Pengembangan Kamus Jenis Kata Yang Dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia

Yoppy Yansyah
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
yoppy787@gmail.com

Ken Kinanti P
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
guinev_gs@yahoo.co.nz

Pada penelitian tentang "Pembangunan Kamus Jenis Kata Sebagai Resource NLP Bahasa Indonesia", resource yang dihasilkan sebanyak 38.870 lema terdiri dari kata dasar dan turunannya, belum menghasilkan bentuk kata majemuk sebagai penambahan sumber daya. Resource kata majemuk sangat dibutuhkan untuk penelitian tentang NLP seperti tokenisasi kata pada POS Tag dan question answering. Berdasarkan kebutuhan akan sumber daya maka dilakukan pengembangan kamus jenis kata yang dilengkapi kata majemuk yang dapat digunakan dalam pengembangan dalam bidang NLP.

Pada penelitian sebelumnya data masukan kbbi berformat .txt, sehingga tidak dapat mendeteksi lema yang dicetak tebal. Oleh sebab itu dilakukan proses konversi format data masukan dari kbbi.pdf menjadi kbbi.doc dan menjadi kbbi.html dan dilakukan dengan 3 proses yaitu 5 tahapan proses merapikan data masukan terdiri dari penghapusan karater html entitas, penghapusan label kecuali label jenis kata, penghapusan nomor makna, penyamaan simbol kata majemuk dan penyatuan kata terpisah. Dalam proses memisahkan kata dilakukan 3 tahapan terdiri dari, pemisahan kata majemuk, pengambil kata dan jenis kata, dan pemisahan kata dasar dengan kata turunannya. Dalam proses mendeteksi kata majemuk dilakukan 4 tahapan terdiri dari penghapusan baris kosong, pengisian kata dan jenis kata untuk kata majemuk, pembersihan dan pengkategorian.

Hasil akhir dari pengembangan kamus jenis kata yang dilengkapi kata majemuk adalah menghasilkan lema sebanyak 51.147 lema terdiri dari 10.609 lema kata majemuk dan 40.439 lema bukan kata majemuk. kamus jenis kata yang dilengkapi kata majemuk sudah tersedia dan dapat dimanfaatkan untuk penelitian NLP (Natural Language Processing).

Kata kunci : Kamus jenis kata, kamus online, resource NLP, kata majemuk.

I. PENDAHULUAN

Kamus merupakan deskripsi kosakata dari suatu bahasa. Kamus menjelaskan definisi bagaimana kata itu bekerja sama untuk membentuk kalimat. Kata adalah kumpulan beberapa huruf yang memiliki makna tertentu. Dalam KBBI (Kamus Besar Bahasa Indonesia) kata adalah unsur bahasa yang diucapkan atau dituliskan yang merupakan perwujudan suatu perasaan dan pikiran yang dapat dipakai dalam berbahasa. [1] Bahasa Indonesia memiliki beberapa ragam kata yang diklasifikasikan berdasarkan bentuknya, diantaranya adalah kata dasar, kata turunan, kata majemuk, dan kata ulang. Kata majemuk adalah gabungan dua kata atau lebih yang membentuk makna baru. Contohnya meja hijau jika kata tersebut digabungkan akan menghasilkan kata majemuk sehingga memiliki makna baru. [2] Kamus Besar Bahasa Indonesia sejak saat diterbitkannya pada edisi pertama tahun 1988 memuat 62.000 lema, pada edisi ke empat tahun 2008 mengalami penambahan hingga 90.049 lema terdiri dari 41.250 lema pokok dan 24.607 kata berimbuhan, 23.536 gabungan kata sebagai sublema.

Perkembangan kamus saat ini tidak hanya dalam bentuk buku saja, namun kamus dapat diakses secara digital. Antara lain; kbbi.web, kamusbahasaindonesia.org, kateglo.com dan lain-lain. Tetapi saat ini pengembangan kamus secara digital tidak memberikan resource (sumber daya) dan hak akses penggunanya, sehingga pengguna hanya bisa menggunakan saja tidak bisa mengubah atau menambahkan. Resource dalam kamus jenis kata yang dilengkapi kata majemuk sangat dibutuhkan untuk pengembangan kamus itu sendiri ataupun dimanfaatkan untuk penelitian tentang NLP (Natural Language Processing) seperti tokenisasi kata pada POS Tag dan question answering.

Pada penelitian tentang "Pembangunan Kamus Jenis Kata Sebagai Resource NLP Bahasa Indonesia" [3], resource yang dihasilkan sebanyak 38.870 lema yang terdiri dari kata dasar dan turunannya, belum menghasilkan bentuk kata majemuk sebagai penambahan sumber daya.

Berdasarkan pada latar belakang masalah yang sudah dipaparkan, penyusun memiliki usulan untuk pengembangan penelitian sebelumnya dengan menambahkan jenis kata dari kata majemuk sebagai resource (sumberdaya) NLP.

II. DATA MASUKAN

Pada peneletian sebelumnya data masukan dengan format .txt tidak dapat mendeteksi lema yang dicetak tebal, sehingga dilakukan konversi format data masukan dari kbbi.pdf diubah menjadi kbbi.doc menggunakan tools Nitro11. Setelah diubah menjadi kbbi.doc dilakukan pemisahan data secara perabjad atau satu persatu dari A sampai Z. Proses terakhir, dilakukan konversi format data dari format .doc ke format .html menggunakan tools html cleaner. Karena data masukan dengan format .html dapat mendeteksi lema yang dicetak tebal.

Tabel 1 Contoh Data Masukan

```
<p><strong>1abid </strong><em>ark a
</em>kekal; abadi; tidak berkesu- dahan</p>
<p><strong>2abid </strong><em>n IsI
</em><strong>1 </strong>orang yg beribadah;
<strong>2 </strong>ber- ibadah; <strong>3
</strong>taat kpd Tuhan; saleh</p>
<p><strong>abilah </strong><em>ark n
</em>cacar; ketumbuhan;</p>
<p><strong>-- peringgi </strong>penyakit raja
singa; sifilis</p>
<p><strong>abece </strong>/ab&eacute;c&eacute;/
<em>n </em>abc (alfabet Latin)</p>
<p><strong>akasia </strong><em>n </em>pohon
atau perdu yg tumbuh di daerah  tropis
 dng  rangkaian  bunga
 ku-</p>
<p>ning atau putih;</p>
<p>-- <strong>kuning </strong>tanaman hias dng
bunga ber- warna kuning pd bulir yg bergantung,
<em>Acecia auriculiformia</em></p>
<p><strong>asap </strong><em>n </em>gas yg
tampak keluar dr sesuatu yg panas, mendidih, atau
dibakar;</p>
<p><strong>mengasap </strong><em>v
</em><strong>1 </strong>menjadi asap; <strong>2
</strong>memasak (mengeringkan, membersihkan,
mengha- rumkan, dsb) dng asap;</p>
<p>~ <strong>daging </strong>memanggang
daging; ~ <strong>pakai-</strong></p>
<p><strong>an </strong>mengharumkan pakaian
dng asap ra- tus  dsb;  ~ 
<strong>nyamuk  </strong>mengusir 
nyamuk dng  asap;  </p>
<p><strong>baca, </strong><strong>membaca
</strong><em>v </em><strong> </strong>melihat
serta memahami isi dr apa yg tertulis (dng
melisankan atau hanya dl hati): <em>jangan
diganggu, ia sedang </em>~ <em>buku</p>
<p>mengeja atau melafalkan apa yg ter-tulis;</p>
<p><strong>~ batin </strong>membaca dl hati;</p>
<p><strong>~ </strong><strong>bibir
</strong><em> </em>mencoba mengerti
pembicaraan lewat gerak bibir</p>
<p><strong>kolaret /</strong>kolart/ <em>n
</em>garis, batas, atau rantai yg mengelilingi
leher</p>
<p><strong>mahabesar </strong><em>a
</em>tidak ada yg melebihi kebe- sarannya; besar
yg tidak ada batasnya</p>
```

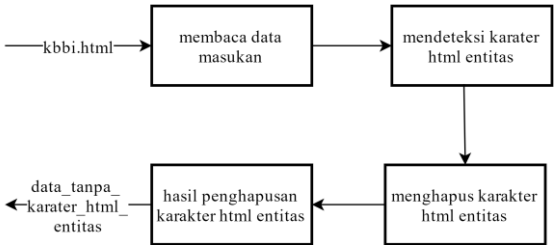
III. METODE DAN IMPLEMENTASI

Metode yang digunakan adalah metode role-based, tahapan proses yang dilakukan dengan 3 proses yaitu 5 tahapan proses merapikan data masukan terdiri dari penghapusan karater html entitas, penghapusan label kecuali label jenis kata, penghapusan nomor makna, penyamaan simbol kata majemuk dan penyatuan kata terpisah. Dalam proses memisahkan kata dilakukan 3 tahapan terdiri dari, pemisahan kata majemuk, pengambil kata dan jenis kata, dan pemisahan kata dasar dengan kata turunannya. Dalam proses mendeteksi kata majemuk dilakukan 4 tahapan terdiri dari penghapusan baris kosong, pengisian kata dan jenis kata untuk kata majemuk, pembersihan dan pengkategorian.

A. Merapikan Data Masukan

1. Penghapusan Karakter Html Entitas

Proses penghapusan karater html entitas seperti   yang berfungsi untuk spasi bertujuan membantu pada proses pemisahan kata majemuk. karena jika tidak dilakukan maka akan dianggap sebagai simbol penanda akhir (;) akan terdeksi.



Gambar 2 Penghapusan Karakter Html Entitas

Contoh seperti dibawah ini.

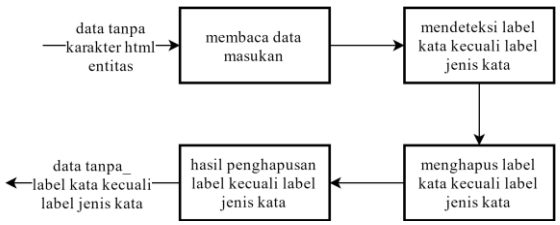
Tabel 2 Penghapusan Karakter Html Entitas

Data Sebelum
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakai-   an mengharumkan pakaian dng asap ra- tus dsb;   ~ nyamuk mengusir   nyamuk dng   asap;
Data Sesudah
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan,

mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap;

2. Proses Penghapusan Label Kecuali Label Jenis Kata

Proses penghapusan label kata seperti ark, Ar, Filol karena tidak diperlukan untuk hasil kamus jenis kata yang dilengkapi kata majemuk.



Gambar 3 Blok Diagram Penghapusan Label Kata Kecuali Label Jenis Kata

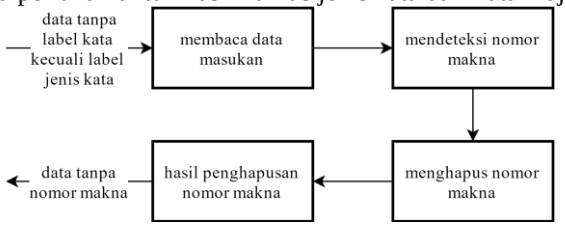
Contoh seperti dibawah ini.

Tabel 3 Penghapusan Label Kata Kecuali Label Jenis Kata

Sebelum
1abid ark akekal; abadi; tidak berkesu- dahan   2abid n IsI 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh
Sesudah
1abid a kekal; abadi; tidak berkesu- dahan   2abid n 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh

3. Penghapusan Nomor Makna

Proses penghapusan nomor bertujuan untuk menghilangkan nomor dari [0-9] karena nomor makna tidak diperlukan untuk hasil kamus jenis kata dari kata majemuk.



Gambar 4 Penghapusan Nomor Makna

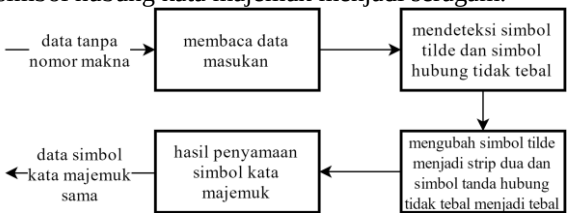
Contoh seperti dibawah ini.

Tabel 4. Penghapusan Nomor Makna

Sebelum
1 abid a kekal; abadi; tidak berkesu- dahan
2 abid n 1 orang yg beribadah; 2 ber- ibadah; 3 taat kpd Tuhan; saleh
Sesudah
abid a kekal; abadi; tidak berkesu- dahan   abid n orang yg beribadah; ber- ibadah; taat kpd Tuhan; saleh

4. Penyamaan Simbol Kata Majemuk

Maksud dari proses penyamaan simbol kata majemuk adalah mengubah simbol tilde (~) menjadi strip dua (--) dan mengubah simbol hubung kata majemuk yaitu simbol (~) dan strip dua (--) yang tidak tebal menjadi tebal bertujuan untuk simbol hubung kata majemuk menjadi seragam.



Gambar 5 Blok Diagram Penyamaan Simbol Kata Majemuk

Contoh seperti dibawah ini

Tabel 5 Penyamaan Simbol Kata Majemuk

Sebelum
akasia n pohon atau perdu yg tumbuh di daerah tropis dng rangkaian bunga ku-   ning atau putih;   -- kuning tanaman hias dng bunga ber- warna kuning pd bulir yg bergantung, Acecia auriculiformia mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap;
Sesudah
akasia n pohon atau perdu yg tumbuh di daerah tropis dng rangkaian bunga ku-   ning atau putih;   -- kuning tanaman hias dng bunga ber- warna kuning pd bulir yg bergantung, Acecia auriculiformia mengasap

v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~</nyamuk> nyamuk mengusir nyamuk dng asap;

5. Penyatuan Kata Terpisah

Maksud dari proses penyatuan kata terpisah adalah kata yang terpisah menjadi kata yang utuh.



Gambar 6 Blok Diagram Penyatuan Kata Terpisah

Contoh seperti dibawah ini.

Tabel 6 Penyatuan Kata Terpisah

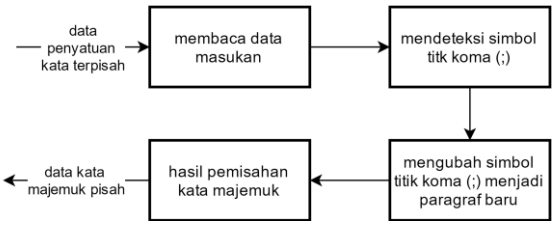
Sebelum
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging; ~pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;
Sesudah
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging;~pakaianmengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;

B. Memisahkan Kata

Proses memisahakan kata dasar dengan kata turunanya dan memisahkan kata majemuk dilakukan dengan 3 tahap yaitu.

1. Pemisahan Kata Majemuk

Proses pemisahan kata majemuk bertujuan untuk kata majemuk dalam satu paragraf menjadi paragraf yang mempunyai satu kata majemuk.



Gambar 7 Blok Diagram Pemisahan Kata Majemuk

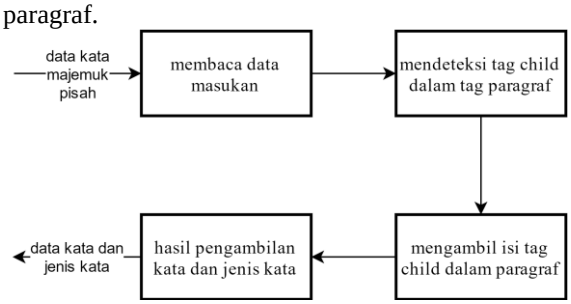
Contoh seperti tabel dibawah ini.

Tabel 7 Pemisahan Kata Majemuk

Sebelum
mengasap v menjadi asap; memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~daging memanggang daging; ~pakaianmengharumkan pakaian dng asap ra- tus dsb; ~nyamuk mengusir nyamuk dng asap;
Sesudah
mengasap v menjadi asap   memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap   ~daging memanggang daging   ~pakaianmengharumkan pakaian dng asap ra- tus dsb   ~nyamuk mengusir nyamuk dng asap

2. Pengambilan Kata dan Jenis Kata

Maksud dari proses pengambila kata dan jenis kata adalah mengabil kata dan jenis katanya dari tag child dalam tag paragraf.



Gambar 8. Blok Diagram Pengambilan Kata dan Jenis Kata

Contoh seperti tabel dibawah ini.

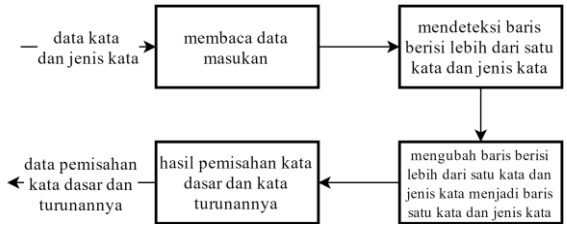
Tabel 8 Pengambilan Kata dan Jenis Kata

Sebelum
mengasap v menjadi asap   memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap

--daging memanggang daging   --pakaian mengharumkan pakaian dng asap ra- tus dsb   --nyamuk mengusir nyamuk dng asap
Sesudah
mengasap v --daging --pakaian --nyamuk

3. Pemisahan Kata Dasar dengan Kata Turunannya

Proses pemisahan kata dasar dengan kata turunanya bertujuan untuk baris yang lebih dari satu kata dan jenis kata di dalam satu baris menjadi satu kata dan jenis kata di dalam baris.



Gambar 9 Blok Diagram Pemisahan Kata Dasar dengan Kata Turunannya

Contoh seperti tabel dibawah ini.

Tabel 9 Pemisahan Kata Dasar dengan Kata Turunannya

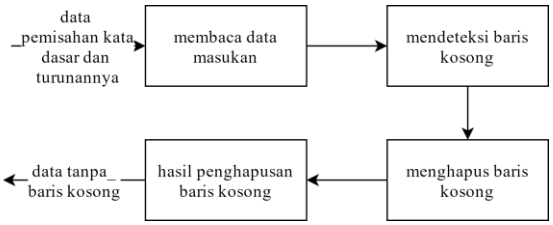
Sebelum
baca, membaca v --batin --bibir
Sesudah
baca, v membaca v --batin --bibir

2.6.3 Mendeteksi Kata Majemuk

Proses melengkapi kata majemuk dilakukan untuk mengisi kata dan jenis kata untuk kata majemuk. Dilakukan dengan 3 tahapan yaitu:

1. Penghapusan Baris Kosong

Proses penghapusan baris kosong bertujuan untuk menghapus baris yang tidak terisi kalimat.



Gambar 10 Penghapusan Bari Kosong

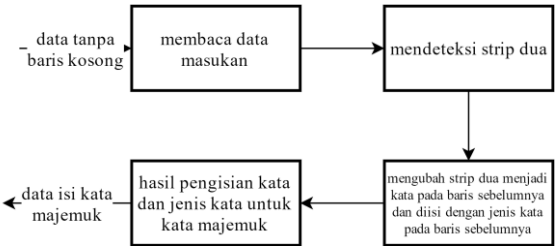
Contoh seperti tabel dibawah ini.

Tabel 10 Penghapusan Baris Kosong

Sebelum
abid a abid n abilah n --peringgi
Sesudah
abid a abid n abilah n --peringgi

2. Pengisian Kata dan Jenis Kata Untuk Kata Majemuk

Maksud dari proses pengisian kata dan jenis kata majemuk adalah mengganti simbol strip dua (--) menjadi kata pada baris sebelumnya dan mengisi jenis kata utama pada baris sebelumnya agar menjadi kata majemuk.



Gambar 11 Blok Diagram Pengisian Kata dan Jenis Kata Untuk Kata Majemuk

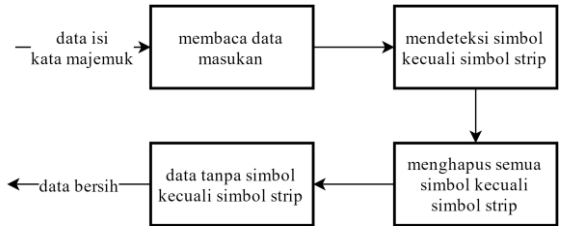
Contoh seperti tabel dibawah ini.

Tabel 11 Pengisian Kata dan Jenis Kata Untuk Kata Majemuk

Sebelum
baca, v membaca v --batin --bibir
Sesudah
baca, v membaca v membaca batin v membaca bibir v

3. Pembersihan

Maksud dari proses pembersihan adalah menghapus semua simbol kecuali strip. Proses ini dilakukan pada proses melengkapi kata majemuk karena jika dilakukan pada saat proses merapikan data masukan tanda tag html akan terhapus.



Gambar 12 Blok Diagram Pembersihan

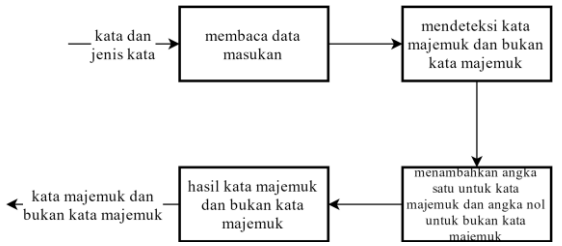
Contoh seperti tabel dibawah ini.

Tabel 12 Pembersihan

Sebelum
baca, v membaca v membaca batin v membaca bibir v koloret /n
Sesudah
baca v membaca v membaca batin v membaca bibir v koloret n
mengasap v mengasap daging v mengasap pakaian v mengasap nyamuk v baca v membaca v membaca batin v membaca bibir v koloret n mahabesar a
Sesudah
abid a 0 abid n 0 abilah n 0 abilah peringgi n 1 abece n 0 akasia n 0 akasia kuning n 1 asap n 0 mengasap v 0 mengasap daging v 1 mengasap pakaian v 1 mengasap nyamuk v 1 baca v 0 membaca v 0 membaca batin v 1 membaca bibir v 1 koloret n 0 mahabesar a 0

4. Pengkategorian

Maksud pengkategorian adalah menentukan kata majemuk dan bukan kata majemuk dengan tada keterangan untuk kata majemuk ditambah dengan angka satu dan bukan kata majemuk ditambah dengan angka nol.



Gambar 13 Blok Diagram Pengkategorian

Contoh seperti tabel dibawah ini.

Tabel 13 Pengkategorian

Sebelum
abid a abid n abilah n abilah peringgi n abece n akasia n akasia kuning n asap n

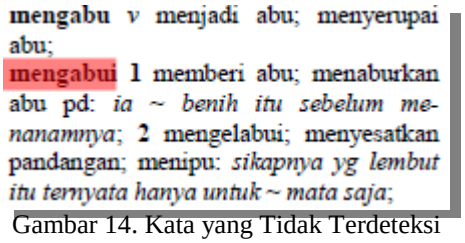
IV. HASIL PENGUJIAN

Hasil pengambilan kata dan jenis kata yang dilengkapi kata majemuk terdiri dari 51.147 lema, dijelaskan pada tabel di bawah ini.

Tabel 14 Hasil Pengambilan Kata dan Jenis Kata

	Lema Pokok	Sub Lema		Total
		Kata Berimbuhan	Kata Majemuk	
Peneltian sebelumnya	38.870		0	38.870
Penelitian sekarang	40439		10609	51147

Dalam penelitian sebelumnya lema yang dihasilkan sebanyak 38.870 lema. Sehingga, dalam penelitian ini jumlah lema meningkat sebanyak 12.277 lema dari 38.870 lema menjadi 51.147 lema terdiri dari kata majemuk 10.609 lema dan bukan kata majemuk (kata dasar dan kata berimbuhan) 40.439 lema. Kata yang tidak terdekteksi adalah kata yang tidak mempunyai jenis kata, seperti gambar di bawah ini.


mengabu *v* menjadi abu; menyerupai abu;
mengabui **1** memberi abu; menaburkan abu pd: *ia ~ benih itu sebelum menanamnya*; **2** mengelabui; menyesatkan pandangan; menipu: *sikapnya yg lembut itu ternyata hanya untuk ~ mata saja*;

Gambar 14. Kata yang Tidak Terdeteksi

V. KESIMPULAN DAN SARAN

Kesimpulan pengembangan kamus jenis kata yang dilengkapi kata majemuk sudah tersedia dan dapat dimanfaatkan untuk penelitian NLP *(natural language processing)*. Dari pendeteksian yang dilakukan, jumlah lema yang dihasilkan sebanyak 51.147 lema terdiri dari 10.609 lema kata majemuk dan 40.439 lema bukan kata majemuk. Hal ini disebabkan pada tahap proses pengambilan kata dan jenis kata terdapat langkah-langkah yang belum terdeksi.

Saran untuk penetian berikutnya adalah peningkatan atau pembaruan langkah-langkah agar dapat mendeteksi kata dan jenis kata dengan baik. Pada langkah pengkategorian kata majemuk yang tidak mempunyai belum bisa terdeteksi sebagai kata majemuk seperti mahabesar dan semifinal sehinnga kata tersebut termasuk kategori bukan kata majemuk. Beberapa kata yang tidak terdeteksi disebabkan oleh kata turunan yang tidak mempunyai jenis kata oleh sebab itu untuk kata turunan yang tidak mempunyai jenis kata, diisi dengan jenis kata pada kata dasarnya.

REFERENCES

[1] Adi Budiwiyanto, “Korpus dalam Penyusunan Kamus,” [Online]. Available: http:// http://badanbahasa.kemdikbud.go.id/. [Diakses 22 Oktober 2016].

[2] Ernawati Waridah, “Ejaan Yang Disempurnakan dan Seputar Kebahasa-Indonesiaan”, Bandung : Ruang Kata, 2016

[3] CG Ceppy Efraim Bolly, “Pembangunnan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia”, Bandung, 2016