Q9

a.

```
$ grep -A 1 '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx | sed -
e 's/<[^>]*>//g' | sed -e 's/--//g' > uncorpus.eng.txt
```

To verify that we did not miss any lines, we can calculate the num of <seg> after the command “grep -A 1 '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx”. And we could also use the command “wc -l” to calculate the line of uncorks.eng.txt which will match the number of <tuv xml:lang="EN"> in the original file.

b.

```
$ wc -w uncorpus.eng.txt
result:2685545
```

c.

```
cat uncorpus.eng.txt |grep -o '\b[A-Za-z0-9\.,]\{1,\}[A-Za-z0-9]\b' |
sort|uniq -c|sort -nr |wc -l
16404
```

In this question, we assumes that words are consisted of A-Z and a-z, numbers are consisted of 0-9, ',', and '.'. But the last letter or digits must be A-Za-z0-9

d.

```
$ cat uncorpus.eng.txt | perl -pe 's/\s/\n/g;'|perl -pe 'tr/A-Z/a-
z/;'|sort|uniq -c|sort -nr |wc -l
result:13711
```

e.

```
$ grep -o '[0-9]\{1,\}[0-9.,]\{1,\}[0-9]\{1,\}' uncorpus.eng.txt | wc
-l
result:43860
```

Examples: 1234 1,234 1.234

f.

```
$ grep -o '\b[0-9]\{1,\}[0-9.,]\{1,\}[0-9]\{1,\}\b' uncorpus.eng.txt |
wc -l
result:43764
```

g.

```
$ grep -o '\b[A-Z]\{1,\}[a-zA-Z]\{0,\}\b' uncorpus.eng.txt |wc -l
result:464237
```

h.

```
$ grep -o '\.\s[A-Z0-9]\{1,\}[A-Za-z0-9]\{0,\}' uncorpus.eng.txt |
sort|uniq -c|sort -nr | head -15
```

Ps: there will be a dot between the number and the word

```
3703 . Requests
2415 . Calls
```

```
2380 . Also
2028 . Welcomes
1941 . Decides
1688 . Urges
1632 . Notes
1607 . Encourages
1482 . Takes
1458 . Reaffirms
1409 . Invites
1044 . Stresses
 890 . Recognizes
 861 . Expresses
 737 . Emphasizes
```

i.

```
$ grep -o '[^.]\W[A-Z]\{1,\}[a-zA-Z]\{0,\}\b' uncorpus.eng.txt| cut -c
3- | sort|uniq -c|sort -nr | head
```

```
19817 United
10112 States
9302 December
8947 Secretary
6313 General
5360 International
5026 Convention
4823 Recalling
4572 Committee
3855 The
```

j.

Roman numerals are consisted of ‘X’ ‘I’ ‘V’ ‘L’ ‘C’ ‘D’ ‘M’

```
grep -o '\b[XIVLCDM]\{1,\}\W' uncorpus.eng.txt |wc -l
```

Result:3880

Ps: there are no 'I am' or 'I'm' in the document