Q8

a.

```
commands: wc -l uncorpora_plain_20090831.tmx
result: 1501316
```

b.

```
command: grep -o '<seg>' uncorpora_plain_20090831.tmx | wc -l
result: 434034
```

c.

From the UNCorpus, we can find the format pattern like this:

```
<tu>
 <tuv>
  <seg>
  </seg>
 </tuv>
</tu>
```

So, for <tuv>

```
command: grep -o '<tuv' uncorpora_plain_20090831.tmx | wc -l
result: 434034
```

For <tu>

```
command: grep -o '<tu ' uncorpora_plain_20090831.tmx | wc -l
result: 72339
```

d.

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

If there is a '<tuv xml:lang="EN">', there is a English segments.

e.

For Arabic

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

For Chinese

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

For French

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

For Russian

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

For Spanish

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

The explanation is the same as the explanation in d.