

BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Sebelum penelitian ini, sudah ada beberapa penelitian serupa mengenai penerapan pengolahan bahasa alami untuk pengolahan data-data teks pada rekam medis. Ratnasari dkk. (2014) pernah melakukan penelitian mengenai pengembangan sebuah model pengolahan bahasa alami untuk perumusan keluhan pasien yang mampu menghasilkan teks dengan makna yang sama dengan teks narasi keluhan pasien, tetapi dalam bahasa yang lebih baku atau dalam bahasa medis yang tepat. Kemudian ada aplikasi pengolahan bahasa alami bernama CLiX ENRICH yang dikembangkan oleh perusahaan Clinithink (*CLiX ENRICH White Paper*, 2014) yang mampu mengekstrak informasi pada teks-teks medis yang ditulis dalam format tidak terstruktur dan kemudian mengodekan ke format SNOMED yang sesuai, aplikasi ini juga memiliki fitur *speech recognition* yang diberi nama Recognosco. The University of Texas Health Science Center at Houston pernah mengembangkan *Clinical Language Annotation, Modeling, and Processing Toolkit* (CLAMP) yaitu sebuah *software* pengolahan bahasa alami klinis yang tersusun dari komponen-komponen pengolahan bahasa alami yang dapat digunakan secara terpisah oleh *user* (CLAMP Team, n. d.). Averbis Health Discovery yang dikembangkan oleh Averbis di Jerman mampu memberikan prediksi diagnosis yang tepat dari masukan teks tidak terstruktur yang menjelaskan kondisi pasien serta mampu memprediksi diagnosis yang tepat untuk penyakit langka dengan menggunakan *machine learning* ("Averbis Health Discovery - Analysis of patient data", n.d.). Canary adalah sebuah aplikasi pengolahan bahasa alami klinis *open source* yang dikembangkan oleh Dr. Alexander Turchin's *research group* di Brigham and Women's Hospital yang didesain agar mudah digunakan oleh *user* yang tidak memiliki latar belakang ilmu komputer atau pemrograman, dan awalnya didesain khusus untuk penggunaan klinis namun dapat dikembangkan untuk penggunaan di bidang lainnya juga (Malmasi dkk., 2017)

Perbandingan kelima aplikasi pada penelitian sebelumnya dengan aplikasi yang dikembangkan pada penelitian ini dapat dilihat pada tabel 2.1. di bawah ini:

Tabel 2.1. Perbandingan Aplikasi yang Dikembangkan pada Penelitian ini dengan Aplikasi pada Penelitian Sebelumnya

	Aplikasi yang Dikembangkan	Aplikasi oleh Ratnasari dkk.	CLiX ENRICH	CLAMP	Averbis	Canary
Fitur <i>Speech recognition</i>	Ada, menggunakan Sphinx4	Tidak Ada	Ada, menggunakan Recognosco	Tidak Ada	Ada	Tidak Ada
Mempelajari Istilah Baru	Bisa	Tidak Bisa	Tidak Bisa	Bisa	Bisa	Bisa
Offline/Online	Offline	Online	Offline dan dapat berbasis Cloud	Offline	Offline	Offline
Lisensi	Closed Source	Closed Source	Closed Source	Gratis untuk akademik. Source Code dapat diminta oleh orang yang memiliki izin.	Closed Source	Gratis dan Open Source
Bahasa Alami Utama	Bahasa Indonesia	Bahasa Indonesia	Bahasa Inggris	Bahasa Inggris	Bahasa Inggris dan Bahasa Jerman	Bahasa Inggris. Namun sudah diujicoba dengan bahasa lainnya.

Berdasarkan perbandingan pada tabel 2.1. di atas dapat disimpulkan bahwa aplikasi yang dikembangkan pada penelitian ini memiliki kelebihan fitur *speech recognition* yang tidak dimiliki oleh beberapa aplikasi yang pernah dikembangkan sebelumnya, yaitu aplikasi yang dikembangkan oleh Ratnasari dkk., CLAMP, dan Canary. Aplikasi yang dikembangkan pada penelitian ini juga mampu mempelajari istilah baru yang tidak dimiliki oleh aplikasi yang dikembangkan oleh Ratnasari dkk. dan CLiX ENRICH, serta komputer *user* tidak perlu terhubung ke internet karena aplikasi ini digunakan secara *offline*. *Source code* aplikasi yang dikembangkan pada penelitian ini bersifat *closed source*, namun *source code* nya suatu saat kemungkinan dapat diberikan kepada seseorang yang telah mendapatkan izin. Beberapa aplikasi yang dikembangkan pada penelitian sebelumnya umumnya menggunakan bahasa alami utama Bahasa Inggris sedangkan bahasa alami

utama pada aplikasi yang dikembangkan pada penelitian ini adalah Bahasa Indonesia, sama seperti aplikasi yang dikembangkan oleh Ratnasari dkk.

2.2. Anamnesis

Anamnesis adalah sebuah kegiatan wawancara antara dokter atau petugas kesehatan lainnya terhadap pasien atau orang-orang yang memiliki hubungan dekat dengan pasien untuk memperoleh keterangan-keterangan mengenai keluhan dan penyakit yang diderita oleh pasien (Wahid, 2012). Tujuan anamnesis adalah untuk mendapatkan informasi menyeluruh dari pasien yang bersangkutan seperti data medis organobiologis, psikososial, serta lingkungan pasien. Dengan menggunakan informasi tersebut, pada akhir anamnesis diharapkan dokter mampu menyimpulkan dugaan organ atau sistem organ yang terganggu, termasuk rumusan masalah kliniknya. Tujuan anamnesis lain yang juga penting adalah untuk menjalin hubungan antara dokter dengan pasien yang profesional dan optimal (Markum, 2007).

Anamnesis terbagi menjadi dua macam, yaitu autoanamnesis dan alloanamnesis atau heteroanamnesis (Markum, 2007). Autoanamnesis adalah anamnesis yang dilakukan secara langsung terhadap pasien itu sendiri karena pasien dianggap mampu untuk melakukan wawancara. Jika pasien dianggap tidak mampu melakukan wawancara maka orang lain yang mengenal pasien dapat menggantikan pasien dalam melakukan wawancara yang disebut sebagai alloanamnesis atau heteroanamnesis. Pada umumnya alloanamnesis atau heteroanamnesis dilakukan terhadap pasien yang masih anak-anak, pasien difabel, pasien yang sedang dalam kondisi tidak sadar, atau pasien yang sedang mengalami gangguan kejiwaan. Dalam praktek sehari-hari tidak jarang juga alloanamnesis dilakukan bersama-sama dengan autoanamnesis.

Sebuah anamnesis yang baik sebaiknya berisi beberapa kelompok data penting sebagai berikut agar dalam melakukan anamnesis tidak ada pertanyaan yang terlewatkan sehingga semua keluhan pasien dapat ditampung (Markum, 2007):

1. Identitas pasien
2. Riwayat penyakit sekarang (didahului keluhan utama)
3. Riwayat penyakit dahulu
4. Riwayat kesehatan keluarga
5. Riwayat pribadi, psikologis, sosial-ekonomi-dan-budaya

Penjelasan dari poin-poin di atas dituliskan di bawah ini.

a. Identitas Pasien

Identitas pasien umumnya meliputi nama lengkap, alamat lengkap, usia, pekerjaan, jenis kelamin, status perkawinan, suku, agama, serta nomor telepon pasien yang dapat dihubungi. Data identitas pasien ini dibutuhkan oleh petugas kesehatan karena beberapa penyakit tertentu terkadang berkaitan dengan usia, jenis kelamin, profesi, status perkawinan, atau suku bangsa tertentu. Data identitas pasien juga berguna untuk memberikan kesan mengenai keadaan sosial ekonomi, budaya, dan lingkungan pasien, serta untuk dapat melakukan komunikasi secara berkesinambungan terhadap pasien. Dengan data identitas pasien, petugas kesehatan dapat merencanakan pengelolaan pasien yang lebih cepat, optimal, dan sesuai dengan kondisi pasien secara menyeluruh.

b. Keluhan Utama

Keluhan utama adalah keluhan yang menyebabkan pasien meminta pertolongan kepada petugas kesehatan. Keluhan utama ini umumnya tertulis secara ringkas beserta lama kemunculannya. Misalnya: “sakit tenggorokan sejak tiga minggu yang lalu.”

c. Riwayat Penyakit Sekarang

Riwayat Penyakit Sekarang (RPS) adalah riwayat mengenai penyakit pasien sekarang, yang dimulai sejak akhir masa sehat. Pada RPS dicatat perkembangan dan perjalanan penyakit yang ditulis secara kronologis sesuai urutan waktunya, termasuk riwayat pengobatan atau perawatan untuk penyakit yang sedang dialami tersebut. Pada RPS juga tertulis deskripsi atau analisis terhadap keluhan utama dan keluhan penting lainnya beserta lokasi dan penjalaran, kualitas, kuantitas (tipe onset, intensitas/keparahan, disabilitas), kronologis, kondisi atau keadaan saat munculnya gejala (onset, durasi, periodisitas, frekuensi), faktor yang mempengaruhi (faktor pencetus, faktor yang memperberat, faktor yang memperingan, kaitannya dengan aktivitas sehari-hari), dan gejala penyerta (Fakultas Kedokteran UI, 2009):

d. Riwayat Penyakit Dahulu

Pada Riwayat Penyakit Dahulu (RPD) dicatat daftar penyakit yang dahulu pernah dialami pasien beserta waktu kemunculannya, termasuk apakah pasien pernah mengalami operasi, kecelakaan, ataupun keadaan alergi. Riwayat penyakit dahulu mencakup anamnesis penyakit sistem pernapasan, sistem kardiovaskuler, sistem muskuloskeletal, saluran cerna, kulit, infeksi, endokrinologi, hematologi, saluran kemih, ginjal, dll. RPD merupakan data penting karena memberikan informasi kepada petugas kesehatan mengenai kemungkinan adanya gejala sisa dan kemungkinan adanya kaitan penyakit dahulu dengan penyakit sekarang, kemungkinan adanya gejala sisa dari penyakit yang dahulu diderita pasien, dan kemungkinan adanya pengaruh terhadap pengelolaan pasien selanjutnya.

e. Riwayat Penyakit Keluarga

Pada Riwayat Penyakit Keluarga, kondisi kesehatan dan umur masing-masing anggota keluarga dituliskan bila masih hidup. Bila anggota keluarga sudah meninggal, dituliskan umur saat meninggal dan penyebab meninggalnya. Anggota keluarga ini meliputi ayah, ibu, saudara laki-laki, saudara perempuan, kakek, nenek, dan anak-anak pasien. Penyakit-penyakit yang kemungkinan disebabkan oleh faktor hereditas atau oleh penularan anggota keluarga dicari dan dicatat pada Riwayat Penyakit Keluarga.

f. Riwayat Pribadi, Psikologis, Sosial Ekonomi dan Budaya

Riwayat Pribadi, Psikologis, Sosial Ekonomi dan Budaya ini berisi data mengenai keterangan kelahiran pasien, peristiwa penting pada masa kecil pasien, sikap pasien terhadap keluarga dekat, keterangan pendidikan pasien, keterangan pekerjaan pasien, aktivitas sehari-hari pasien, lingkungan tempat tinggal pasien, gaya hidup pasien, kondisi perkawinan pasien, masalah-masalah yang dihadapi pasien, dll. Data riwayat pribadi, psikososial, ekonomi, budaya dan keluarga ini merupakan informasi penting yang dapat digunakan sebagai bahan pertimbangan dalam pengelolaan yang optimal dan menyeluruh terhadap pasien untuk saat ini maupun selanjutnya.

Data-data anamnesis nantinya akan tersimpan di dalam rekam medis (Kemenkes RI, 2008). Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia no. 269/MENKES/PER/ III/2008 tentang Rekam Medis, rekam medis adalah berkas yang berisikan catatan dan dokumen tentang identitas pasien, pemeriksaan, pengobatan, tindakan dan pelayanan lain yang telah diberikan kepada pasien. Menurut Peraturan

Menteri Kesehatan Republik Indonesia no. 269/MENKES/PER/ III/2008 tentang Rekam Medis, rekam medis dapat dipakai untuk hal-hal sebagai berikut:

- a. Pemeliharaan kesehatan dan pengobatan pasien;
- b. Alat bukti dalam proses penegakan hukum, disiplin kedokteran dan kedokteran gigi dan penegakan etika kedokteran dan etika kedokteran gigi;
- c. Keperluan pendidikan dan penelitian;
- d. Dasar pembayaran biaya pelayanan kesehatan
- e. Data statistik kesehatan

Rekam medis berbasis elektronik atau disebut RME menyimpan informasi-informasi pasien pada komputer bukan pada kertas. Rekam medis berbasis kertas memiliki beberapa kelemahan seperti mudah rusak, mudah hilang, membutuhkan tempat penyimpanan yang besar, dan tulisannya sulit dibaca oleh sebagian orang (Ajami dan Bagheri-Tadi, 2013). Karena pada rekam medis elektronik data medis dicatat dan disimpan menggunakan komputer, maka kelemahan-kelemahan pada rekam medis berbasis kertas tersebut dapat diminimalisasi. Rekam medis elektronik mampu mengurangi biaya penyimpanan dan penyalinan rekaman medis, meningkatkan dokumentasi, memberikan efisiensi dalam penanganan rekaman medis, serta meningkatkan kecepatan untuk mengakses rekaman medis (Bilimoria, 2007)

2.3. Bahasa Alami

Bahasa adalah sebuah cara berkomunikasi secara lisan maupun tulisan yang terdiri dari kata-kata yang terstruktur, kumpulan-kumpulan simbol (seperti huruf, angka, dan karakter khusus), serta kumpulan-kumpulan aturan yang mengatur penyusunan dan penggunaan dari kata-kata dan simbol-simbol tersebut secara konvensional (Iroju dan Olakeke, 2015). Bahasa alami adalah bahasa yang manusia pelajari dari lingkungan mereka dan digunakan untuk berkomunikasi dengan sesamanya. Bagaimana pun bentuk komunikasinya, bahasa-bahasa alami digunakan untuk mengekspresikan pengetahuan dan emosi seseorang, serta untuk menyampaikan respons ke orang-orang lain atau lingkungan sekitarnya. Bahasa alami biasanya dipelajari pada awal masa kanak-kanak dari orang-orang sekitar nya (Reshamwala dkk., 2013). Contoh bahasa alami adalah bahasa Indonesia, Bahasa Jawa, dan Bahasa Inggris. Bahasa alami berbeda dengan bahasa-bahasa buatan dan

bahasa formal seperti bahasa-bahasa yang digunakan untuk menuliskan program komputer atau bahasa-bahasa yang digunakan untuk mempelajari ilmu logika (Lyons, 1991).

Bahasa alami berkembang secara alamiah untuk memenuhi kebutuhan komunikasi antar manusia dan tidak secara sengaja dirancang untuk kemudahan pemrosesan oleh komputer. Oleh karena itu, pengolahan bahasa alami oleh komputer jauh lebih sulit dilakukan dibandingkan dengan pengolahan bahasa buatan oleh komputer. Beberapa karakteristik umum bahasa alami yang menyebabkannya sulit diproses oleh komputer antara lain (Utami dan Hartati, 2007):

1. Sering terjadi ambiguitas dalam bahasa alami.
2. Jumlah kosa kata dalam bahasa alami sangat besar dan berkembang dari waktu ke waktu.

Ambiguitas dalam bahasa alami adalah adanya lebih dari satu kemungkinan arti pada teks maupun ucapan bahasa alami. Jenis-jenis ambiguitas dalam bahasa alami antara lain (Reshamwala dkk., 2013).

1. Ambiguitas Sintaksis

Ambiguitas sintaksis terjadi ketika adanya lebih dari satu cara untuk melakukan *parsing* pada sebuah kalimat. Contohnya adalah pada kalimat "Polisi menangkap penjahat dengan topi merah". Frasa "topi merah" bisa memiliki kemungkinan arti sebagai bagian dari frasa nomina yang menjelaskan si penjahat, namun bisa juga memiliki kemungkinan arti sebagai frasa preposisi yang menjelaskan cara polisi menangkap si penjahat tersebut, yaitu dengan menggunakan topi merah.

2. Ambiguitas Semantik

Ambiguitas Semantik terjadi jika sebuah teks mengandung kata atau frasa yang dapat memiliki dua arti. Contohnya pada kalimat "Dia sakit hati.". Kata "hati" pada kalimat tersebut bisa memiliki arti salah satu organ tubuh manusia atau juga bisa memiliki arti perasaan manusia.

3. Ambiguitas Referensial

Ambiguitas referensial terjadi ketika suatu kalimat mengacu pada suatu kata tanpa secara eksplisit menyebut kata tersebut dan menggunakan kata ganti. Contohnya pada kalimat "Andi bertengkar dengan Budi hingga dia terluka". Kata "dia" pada kalimat tersebut memiliki kemungkinan mengacu kepada "Andi" atau "Budi".

4. Ambiguitas Lokal

Ambiguitas lokal terjadi ketika suatu bagian dari sebuah kalimat kurang jelas artinya secara lokal namun keseluruhan kalimat membuatnya menjadi lebih jelas artinya. Contohnya adalah pada bagian kalimat ini "Dia datang bulan...". Penggalan kalimat tersebut memiliki arti kurang jelas, namun setelah dilengkapi menjadi kalimat "Dia datang bulan Januari" atau menjadi kalimat "Dia datang bulan tidak teratur" penggalan kalimat tersebut menjadi lebih jelas artinya. Pada contoh kalimat pertama "datang bulan" artinya adalah "datang pada saat bulan...", sedangkan pada contoh kalimat kedua "datang bulan" artinya adalah "menstruasi".

2.3.1. Tata Bahasa

Tata bahasa terbagi menjadi dua macam pengertian menurut Keraf (1999). Pertama, cabang ilmu bahasa yang mempelajari dan mendeskripsikan kaidah-kaidah yang menjadi dasar bentuk sebuah bahasa; dan kedua, semacam buku yang memuat himpunan kaidah dan patokan umum mengenai struktur suatu bahasa. Tata Bahasa mencakup beberapa aspek antara lain fonologi, morfologi, sintaksis, dan semantik (Kridalaksana, 2001). Berikut ini dijelaskan hal-hal yang berkaitan dengan aspek-aspek tersebut.

a. Fonologi

Fonologi adalah bidang linguistik yang mempelajari, menganalisis, dan membicarakan runtunan bunyi-bunyi bahasa (Achmad dan Abdullah, 2012). Fonologi dapat dibagi menjadi fonetik dan fonemik. Fonetik adalah cabang studi fonologi yang mempelajari bunyi bahasa tanpa memperdulikan apakah bunyi-bunyi tersebut berfungsi sebagai pembeda makna atau tidak. Objek kajian fonetik adalah fon, yaitu bunyi yang tidak memperhatikan fungsinya sebagai pembeda makna. Fonemik adalah cabang studi fonologi yang mempelajari bunyi bahasa dengan memperhatikan fungsi bunyi tersebut sebagai pembeda makna. Objek kajian fonemik adalah fonem, yaitu bunyi bahasa yang membedakan makna suatu kata, jika suatu bunyi tidak membedakan makna maka bunyi tersebut bukanlah fonem. Sebagai contoh huruf 'k' pada kata "botak" pada Bahasa Indonesia dapat diucapkan dengan bunyi *glottal stop* maupun dengan bunyi *voiceless velar stop*, namun kedua bunyi tersebut tidak membedakan makna suatu kata, sehingga kedua bunyi tersebut tidak dianggap sebagai fonem berbeda. Sedangkan huruf 'e' pada kata "apel" memiliki arti berbeda jika huruf 'e' tersebut diucapkan sebagai e pepet dan jika

huruf ‘e’ tersebut diucapkan sebagai e taling, jika huruf ‘e’ pada kata “apel” diucapkan sebagai e pepet maka makna kata “apel” tersebut adalah sejenis buah-buahan, sedangkan jika huruf ‘e’ pada kata “apel” diucapkan sebagai e taling maka makna kata “apel” tersebut adalah upacara. Oleh karena itu e pepet dan e taling dianggap sebagai fonem berbeda karena dapat membedakan makna. Setiap bahasa memiliki jumlah fonem yang berbeda-beda, menurut Achmad dan Abdullah (2012) jumlah fonem dalam Bahasa Indonesia ada 24 buah, yang terdiri dari 6 buah fonem vokal (yakni a, i, u, e, o, dan ə) dan 18 buah fonem konsonan (yakni, p, t, c, k, b, d, j, g, m, n, ŋ, ɲ, s, h, r, l, w, dan y). Bunyi-bunyi yang merupakan realisasi dari sebuah fonem, seperti bunyi [k] dan [ʔ] untuk fonem Bahasa Indonesia /k/ pada kata “bakti” dan kata “bapak” disebut sebagai alofon-alofon dari fonem /k/. Fonem dapat dibagi menjadi fonem segmental dan fonem suprasegmental. Fonem segmental terdiri atas fonem vokal dan fonem konsonan, sedangkan fonem suprasegmental terdiri atas tekanan, durasi, dan nada.

b. Morfologi

Morfologi adalah cabang ilmu linguistik yang menyelidiki struktur kata, bagian-bagiannya, serta cara pembentukannya (Achmad dan Abdullah, 2012). Morfem adalah satuan bahasa terkecil dalam tataran morfologis yang tidak dapat dibagi lagi menjadi bagian bermakna yang lebih kecil. Bentuk-bentuk realisasi yang berbeda dari sebuah morfem yang sama disebut alomorf. Misalnya morfem me- memiliki bentuk realisasi mem- pada kata "membawa", men- pada kata "mencegah", meny- pada kata "menyuap" dan meng- pada kata "menggunting". Morfem dapat dibedakan menjadi dua macam yaitu morfem bebas dan morfem terikat. Morfem bebas mampu berdiri sendiri sebagai kata, sedangkan morfem terikat tidak mampu untuk berdiri sendiri dan harus selalu disatukan dengan morfem lainnya melalui proses morfologis atau pembentukan kata. Semua imbuhan dalam Bahasa Indonesia adalah morfem terikat. Kata-kata seperti "renta" yang hanya muncul dalam "tua renta" dan "kerontang" yang hanya muncul dalam "kering kerontang" juga termasuk morfem terikat. Morfem dapat berbentuk morfem utuh maupun morfem terbagi. Morfem terbagi contohnya adalah konfiks seperti ke-...-an pada kata "kesatuan" dan konfiks per-...-an pada kata perbaikan. Infiks dapat mengubah morfem utuh menjadi morfem terbagi, misalnya morfem “getar” jika diberi infiks “-em-” maka menjadi kata “gemetar” yang membagi morfem “getar” tersebut. Pada beberapa bahasa terdapat morfem suprasegmental, yaitu morfem yang dibentuk oleh unsur-unsur suprasegmental seperti nada, tekanan, dll. Pada Bahasa Mandarin misalnya, kata "ma" dapat memiliki arti

"kuda", "ibu", atau "memarahi" tergantung nada dalam pengucapannya, sehingga kata "ma" dalam Bahasa Mandarin merupakan morfem suprasegmental karena mengantung unsur suprasegmental yaitu nada. Bahasa Indonesia sendiri tidak memiliki morfem suprasegmental.

Menurut Kamus Besar Bahasa Indonesia (2008), kata adalah satuan (unsur) bahasa yang terkecil yang dapat diujarkan sebagai bentuk yang bebas; satuan (unsur) bahasa yg berupa morfem bebas. Kata adalah satuan bahasa terbesar dalam tataran morfologis. Proses pembentukan kata-kata dari satuan lain yang merupakan dasarnya sering disebut juga proses morfologi. Ada tiga macam proses morfologi dalam Bahasa Indonesia yaitu afiksasi, reduplikasi, dan komposisi (Ramlan, 1985). Afiksasi adalah proses penambahan imbuhan pada kata dasar yang dapat dibedakan menjadi prefiks, infiks, sufiks, dan konfiks atau simulfiks. Reduplikasi adalah pengulangan bentuk dasar atau satuan gramatik menjadi sebuah kata ulang. Komposisi adalah penggabungan kata-kata yang membentuk kata majemuk.

Kelas kata ialah pengelompokan kata berdasarkan perilaku atau sifat kata tersebut dalam kalimat. Secara umum, kelas kata terbagi menjadi berikut ini:

1. Nomina

Nomina atau kata benda adalah kata yang mengacu kepada suatu benda yang bersifat konkret maupun abstrak. Contoh nomina adalah rumah, mobil, akal, dan hati.

2. Verba

Verba atau kata kerja adalah kata-kata yang menyatakan perbuatan atau tindakan. Contoh verba adalah berlari, memakan, dan tertidur.

3. Adjektiva

Adjektiva atau kata sifat merupakan kata yang menyatakan sifat atau keadaan dari suatu nomina (kata benda) atau suatu pronomina (kata ganti). Contoh adjektiva adalah merah, bulat, kasar, dan besar.

4. Adverbia

Adverbia atau kata keterangan adalah kata yang memberi keterangan pada verba, adjektiva, nomina predikatif, atau kalimat. Contoh adverbia adalah sangat, terlalu, agak, dan cukup.

5. Pronomina

Pronomina atau kata ganti adalah kata yang digunakan untuk menggantikan benda atau sesuatu yang dibendakan. Contohnya adalah aku, kita, ini, dan sesuatu.

6. Numeralia

Numeralia atau kata bilangan adalah kata yang menyatakan jumlah, kumpulan, dan urutan atau tingkatan suatu benda atau sesuatu yang dibendakan. Contohnya adalah satu, seratus, kedua, dan sebagian.

7. Kata Tugas

Kata tugas hanya mempunyai arti gramatikal dan tidak memiliki arti leksikal. Kata tugas umumnya merupakan kelas kata yang keanggotaannya tertutup yaitu hampir tidak pernah ada penambahan kata baru pada kelas kata ini, dan kebanyakan tidak mengalami proses morfologis. Kata tugas terbagi menjadi artikel (kata sandang), preposisi (kata depan), konjungsi (kata hubung), partikel, dan interjeksi (kata seru). Artikel atau kata sandang adalah kata yang mendampingi kata benda atau yang membatasi makna jumlah orang atau benda seperti sang dan si. Preposisi atau kata depan adalah kata yang selalu berada di depan kata benda, kata sifat, atau kata kerja untuk membentuk frasa preposisional seperti ke, dari, dan sekitar. Konjungsi atau kata hubung adalah kata tugas yang menghubungkan antar klausa, antar kalimat, atau antar paragraf seperti karena, tetapi, dan sehingga. Partikel adalah kata yang bertugas memulai, mempertahankan, atau mengukuhkan sebuah kalimat dalam komunikasi seperti lah, dong, dan deh. Dan terakhir yaitu interjeksi atau kata seru ialah kata yang mengungkapkan perasaan seperti wah, hai, dan yes.

c. Sintaksis

Sintaksis mempelajari kata, frasa, klausa, dan kalimat, serta hubungannya dan cara penyusunannya hingga menjadi satu ujaran yang tepat (Achmad dan Abdullah, 2012). Seperangkat aturan yang mengatur deretan kata-kata yang membentuk kalimat disebut sebagai alat sintaksis. Alat sintaksis yang mengatur unsur-unsur bahasa sehingga terbentuk satuan bahasa yang disebut kalimat antara lain urutan kata, kata tugas, bentuk kata, dan intonasi.

Bidang kajian sintaksis adalah frasa, klausa, dan kalimat. Frasa adalah satuan konstruksi yang terdiri dari dua atau lebih kata yang membentuk satu kesatuan (Keraf, 1984). Sebanyak apapun kata penyusun frasa, selama tidak melampaui batas fungsi atau jabatannya baik sebagai subjek, predikat, objek, pelengkap, maupun keterangan, maka masih bisa disebut sebagai frasa (Ramlan, 2001). Menurut Keraf (1984), klausa adalah sebuah konstruksi yang di dalamnya terdapat beberapa kata yang mengandung unsur predikatif. Kalimat adalah satuan bahasa terkecil yang merupakan kesatuan pikiran

(Widjono, 2007). Manaf (2009) menjelaskan bahwa yang membedakan klausa dan kalimat adalah klausa tidak diakhiri intonasi final, sedangkan kalimat diakhiri dengan intonasi final yang dapat berupa intonasi berita, intonasi tanya, intonasi perintah, atau intonasi kagum.

Fungsi sintaksis adalah wadah yang dapat diisi oleh bentuk bahasa tertentu (Manaf, 2009). Subjek dan predikat adalah fungsi sintaksis yang wajib ada pada setiap kalimat, sedangkan fungsi sintaksis objek, pelengkap, dan keterangan tidak wajib ada pada setiap kalimat. Subjek merupakan unsur utama suatu kalimat yang dijelaskan oleh fungsi sintaksis predikat. Predikat merupakan unsur kalimat yang menjelaskan subjek. Objek adalah unsur kalimat yang melengkapi verba transitif pada suatu kalimat aktif. Pelengkap berfungsi melengkapi predikat dan tidak dapat diubah menjadi subjek. Keterangan adalah unsur tambahan dalam kalimat yang berfungsi memperjelas atau memberi keterangan pada seluruh kalimat.

d. Semantik

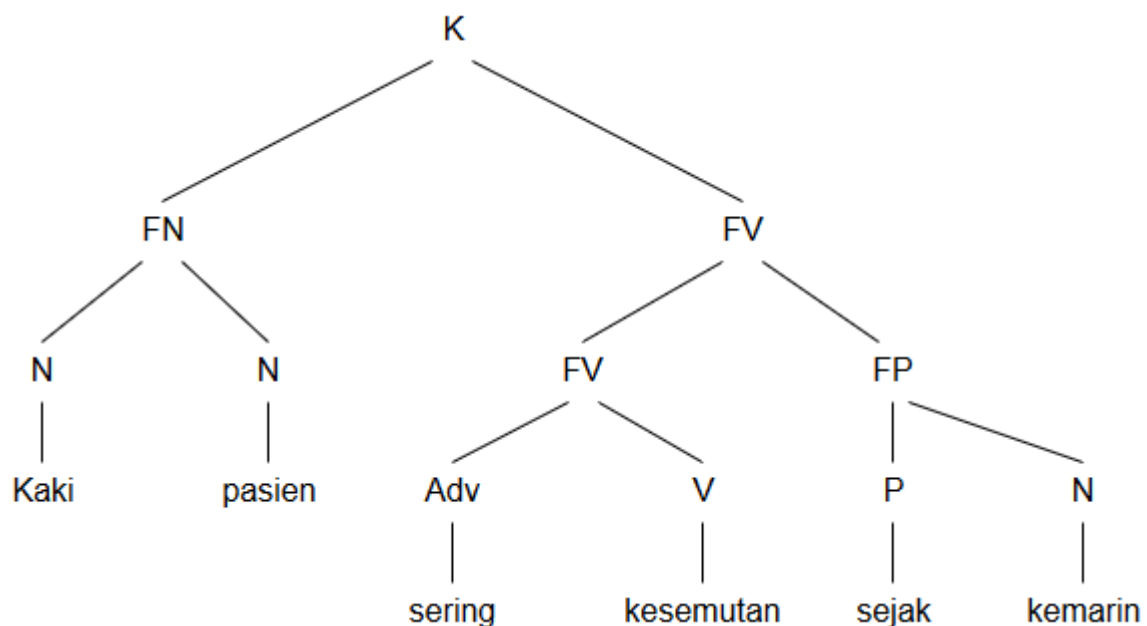
Semantik adalah cabang ilmu linguistik yang mempelajari arti atau makna suatu kata (Achmad dan Abdullah, 2012). Arti kata dapat dibagi menjadi tiga yaitu arti leksikal, arti gramatikal, dan arti idiomatik. Arti leksikal adalah arti kata menurut kamus, yang umumnya dianggap sejajar dengan arti denotatif. Arti gramatikal muncul setelah proses gramatika, misalnya morfem ber- jika berdiri sendiri tidak memiliki makna, morfem tersebut baru memiliki makna setelah bergabung dengan morfem lainnya. Arti idiomatik muncul karena beberapa kata bergabung membentuk sebuah makna baru yang tidak dapat ditelusuri dari kata-kata pembentuknya, contohnya adalah "buah tangan" yang memiliki arti "oleh-oleh" dan "besar kepala" yang memiliki arti "sombong".

2.3.2. Parsing

Parsing adalah proses pengonversian sebuah kalimat menjadi sebuah pohon (dinamakan *parse tree*) yang merepresentasikan struktur sintaksis dari kalimat tersebut (Reshamwala dkk., 2013). *Parse tree* dapat dihasilkan dari kalimat-kalimat pada bahasa alami maupun dari pemrosesan bahasa komputer seperti bahasa pemrograman. Kalimat "Mobil hitam berhenti di jalan" terdiri dari frasa nominal "Mobil hitam" dan frasa verbal "berhenti di jalan". Pada *top-down parsing*, pohon kalimat tersebut memiliki akar kalimat tersebut dan kemudian bercabang menjadi frasa nominal dan frasa verbal, dan kemudian akhirnya bercabang kembali menjadi nomina-nomina, adjektiva, verba, dan preposisi.

Parsing menentukan apakah sebuah kalimat adalah valid menurut aturan-aturan tata bahasa sebuah bahasa. Ada dua macam teknik *parsing*, yaitu *top-down* dan *bottom-up*.

Top down parser memulai pemeriksaan dari tingkatan tertinggi pada *parse tree* dan bekerja ke bawah mencari *node* terminal sesuai dengan kelas kata dari kata-kata pada kalimat masukan. Gambar 2.1. di bawah ini adalah contoh *parse tree* yang dibentuk oleh *top down parser*. Awalnya kalimat "Kaki pasien sering kesemutan sejak kemarin" bercabang menjadi frasa nominal "Kaki pasien" dan frasa verbal "sering kesemutan sejak kemarin". Kemudian frasa nominal "Kaki pasien" bercabang kembali menjadi nomina "Kaki" dan nomina "pasien", dan frasa verbal "sering kesemutan sejak kemarin" bercabang kembali menjadi frasa verbal "sering kesemutan" dan frasa preposisional "sejak kemarin". Dan terakhir frasa verbal "sering kesemutan" bercabang menjadi adjektiva "sering" dan verba "kesemutan", dan frasa preposisional "sejak kemarin" bercabang menjadi preposisi "sejak" dan nomina "kemarin".



Gambar 2.1. Contoh *Parse Tree*.

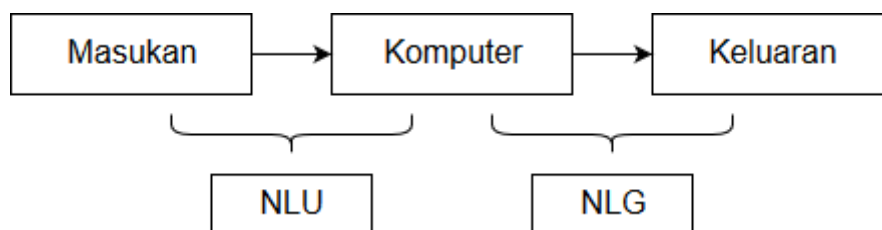
Bottom up parser merupakan kebalikan dari top down parser, yaitu memulai pemeriksaan dari *node-node* terminal dan bekerja ke atas mencari *node* akar dari *parse tree*. Pada contoh gambar 4 di atas, bottom up parser melakukan pencarian dari nomina "Kaki", nomina "pasien", adjektiva "sering", verba "kesemutan", preposisi "sejak" dan verba

“kesemutan” menuju ke arah pembentukan frasa-frasa yang membentuk kalimat pada akar, hingga ke pembentukan kalimat pada akarnya.

2.4. Pengolahan Bahasa Alami

Pengolahan Bahasa Alami atau juga dikenal sebagai Komputasi Linguistik adalah sebuah cabang Ilmu Komputer yang mengurus mengenai penggunaan teknik-teknik komputasi untuk mempelajari, memahami, dan memproduksi bahasa alami manusia (Hirschberg dan Manning, 2015). Pengolahan Bahasa Alami memiliki beberapa tujuan, yaitu membantu komunikasi antar manusia seperti pada *Machine Translation*, membantu komunikasi antara manusia dengan mesin misalnya pada *speech recognition* dan *speech synthesis*, atau memudahkan manusia dan mesin dengan belajar dan menganalisis dari konten-konten bahasa alami manusia yang saat ini tersedia dalam jumlah besar secara *online*. Pemrosesan Bahasa Alami merupakan bidang ilmu multidisipliner yang merupakan gabungan dari bidang ilmu Ilmu Komputer, Linguistik, Sains Kognitif, Psikologi, Filsafat, dan Logika Matematika (Iroju dan Olakeke, 2015).

Pemrosesan Bahasa Alami terbagi menjadi dua tugas utama (Regina dan Michael, 2005), yaitu *Natural Language Understanding* (NLU) dan *Natural Language Generation* (NLG) yang diilustrasikan pada gambar 2.2. di bawah. Sesuai namanya, NLU adalah sebuah proses pemahaman bahasa alami oleh komputer. NLU mengonversi sebuah masukan bahasa alami manusia ke bentuk yang dapat dipahami oleh komputer (Saetre, 2004). NLG adalah proses pengkonstruksian sebuah bahasa alami oleh komputer untuk tujuan komunikasi tertentu (Dale, 2006).



Gambar 2.2. Tugas utama pengolahan bahasa alami (Regina dan Michael, 2005)

Beberapa tahun belakangan ini Pengolahan Bahasa Alami sudah berkembang menjadi area riset yang menarik dan beberapa teknologi terapannya telah digabungkan pada produk-produk konsumen seperti Siri pada produk-produk Apple dan *Skype*

Translator. Empat faktor kunci menjadi penyebab perkembangan pesat Pengolahan Bahasa Alami beberapa tahun belakangan ini (Hirschberg dan Manning, 2015), yaitu peningkatan kemampuan komputer yang pesat, ketersediaan data-data linguistik yang melimpah yang salah satunya disebabkan oleh berkembangnya internet, kesuksesan pengembangan metode-metode *machine learning*, dan pemahaman yang lebih mendalam dari struktur bahasa manusia. Salah satu keterbatasan Pengolahan Bahasa Alami saat ini adalah sebagian besar sumber daya dan sistem Pengolahan Bahasa Alami hanya tersedia untuk *high-resource languages* (HRL), seperti Bahasa Inggris, Bahasa Prancis, Bahasa Spanyol, Bahasa Jerman, dan Bahasa Mandarin. Banyak *low-resource languages* (LRL) seperti Bahasa Indonesia, Bahasa Bengali, Bahasa Cebuano, dan Bahasa Swahili tidak memiliki sistem atau sumber daya yang cukup meskipun diucapkan dan ditulis oleh jutaan orang (Hirschberg dan Manning, 2015). Di masa depan ilmuwan berharap dapat mengembangkan sumber daya dan sistem untuk ratusan ribu atau jutaan bahasa, tidak hanya beberapa saja.

2.4.1. Tingkatan Analisis Pengolahan Bahasa Alami

Pengolahan Bahasa Alami biasanya dikerjakan pada tingkatan-tingkatan yang dideskripsikan di bawah ini (Kumar, 2012):

1. Analisis Fonologi

Tingkatan ini menangani interpretasi suara-suara pembicaraan di dalam dan antar kata-kata. Ada tiga aturan yang digunakan dalam analisis fonologi, yaitu aturan fonetik, aturan fonemik, dan aturan prosodi. Aturan fonetik menangani tentang suara-suara di dalam kata. Aturan fonemik menangani tentang variasi-variasi dari pengucapan ketika kata-kata diucapkan bersama-sama. Sedangkan aturan prosodi menangani fluktuasi pada tekanan dan intonasi dalam sebuah kalimat. Dalam sebuah sistem pengolahan bahasa alami yang menerima masukan suara, gelombang suara masukan dianalisis dan dikodekan ke bentuk sinyal digital untuk penginterpretasian menggunakan berbagai aturan tersebut.

2. Analisis Morfologi

Kata-kata dalam sebuah teks dapat tersusun dari satu atau lebih morfem. Morfem adalah unit terkecil dalam sebuah teks yang memiliki arti. Sebagai contoh, kata “menerjemahkan” dapat dianalisis secara morfologis menjadi tiga morfem terpisah, yaitu awalan “me-“, kata dasar “terjemah”, dan akhiran “-kan”. Aplikasi pengolahan bahasa

alami hanya perlu mengenali arti dari seluruh morfem saja dan aturan pembentukan kata-kata dari morfem-morfem tersebut untuk dapat mengenali arti dari kata-kata yang tersusun dari lebih dari satu morfem, dan tidak perlu mengenali seluruh kata yang ada pada kamus.

3. Analisis Leksikal

Pada tingkatan analisis ini, sistem pengolahan bahasa alami menginterpretasikan arti dari kata-kata individu. Sistem pengolahan bahasa alami memecah masukan menjadi unit-unit kecil bernama *token* pada tingkatan analisis ini. *Token* ini dapat berupa kata, angka, atau tanda baca. Proses pemecahan masukan menjadi *token-token* ini disebut juga *tokenization*. Proses *tokenization* pada Bahasa Indonesia atau Bahasa Inggris cukup sederhana karena setiap kata umumnya dipisahkan oleh spasi atau tanda baca saja. Pada beberapa bahasa, seperti Bahasa Mandarin, kata-kata tidak dipisahkan oleh spasi sehingga proses *tokenization* menjadi lebih sulit.

4. Analisis Sintaksis

Tingkatan ini berfokus pada penganalisisan kata-kata dalam sebuah kalimat untuk menemukan struktur tata bahasa dari kalimat tersebut. Tingkatan ini membutuhkan sebuah aturan tata bahasa dan sebuah *parser*. Keluaran dari tingkat pemrosesan ini adalah sebuah representasi dari kalimat tersebut yang mengungkapkan hubungan ketergantungan struktural antara kata-kata. Sintaksis menyampaikan makna pada sebagian besar bahasa karena urutan dan ketergantungan kata-kata berkontribusi pada makna suatu kalimat. Misalnya kalimat “polisi menembak penjahat” dan kalimat “penjahat menembak polisi” hanya berbeda dalam hal sintaksis, namun memiliki arti yang cukup berbeda.

5. Analisis Semantik

Tingkatan analisis semantik berusaha mencari arti sebuah kata, frasa, atau kalimat yang paling tepat. Jika suatu kata, frasa, atau kalimat memiliki kemungkinan makna lebih dari satu, analisis semantik akan mencari makna yang paling sesuai dari kemungkinan-kemungkinan makna tersebut. Misal kata “rapat” dapat memiliki makna “pertemuan” atau memiliki makna “tidak renggang”, analisis semantik akan menentukan kemungkinan makna dari kata tersebut dengan melihat konteks penggunaan kata tersebut.

6. Analisis Wacana

Tingkatan analisis wacana menganalisis unit yang lebih besar dari kalimat dan hubungannya dengan konteks yang sedang dibahas. Tingkatan ini berusaha menemukan keteraturan linguistik dalam sebuah wacana menggunakan kriteria-kriteria tata bahasa, fonologi, dan semantik. Misalnya kata “dia” pada kumpulan kalimat “Danu menabrak mobil Dina. Akibat kecelakaan tersebut dia tewas” dapat berarti si “Danu” atau si “Dina”. Analisis wacana berusaha mencari tahu siapa yang dimaksud dengan “dia” pada kalimat tersebut dengan menganalisis kalimat-kalimat lainnya baik kalimat sesudah maupun sebelumnya.

7. Analisis Pragmatik

Tingkatan analisis ini menganalisis bagaimana mencari tahu makna tersembunyi dalam suatu teks meski tidak tertulis secara eksplisit. Tingkatan analisis ini membutuhkan pengetahuan umum yang luas mencakup pemahaman dari maksud, rencana, dan tujuan dari suatu teks. Misalnya kata “akar” dapat memiliki kemungkinan arti bagian dari tumbuhan atau memiliki arti salah satu operator matematika, pengetahuan pragmatis bahwa teks tersebut membahas adalah mengenai tumbuhan atau mengenai matematika dapat digunakan untuk mengetahui arti dari kata “akar” tersebut.

2.4.2. Pendekatan Pengolahan Bahasa Alami

Berdasarkan pendekatannya, Pengolahan Bahasa Alami dibagi menjadi *rule-based NLP*, *statistical NLP*, dan *hybrid NLP* (Wolniewicz, 2015). Pada *Rule-Based NLP* sekelompok pakar-pakar menulis aturan-aturan deterministik untuk mengimplementasikan pemetaan-pemetaan dalam komponen-komponen Pengolahan Bahasa Alami tersebut. Sebagian besar sistem Pengolahan Bahasa Alami menggunakan pendekatan *rule-based* untuk *tokenization* karena bekerja sangat baik untuk langkah tersebut dalam proses Pengolahan Bahasa Alami. Aturan-aturan adalah sebuah cara bagus untuk mengenkapsulasi pengetahuan-pengetahuan pakar tertentu. Namun *parsing* sulit untuk dikodekan, sehingga sebagian besar *parser* modern adalah berbasis statistik.

Statistical NLP artinya sistem mempelajari pemetaan terhadap komponen-komponen Pengolahan Bahasa Alami sebagai hubungan-hubungan statistik dengan memproses banyak contoh. Contoh dari *statistical NLP* adalah pemfilter *spam* pada

aplikasi *email*. Sebuah pemfilter *spam* melihat pada statistik keseluruhan sebuah kelompok besar dari masukan konsumen, sehingga keterlibatan pakar sangat minimal. Tingkat akurasi dari model statistik meningkat seiring dengan jumlah data yang tersedia untuk pembelajaran dan model statistik ini dapat terus belajar dengan ongkos yang lebih murah dibandingkan dengan sistem *rule-based*. Metode statistik membutuhkan data-data dalam jumlah sangat banyak untuk melatihnya agar menjadi lebih handal, sehingga metode ini tidak cocok digunakan jika jumlah data sebagai contohnya sangat terbatas.

Saat ini sebagian besar sistem Pengolahan Bahasa Alami menggunakan kombinasi kedua model, atau yang biasa disebut dengan pendekatan *hybrid*. *Hybrid NLP* mencoba menyatukan keunggulan dari kedua metode. Contohnya *tokenizer* lebih cocok menggunakan metode *rule-based*, dan jika tersedia data pembelajaran dalam jumlah yang sangat banyak, metode *statistical NLP* cenderung lebih mudah dan lebih murah untuk digunakan dibandingkan metode *rule-based*.

Pengembangan aplikasi pada penelitian ini menggunakan kedua pendekatan Pengolahan Bahasa Alami, yaitu *statistical NLP* pada proses penerjemahan menggunakan data pembelajaran statistik serta pada proses *speech recognition*, dan *rule-based NLP* pada proses penerjemahan menggunakan data pembelajaran berdasarkan contoh, proses penerjemahan menggunakan data pembelajaran berdasarkan *keyword*, proses penerjemahan menggunakan data *default*, serta pada proses *speech recognition*. Pengembangan aplikasi pada penelitian ini menggunakan kedua pendekatan Pengolahan Bahasa Alami karena kedua pendekatan Pengolahan Bahasa Alami tersebut memiliki keunggulan masing-masing. *Rule-based NLP* dapat dengan mudah diprogram untuk melakukan proses Pengolahan Bahasa Alami tanpa perlu memasukkan data *training* dalam jumlah yang sangat banyak, namun sulit dilatih untuk mempelajari masukan baru seperti pada *statistical NLP*. Meskipun *statistical NLP* mudah dilatih untuk mempelajari masukan baru, *statistical NLP* membutuhkan data *training* yang sangat banyak untuk dapat melakukan proses Pengolahan Bahasa Alami dengan hasil yang cukup memuaskan.

2.4.3. Aplikasi Pengolahan Bahasa Alami

Ada banyak aplikasi dari Pengolahan Bahasa Alami, di bawah ini adalah daftar enam aplikasi Pengolahan Bahasa Alami yang umum dan menjadi topik penelitian yang populer:

1. *Machine Translation.*

Machine Translation mengkonversi teks dalam satu bahasa menjadi bahasa lainnya. Contoh aplikasi ini adalah layanan penerjemah *online Google Translate*. Pada awal perkembangannya, *Machine Translation* dikembangkan menggunakan metode *rule-based* yang kurang menghasilkan kesuksesan. Bidang ini bertransformasi pada awal tahun 1990-an ketika peneliti-peneliti dari IBM memperoleh teks paralel Bahasa Inggris dan Bahasa Prancis dalam jumlah besar dan menggunakannya untuk mengembangkan sistem *Machine Translation* Bahasa Prancis-Bahasa Inggris menggunakan metode statistik (Brown dkk., 1993). Pada *Machine Translation* dengan metode statistik, sebuah dokumen ditranslasikan berdasarkan distribusi probabilitas $p(i/e)$ bahwa sebuah *string* i di bahasa tujuan (misal Bahasa Indonesia) adalah hasil translasi dari *string* e dalam bahasa sumber (misalnya Bahasa Inggris). Salah satu pendekatan untuk pemodelan distribusi probabilitas $p(i/e)$ adalah dengan penerapan Teorema Bayes, yaitu bahwa nilai $p(i/e)$ proporsional dengan nilai $p(e/i) * p(i)$, dimana $p(e/i)$ adalah kebalikan dari $p(i/e)$, yaitu probabilitas *string* e adalah hasil translasi dari *string* i , dan $p(i)$ adalah probabilitas kemunculan *string* i pada teks yang tertulis dalam bahasa tujuan. Pencarian hasil translasi terbaik dilakukan dengan mencari nilai i yang memaksimalkan nilai $p(e/i) * p(i)$ (Brown dkk., 1993). Beberapa tahun belakangan ini berkembang pendekatan *Machine Translation* yang cukup menjanjikan dengan penggunaan *deep learning* (Hirschberg dan Manning, 2015).

2. *Speech recognition.*

Speech recognition adalah sistem Pengolahan Bahasa Alami yang mengkonversi suara manusia menjadi bentuk teks. Pendekatan *speech recognition* yang paling umum dan yang terbaik saat ini adalah menggunakan pendekatan probabilistik. *Speech recognition* berbeda dengan *voice recognition*. Pada *voice recognition*, mesin berusaha menebak identitas sang pemilik suara, sedangkan pada *speech recognition* mesin berusaha memahami maksud dari ucapan seseorang.

3. *Speech synthesis*

Speech synthesis adalah kebalikan dari *speech recognition*. Pada *speech synthesis*, mesin menghasilkan suara dengan makna yang sama dengan masukan teks dari *user*. Dua teknik yang paling sering digunakan pada *speech synthesizer* adalah *formant synthesis* dan *concatenation synthesis*. Pada *concatenation synthesis*, mesin membutuhkan *database* contoh suara manusia untuk menghasilkan *output* suara, sedangkan pada *formant synthesis*

tidak dibutuhkan contoh suara manusia, sehingga teknik *concatenation synthesis* menghasilkan *output* suara yang lebih alamiah.

4. *Question answering.*

Aplikasi Pengolahan Bahasa Alami ini mengekstrak informasi relevan dari kumpulan teks untuk menjawab pertanyaan yang diajukan dalam bahasa alami. Contoh penerapan yang paling terkenal adalah IBM Watson yang mampu mengalahkan juara manusia pada acara kuis Jeopardy!

5. *Classification.*

Aplikasi Pengolahan Bahasa Alami ini mengurutkan dan mengorganisasikan informasi ke kelompok kategori-kategori yang sesuai (Wolniewicz, 2015). Contoh aplikasinya adalah filter *spam* pada aplikasi *email* dan Google News yang mengelompokkan teks ke dalam kategori-kategori yang sesuai seperti *headline* olahraga, skor sepakbola, ulasan film, berita mengenai perang di Suriah, dan lain-lain.

6. *Text Summarization.*

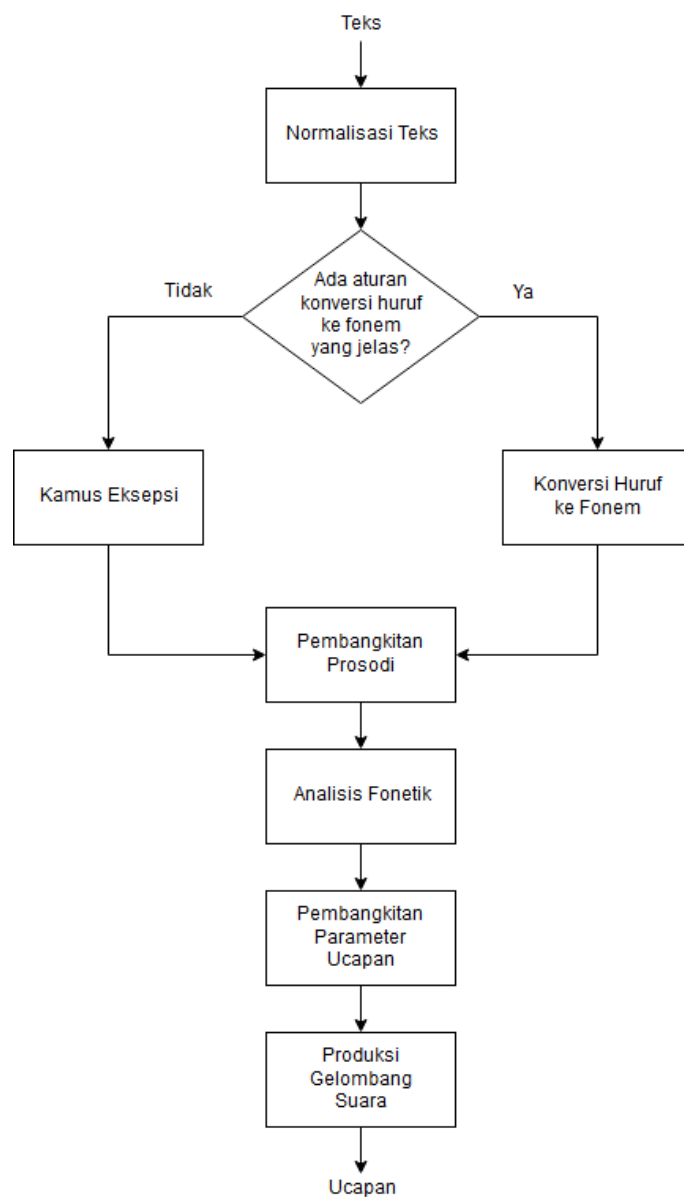
Sistem Pengolahan Bahasa Alami ini berusaha membuat ringkasan dari sebuah teks. Ringkasan dapat didefinisikan sebagai sebuah teks yang diproduksi dari satu atau lebih teks yang menyampaikan informasi-informasi penting dalam teks sumber dan umumnya tidak lebih panjang dari setengah teks sumbernya (Radev dkk., 2002). *Text summarization* biasanya dilakukan pada tingkat analisis wacana.

Pada subbab berikut ini dijelaskan lebih detail mengenai *speech synthesis* dan *speech recognition*, karena kedua aplikasi Pengolahan Bahasa Alami tersebut yang digunakan pada sistem yang dikembangkan pada penelitian ini.

2.4.4. *Speech Synthesis*

Sistem *speech synthesis* terdiri dari dua bagian, yaitu bagian konverter teks ke fonem yang mengubah teks masukan menjadi rangkaian kode-kode bunyi yang direpresentasikan dengan kode fonem, *pitch*, dan durasinya serta bagian konverter fonem ke ucapan yang menerima masukan berupa kode-kode bunyi yang dihasilkan oleh bagian sebelumnya untuk menghasilkan ucapan yang sesuai dengan teks masukan (Arman, 2003).

Ada dua teknik yang banyak digunakan untuk mengimplementasikan konverter fonem ke ucapan, yaitu *formant synthesizer* dan *diphone concatenation*. *Formant synthesizer* melakukan komputasi untuk menghasilkan sinyal ucapan berdasarkan suatu model matematis, sedangkan *diphone concatenation* menghasilkan sinyal ucapan dengan cara menggabung-gabungkan segmen-segmen bunyi berupa *diphone* (gabungan dua buah fonem) yang telah direkam sebelumnya dan tersimpan di dalam *diphone database*. *Diphone concatenation* menghasilkan bunyi yang terdengar lebih alamiah dibandingkan *formant synthesizer*.



Gambar 2.3. Urutan proses konversi dari teks ke ucapan (Arman, 2003)

Tahapan-tahapan utama konversi dari teks menjadi ucapan dapat dilihat pada gambar 2.3. di atas. Tahap normalisasi teks mengubah semua teks masukan menjadi teks yang secara lengkap menunjukkan cara pengucapannya. Tahap berikutnya melakukan konversi dari teks yang dihasilkan pada tahap sebelumnya menjadi kode-kode fonem. Konversi teks menjadi kode-kode fonem dapat dilakukan dengan konversi sederhana dan berlaku umum untuk berbagai kondisi, atau dengan aturan konversi yang bersifat kondisional tergantung dari huruf-huruf atau fonem-fonem tetangganya, bahkan terdapat bentuk-bentuk translasi yang tidak dapat ditemukan keteraturannya. Konversi sederhana dapat diimplementasikan dengan tabel konversi yang berisi pasangan antara urutan huruf dan urutan fonem. Aturan konversi yang bersifat kondisional biasanya diimplementasikan dengan tabel konversi yang diterapkan jika kondisi rangkaian huruf tetangga kiri dan kanannya terpenuhi. Bahasa Indonesia termasuk contoh bahasa yang sebagian besar kata-katanya dapat dikonversikan menjadi fonem dengan aturan yang sederhana. Namun ada kondisi-kondisi yang tidak dapat ditemukan keteraturannya, seperti simbol huruf e yang dapat diucapkan sebagai e pepet atau e taling. Pada diagram gambar 2.3. di atas, konversi yang memiliki aturan yang jelas ditangani oleh bagian Konversi Huruf ke Fonem, sedangkan konversi yang tidak memiliki keteraturan ditangani oleh bagian Kamus Eksepsi. Setelah rangkaian fonem yang merepresentasikan bunyi kalimat telah dihasilkan, bagian Pembangkitan Prosodi melengkapi setiap unit fonem yang dihasilkan dengan data *pitch* serta durasi pengucapannya. Tahap berikutnya adalah tahap Analisis Fonetik yang melakukan perbaikan di tingkat bunyi, misalnya seperti memberi sisipan fonem /y/ dalam pengucapan kata ilmiah di antara fonem /i/ dan /a/.

Pada penelitian ini digunakan sebuah aplikasi *open source* yang menerapkan *speech synthesis* yang bernama eSpeak. Secara *default* eSpeak menggunakan teknik *formant synthesis* dalam menghasilkan suara dan untuk saat ini mampu menghasilkan suara untuk 100 bahasa dan dialek berbeda. Suara yang dihasilkan cukup jernih dan cukup cepat dalam menghasilkan suara, namun tidak menghasilkan suara se-alamiah suara yang dihasilkan aplikasi *speech syntesis* lainnya yang menghasilkan suara dengan menggunakan teknik *diphone concatenation* ("eSpeak text to speech", n.d.).

2.4.5. *Speech recognition*

Speech recognition adalah sistem yang mengkonversi suara dari sebuah rekaman sinyal audio menjadi teks. Salah satu metode *Speech recognition* yang paling umum dan

yang terbaik saat ini adalah menggunakan pendekatan probabilistik. Pada pendekatan ini, setiap *input* suara yang dimasukkan dicari hasil transkripsinya dengan mencari kata-kata atau kumpulan-kumpulan pada kamus yang memiliki probabilitas tertinggi merupakan hasil transkripsi dari *input* suara tersebut (Gruhn dkk., 2011). Masalah-masalah yang bisa muncul pada *speech recognition* adalah gangguan kebisingan latar belakang, perbedaan logat *user*, jumlah kosakata sebuah bahasa yang sangat besar, pendeteksian suara-suara yang bukan merupakan kata seperti suara batuk, suara bersin, dan lain lain.

Tingkat akurasi *speech recognition* dapat bervariasi karena beberapa alasan sebagai berikut (Tebelskis, 1995):

1. Ukuran kosakata

Sistem lebih mudah mengenali kata menggunakan kamus dengan jumlah kosakata yang sedikit. Tingkat akurasi berkurang seiring dengan penambahan kosakata pada kamus. Hal ini disebabkan karena dengan bertambahnya kosakata semakin membuat sistem kebingungan memilih kata yang sesuai terutama kata-kata yang memiliki ucapan yang sangat mirip, misalnya kata "tubuh" dengan kata "tuduh", sehingga sistem kadang memberikan *output* kata yang tidak sesuai namun mirip pengucapannya.

2. *Speaker dependence* atau *speaker independence*

Sistem *speaker dependent* berarti sistem *speech recognition* tersebut didesain hanya mampu mengerti ucapan dari orang-orang tertentu saja, biasanya orang-orang yang suaranya menjadi sumber latihan sistem tersebut. Sedangkan sistem *speaker independent* didesain agar mampu mengerti ucapan siapapun dengan tingkat akurasi yang cukup memuaskan. Sistem *speaker dependent* biasanya memiliki tingkat akurasi yang lebih baik dibandingkan sistem *speaker independent* meski jumlah *user* yang dapat menggunakannya hanya terbatas.

3. *Isolated, discontinuous*, atau *continuous speech*

Beberapa sistem *speech recognition* didesain untuk hanya menerima masukan *isolated speech* maksudnya hanya mencoba mengenali satu kata dalam satu waktu saja. Sistem *speech recognition* lainnya didesain untuk hanya menerima masukan *discontinuous speech*, yaitu mampu mengenali beberapa kata dalam satu waktu, namun antar kata yang berbeda harus dipisahkan dengan keheningan selama

beberapa waktu tertentu. Sistem *speech recognition* yang didesain untuk menerima masukan *continuous speech* mampu mengenali rangkaian beberapa kata dalam satu waktu yang diucapkan secara kontinyu tanpa terputus-putus dan sistem sendiri lah yang mencoba menebak batas antar kata. Sistem *speech recognition* yang didesain untuk hanya menerima masukan *isolated speech* dan *discontinuous speech* memberikan tingkat akurasi yang lebih tinggi namun *user* tidak dapat berbicara secara alamiah, sedangkan *user* dapat berbicara secara alamiah tanpa terputus-putus pada *speech recognition* yang didesain untuk menerima masukan *continuous speech*.

4. Batasan bahasa

Maksud batasan bahasa disini adalah sistem hanya didesain untuk mengenali kalimat-kalimat tertentu saja yang sudah dilatihkan kepadanya. Misalnya sistem *speech recognition* untuk menyalakan lampu hanya menerima ucapan "lampu menyala" dan menolak ucapan "dimana dompet saya". Batasan lainnya bisa berupa semantik, misal menolak kalimat "motor saya tertawa" atau bisa juga sintaksis, misalnya menolak kalimat "saya tertawa motor". Batasan direpresentasikan oleh *grammar* yang memfilter kalimat-kalimat yang tidak masuk akal. *Grammar* biasanya dinilai berdasarkan *perplexity*, yaitu angka yang menunjukkan faktor percabangan rata-rata dari *grammar*. Tingkat *perplexity* suatu *grammar* menentukan tingkat kesulitan suatu sistem dalam mengenali suara.

5. Ucapan spontan atau hasil membaca

Sistem lebih mudah memahami ucapan yang dihasilkan dari membaca ketimbang ucapan yang dihasilkan secara spontan karena ucapan-ucapan seperti "uh", "um", suara batuk, suara tertawa, dll. dapat muncul.

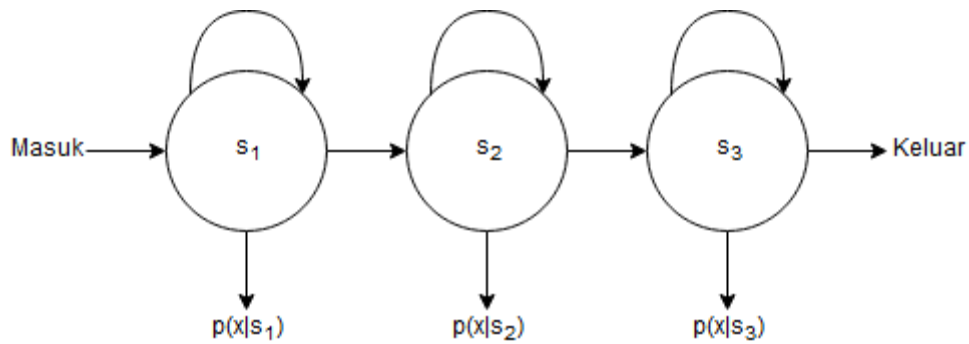
6. Kualitas perekaman

Performa sistem *speech recognition* dapat dipengaruhi oleh kebisingan latar belakang, gaung, gema, jenis mikrofon, dan lain-lain.

Pada penelitian ini digunakan Sphinx4 sebagai fitur *speech recognition* pada aplikasi yang dikembangkan. Sphinx4 bersifat *speaker independent* dan dibangun menggunakan *Gaussian Mixture Models* bersama-sama dengan *Hidden Markov Models*

(GMM-HMM) ("*Basic concepts of speech recognition*", n.d.). Sphinx4 dapat dilatih untuk mengenali berbagai macam bahasa. Tiga model digunakan dalam Sphinx4 untuk melakukan *speech recognition*, yaitu model akustik, kamus fonetik, dan model bahasa ("*Basic concepts of speech recognition*", n.d.). Model akustik berisi sifat-sifat akustik untuk tiap-tiap *senone*. Kamus fonetik berisi pemetaan dari kata-kata ke fon-fon penyusunnya. Model bahasa digunakan untuk membatasi pencarian kata-kata, yaitu dengan menebak kata-kata apa yang dapat muncul setelah kata-kata yang sudah dikenali sebelumnya, sehingga kata-kata yang tidak mungkin muncul diabaikan dalam pencarian. Model bahasa dapat menggunakan pendekatan *grammar* maupun statistik ("*Building a language model*", n.d.). *Grammar* untuk model bahasa dapat dibuat menggunakan Java Speech Grammar Format (JSGF). *Grammar* memungkinkan pengembang untuk menentukan masukan yang sesuai dengan sangat tepat, misalnya kata A hanya bisa ditulis setelah kata B, kata C hanya dapat diulang sebanyak dua kali, dan sebagainya. Namun, dalam percakapan alami sehari-hari seseorang sering tidak berkata sesuai dengan aturan-aturan tersebut, sehingga saat ini pengembang aplikasi lebih suka menggunakan model bahasa dengan pendekatan statistik ketimbang dengan pendekatan *grammar* ("*Building a language model*", n.d.). Model bahasa dengan pendekatan statistik mengandung probabilitas dari kata-kata dan kombinasi-kombinasi kata. Probabilitas-probabilitas tersebut dihitung dari data-data contoh berupa kumpulan *corpus* pada suatu bahasa. Model bahasa menggunakan statistik direkomendasikan untuk masukan yang lebih bebas. Model bahasa memiliki masalah pada pengenalan nama. Untuk mengatasi hal ini, sebuah model bahasa dapat berisi potongan lebih kecil seperti suku kata atau bahkan fon, namun tingkat akurasi *speech recognition* nya menjadi lebih rendah dibandingkan dengan model bahasa yang berbasis kata.

HMM adalah proses stokastik ganda dengan sebuah proses stokastik yang mendasarinya tidak dapat diobservasi langsung dan hanya dapat diobservasi melalui himpunan proses stokastik lainnya yang memproduksi urutan dari simbol-simbol yang diobservasi (Singh dkk., 2012). Proses stokastik atau proses acak adalah sebuah objek matematika yang biasanya didefinisikan sebagai sebuah kumpulan dari variabel-variabel acak. Gambar 2.4. dibawah mengilustrasikan sebuah HMM tiga *state* dimana tiap-tiap *state* s_i memiliki sebuah kepadatan probabilitas $p(x|s_i)$ yang menyatakan kepadatan probabilitas untuk observasi akustik x untuk *state* s_i . Tiga *state* s_1 , s_2 , dan s_3 membentuk kata S.



Gambar 2.4. Hidden Markov Model tiga *state*.

Gaussian Mixture Model (GMM) adalah fungsi kepadatan probabilitas parametrik yang direpresentasikan sebagai penjumlahan berbobot dari kepadatan komponen *Gaussian* (Vyas, 2013). GMM umumnya digunakan sebagai sebuah model parametrik dari distribusi probabilitas pada pengukuran kontinyu atau fitur-fitur dalam sebuah sistem biometrik, seperti fitur spektral dalam sistem *speech recognition*. Parameter-parameter GMM diperkirakan dari data pelatihan menggunakan algoritma Ekspektasi-Maksimalisasi (EM) iteratif atau estimasi *Maximum A Posteriori* (MAP) dari model sebelumnya yang telah dilatih. GMM yang lengkap diparameterisasi oleh vektor rata-rata, matriks kovariansi dan bobot campuran dari semua kepadatan komponen.