a)How many lines does the UNCorpus file have?

```
wc -l uncorpora_plain_20090831.tmx
###1501316 uncorpora_plain_20090831.tmx###
```

b)How many segments <seg>?

```
grep -o '<seg>' uncorpora_plain_20090831.tmx | wc -l
###434034###
```

c)How many non-segments? As in tags that are not <seg> like <tuv>?

```
egrep -o '<\/[a-z]+>' uncorpora_plain_20090831.tmx | wc -l
```

You can subtract the result from this command from the one above it yeilding the result. This assumes that <tuv></tuv> are counted at 1 tag, not 2.

```
###994927-434034=560893###
```

d)How many English segments does the text have?

```
grep -o '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx | wc -l
###72339###
```

e)How many segments exist for each language (done in one command)?

```
egrep -o '<tuv xml:lang="[A-Z]+">' uncorpora_plain_20090831.tmx | sort
| uniq -c
###72339 <tuv xml:lang="AR">###
###72339 <tuv xml:lang="EN">###
###72339 <tuv xml:lang="ES">###
###72339 <tuv xml:lang="FR">###
###72339 <tuv xml:lang="RU">###
###72339 <tuv xml:lang="ZH">###
```