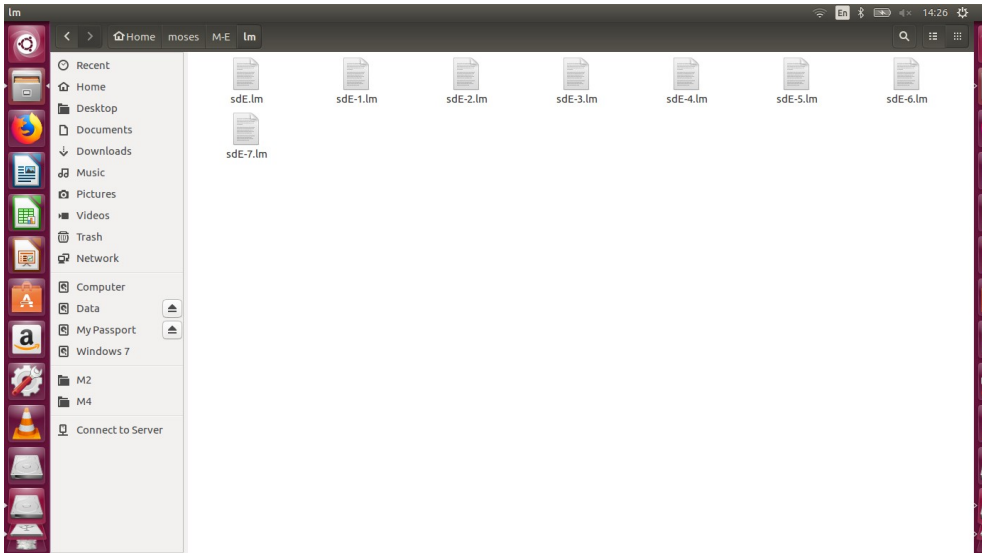
Gambar 4.23 Tampilan Model Bahasa M-B



Gambar 4.24 Tampilan Model Bahasa M-C



Gambar 4.25 Tampilan Model Bahasa M-D



Gambar 4.26 Tampilan Model Bahasa M-E

Dokumen yang telah dilakukan proses penambahan kuantitas korpus monolingual bahasa Sunda telah siap untuk dilakukan proses implementasi ke dalam mesin penerjemah bahasa Indonesia ke bahasa Sunda. Implementasi meliputi proses pemodelan bahasa oleh SRLIM untuk selanjutnya dilakukan pengujian ulang hasil terjemahan mesin translasi oleh BLEU.

4.1.8. Pengujian Ulang Hasil Terjemahan Mesin Translasi Oleh BLEU

Setelah mendapatkan model bahasa, langkah berikutnya adalah melakukan pengujian kembali hasil terjemahan mesin translasi bahasa Indonesia ke bahasa Sunda yang telah melewati proses penambahan kuantitas korpus monolingual. Langkah pengujian yang dilakukan sama halnya dengan langkah pengujian sebelumnya, yakni dengan cara melakukan pengujian otomatis dengan menggunakan metode *K-Fold Cross-Validation* yang akan memberikan *output* berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin dan pengujian oleh ahli bahasa.

Pengujian dilakukan dengan cara membandingkan nilai BLEU hasil terjemahan otomatis dari mesin penerjemah bahasa Indonesia - bahasa Sunda sebelum dan setelah melewati tahap penambahan kuantitas korpus monolingual. Apabila terdapat peningkatan dari nilai BLEU, maka perhitungan persentase peningkatan dihitung dengan persamaan 4.1

$$P = \frac{b-a}{a} x\,100\,\% \qquad (4.1)$$

Keterangan:

P = Peningkatan persentase

b = Nilai setelah

a = Nilai sebelum

Tabel 4.4 Tabel Perbandingan Nilai BLEU Setelah Penambahan Kuantitas Korpus Monolingual

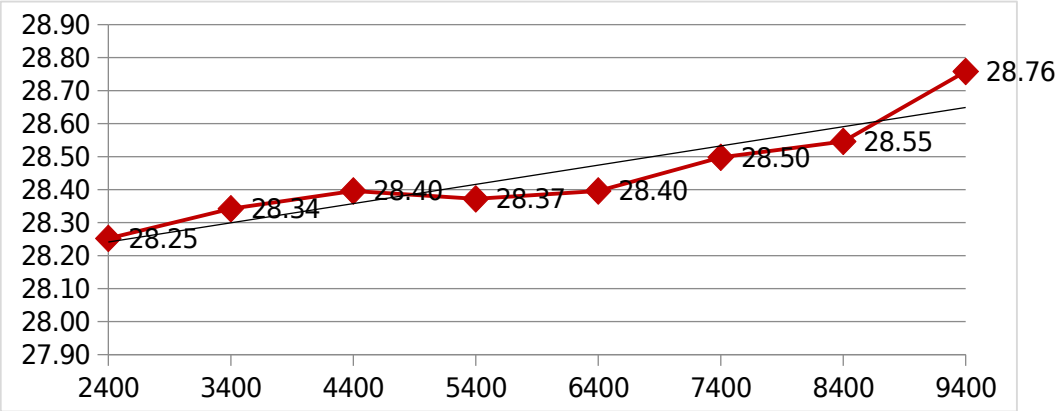
No	Mesin	CP	CM1	CM2	CM3	CM4	CM5	CM6	CM7
1	M-A	25.40	25.40	25.92	25.77	25.65	25.74	25.62	25.99
2	M-B	27.02	25.92	26.08	25.88	25.71	25.94	26.20	26.21
3	M-C	30.34	30.67	30.27	29.92	29.42	29.80	29.95	30.35
4	M-D	27.93	27.86	27.88	28.41	29.06	28.80	28.63	28.69
5	M-E	30.57	31.86	31.83	31.88	32.14	32.21	32.33	32.55
BLEU		28.2	28.3	28.4	28.3	28.4	28.5	28.5	28.7

	5	4	0	7	0	0	5	6

Keterangan :

- CP = Korpus Paralel
- CM = Korpus Monolingual

Tabel 4.4 merupakan Nilai BLEU Sebelum dan Sesudah dilakukan penambahan korpus monolingual bahasa Sunda. Gambar 4.27 merupakan tampilan grafik peningkatan nilai BLEU berdasarkan penambahan kuantitas korpus monolingual.



Gambar 4.27 Tampilan grafik peningkatan nilai BLEU berdasarkan penambahan kuantitas korpus monolingual

Tabel 4.5 Tabel Peningkatan Persentase Nilai BLEU

Korpus Monolingual	Nilai BLEU	Peningkatan Persentase
CM1 (3400 Kalimat)	28.34%	0.32%
CM2 (4400 Kalimat)	28.40%	0.51%
CM3 (5400 Kalimat)	28.37%	0.42%
CM4 (6400 Kalimat)	28.40%	0.51%
CM5 (7400 Kalimat)	28.50%	0.87%
CM6 (8400 Kalimat)	28.55%	1.04%
CM7 (9400 Kalimat)	28.76%	1.79%

Berdasarkan Tabel 4.4 dan 4.5 dapat dilihat bahwa terjadi peningkatan nilai BLEU sebelum dilakukan penambahan kuantitas korpus monolingual dan setelah dilakukan penambahan kuantitas korpus monolingual, nilai BLEU sebelum dilakukan penambahan kuantitas korpus monolingual pada korpus paralel 2.400 kalimat sebesar 28.25% dan setelah dilakukan penambahan kuantitas

korpus monolingual dengan korpus 3.400 sebesar 28.34% meningkat 0,32%, dengan korpus 4.400 sebesar 28.40% meningkat 0,51%, dengan korpus 5.400 sebesar 28.37% meningkat 0.42%, dengan korpus 6.400 sebesar 28.40% meningkat 0.51%, dengan korpus 7.400 sebesar 28.50% meningkat 0.87%, dengan korpus 8.400 sebesar 28.55% meningkat 1.04% dan dengan korpus 9.400 sebesar 28.76% meningkat 1.79% berdasarkan Persamaan 4.1 dilihat dari sebelum dan sesudah mengalami penambahan kuantitas korpus monolingual.

4.1.9. Pengujian Ahli Bahasa

Pengujian ahli bahasa dilakukan terhadap hasil terjemahan mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda. Pengujian dilakukan dengan mengambil kalimat yang mengalami perubahan pada hasil terjemahan otomatis yang terdapat pada korpus Paralel sebelum dan sesudah dilakukan penambahan kuantitas korpus monolingual sebanyak 20 kalimat. Tabel 4.6, 4.7, 4.8 merupakan kalimat sumber dan kalimat hasil terjemahan sebelum dan setelah dilakukan penambahan kuantitas korpus monolingual.

Tabel 4.6 Kalimat Sumber

No	Kalimat Sumber
1	handuk dimasukan ke koper
2	kang inu mana si kang inu , tidak salah , sambil sorak
3	tidak tahu sama awit mungkin ke pembantu
4	padahal cuma sahabat waktu zaman sekolah
5	tak terdengar suara lagi
6	kasihan kalau benar tuh
7	alhamdulillah , terima kasih bu uti , gertak hati awit
8	tapi apa artinya itu tatapan kalau untuk menyakiti hatinya
9	tidak tentu harus bagaimana menjawab , kata inu
10	tidak kalah mila punya yang lain , saya juga harus punya
11	sekali lagi lubang botol susu tuh disentuhkan pada bibir keran
12	enden , jangan terlalu lama , nanti turun darah , suara pengasuhnya
13	syukur lah , cuma jangan sampai ke sakit berdiri ih
14	saya tuh mau kuliah
15	saya tidak kuat menahan panas hati , kata mila
16	karena kang inu tidak mau terikat tante mila

17	kang inu awit harus hidup
18	sampai ke halaman , inu yakin awit akan mengejar sambil meneriakan namanya
19	kalau ke dokter itu kan harus punya uang
20	seperti inu waktu datang minggu lalu menyangka dirinya tuh perempuan nakal

Tabel 4.7 Kalimat Terjemahan Sebelum Penambahan Kuantitas Korpus Monolingual

No	Kalimat Terjemahan
1	anduk disesepkeun kana koper
2	kang inu mana si kang inu , teu salah , bari sorak
3	teuing ka awit meureun ka pembantu
4	padahal ngan sabot basa zaman sakola
5	teu kadenge gerung deui
6	deudeuh mun enya teh
7	alhamdulillah , nuhun ma uti , gerentes hate awit
8	tapi naon hartina eta tatapan ari keur ngaraheutan hatena
9	teu pupuguh kudu kumaha nembalan , cek inu
10	teu eleh mila teh boga nu sejen , abdi ge kudu ku
11	pisan deui liang botol susu teh disentuhkan kana biwir keran
12	enden , ulah teuing lila , ke turun darah , sora asuhanana teh
13	sukur we atuh , ngan ulah tepi ka nyeri nangtung euy
14	abdi teh hayang kuliah
15	abdi teukiat nahan panas hate , cek mila
16	enyaan kang inu teh hayang kabeulit ceu mila
17	kang inu awit mesti hidup
18	tepi ka pakarangan , inu awit yakin bakal mengejar bari ngagorowokkeun jenengan teh
19	mun ka dokter eta da kudu boga artos
20	asa inu basa datang minggu gek nyangka dirina teh awewe bangor

Tabel 4.8 Kalimat Terjemahan Sesudah Penambahan Kuantitas Korpus Monolingual

No	Kalimat Terjemahan Sesudah *Tagging*
1	anduk diasupkeun ka koper
2	kang inu mana si kang inu , moal salah , bari sorak
3	teu nyaho ka awit meureun ka pembantu
4	padahal ngan sobat basa zaman sakola
5	teu kadenge sora deui
6	karunya mun enya teh
7	alhamdulillah , hatur nuhun ma uti , gerentes hate awit
8	tapi naon hartina eta tatapan mun keur ngaraheutan hatena
9	teu puguh kudu kumaha nembalan , cek inu

10	teu eleh mila boga nu sejen , abdi ge kudu boga
11	sakali deui liang botol susu teh disentuhkan kana biwir keran
12	enden , ulah lila teuing , engke turun darah , sora pangasuhna
13	sukur we atuh , ngan ulah tepi ka nyeri nangtung euy
14	urang mah rek kuliah
15	aing teu kuat nahan panas hate , cek mila
16	lantaran kang inu teh hayang patali ceu mila
17	kang inu awit kudu hirup
18	nepi ka buruan , inu yakin awit bakal mengejar bari ngagorowokkeun ngaranna
19	ari ka dokter teh da kudu boga duit
20	kawas inu basa datang minggu tuluy nyangka dirina teh awewe bangor

Proses penilaian dilakukan oleh ahli bahasa Sunda. Ahli bahasa dipilih penulis karena merupakan keturunan asli Sunda dan paham mengenai bahasa Sunda. Ahli bahasa melakukan penilaian yang sesuai dengan pengetahuan dan pemahaman mereka masing-masing. Penilaian yang dilakukan ahli bahasa adalah dengan membandingkan hasil terjemahan sebelum dan sesudah dilakukan Penambahan Kuantitas Korpus Monolingual dengan terjemahan menurut ahli bahasa secara pribadi.

Tabel 4.9 Tabel Contoh Hasil Terjemahan Sebelum dan Sesudah Penambahan Kuantitas Korpus Monolingual

Kalimat Sumber	Kalimat Referensi	Terjemahan Sebelum Ditambahkan Kuantitas Korpus Monolingual	Terjemahan Setelah Ditambahkan Kuantitas Korpus Monolingual
handuk dimasukan ke koper	anduk diasupkeun ka koper	anduk **disesepkeun** kana koper	anduk diasupkeun ka koper
tak terdengar suara lagi	teu kadenge sora deui	teu kadenge **gerung** deui	teu kadenge sora deui
kasihan kalau benar tuh	karunya mun enya teh	**deudeuh** mun enya teh	karunya mun enya teh
tapi apa artinya itu tatapan kalau untuk menyakiti hatinya	tapi naon hartina eta teuteupan mun keur ngaraheutan hatena	tapi naon hartina eta **tatapan ari** keur ngaraheutan hatena	tapi naon hartina eta **tatapan** mun keur ngaraheutan hatena
sekali lagi lubang botol susu tuh disentuhkan pada bibir keran	sakali deui liang botol susu teh diantelkeun kana biwir keran	**pisan** deui liang botol susu teh **disentuhkan** kana biwir keran	sakali deui liang botol susu teh **disentuhkan** kana biwir keran

Penilaian terhadap hasil terjemahan oleh ahli bahasa terdapat pada bagian Lampiran E. Perhitungan akurasi dilakukan dengan Persamaan 4.2 :

$$P = \frac{C}{R} 100\% \quad (4.2)$$

P = Persentase akurasi

C = Jumlah kata yang diterjemahkan dengan tepat menurut penilaian dari ahli bahasa

R = Jumlah kata hasil terjemahan

Tabel 4.9 terdapat 5 buah kalimat yang berisi kalimat sumber, referensi, kalimat terjemahan sebelum dan sesudah penambahan kuantitas korpus monolingual. Kata yang ditebalkan merupakan kata yang salah dalam penerjemahannya. Kalimat pertama dengan kalimat sumber “handuk dimasukan ke koper” di terjemahkan manual menghasilkan kalimat referensi “anduk diasupkeun ka koper”, sebelum dilakukan penambahan kuantitas korpus monolingual menjadi “anduk disesepkeun kana koper” dan setelah penambahan kuantitas korpus monolingual menjadi “anduk diasupkeun ka koper”. Setelah di lakukan penambahan kuantitas korpus monolingual terjadi perbaikan kata yang sebelumnya disesepkeun menjadi diasupkan. Menurut ahli bahasa kata disesepkan tersebut salah karena terjemahan kata tersebut adalah “disisipkan” sedangkan dalam kalimat sumber adalah “dimasukan”. Kalimat kedua dengan kalimat sumber “tak terdengar suara lagi” di terjemahkan manual menghasilkan kalimat referensi “teu kadenge sora deui”, sebelum dilakukan penambahan kuantitas korpus monolingual menjadi “teu kadenge gerung deui” dan setelah penambahan kuantitas korpus monolingual menjadi “teu kadenge sora deui”. Setelah di lakukan penambahan kuantitas korpus monolingual terjadi perbaikan kata yang sebelumnya gerung menjadi sora. Menurut ahli bahasa kata gerung tersebut salah karena terjemahan kata gerung adalah “geram” sedangkan dalam kalimat sumber adalah “suara”. Kalimat ketiga dengan kalimat sumber “kasihan kalau benar tuh” di terjemahkan manual menghasilkan kalimat referensi “karunya mun enya teh”, sebelum dilakukan penambahan kuantitas korpus monolingual menjadi “deudeuh mun enya teh” dan setelah penambahan kuantitas korpus monolingual menjadi

“karunya mun enya teh”. Setelah di lakukan penambahan kuantitas korpus monolingual terjadi perbaikan kata yang sebelumnya deudeuh menjadi karunya. Menurut ahli bahasa kata deudeuh tersebut salah karena terjemahan kata deudeuh adalah “sayang” sedangkan dalam kalimat sumber adalah “kasihan”. Kalimat keempat dengan kalimat sumber “tapi apa artinya itu tatapan kalau untuk menyakiti hatinya” di terjemahkan manual menghasilkan kalimat referensi “tapi naon hartina eta teuteupan mun keur ngaraheutan hatena”, sebelum dilakukan penambahan kuantitas korpus monolingual menjadi “tapi naon hartina eta tatapan ari keur ngaraheutan hatena” dan setelah penambahan kuantitas korpus monolingual menjadi “tapi naon hartina eta tatapan mun keur ngaraheutan hatena”. Setelah di lakukan penambahan kuantitas korpus monolingual terjadi perbaikan kata yang sebelumnya ari menjadi mun. Menurut ahli bahasa kata ari sebenarnya benar karena terjemahan kata ari dan mun itu adalah “kalau”. Tetapi yang membuat salah adalah cara pemakaian kata tersebut dalam kalimat, ahli bahasa menganggap pemakaian kata tersebut tidak cocok dan lebih cocok menggunakan kata mun. Kalimat kelima dengan kalimat sumber “sekali lagi lubang botol susu tuh disentuhkan pada bibir keran” di terjemahkan manual menghasilkan kalimat referensi “sakali deui liang botol susu teh diantelkeun kana biwir keran”, sebelum dilakukan penambahan kuantitas korpus monolingual menjadi “pisan deui liang botol susu teh disentuhkan kana biwir keran” dan setelah penambahan kuantitas korpus monolingual menjadi “sakali deui liang botol susu teh disentuhkan kana biwir keran”. Setelah di lakukan penambahan kuantitas korpus monolingual terjadi perbaikan kata yang sebelumnya pisan menjadi sakali. Menurut ahli bahasa kata pisan sebenarnya benar karena terjemahan kata pisan dan sakali itu adalah “sekali”. Tetapi yang membuat salah adalah cara pemakaian kata tersebut dalam kalimat, ahli bahasa menganggap pemakaian kata pisan tidak cocok digunakan di awal kalimat dan lebih cocok menggunakan kata sakali untuk digunakan dalam kalimat tersebut.

pada kalimat pertama terdapat 4 kata, kalimat kedua terdapat 4 kata, kalimat ketiga terdapat 4 kata, kalimat keempat memiliki 9 kata dan kalimat kelima memiliki 10 kalimat. Total kata terjemahan dari 5 buah korpus tersebut adalah 31 kata R = 31. Jumlah kata terjemahan yang tepat sebelum penambahan

kuantitas korpus monolingual adalah 24 kata C = 24. Jumlah kata yang tepat sesudah penambahan kuantitas korpus monolingual adalah 29 kata C = 29. Persentase akurasi terjemahan sebelum penambahan kuantitas korpus monolingual sebesar 77.41%, persentase akurasi terjemahan setelah penambahan kuantitas korpus monolingual sebesar 93.55%. Penilaian berdasarkan persamaan 4.2. Dari hasil persentasi tersebut terjadi peningkatan akurasi penerjemahan sebesar 20.85% berdasarkan persamaan 4.1. Tabel 4.10 menunjukkan tabel penilaian ahli bahasa dengan jumlah 20 kalimat.

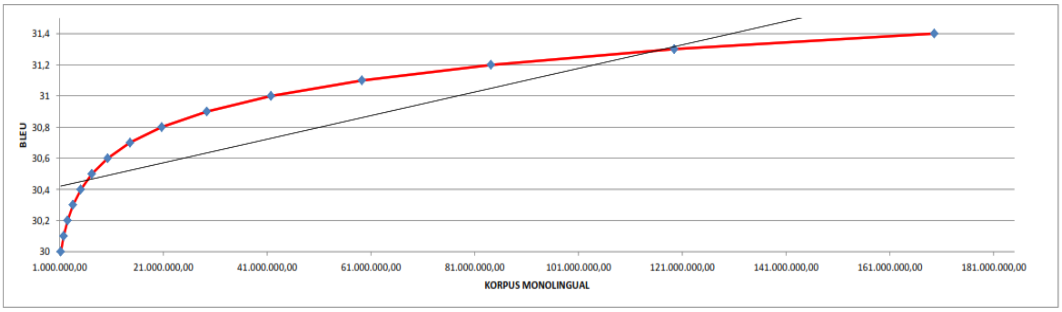
Tabel 4.10 Tabel Akurasi Ahli Bahasa

Kalimat Hasil Terjemahan	**Ahli Bahasa**	**C,R**	$\boldsymbol{P = \frac{C}{R} 100\%}$
Sebelum Penambahan Kuantitas Korpus Monolingual	Bella Yuda	C = 115, R = 153	75.16%
Setelah Penambahan Kuantitas Korpus Monolingual	Bella Yuda	C = 137, R=153	89.54%

Tabel 4.10 dihasilkan dari perhitungan total kata yang di ujikan oleh ahli bahasa sesuai lampiran D. Total keseluruhan kata yang diujikan sebesar R = 153 kata. Sebelum penambahan kuantitas korpus monolingual jumlah kata yang benar sebesar C = 115 kata. Berdasarkan persamaan 4.2, persentase akurasi oleh ahli bahasa sebesar 75.16%, setelah dilakukan penambahan kuantitas korpus monolingual terjadi peningkatan kata yang benar menjadi C = 137. Persentase akurasi oleh ahli bahasa setelah dilakukan penambahan kuantitas korpus monolingual adalah sebesar 89.54%. Dari hasil tersebut dihasilkan persentase peningkatan akurasi sebelum dan sesudah penambahan korpus monolingual sebesar 19.13% sesuai persamaan 4.1.

4.1.10. Perkiraan Jumlah Korpus Berdasarkan Penambahan Kuantitas Korpus Monolingual

Pada Gambar 4.28 dapat dilihat bahwa untuk mendapatkan nilai BLEU hingga 31.40% membutuhkan penambahan kuantitas korpus monolingual kurang lebih sebesar 169.574.400. lihat pada lampiran F.



Gambar 4.28 Tampilan grafik perkiraan jumlah korpus berdasarkan penambahan kuantitas korpus monolingual

Perkiraan jumlah korpus pada mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda dapat dihitung berdasarkan fungsi logaritma. Adapun nilai dari fungsi logaritma diperoleh dari grafik uji akurasi terhadap kuantitas korpus yang terdapat pada Gambar 4.28. Berdasarkan nilai dari grafik tersebut didapatkan persamaan y = a + b ln x. Nilai a dan b diperoleh dengan persamaan berikut:

$$b = n\Sigma ¿¿$$

$$¿ \frac{(8)(1956{,}21)-(227{,}56)(68{,}76)}{(8)(592{,}47)-(4727{,}39)}$$

$$¿ 0{,}29$$

$$a = \Sigma y - b\Sigma ¿¿$$

$$¿ \frac{(227{,}56)-(0{,}29)(68{,}76)}{8}$$

$$¿ 25{,}99$$

Sehingga nilai dari x dari y = 31,40 dapat di hitung dengan persamaan:

$$y = a + b \ln x$$

$$x = e^{\frac{y-a}{b}}$$

$$x = e^{\frac{31,40-25,99}{0,29}}$$

¿169.574 .400

4.2. Analisis Hasil Pengujian

Berikut merupakan analisis terhadap hasil pengujian yang telah dilakukan.

1. Penilaian otomatis terhadap hasil terjemahan pada mesin penerjemah statistik bahasa Indonesia - bahasa Sunda sebelum dilakukan penambahan kuantitas korpus monolingual dengan 2.400 korpus paralel didapatkan nilai BLEU sebesar 28.25%, setelah dilakukan penambahan kuantitas korpus monolingual 3.400 sebesar 28.34% meningkat 0,32%, dengan korpus 4.400 sebesar 28.40% meningkat 0,51%, dengan korpus 5.400 sebesar 28.37% meningkat 0.42%, dengan korpus 6.400 sebesar 28.40% meningkat 0.51%, dengan korpus 7.400 sebesar 28.50% meningkat 0.87%, dengan korpus 8.400 sebesar 28.55% meningkat 1.04% dan dengan korpus 9.400 sebesar 28.76% meningkat 1.79%
2. Penilaian yang dilakukan ahli bahasa seperti terdapat pada Tabel 4.11 pada kalimat yang diterjemahkan dengan tepat sebelum penambahan kuantitas korpus monolingual mengalami peningkatan dengan persentase sebesar 83.77% menjadi 96.75% setelah dilakukan penambahan kuantitas korpus monolingual, sehingga peningkatan akurasi dapat dihitung C = $\frac{b-a}{a}$ 100% , sehingga C = $\frac{89.54-75.16}{75.16}$ 100% dan didapat nilai peningkatan sebesar 19.13%.
3. Peningkatan penilaian otomatis sangat berbeda dengan penilaian oleh ahli bahasa, penilaian otomatis sangat tergantung pada kalimat rujukan dan jumlah korpus paralel yang digunakan dan sedangkan penilaian langsung oleh ahli bahasa dipengaruhi oleh pengetahuan dan pemahaman sehingga lebih baik dalam melakukan penilaian.

4. Dari hasil uji mesin penerjemah statistik didapatkan nilai BLEU yang digunakan untuk menghitung persamaan logaritma. Dan didapatkan bahwa untuk mencapai nilai BLEU 31,40% membutuhkan paling kurang .169.574.400 korpus monolingual.
5. Penurunan atau peningkatan nilai BLEU ketika di lakukan penambahan kuantitas korpus monolingual di pengaruhi oleh kualitas korpus yang di tambahkan.

BAB V
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil analisis dan pengujian, maka kesimpulan yang dapat diambil sebagai berikut.

1. Berdasarkan hasil penelitian, penambahan kuantitas korpus monolingual dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Sunda.
2. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Sunda dengan korpus 2.400 kalimat sebesar 28.25% dan setelah dilakukan penambahan kuantitas korpus monolingual 3.400 sebesar 400 sebesar 28.34% meningkat 0,32%, dengan korpus 4.400 sebesar 28.40% meningkat 0,51%, dengan korpus 5.400 sebesar 28.37% meningkat 0.42%, dengan korpus 6.400 sebesar 28.40% meningkat 0.51%, dengan korpus 7.400 sebesar 28.50% meningkat 0.87%, dengan korpus 8.400 sebesar 28.55% meningkat 1.04% dan dengan korpus 9.400 sebesar 28.76% meningkat 1.79%
3. Penilaian yang dilakukan oleh ahli bahasa menghasilkan persentase peningkatan sebelum dilakukan penambahan kuantitas korpus monolingual sebesar 83.77% dan setelah dilakukan penambahan kuantitas korpus monolingual menjadi 96.75%. Dari hasil tersebut terjadi peningkatan hasil terjemahan sebesar 19.13%.
4. Untuk mencapai nilai BLEU hingga 31,40% dibutuhkan setidaknya 169.574.400 korpus monolingual yang ditambahkan kedalam mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.
5. Penurunan atau peningkatan nilai BLEU ketika di lakukan penambahan kuantitas korpus monolingual dipengaruhi oleh kualitas korpus yang ditambahkan.

5.2. Saran

Beberapa saran yang dapat diberikan sebagai pengembangan dari penelitian ini adalah sebagai berikut.

1. Perlu penambahan jumlah korpus untuk meningkatkan kualitas terjemahan mesin penerjemah statistik.
2. Perlu menggunakan korpus yang berkualitas agar mendapatkan nilai akurasi yang lebih tinggi.
3. Perlu dilakukan penelitian lanjutan untuk melakukan analisis dalam menghasilkan terjemahan bahasa Indonesia – bahasa Sunda dengan menggunakan metode penelitian yang lain.
4. Melakukan implementasi mesin penerjemah statistik ke dalam bahasa daerah yang lain dengan metode penambahan kuantitas korpus monolingual.
5. Perlu dilakukan pengecekan ulang terhadap korpus teks paralel untuk mencegah kesalahan penulisan (*typo*).