

Bahasa
Indonesia

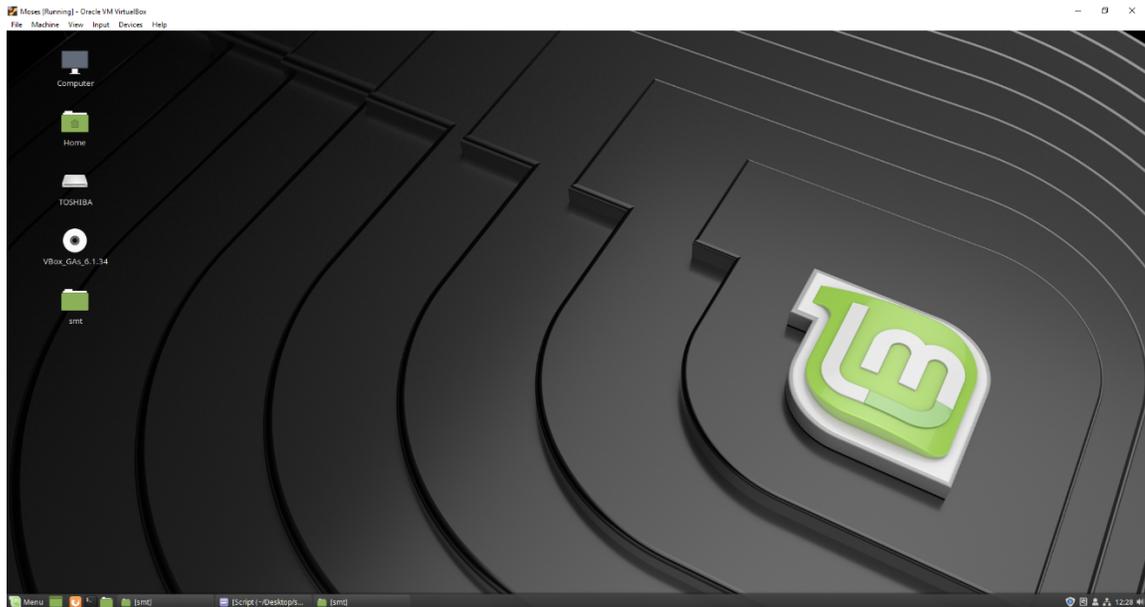
**MESIN PENERJEMAH STATISTIK
PADA BAHASA INDONESIA KE BAHASA
MELAYU BENGKULU**

**Dewi Soyusiawaty
Bella Okta Sari Miranda**

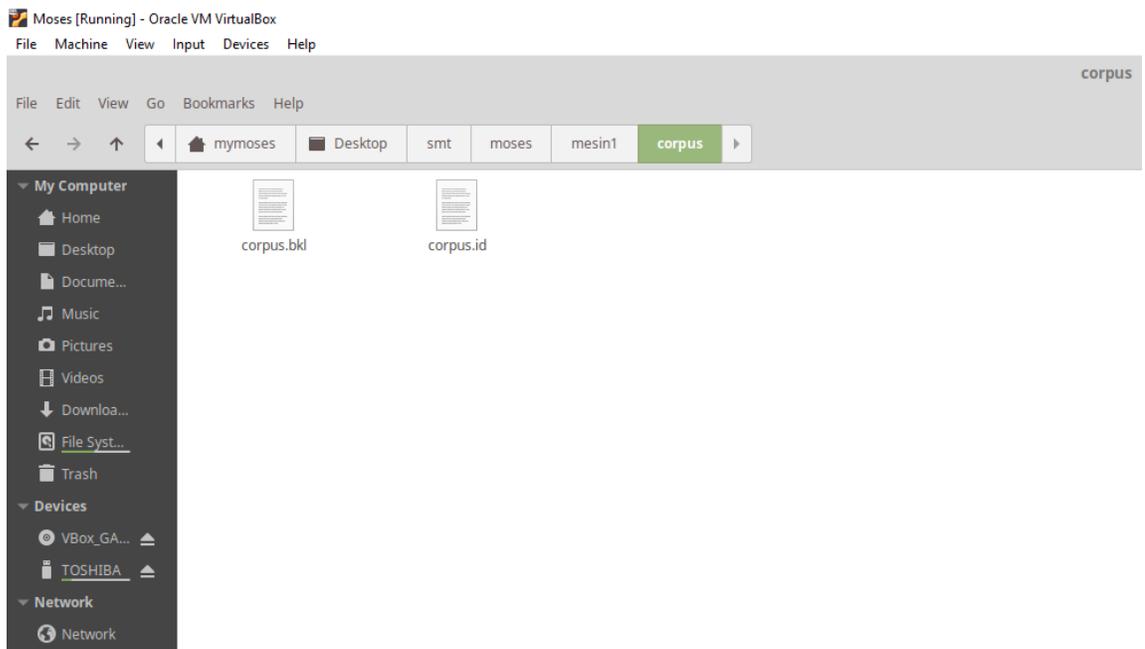


Berikut tahapan penggunaan Mesin Penerjemah Statistik pada Bahasa Indonesia ke Bahasa Melayu Bengkulu:

1. Membuka aplikasi virtual box lalu klik start pada tools linux mint. Tahapan ini untuk menjalankan linux mint.



2. Buat corpus.id dan corpus.bkl pada folder corpus yang pada folder SMT, dengan memasukkan jumlah corpus yang akan di lakukan translasi.



3. Lalu lakukan cleaning pada corpus.id dan corpus.bkl, dengan script code berikut pada terminal:

```
~/Desktop/smt/moses/scripts/training/clean-corpus-n.perl corpus/corpus id bkl corpus/clean
1 3000
```

4. Hasil cleaning korpus dapat di lihat pada folder corpus, berikut adalah tampilan hasil cleaning.

```
Moses [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help

clean.id (~/Desktop/smt/moses/mesin1/corpus)

File Edit View Search Tools Documents Help

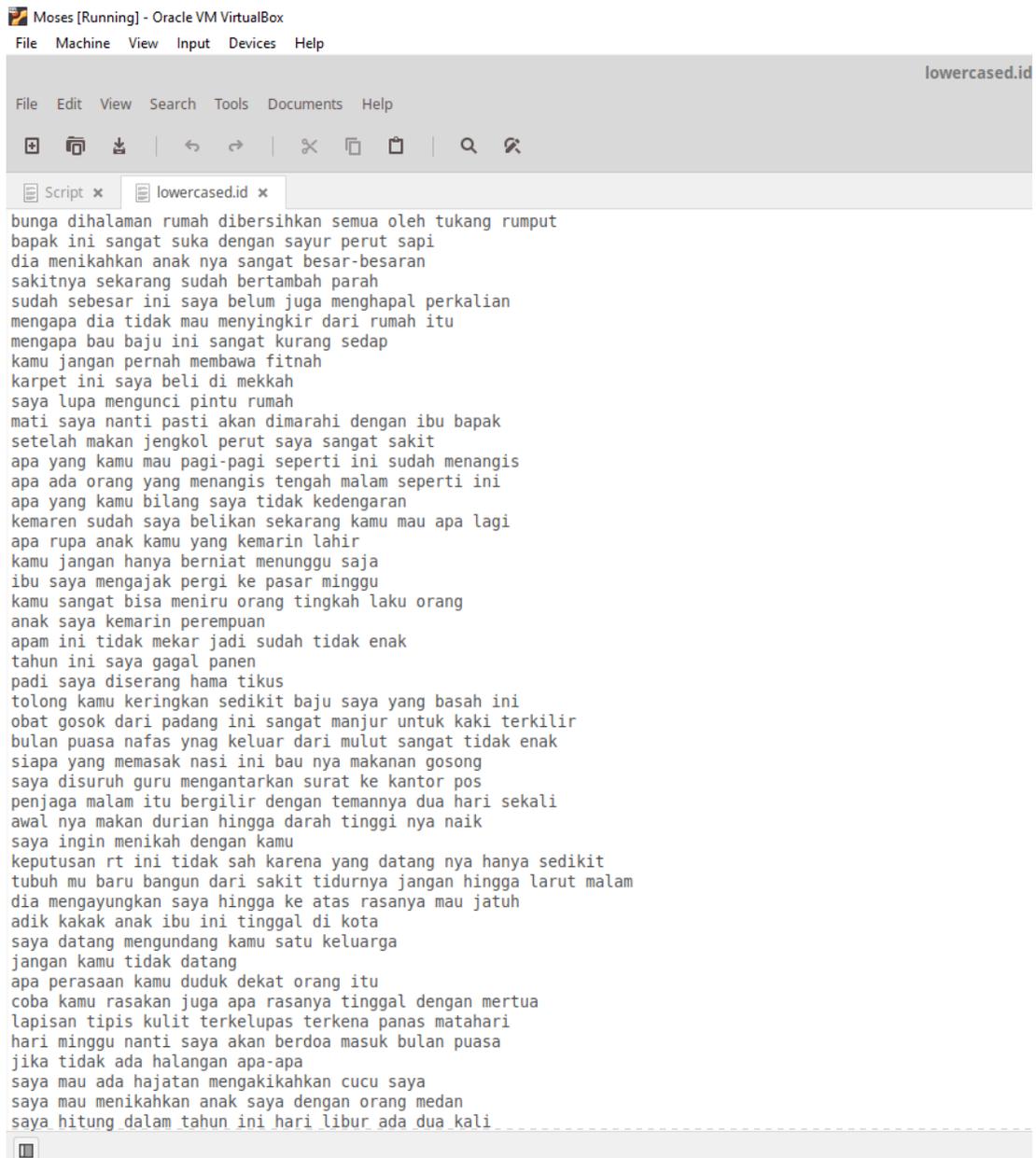
Script x clean.id x

Bunga di halaman rumah dibersihkan semua oleh tukang rumput
Bapak ini sangat suka dengan sayur perut sapi
Dia menikahkan anak nya sangat besar-besaran
Sakitnya sekarang sudah bertambah parah
Sudah sebesar ini saya belum juga menghafal perkalian
Mengapa dia tidak mau menyingkir dari rumah itu
Mengapa bau baju ini sangat kurang sedap
Kamu jangan pernah membawa fitnah
Karpit ini saya beli di mekkah
Saya lupa mengunci pintu rumah
Mati saya nanti pasti akan dimarahi dengan ibu bapak
Setelah makan jengkol perut saya sangat sakit
Apa yang kamu mau pagi-pagi seperti ini sudah menangis
Apa ada orang yang menangis tengah malam seperti ini
Apa yang kamu bilang saya tidak kedengaran
Kemaren sudah saya belikan sekarang kamu mau apa lagi
Apa rupa anak kamu yang kemarin lahir
Kamu jangan hanya berniat menunggu saja
Ibu saya mengajak pergi ke pasar minggu
Kamu sangat bisa meniru orang tingkah laku orang
Anak saya kemarin perempuan
Apam ini tidak mekar jadi sudah tidak enak
Tahun ini saya gagal panen
Padi saya diserang hama tikus
Tolong kamu keringkan sedikit baju saya yang basah ini
Obat gosok dari Padang ini sangat manjur untuk kaki terkilir
Bulan puasa nafas ynag keluar dari mulut sangat tidak enak
Siapa yang memasak nasi ini bau nya makanan gosong
Saya disuruh guru mengantarkan surat ke kantor pos
Penjaga malam itu bergilir dengan temannya dua hari sekali
Awal nya makan durian hingga darah tinggi nya naik
Saya ingin menikah dengan kamu
Keputusan RT ini tidak sah karena yang datang nya hanya sedikit
Tubuh mu baru bangun dari sakit tidurnya jangan hingga larut malam
Dia mengayunkan saya hingga ke atas rasanya mau jatuh
Adik kakak anak ibu ini tinggal di kota
Saya datang mengundang kamu satu keluarga
Jangan kamu tidak datang
Apa perasaan kamu duduk dekat orang itu
Coba kamu rasakan juga apa rasanya tinggal dengan mertua
Lapisan tipis kulit terkelupas terkena panas matahari
Hari minggu nanti saya akan berdoa masuk bulan puasa
Jika tidak ada halangan apa-apa
Saya mau ada hajatan mengakikahkan cucu saya
Saya mau menikahkan anak saya dengan orang medan
Saya hitung dalam tahun ini hari libur ada dua kali
```

5. Lakukan lowercase pada fle corpus.id dam corpus.bkl, dengan mengetik source code berikut:

```
~/Desktop/smt/moses/scripts/tokenizer/lowercase.perl < corpus/clean.id >
corpus/lowercased.id
~/Desktop/smt/moses/scripts/tokenizer/lowercase.perl < corpus/clean.bkl >
corpus/lowercased.bkl
```

6. Hasil lowercase dapat dilihat pada folder corpus, berikut tampilan hasil lowercase:



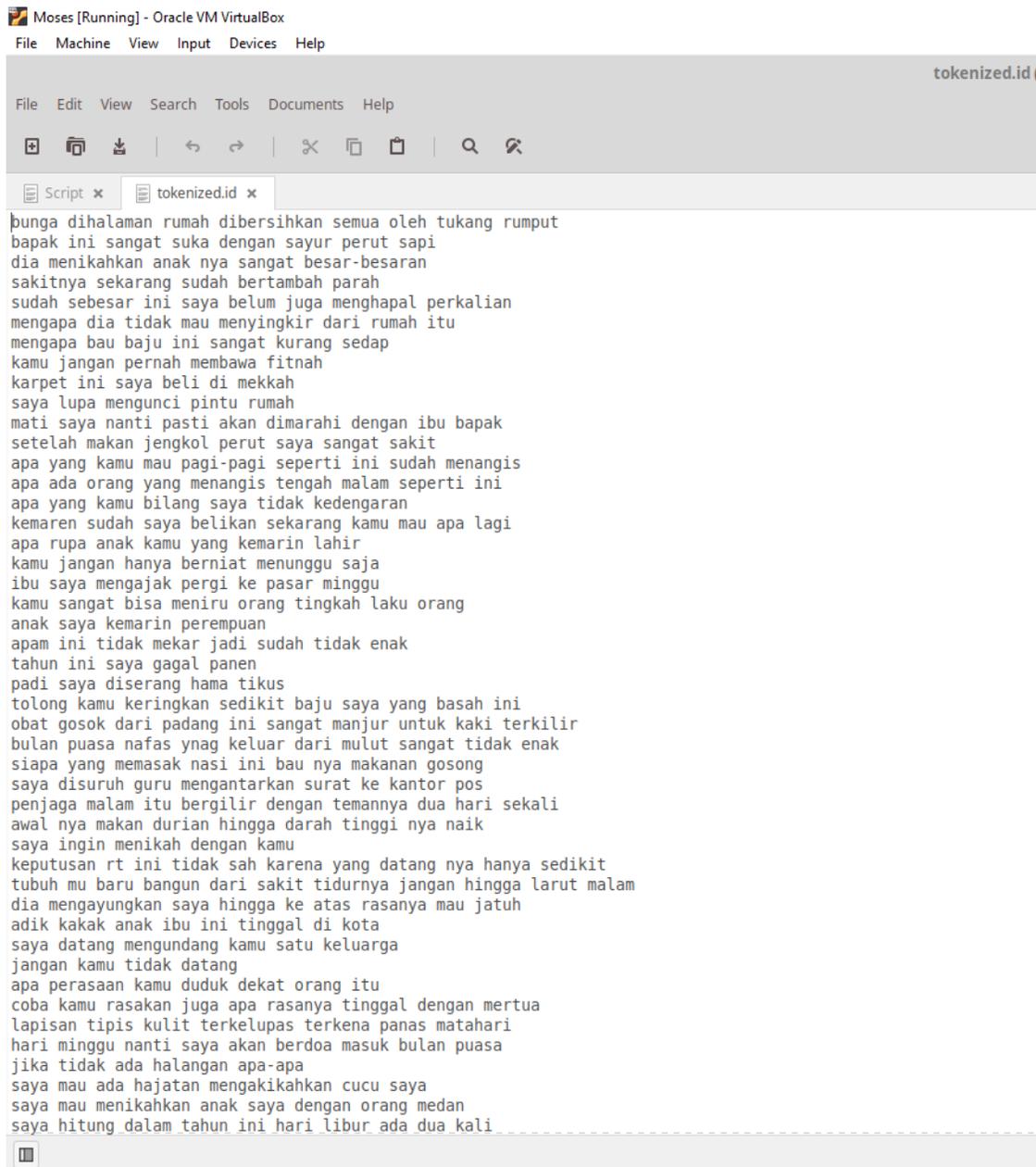
The screenshot shows a text editor window titled 'lowercased.id' with a menu bar (File, Edit, View, Search, Tools, Documents, Help) and a toolbar. The text content is as follows:

```
bunga di halaman rumah dibersihkan semua oleh tukang rumput
bapak ini sangat suka dengan sayur perut sapi
dia menikahkan anak nya sangat besar-besaran
sakitnya sekarang sudah bertambah parah
sudah sebesar ini saya belum juga menghafal perkalian
mengapa dia tidak mau menyingkir dari rumah itu
mengapa bau baju ini sangat kurang sedap
kamu jangan pernah membawa fitnah
karpet ini saya beli di mekkah
saya lupa mengunci pintu rumah
mati saya nanti pasti akan dimarahi dengan ibu bapak
setelah makan jengkol perut saya sangat sakit
apa yang kamu mau pagi-pagi seperti ini sudah menangis
apa ada orang yang menangis tengah malam seperti ini
apa yang kamu bilang saya tidak kedengaran
kemaren sudah saya belikan sekarang kamu mau apa lagi
apa rupa anak kamu yang kemarin lahir
kamu jangan hanya berniat menunggu saja
ibu saya mengajak pergi ke pasar minggu
kamu sangat bisa meniru orang tingkah laku orang
anak saya kemarin perempuan
apam ini tidak mekar jadi sudah tidak enak
tahun ini saya gagal panen
padi saya diserang hama tikus
tolong kamu keringkan sedikit baju saya yang basah ini
obat gosok dari padang ini sangat manjur untuk kaki terkilir
bulan puasa nafas ynag keluar dari mulut sangat tidak enak
siapa yang memasak nasi ini bau nya makanan gosong
saya disuruh guru mengantarkan surat ke kantor pos
penjaga malam itu bergilir dengan temannya dua hari sekali
awal nya makan durian hingga darah tinggi nya naik
saya ingin menikah dengan kamu
keputusan rt ini tidak sah karena yang datang nya hanya sedikit
tubuh mu baru bangun dari sakit tidurnya jangan hingga larut malam
dia mengayunkan saya hingga ke atas rasanya mau jatuh
adik kakak anak ibu ini tinggal di kota
saya datang mengundang kamu satu keluarga
jangan kamu tidak datang
apa perasaan kamu duduk dekat orang itu
coba kamu rasakan juga apa rasanya tinggal dengan mertua
lapisan tipis kulit terkelupas terkena panas matahari
hari minggu nanti saya akan berdoa masuk bulan puasa
jika tidak ada halangan apa-apa
saya mau ada hajatan mengakikahkan cucu saya
saya mau menikahkan anak saya dengan orang medan
saya hitung dalam tahun ini hari libur ada dua kali
```

7. Lakukan tokenisasi pada file corpus.id dan corpus.bkl dengan script code berikut:

```
~/Desktop/smt/moses/scripts/tokenizer/tokenizer.perl <
corpus/lowercased.id > corpus/tokenized.id
~/Desktop/smt/moses/scripts/tokenizer/tokenizer.perl <
corpus/lowercased.bkl > corpus/tokenized.bkl
```

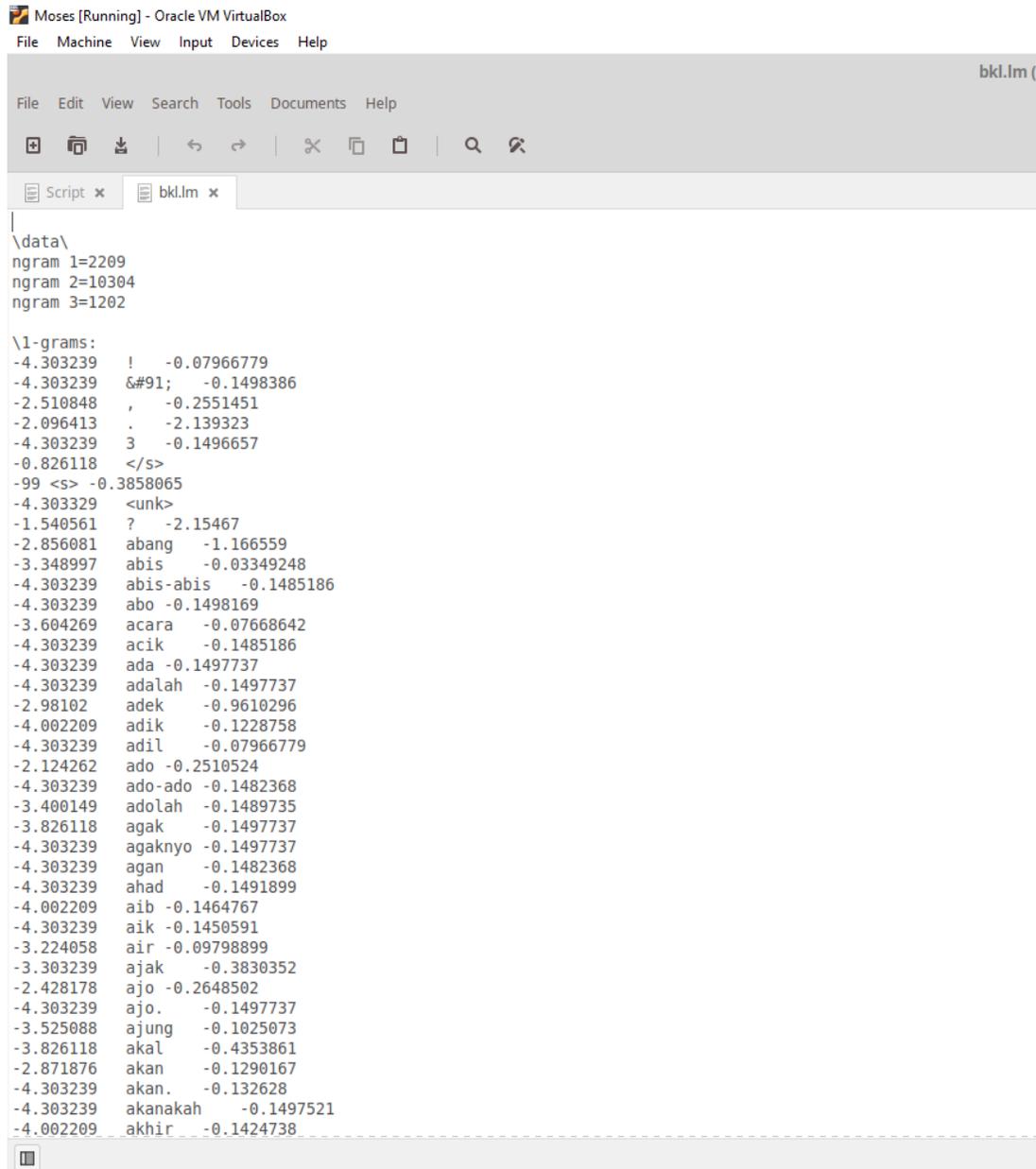
8. Hasil tokenisasi dapat dilihat pada folder corpus, berikut tampilan hasil tokenisasi:



9. Lakukan tahapan language model, dengan mengetik kan source code berikut pada terminal:

```
~/Desktop/smt/srilm/bin/i686-m64/ngram-count -order 3 interpolate -unk -text  
corpus/tokenized.bkl -lm lm/bkl.lm
```

10. Pada tahapan language model menghasilkan file n-gram dan probabilitas, berikut tampilan:



The screenshot shows a terminal window titled "Moses [Running] - Oracle VM VirtualBox". The window contains the following text:

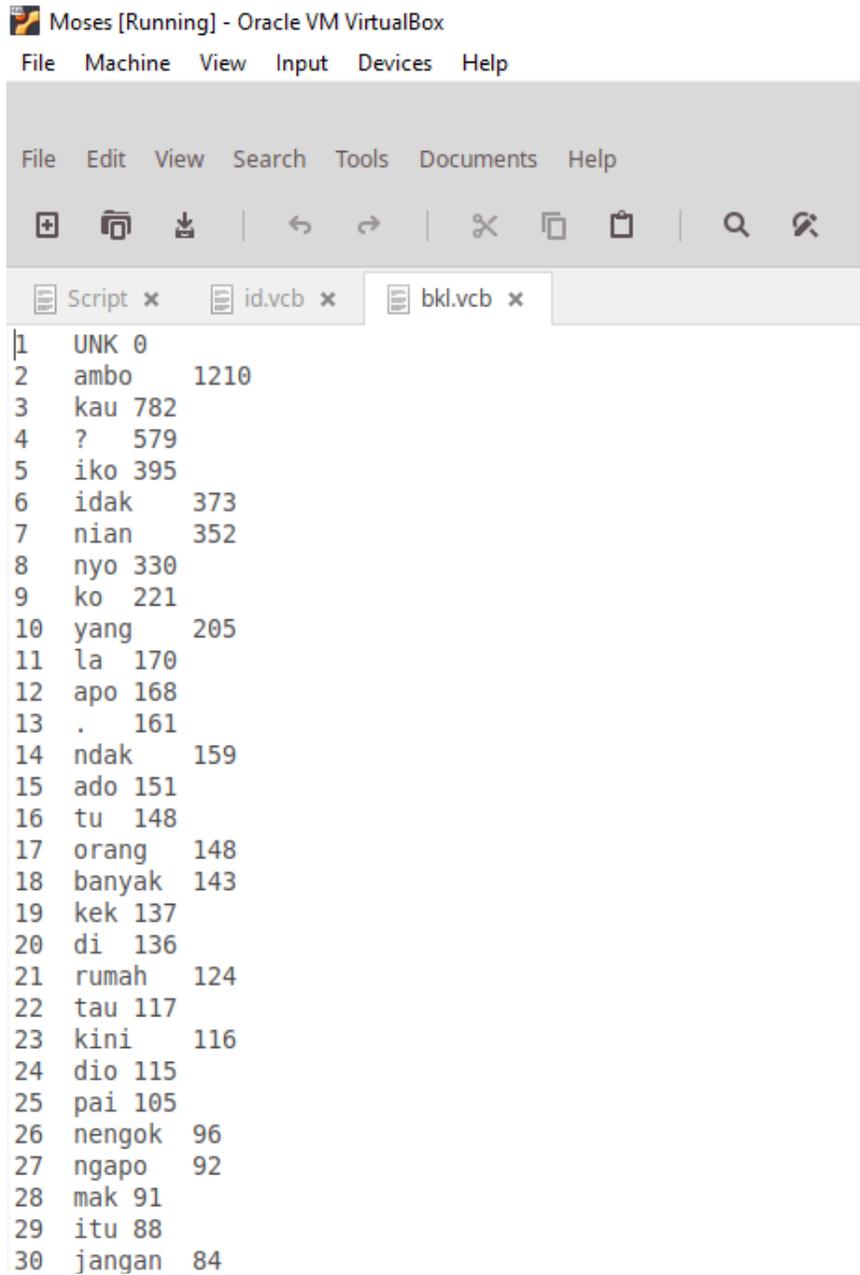
```
File Machine View Input Devices Help  
File Edit View Search Tools Documents Help  
Script x bkl.lm x  
\data\  
ngram 1=2209  
ngram 2=10304  
ngram 3=1202  
  
\1-grams:  
-4.303239 ! -0.07966779  
-4.303239 &#91; -0.1498386  
-2.510848 , -0.2551451  
-2.096413 . -2.139323  
-4.303239 3 -0.1496657  
-0.826118 </s>  
-99 <s> -0.3858065  
-4.303329 <unk>  
-1.540561 ? -2.15467  
-2.856081 abang -1.166559  
-3.348997 abis -0.03349248  
-4.303239 abis-abis -0.1485186  
-4.303239 abo -0.1498169  
-3.604269 acara -0.07668642  
-4.303239 acik -0.1485186  
-4.303239 ada -0.1497737  
-4.303239 adalah -0.1497737  
-2.98102 adek -0.9610296  
-4.002209 adik -0.1228758  
-4.303239 adil -0.07966779  
-2.124262 ado -0.2510524  
-4.303239 ado-ado -0.1482368  
-3.400149 adolah -0.1489735  
-3.826118 agak -0.1497737  
-4.303239 agaknyo -0.1497737  
-4.303239 agan -0.1482368  
-4.303239 ahad -0.1491899  
-4.002209 aib -0.1464767  
-4.303239 aik -0.1450591  
-3.224058 air -0.09798899  
-3.303239 ajak -0.3830352  
-2.428178 ajo -0.2648502  
-4.303239 ajo. -0.1497737  
-3.525088 ajung -0.1025073  
-3.826118 akal -0.4353861  
-2.871876 akan -0.1290167  
-4.303239 akan. -0.132628  
-4.303239 akanakah -0.1497521  
-4.002209 akhir -0.1424738
```

11. Lakukan training pada corpus.id dan corpus.bkl dengan mengetikkan source code berikut:

```
~/Desktop/smt/moses/scripts/training/train-model.perl -root-dir . --corpus  
corpus/tokenized --f id --e bkl --lm 0:3:$PWD/lm/bkl.lm -external-bin-dir  
~/Desktop/smt/training-tools
```

12. Tahapan training menghasilkan dokumen vocabulary, word alignment dan tabel model translasi. Hasil training dapat dilihat pada folder SMT, Berikut tampilan hasil training:

a. Vocabulary



The screenshot shows a terminal window titled "Moses [Running] - Oracle VM VirtualBox". The window has a menu bar with "File", "Machine", "View", "Input", "Devices", and "Help". Below the menu bar is a toolbar with icons for file operations and search. The terminal content shows a list of words and their frequencies, numbered 1 to 30. The words are: UNK 0, ambo 1210, kau 782, ? 579, iko 395, idak 373, nian 352, nyo 330, ko 221, yang 205, la 170, apo 168, . 161, ndak 159, ado 151, tu 148, orang 148, banyak 143, kek 137, di 136, rumah 124, tau 117, kini 116, dio 115, pai 105, nengok 96, ngapo 92, mak 91, itu 88, jangan 84.

```
1 UNK 0
2 ambo 1210
3 kau 782
4 ? 579
5 iko 395
6 idak 373
7 nian 352
8 nyo 330
9 ko 221
10 yang 205
11 la 170
12 apo 168
13 . 161
14 ndak 159
15 ado 151
16 tu 148
17 orang 148
18 banyak 143
19 kek 137
20 di 136
21 rumah 124
22 tau 117
23 kini 116
24 dio 115
25 pai 105
26 nengok 96
27 ngapo 92
28 mak 91
29 itu 88
30 jangan 84
```


13. Mencoba masukan kalimat dan melihat hasil terjemahan dengan mengetikkan source code berikut:

```
~/Desktop/smt/moses/bin/moses -f model/moses.ini
```

14. Contoh dari kalimat bahasa Indonesia ke bahasa melayu Bengkulu yang telah diterjemahkan:

```
kamu sekarang dimana
Translating: kamu sekarang dimana
Line 1: Initialize search took 0.000 seconds total
Line 1: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 1: Search took 0.000 seconds /ngram-count -order 3 interpolate -unk -text
kau kini dimano=d.bkl -lm lm/bkl.lm
BEST TRANSLATION: kau kini dimano [111] [total=-6.097] core=(0.000,-3.000,3.000
,-0.800,-0.539,-0.074,-0.208,0.000,-18.745)
Line 1: Decision rule took 0.000 seconds total del.perl -root-dir . --corpus
Line 1: Additional reporting took 0.000 seconds total lm -external-bin-dir ~/
Line 1: Translation took 0.001 seconds total
kamu sedang apa
Translating: kamu sedang apa
Line 2: Initialize search took 0.000 seconds total
Line 2: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 2: Search took 0.000 seconds
kau lagi apo ~/Desktop/smt/moses/bin/moses -f model/moses.ini < corpus/tokenized.id > out
BEST TRANSLATION: kau lagi apo [111] [total=-5.988] core=(0.000,-3.000,3.000,-1
.742,-1.705,-0.074,-0.074,0.000,-17.738)
Line 2: Decision rule took 0.000 seconds total
Line 2: Additional reporting took 0.000 seconds total
Line 2: Translation took 0.000 seconds total ratio=1.001 hyp len=17180
```

15. Pengujian moses decoder yang dilakukan secara otomatis dengan mengetikkan source code berikut:

```
~/Desktop/smt/moses/bin/moses -f model/moses.ini <
corpus/tokenized.id > out
~/Desktop/smt/moses/scripts/generic/multi-bleu.perl
corpus/tokenized.bkl < out
```

16. Hasil pengujian yaitu suatu nilai akurasi yang dihasilkan menggunakan BLEU pada moses decoder.

Berikut tampilan hasil pengujian:

```
File Edit View Search Terminal Help
Line 2998: Initialize search took 0.001 seconds total
Line 2998: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 2998: Search took 0.062 seconds
BEST TRANSLATION: apo alat nyo mangko anai ko ilang ? [11111111] [total=-8.684]
core=(0.000,-8.000,7.000,-0.773,-0.879,-0.022,-3.726,0.000,-34.007)
Line 2998: Decision rule took 0.000 seconds total
Line 2998: Additional reporting took 0.005 seconds total
Line 2998: Translation took 0.071 seconds total
Translating: sebanyak apa rayap dirumah kamu ?
Line 2999: Initialize search took 0.001 seconds total
Line 2999: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 2999: Search took 0.021 seconds
BEST TRANSLATION: sebanyak apo anai dirumah kau ? [111111] [total=-8.224] core=
(0.000,-6.000,6.000,-0.143,-0.164,-0.075,-0.270,0.000,-30.587)
Line 2999: Decision rule took 0.000 seconds total
Line 2999: Additional reporting took 0.001 seconds total
Line 2999: Translation took 0.025 seconds total
Name:moses      VmPeak:80020 kB VmRSS:47976 kB  RSSMax:47976 kB user:34.025  s
ys:1.513      CPU:35.538      real:92.780
mint@mint-VirtualBox:~/Desktop/smt/moses/uji-mesin$ ~/Desktop/smt/moses/scripts/
generic/multi-bleu.perl corpus/tokenized.bkl < out
BLEU = 92.55, 97.5/93.7/90.9/88.3 (BP=1.000, ratio=1.001, hyp_len=17100, ref_len
=17089)
```