a) Below are the four commands used to get the top 20 words in each language specified. The words for each language are listed in the table below. I am changing upper-case words to lower-case to get a more accurate word count.

For English: grep -A1 '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<\/seg>//' | perl -pe 's/<tuv xml:lang="EN">//'| perl -pe 's/--//' | grep '[^[:blank:]]' | perl -pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq -c | sort -nr | head -20

For Arabic: grep -A1 '<tuv xml:lang="AR">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<\/seg>//' | perl -pe 's/<tuv xml:lang="AR">//'| perl -pe 's/--//' | grep '[^[:blank:]]' | perl -pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq -c | sort -nr | head -20

For Spanish: grep -A1 '<tuv xml:lang="ES">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<\/seg>//' | perl -pe 's/<tuv xml:lang="ES">//'| perl -pe 's/--//' | grep '[^[:blank:]]' | perl -pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq -c | sort -nr | head -20

For Russian: grep -A1 '<tuv xml:lang="RU">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<\/seg>//' | perl -pe 's/<tuv xml:lang="RU">//'| perl -pe 's/--//' | grep '[^[:blank:]]' | perl -pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq -c | sort -nr | head -20

ENGLISH		ARABIC		SPANISH		RUSSIAN	
578773		578717		578719		578727	
272274	the	92683	313424	في	de	135693	и
175498	of	45238	180894	من	la	100876	в
136608	and	38530	131116	على	y	37886	по
101442	to	34817	-	93943	en	37178	на
67237	in	34466	87265	إلى	los	28031	с
35911	on	24225	83506	أن	el	19152	Объединенных
32424	for	18986	78253	التي	las	18190	Организации
22567	that	18947	77417	وإذ	a	18148	о
21288	its	17776	69361	الأمم	que	15609	для
20916	a	16764	52784	المتحدة	del	14885	от
20383	with	15046	37384	عن	para	14180	что
20010	united	11324	28354	الدول	con	13804	к
19994	as	10979	22762	أو	su	13361	Наций
17748	international	10175	21725	المؤرخ	por	11698	также
17262	by	9476	21503	كانون	al	10911	года,
17189	nations	9413	20267	مع	sobre	9245	декабря
14018	at	9163	18669	جميع	naciones	9233	года
13408	all	8916	15773	بما	se	8520	их
12037	states	8800	14139	العام	estados	8379	призывает

b) Below is the command for fetching the bottom 20 english words for the comparison

```
grep -A1 '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<
\/seg>//' | perl -pe 's/<tuv xml:lang="EN">//'| perl -pe 's/--//' | grep '[^[:blank:]]' | perl -
pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq -c | sort | head -20
```

Chart comparing top 20 english words to other languages

ENGLISH	ARABIC	SPANISH	RUSSIAN
the	at	from	and
of	from	the	at
and	on	and	by
to	-	in	on
in	to me	the	with
on	that	the	united
for	which	the	organizations
that	taking	to	about
its	nations	what	for
a	united	the	from
with	about	for	what
united	countries	with	to
as	or	his	nations
international	of	by	also
by	December	to the	of the year,
nations	with	on	December
at	all	nations	of the year
all	including	he	their
states	year	state	urges

Chart comparing bottom 20 english words to other languages (top english word in list is bottom most english word)

ENGLISH	ARABIC	SPANISH	RUSSIAN
"22.1			
"200"	at	from	and
"100	from	the	at
"10	on	and	by
"(v)	-	in	on
"(p)	to me	the	with
"(o)	that	the	united
"(n)	which	the	organizations
"(m)	taking	to	about
"(l)	nations	what	for

"(k)	united	the	from
"(ix)	about	for	what
"(ii)"	countries	with	to
"(i)"	or	his	nations
"(e.g.,	of	by	also
"(d)";	December	to the	of the year,
"(d)"	with	on	December
"(conclusion	all	nations	of the year
"(b)"	including	he	their
"	year	state	urges

Analysis: for all languages, the top word was nothing. This is due to the issue explained in question 10. Looking at the similarities in the languages, we can see that all languages contain quite a few filler words (a, an, the, to, from, what, etc) as we would expect. However, Arabic notably has the least amount of these words. This can be attributed to the fact that Arabic is a morphologically rich language, and articles and filler words are represented by the endings on words as opposed to individual words themselves. It's also interesting to see that some words are repeated mutliple times in the same language (the in spanish). This is due to Spanish's use of gender and number playing a role in how the appears. For example, the boy, the girl, the apples all would use different words for the. When looking at the bottom words of english comparing to the top words in the other languages, nothing overlaps (and this is expected). All of the bottom english words clearly stem from the fact that we tokenized on whitespace instead of a more complex regular expression. Therefore, the bottom most words contain quotes and mostly numbers and special characters.