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 578777 578777 578777 578777 578777 69361 61 61 61 62 62 62 63 64 64 64 64 64 64 64	+		+		+		+	+
272274 the 92683 313424 de 135693 и 175498 of 45238 180894 ta 100876 в 136608 and 38530 131116 y 37886 по 101442 to 34817 - 93943 en 37178 на 67237 in 34466 87265 los 28031 с 35911 on 24225 83506 las 18190 Организации 22424 for 18986 78253 las 18190 Организации 22567 that 18947 77417 g a 18148 o 21288 its 17776 69361 que 15609 для 20916 a 16764 52784 las 14180 что 20383 with 15046 37384 para 14180 что 20010 united 11324 28354 g g su 13361 Haций 17748 international 10175 21725 su 13361 Haций 17748 nations 9413 20267 sobre 9245 декабря 14018 at 9163 18669 se 8520 их	ENGLISH		ARABIC		SPANISH		RUSSIAN	
175498 of 45238 180894 оо la 100876 в 136608 and 38530 131116 оо оо у 37886 по 101442 to 34817 - 93943 en 37178 на 67237 in 34466 87265 оо los 28031 с 35911 on 24225 83506 оо el 19152 Объединенных 32424 for 18986 78253 оо оо оо оо оо оо оо	578773		578717	<u> </u>	578719		578727	<u> </u>
136608 and 38530 131116	272274	the	92683	313424	فـي	de	135693	и
101442 to 34817 - 93943 en 37178 на 67237 in 34466 87265	175498	of	45238	180894	 مـن	la	100876	B
10 34466 87265 10 10 28031 С 35911 00 24225 83506 10 el 19152 Объединенных 32424 for 18986 78253 12 las 18190 Организации 22567 that 18947 77417 21288 its 17776 69361 21288 its 17776 69361 21288 its 17776 69361 21288 14885 0 21288 its 16764 52784 21288 14885 0 20916 a 16764 52784 212888 212888 212888 212888 212888 212888 212888 212888 212888 212888 21	136608	and	38530	131116	على على	у	37886	по
35911 on 24225 83506 oi el 19152 Объединенных 32424 for 18986 78253 от las 18190 Организации 22567 that 18947 77417 ој a 18148 o 21288 its 17776 69361 ој que 15609 для от 20916 a 16764 52784 ој от он от от от от от от	101442	to	34817		93943	en	37178	на
32424 for 18986 78253 от организации 22567 that 18947 77417 организации орган	67237	in	34466	87265	إلى	los	28031	c
22567 that 18947 77417 319 a 18148 0 21288 its 17776 69361 еди que 15609 для 20916 a 16764 52784 потемен 14885 от 20383 with 15046 37384 о рага 14180 что 20010 united 11324 28354 о о о о о о о о о о о о о о о о о о	35911	on	24225	83506	ان أن	el	19152	Объединенных
21288 its 17776 69361 віть que 15609 для 20916 а 16764 52784 вітьські del 14885 от 20383 with 15046 37384 вітьські рага 14180 что 20010 united 11324 28354 вітьські соп 13804 к 19994 аѕ 10979 22762 вітьські рог 11698 также 17748 international 10175 21725 вітьські рог 11698 также 17262 ву 9476 21503 вітьські а вітьські 10911 года, 17189 nations 9413 20267 вітьські зовітье 9245 декабря 14018 ат 9163 18669 вітьські яку ве 8520 их 13408 ав вітьські віть	32424	for	18986	78253	 التي	las	18190	Организации
20916 a 16764 52784 Полтана 14885 От 20383 With 15046 37384 арта 14180 что 20010 United 11324 28354 от соп 13804 к 19994 аѕ 10979 22762 ој ѕи 13361 Наций 17748 international 10175 21725 ој рог 11698 также 17262 ру 9476 21503 ој аl 10911 года, 17189 nations 9413 20267 од ѕорге 9245 декабря 14018 аt 9163 18669 еди ѕе 8520 их 13408 all 8916 15773 од ѕе 8520 их	22567	that	18947	77417	 وإذ	a	18148	0
20383 with 15046 37384 عن para 14180 что 20010 united 11324 28354 الدول con 13804 к 19994 as 10979 22762 اول المؤرخ su 13361 Наций 17748 international 10175 21725 المؤرخ por 11698 также 17262 by 9476 21503 المؤرخ al 10911 года, 17189 nations 9413 20267 عن sobre 9245 декабря 14018 at 9163 18669 جميع naciones 9233 года 13408 all 8916 15773 يا العواد se 8520 их 13408 all 8916 15773 العواد se 8520 их 1	21288	its	17776	69361	ا الأمع	que	15609	 для
20010 united 11324 28354 Порто 13804 к 19994 as 10979 22762 ы su 13361 Наций 17748 international 10175 21725 Порто 11698 также 17262 by 9476 21503 Порто аl 10911 года, 17189 nations 9413 20267 ы sobre 9245 декабря 14018 at 9163 18669 елде naciones 9233 года 13408 all 8916 15773 ы se 8520 их 13408 аll 8916 15773 ы se 8520 их 18609	20916	a	16764	52784	المتحدة	del	14885	от
19994 as 10979 22762 ј su 13361 Наций 17748 international 10175 21725 рог 11698 также 17262 by 9476 21503 كانون al 10911 года, 17189 nations 9413 20267 ав sobre 9245 декабря 14018 at 9163 18669 елге пасіопез 9233 года 13408 all 8916 15773 леге se 8520 их	20383	with	15046	37384	عن	para	14180	что
17748 international 10175 21725 المؤرخ рог 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 ما	20010	united	11324	28354	الدول	con	13804	K
17262 by 9476 21503 كانون al 10911 года, 17189 nations 9413 20267 ав sobre 9245 декабря 14018 at 9163 18669 елгер naciones 9233 года 13408 all 8916 15773 ав se 8520 их	19994	as	10979	22762	أو أو	su	13361	Наций
17189 nations 9413 20267 декабря sobre 9245 декабря 14018 at 9163 18669 декабря 13408 all 8916 15773 декабря se 8520 их	17748	international	10175	21725	المؤرخ	por	11698	также
14018 at 9163 18669 جميع naciones 9233 года 13408 all 8916 15773 ь se 8520 их	17262	by	9476	21503	- کانون	al	10911	года,
	17189	nations	9413	20267	مع	sobre	9245	декабря
	14018	at	9163	18669	 جمیع '	naciones	9233	года
	13408	all	8916	15773	بما		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

	L		L
ENGLISH	ARABIC	SPANISH	RUSSIAN
İ	 		
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) ${\sf vol}$

L	L	L	L
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
T			r

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.