Quality Assessment of General and Categorized Arabic Text Corpora

Ali Salhi
Department of Computer Systems Engineering
Birzeit University
Birzeit, Palestine.
eng.salhi.ali@gmail.com

Adnan Yahya
Department of Computer Systems Engineering
Birzeit University
Birzeit, Palestine.
yahya@birzeit.edu

***Abstract*— Many Natural Language Processing and Information Retrieval methods are based on the extensive use of text corpora. The credibility of the results can be heavily influenced by the underlying corpus quality. Much research has been utilizing Arabic corpora into various tasks of Arabic Information Processing. In this paper we discuss a suite of metrics that can be used to ascertain the quality of Arabic corpora. We borrow heavily from the extensive work on corpora quality for other languages and try to adapt some of them for Arabic. We also apply these measures to sample corpora, including categorized corpora and report on the results. The main corpora we experiment with are: a general corpus extracted from newspaper (AlQuds newspaper) articles covering the years 2009 -2012 and a highly categorized (split into 9 major and 25 minor categories) corpus built from Arabic Wikipedia. We employ different filtration methods, discuss and examine different corpora features as quality metrics. The metrics are based on parameters such as character/word N-gram frequencies and Zipf's law applicability. We also study error rates, vocabulary properties, monolinguality, the effects of normalization, as well as corpora stability with size growth. It is our intention to make our corpora quality assessment tools available online for possible use by other researchers[1].**

Keywords*— *Arabic Text Corpora, Corpus Quality Measures, Arabic Information Retrieval, Zipf's Law, Statistical Analysis of Text, Arabic NLP

I. INTRODUCTION

The Arabic online content has increased noticeably during the last period[13]. While still far from sufficient, this reflects awareness among Arab writers and users about the importance of creating and publishing on the web. Content increase comes with a great need for tools to overcome the many challenges of information processing and retrieval. The challenges include understating Arabic content, efficient retrieval of useful information through quality and efficient search, and providing tools to facilitate content generation and processing such as spell-checking, named entities extraction, document categorization, query optimization and filtration, text to speech/speech to text systems and much more. Some of tools also come handy in processing the massive increase of Arabic content in social networks, where the need for filtration systems and useful content extraction are a must due to the great mix of languages, dialects and scripts.

Arabic corpora are essential when building language processing tools. For example in our earlier work we relied on the statistical analysis of corpora obtained from different sources such as Arabic newspapers and Arabic Wikipedia[21,22]. The availability of high quality corpora is very important for researchers, learners, and language based application builders as well as for tasks like text segmentation/classification[12]. Corpus Quality assessment can also be used as a tool to assert the adequateness of a collected corpus for the intended purposes[4]. Quality has strong association with the statistical properties of the corpus[4] and to be consistent one needs to be precise about the definitions used in quality assessment, also to make comparisons meaningful[7]. Definition of concepts like word, sentence, stem and the likes may affect the final results, as may corpus collection method with the associated scope, duplication level, currency and other factors[4,7]. Generating a quality corpus may require cleaning of the original material, especially if the source is general web data. Issues like removal of (near) duplicates (say multiple quotes of a newswire story), preprocessing of non-words and non-sentences, say by splitting and spelling correction may be important to achieving quality corpora[7,9] acquire added importance. Dealing with text corpora is no more limited to linguists but are of interest to researchers in information retrieval and data mining in health, finance, literature, social sciences, commerce, and many more. Much of the research with text corpora has been conducted on industrial nations' languages and their corpora, but in recent years much more interest was exhibited in Arabic information processing/retrieval, both in the industrial countries and in the Arab region. Major companies marketing IR tools (e.g. Google, Microsoft, Yahoo, IBM) as well as governmental bodies and academic institutions have been engaging in major efforts in this area.

This paper is about quality estimators for Arabic corpora. The presentation heavily utilizes two text corpora we built: the first is a general purpose corpus extracted from a AlQuds (a Palestinian newspaper) and covering 4 years: (2009 -2012), and the second is a categorized (split into more than 25 sub categories which are grouped into 9 major categories) corpora that we built using Arabic Wikipedia using an in-house developed extraction algorithm[23].

The paper is organized as follow: In Section 2 we give a general background about corpora quality and discuss the potential applicability to Arabic texts. We will address some definitions regarding the statistical characteristics of corpora

[1] For more information about the tools and their availability for online use, please contact the authors.

and highlight available resources. In section 3 we discuss the quality assessment experiments we applied to our corpora with different quality measurements from earlier related work and report the results. Finally in section 4 we provide a short discussion and some conclusions and potential future work.

II. Background and Related work

The quality of web based material is acquiring special importance due to the volume of electronic documents and the accessibility of the Internet for all. Users would like to get assurances about the quality of the information being accessed to help determine the degree of trust. Of interest here are two concepts: one is the quality of web documents, an important issue that is not the topic of this paper. The other is the overall quality of a corpus: a collection of articles serving as a representative of the genre or language and possibly a main infrastructure for many IR tasks. It is corpora quality that we discuss here. Clearly, a corpus may include documents and there may very well be a relationship between the quality of a corpus and the quality of constituent documents, but the relation need not be one-to-one. One can think of a high quality corpus with some low quality documents, and good documents may not necessarily produce a good corpus representing the language or part of it, say due to topic bias. So our concern here is Arabic corpora quality. We discuss measures of corpora quality and quality metrics, and their applicability to Arabic.

A. Arabic Language writing system:

Arabic is a Semitic language spoken in about 24 countries and by about 300 million people mostly in Asia and Africa. Most Arabic writing is in Modern Standard Arabic (MSA): the language of education and formal communications in the Arab World. MSA coexists with a large number of dialects which may vary substantially even within individual countries, though some are more dominant than others. In the written form, dialects are mostly used in social media, generally using Arabic alphabet, but also Latin alphabet. The following properties of Arabic writing system are of relevance to the topic of this paper:

- The Arabic alphabet has 28 letters and is used for several other languages like Farsi and Urdu as well as much of dialect material dominant in social media.
- Formally, almost every written Arabic letter should have a diacritical mark (short vowel) for proper pronunciation. However, most Arabic writing is without these short vowels, resulting in added ambiguity. Arabic readers have to rely on the context and their knowledge to reconstruct the short vowels.
- Arabic writing is also tolerant of some spelling errors mostly in what we call *Confusion Letters*: 1. the different shapes of the first letter of the alphabet (Hamza/Alef): { ا,أ,إ,آ,ء,ؤ,ئ }; 2. the "Alef Maqsoura" and "Ya" (ى, ي) no relation, just common shape: differ in dotting; and 3. the "Ha" and "Ta Marbouta" (ه,ة), no relation, just common shape: differ in dotting. While the rules for the correct selection from each group are clear, many tend to pay little attention to that resulting in quite a number of tolerated spelling errors. So much so that many IR Systems resort to normalization: representing each group by a single letter (along the lines of normalizing to all lower case in English). Well written texts do not make these mistakes and thus require no normalization.
- While Arabic writing separates text into words, Arabic is an inflectional language and a large number of affixes attach to a base word, resulting in longer words and even single word sentences. A word for us here is a white space or special character delimited string. Also, some articles are attached to the following word increasing word length. So while the removal of short vowels tend to shorten words and reduce the number of distinct words –types-, the large number of affixes has the opposite effect: increasing word length increase the number of distinct words -types- count of the text.

B. Arabic Corpora – Available Resources

Looking at earlier work on Arabic corpora we can detect some resources: both free and for a fee. [1] introduces a free corpus of about 5000 articles with around 3 million words split into 4 different categories. [18] uses seven different corpora for work on Arabic classification. We were able to retrieve 5 of them for our earlier work[23]. [8] reports on an Arabic corpus using around 4000 articles from Al-Hayat newspaper archive of the year 1998. The Arabic Contemporary Corpus is another free corpus is split over 16 different categories[3]. Also, the Six-Language Parallel Corpus of the United Nations published documents which can be accessed online.[2] LDC provides a wide range of corpora for researchers with different properties and features and in different languages including Arabic, however LDC corpora are not free.[3] Ref. [3] and [2] summarize some of the current corpora such as the Qur'anic Arabic (77,430 words), QAMUS-Backwalter Arabic Corpus (2.5–3 billion words, found on LDC site), CLARA (50 million words), Agence France Presse (Arabic Newswire Corpus with 80 million words) and much more which can be found with more details online.[4]

So the need for free Arabic corpora with large processed terms, with a broad variety of words in many categories is, in our view, essential and can come handy for building different language processing tools. The availability of categorized corpora with variety of categories and with a large enough number of documents in each can be quite useful for building, testing and experimenting with different fine-tuned categorizing algorithms.

C. Corpora Quality Measures:

Generally, the quality of a corpus is associated with the degree to which it represents the properties of the collection of its text class. It is generally expected that other documents in the class exhibit behavior close to that of the corpus. One may talk about a general corpus in terms of topic span with documents from a large number of topics or a specialized/categorized corpus with documents on a specific

[2] http://www.uncorpora.org/
[3] http://www.ldc.upenn.edu/
[4] http://www.comp.leeds.ac.uk/latifa/arabic_corpora.htm

topic. One can also think of corpora representing styles: e.g. a corpus of highly specialized academic material, or of short texts like tweets or SMS. Also single and multilingual corpora are of importance. Here, we concentrate on MSA corpora, general and specialized.

There is no consensus on a single criterion for corpus quality and that explains the wide variety of measures for assessing corpora quality[5,6]. Next we list some of the corpora assessment methods discussed in the literature. Basically these are intrinsic measures in the sense that they relate to the corpus/document textual content, but not necessarily to its meta data: geo-origin, author, date, and so on. Rather than discussing each measure here and applying it to our corpora later, we opted to list many of the assessment measures here and provide the details together with the application to our two corpora in the next section.

Among the measure used in the literature and applied to our corpora are the following:

- N-gram frequencies of corpus content: word and character including punctuation marks and confusion letters.
- Word behavior including issues like token-to-type ratio (TTR), adherence to Zipf's law, vocabulary growth, PoS behavior, corpus homogeneity, domain broadness and specialized knowledge content.
- Length of words, sentences and related constructs and word length and sentence length distributions.
- Corpus cleanliness: including issues like monolinguality (purity), error rate and Out of Vocabulary occurrences (OOVs).

III. Experimentation on Corpora Quality

A. *Character Frequency, Normalization and Confusion Letters*

One of the problematic issues of everyday Arabic writing is tolerance for certain spelling errors, basically dealing with the confusion letters: various forms of Hamza and Alef (basically dropping the Hamza in favor of an Alef), the dotting of Ha and Ta Marbouta and also the dotting of Alef Maqsoura and Ya. The latter pairs differ only in dotting (absent in the first element, present in the second). Despite the resulting ambiguity the problem is so widespread that many resort to normalization: having a single form for each of the three classes. We believe that the writing system should be less tolerant of such errors. However, we would like to study the proportion of each of these forms in the correctly written spelled Arabic text or texts with limited such errors. For that we processed the Articles of the Arabic Wikipedia after some cleaning (removal of non-Arabic text, links, and ignoring articles of less than 50 words. It is the authors' view that the Wikipedia is almost free of such errors, because of the editing that goes into it and based on random inspections. Table I. gives the relative frequencies of the Arabic letters. The confusion groups are given both as individual shapes and as a single group. The group frequency reflects the standard frequency of the text with normalization and the individual shapes frequency reflects the standard frequencies in a well written Arabic text/corpus uses. For us the deviation of individual shapes from these frequencies is a corpus quality measures. The larger the sum of absolute differences, the worse the quality of the text. To see that we can compare the two corpora (AlQuds and Wikipedia), as in Table I.

From Table I, we notice that little differences exist in character frequencies for formal texts (Arabic Wikipedia and AlQuds) with minor variations for different text categories as reflected by the small SD.

However, that may not apply to arbitrary texts, and in particular those not undergoing formal editing. Also that is not the case for the Egyptian Wikipedia (ARZ), which seems to be less strict with the confusion letters despite being edited.

TABLE I. Letters Relative Frequency in our Corpora

Letter	Relative Frequency			Standard Deviation for all 9 categories
	Arabic Wikipedia (AR)	AlQuds	Egyptian Wikipedia (ARZ)	
ا	0.118584	0.127587	0.118659	0.004352
أ	0.014788	0.010328	0.006423	0.001139
إ	0.006319	0.004022	0.002359	0.000655
ء	0.002469	0.003007	0.000953	0.000428
ؤ	0.000694	0.001151	0.000518	0.000216
آ	0.000868	0.000652	0.000369	0.000235
ئ	0.00293	0.003865	0.001283	0.000429
ا+أ+إ+ء+ؤ+ئ	**0.146652**	**0.150612**	**0.130565**	**0.00444**
ه	0.01985	0.018656	0.034016	0.003337
ة	0.026745	0.026671	0.012304	0.002534
ه+ة	**0.046595**	**0.045327**	0.046319	0.001984
ي	0.067045	0.060906	0.057745	0.003446
ى	0.006594	0.006753	0.020583	0.000387
ى+ي	**0.073638**	**0.067659**	**0.078327**	**0.003242**
ب	0.030447	0.026228	0.03323	0.002952
ت	0.035278	0.03427	0.032112	0.005632
ث	0.005238	0.003978	0.002314	0.000505
ج	0.011721	0.011176	0.010865	0.001041
ح	0.014211	0.014897	0.014645	0.001223
خ	0.006501	0.006189	0.005497	0.000719
د	0.025613	0.024578	0.025381	0.001476
ذ	0.004652	0.004502	0.001173	0.00069
ر	0.039362	0.03662	0.041207	0.00275
ز	0.005656	0.005545	0.005582	0.000676
س	0.021293	0.021499	0.024146	0.001523
ش	0.008068	0.007793	0.009297	0.000688
ص	0.007289	0.007013	0.008757	0.000861
ض	0.004532	0.005392	0.003503	0.000866
ط	0.007522	0.008769	0.006828	0.000995
ظ	0.001742	0.001699	0.00153	0.000463
ع	0.025848	0.026519	0.023836	0.001478
غ	0.004311	0.002932	0.003496	0.000519
ف	0.020782	0.020321	0.021098	0.001936
ق	0.016159	0.01727	0.014118	0.001288
ك	0.017869	0.013614	0.019293	0.00144
ل	0.091042	0.093332	0.082373	0.003424
م	0.051613	0.05056	0.051839	0.001978
ن	0.042772	0.039874	0.044429	0.00208
و	0.046473	0.043316	0.049382	0.001513
The rest	**0.545994**	**0.527886**	**0.53593**	**0.002919**
Corpus size: Characters	**331,209K**	**871,081K**	**9,402K**	

We can talk about a quality measure reflecting attention to confusion letters or following writing rules in corpus A, let's call it Strictness S, where:

$$S_A = 1 - \Sigma_i |F_i - F_i^A| \quad (1)$$

Where "*i*" is in {ا أ إ ء ؤ ئ ه ة ى ي } and F_i is the reference frequency of the letter shape and F_i^A is the actual frequency of that shape in corpus A. For us, we take the reference frequency to be that of the Wikipedia.

For comparison, we calculate the S parameter for the Egyptian Wikipedia and AlQuds corpus as well, with the given frequencies of confusion letters in Table I.

$$S = 1 - |ا^W - ا| + |أ^W - أ| + |إ^W - إ| + |آ^W - آ| + |ء^W - ء| + |ئ^W - ئ| + |ؤ^W - ؤ| + |ة^W - ة| + |ه^W - ه| + |ي^W - ي| + |ى^W - ى|$$

$S_{(Arabic\ Wikipedia)} = 1.0$ $S_{(AlQuds)} = 1 - 0.025472 = .974528$

$S_{(Egyptian\ Wikipedia)} = 1 - 0.068134 = 0.931866$

B. Stylistic Measures: (punctuation, sentence length, stop word behavior)

1) Punctuation:

Fig. 1, and Fig. 2, show the relative frequency of the punctuation signs while processing the Arabic Wikipedia and AlQuds Newspaper.

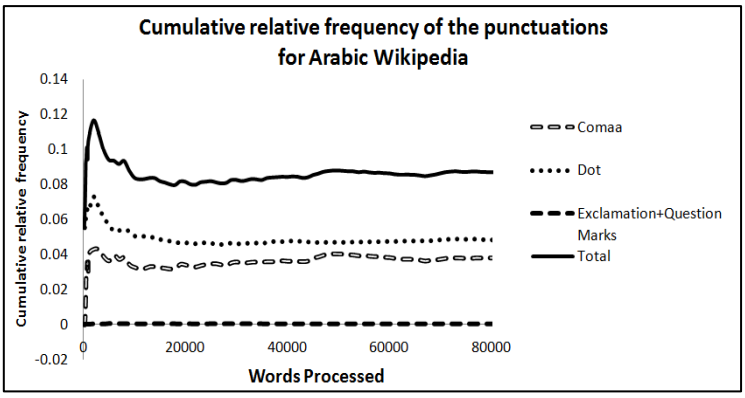

Fig. 1. Cumulative relative frequency of the punctuations for Wikipedia

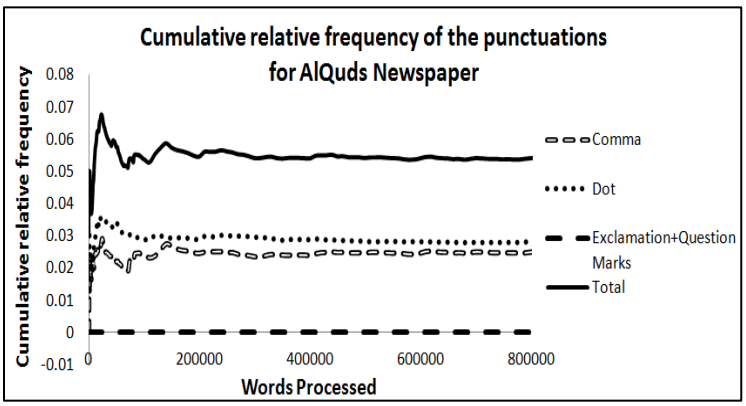

Fig. 2. Cumulative relative frequency of punctuation characters for AlQuds

The usage stabilizes early relative to the size. However, we note the difference between the two corpora in terms of employing punctuation in favor of the Wikipedia (See Fig 3.) . We believe that more usage of punctuation is generally a quality indicator in a corpus.

2) Sentence length:

Fig. 4, and Fig. 5, represent the distribution of sentences based on number of words/characters in the Arabic Wikipedia and AlQuds.

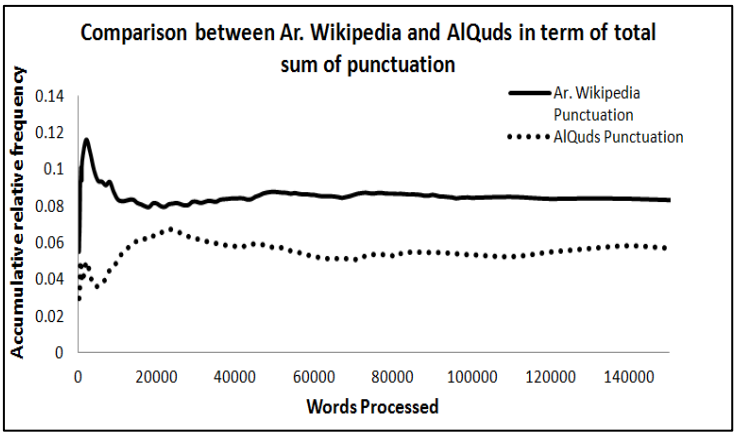

Fig. 3. Cumulative relative frequency: punctuations for AlQuds, Wikipedia

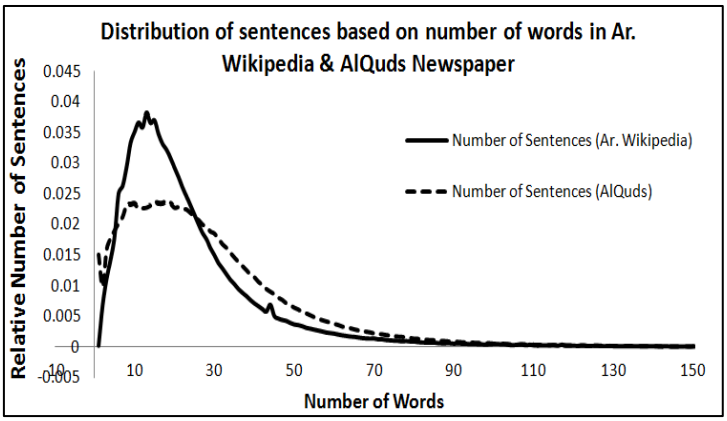

Fig. 4. Distribution of sentences length (in words) for Wikipedia, AlQuds

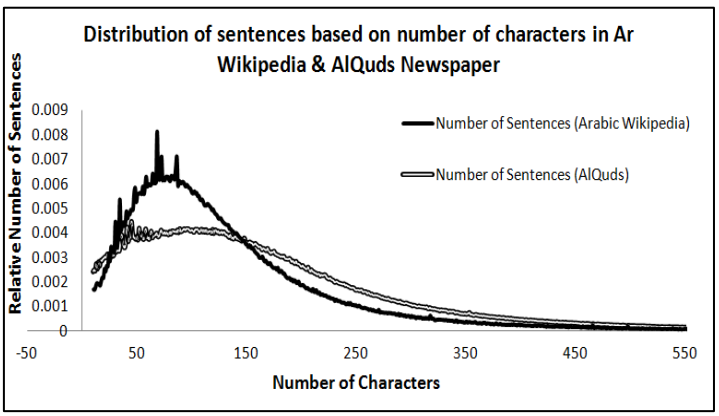

Fig. 5. Distribution sentence length (in characters) for Wikipedia and AlQuds

The peak sentence length in words for Arabic Wikipedia is 13 words and the average sentence length is 25.6 words. For AlQuds Newspaper the peak sentence length is 18 and the average sentence length is 29.5 words.

Regarding characters, the peak sentence length in characters for Arabic Wikipedia is 68 and average sentence length is 149.5 characters. For AlQuds Newspaper he peak sentence length in is 45 and average sentence length is 178.5 characters.

3) Stop Words Behavior:

We studied the behavior of stop words frequencies for the corpora, general and categorized, and how they change as the corpus size grows. We omit the details for space considerations but we can state that the relative frequencies of these words stabilize as early as we process 90K words or even less.

C. Word Statistical Behavior Patterns:

1) Word length distribution:

For Arabic one can anticipate longer words compared to English, with topical variations[4,8,19,20]. Fig. 6, shows the relative frequency of tokens distributed by the number of characters for both corpora. The distributions are quite close with both peaking at about 5 characters. Fig. 7, shows the same but for types (distinct words), and we added the graph for a list of 9 million unique Arabic words that we use later as our word look-up dictionary.[5] This is more a characterization of the corpus vocabulary. The Wikipedia looks more normal with a peak at 6-7 characters (the same for AlQuds), while the dictionary words are skewed towards longer (and it seems less used) words with the peak at 8 characters. AlQuds corpus seems to have a high proportion of longer words maybe reflecting a larger percentage of concatenated words with low frequency. This was borne out by distributions with least frequent words, probably concatenated words errors, removed.

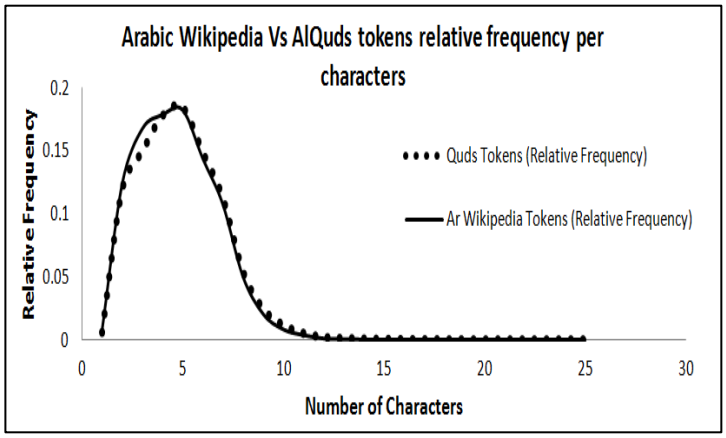

Fig. 6. Relative frequency for token length: Arabic Wikipeida

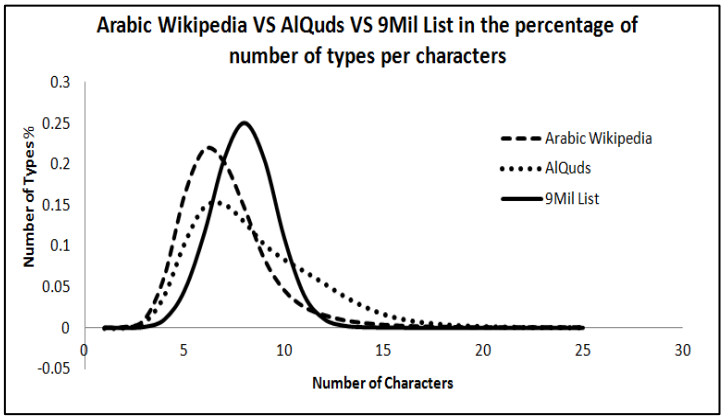

Fig. 7. Disribution of Type length for Wikipedia, AlQuds, 9Mil dictionary

2) Part of Speech Statistics:

We used Stanford PoS Tagger to label our cleaned corpora text[6]. We split each corpus to sentences, and then we input those sentences (each corpus alone) to the tagger. The output of the tagger is a tag for each word in each sentence that represents the word part of speech. Fig. 8 shows the results.

3) Type-to-Token Ratio (TTR):

TTR is the number of tokens processed over the number of distinct types found. Of interest here are both the values and the way the number changes as the size of the text grows. Here too one is expected to notice substantial difference between Arabic and English due to the writing rules (affixes): one can expect a consistently larger TTR for Arabic[19,20].

The Wikipedia (AlQuds) text corpus has 51,754,172 (125,225,339) tokens, 1,062,486 (1,749,247) types and a final TTR of 48.7 (71.6), 97.04 (172.7) when we consider only tokens with a frequency of more than 1 and 288 (525) when we consider only tokens with a frequency of more than 10.

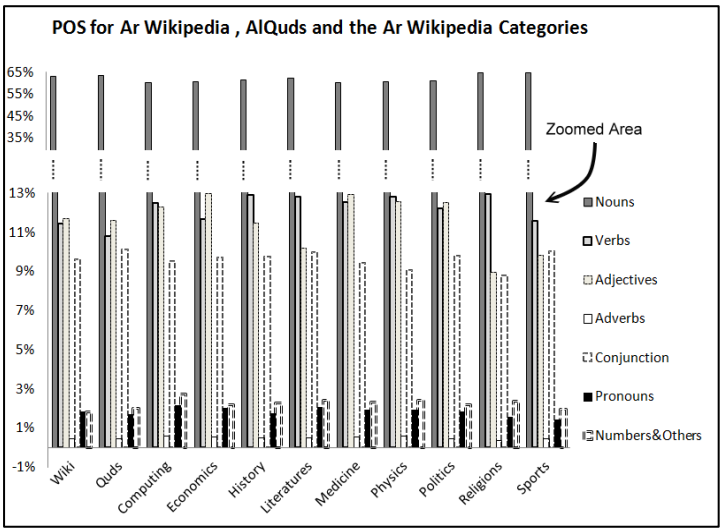

Fig 8. PoS tags for Wikipedia, AlQuds and Wikipedia Categories

[5] http://sourceforge.net/projects/arabic-wordlist/

[6] http://nlp.stanford.edu/software/tagger.shtml

It is of interest to see the behavior of specialized vs general corpora, say by comparing AlQuds or General Arabic Wikipedia with Categorized Arabic Wikipedia sub-coropra at a given point (where the texts are equal: for example by considering the shortest specialized corpus and having TTR at that size for all corpora). These results are given in Fig. 9.

Given that for word count *i*, *TTR(i)= Tk(i)/Ty(i), where Tk(i)* is the token count and *Ty(i)* is the type count, Fig. 9, shows the TTR(i) for AlQuds vs Arabic Wikipedia. Note that the lower graph reflects a richer vocabulary for the same corpus size. This is in line with the richer knowledge content of the Wikipedia. Fig. 10 shows the TTR for Wikipedia Categorized Corpora.

Fig. 11 and Fig. 12 show the TTR calculations for AlQuds vs the Arabic Wikipedia and the Arabic Wikipedia Categories based on word stems using Khoja Stemmer [7].

Note the larger difference in vocabulary richness in favor of the Wikipedia in Fig. 11, probably reflecting diversity in topics and the vocabulary of the Wikipedia.

In Fig 13 we give the graph of the TTR change with the corpus size increase computed as TTR'(i)=(TTR(i+1)-TTR(i))/(Tk(i)-Tk(i-1)). It is more of the slope at point i.

One can observe better stability in AlQuds, maybe reflecting topic changes in the Wikipedia corpus. Stabilization seems to take place around 25K words.

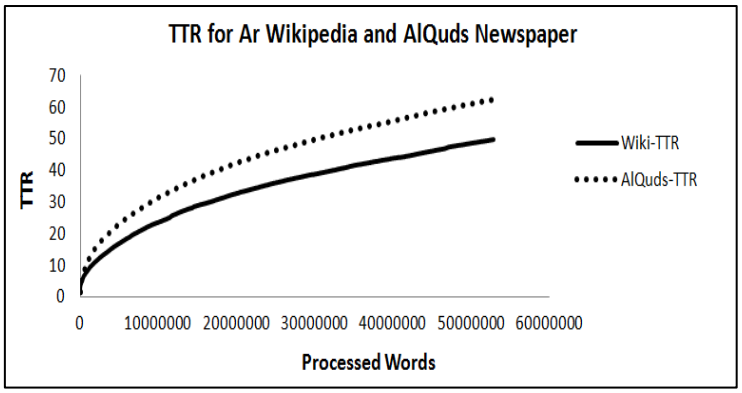

Fig 9. TTR For Ar. Wikipedia and AlQuds Newspaper: Tokens/Types

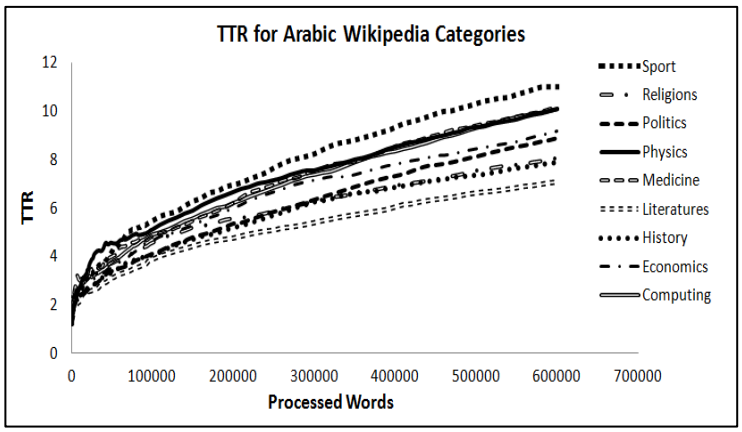

Fig 10. TTR for Ar. Wikipedia categories: Tokens/Types

[7] http://zeus.cs.pacificu.edu/shereen/research.htm

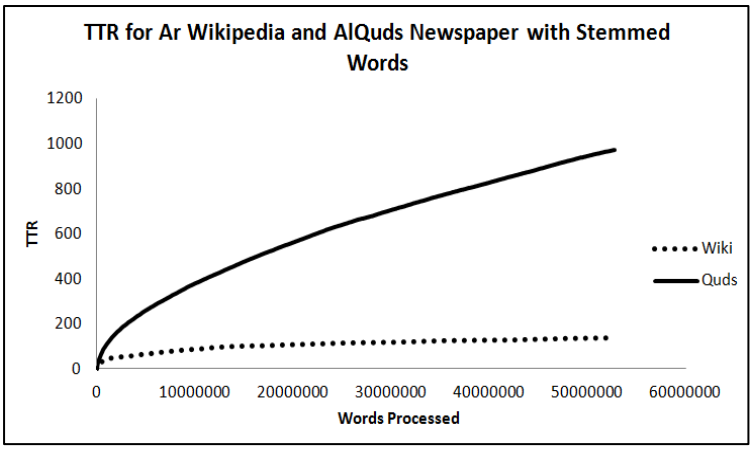

Fig 11. TTR for Ar. Wikipedia, AlQuds: Tokens/ Stemmed Types

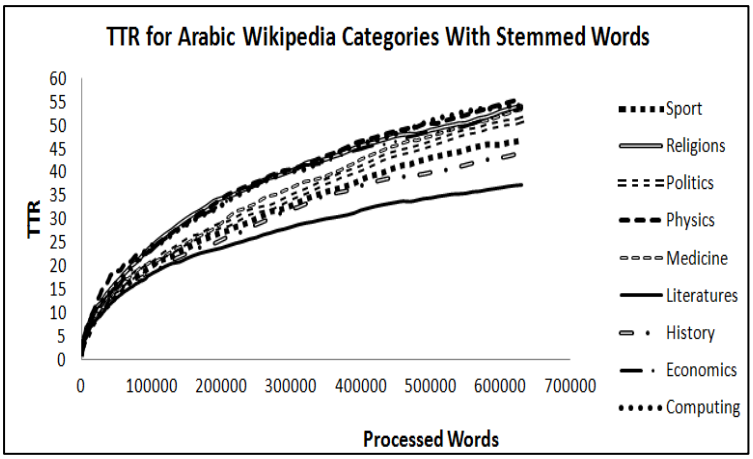

Fig 12. TTR for Ar. Wikipedia categories: Tokens/ Stemmed Types

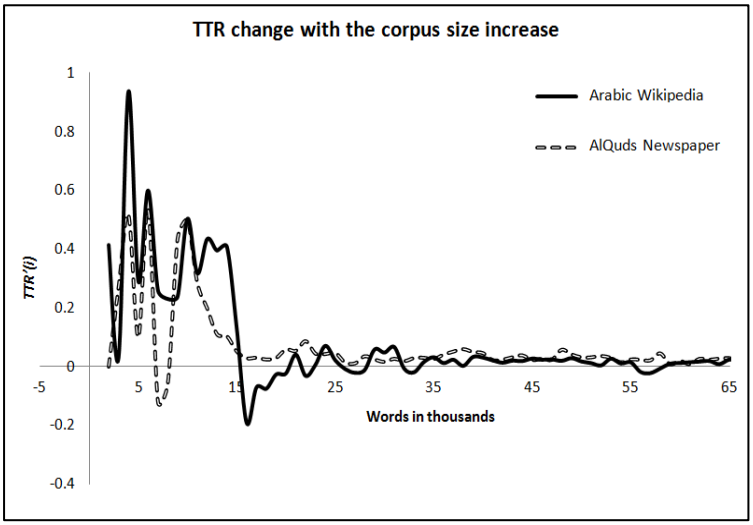

Fig 13. TTR change with the corpus size increase

4) Variety and Complexity:

Two related simple measures are[17]:

- Variety:

$$V = n/\log(N) \qquad (2)$$

Where *n* is the number of types and *N* is the overall number of tokens. Using our notation, (2) will become (2') :

$$V(i) = T_y(i)/log(T_k(i)) \quad (2')$$

- Complexity:

$$C = W*log(S) \quad (3)$$

Where *W* is the average word length in characters and *S* is the average sentence length in words.

Table II. shows the values for Variety and Complexity for our corpora, in general for Wikipedia, AlQuds and AlQuds at Wikipedia size (51,754,172 words of AlQuds), and for the first two cases when less frequent words (frequency 1 and 10 and less) are discarded, given as Variety@0, Variety@1 and Variety@10, respectively.

5) Homogeneity:

Given that a corpus is based on documents from various sources, homogeneity is the degree to which different parts of the corpus exhibit similar behavior (in terms of frequency distributions). The schemes for calculating this parameter may be elaborate[19,11]. Here we limit ourselves to a crude estimation of homogeneity by comparing the (relative) frequency distributions for the top 1000 words of text chunks of sufficient size each (1/10 of the corpus) to the entire corpus. We calculate Kullback-Leibler (KL) distance measure[17,15] (DKL). The metric is the sum of absolute differences between compared corpora relative word frequencies (Equation (4)) and smaller values reflect more similarity.

$$D_{KL}(P,Q) = \sum_i P(i).log\frac{P(i)}{Q(i)} \quad (4)$$

Where *P* is sub corpus *(chunk)* relative frequency distribution and Q is the entire corpus distribution. The sum is over the top 1000 words in the entire corpus.

Table III holds the DKL calculations for the Arabic Wikipedia and AlQuds newspaper.

6) Spelling Errors:

Error Rate, is the percentage of the language elements found in the text but not in the standard vocabulary[9]. Reference [9] gives 5% as the max error rate acceptable for corpus certification.

Error Dispersion, specifies the repetition of errors in text.

$$Error\ Rate = (EW * 100)/TW \quad (5)$$

Where *EW* is the number of error tokens and *TW* is the total number of tokens.

$$Dispersion = 100 - (ER/TE)*100 \quad (6)$$

Where *ER* is the number of repeated errors and *TE* is the total number of errors.

TABLE II. VARITY AND COMPLEXITY

Corpus	Tokens	Types	Variety @0	Variety @1	Variety @10	Comp-lexity
Wikipe-dia	51,754,172	1,062,486	137,735	68,642	23777	6.98
AlQuds	125,225,539	1,749,247	216,018	88,849	30396	6.90
AlQuds at Wiki	51,754,172	828,066	107347	-	-	-

TABLE III. DKL VALUES FOR AR. WIKIPEDIA AND ALQUDS

Arabic Wikipedia		AlQuds Newspaper	
Chunk	DKL	Chunk	DKL
C1	0.002346	C1	-0.03120
C2	0.009445	C2	-0.03447
C3	-0.003152	C3	-0.03356
C4	0.001856	C4	-0.03327
C5	-0.001910	C5	-0.03376
C6	0.001853	C6	-0.03462
C7	0.001158	C7	-0.03491
C8	0.001675	C8	-0.03357
C9	-0.011570	C9	-0.03529
C10	-0.003096	C10	-0.03467

Example: If we have a corpus of 10,000 words and 128 errors of which only 32 are distinct (and thus 96 are repeated) then:

Error Rate is: 128/10,000=1.28% in all cases.

Dispersion is: 100- ((128-32)/128)*100)= 100-25=75%

If the errors are all different then the repeated errors are 0 and dispersion is 100-0/128*100 = 100%.

If the errors are all the same word then the repeated errors are 128 dispersion is 100-128/128*100= 0%.

To detect how many errors there are in our corpora we need a general words dictionary. For that we used a 9 million Arabic word list validated against Microsoft Word spell checker[8].

For double checking we extracted random samples from the 9 million list and ran it against Microsoft word 2010. All samples passed the test. We also generated a list of words with errors according to MSWord and ran them against the list. The words failed the test. We will be using this list as our reference for correct general words.

Table IV shows the result of running our corpora against the 9 million list. Note that we did this experiment twice, once *with* the neutralizing of the confusion letters so that إبراهيم and ابراهيم are treated as same word. And once *without* neutralizing the effect of confusion letters. In all cases we think that the neutralizing of confusion letters (normalization) is essential step for improving quality of any corpora.

It is worth mentioning here is that the error rate in the corpora will include the OOV (Out Of Vocabulary Words) rate, discussed next.

TABLE IV. ERROR RATES IN ARABIC WIKIPEDIA AND ALQUDS

	AlQuds Newspaper		Arabic Wikipedia	
	With	Without	With	Without
Correct Words	652,068	572843	560,072	487,460
Correct Freq	121,759,692	113,973,289	49,380,408	48,456,135
Error words	1,097,179	1,176,404	502,414	575,026
Error Freq	3,465,847	11,252,250	2,373,764	3,298,037
Error Rate	2.77%	9%	4.56%	6.37%
Dispersion	31.66%	10.45%	21.17%	17.43%

[8] http://sourceforge.net/projects/arabic-wordlist/

7) Specialized Knowledge Estimates:

One may use reference corpora also to estimate the specialized knowledge content of manuscripts/corpora through the study of their statistical characteristics[10]. If one uses a general purpose dictionary, the assumption is that highly specialized documents/corpora tend to have high out of vocabulary (OOV) words reflecting knowledge rich content, a formal language and low readability[10]. General texts on the other hand tend to have more common words and thus low OOV reflecting poorer specialized knowledge and maybe a less formal language. There are more factors to assess knowledge content and language formality like percentage of longer words, sentence statistics, use of passive voice and pronouns, and density of some knowledge patterns: expressions characterizing formal writing style and use of moderate amounts of English in Arabic texts [10] but we limit ourselves here to the basics.

To assess the density of general vocabulary we ran the same mechanism that we used to detect errors rate earlier, however here we will focus on the top words (starting with 1000 words and so on), we think it's reasonable to assume that any word that can't be found in the 9 million list and holds a high frequency in its corpus is not an error word, but an out of vocabulary word, since it is rare to have an error word with high repetition in a corpus. So we are assuming that we will have the lowest value for OOV if we tested, say top 1000 words, and we will have the highest value of OOV if we tested all the corpus (in this case OOV equals error rate that we found earlier).

Table V shows the OOV rates for the top N words (types) in both Arabic Wikipedia and AlQuds Newspaper. As long as we process more words it's likely to have a higher OOV and the rate begins to look more as an Error rate rather than an OOV rate. Please note that the values in Table V is calculated with neutralization of confusion letters.

8) Monolinguality:

The presence of foreign language content in corpora may cause problems in the way the corpus represents its language class. However, foreign language text may occur naturally in a corpus: entity names, quotes and plain confusion, say in social media. For Arabic, isolating foreign text is an easy issue for languages using other alphabets. For languages with Arabic alphabet the issue is not really problematic, and many of the shared words form Persian are already part of Arabic.

In Arabic, things may get problematic in the dialect content in Arabic corpora. This will be the focus when assessing the corpus quality using the approach in [16] to estimate the monolinguality (purity) of English Corpora relative, say, to German content. The idea is not to worry about individual words (they can be names: people's, songs,…) but rather about full sentences.

If the amount of foreign content is large then one can assume that the foreign content (if separated) will have the word distribution of a general corpus of the foreign language; more so for the most frequent words of that language.

TABLE V. OOV RATE WITH CORPUS SIZE GROWTH

Number of Types	OOV Rate → Error Rate	
	AlQuds	**Arabic Wikipedia**
1000	0.0%	0.062%
2000	0.038%	0.10%
3000	0.052%	0.15%
5000	0.10%	0.22%
10000	0.20%	0.38%
20000	0.39%	1.15%
30000	0.52%	1.54%
40000	0.62%	1.80%
50000	0.79%	2.07%
75000	0.89%	2.47%
100000	1.03%	2.73%
All Types	2.77%	4.56%
Mid-Range	1.38%	2.31%

Applying this to Arabic, if the MSA corpus has a high enough percentage (of, say, Egyptian Arabic (ARZ) then the ARZ most common words will have the distribution found in a general ARZ corpus. If we consider the most common words of ARZ, which are considered noise in an MSA corpus, we can have three cases[16]:

a) The word w in ARZ is also a word in MSA with a similar relative frequency (w has the same frequency in AR and ARZ). Then w will keep its distribution in the noisy corpus when compared to a clean MSA corpus. Example of such words are the common stop words: In إن , Ana أنا (stop words in both MSA and ARZ).

b) The word *w* in ARZ is also a word in MSA but *w* has a much lower frequency (in MSA) in which case *w* will have a higher distribution in the noisy corpus when compared to a clean MSA corpus but of course much lower than *w* frequency in ARZ. Example: Kida كده , Omal امال; Tamalli تملي (with different vocalization they translate into: "So" in ARZ and "his labor" in MSA, "how else" in ARZ and "hopes" in MSA, "still" in ARZ and "she dictates" in MSA, respectively).

c) The word w in ARZ is not a word in MSA. w will have a much higher frequency in the noisy corpus than its 0 value in the clean MSA corpus. Example: Izzai إزاي.

In case c, it is possible to use the frequencies of the noise language frequent words to estimate the amount of noise in the main corpus. Here is how it works: if we know that the word "Izzai إزاي " (ARZ for "how") only occurs in ARZ, and its frequency in ARZ corpus is x% and it shows up in the MSA corpus at relative frequency y% then the amount of ARZ in the MSA corpus is y/x. In case b we need to account for the original ARZ content in the clean MSA text in obvious ways.

For example, given a 20M word MSA corpus. If clean, "izzai إزاي" will appear zero times. In an ARZ pure corpus "izzai إزاي " appears 1% of the text: we have 100 words of ARZ for every occurrence of "izzai إزاي ". Now if "izzai" in the noisy 20M Word MSA corpus has 0.0001=0.01% frequency then the corpus has 0.0001*20,000,000=2,000 occurrences of "izzai إزاي " and therefore 200,000 ARZ words. The percentage of ARZ in the corpus is 200,000 /20000000=1/100=1%. That is 0.0001/0.01=0.01.

We tested this for the Arabic Wikipedia through adding content from the Egyptian Wikipedia at the 0.9% noise level and at the 10% noise level. Our computed noise level using

the above approach were 0.8% and 8.94% which, we believe, are close enough to render such an approach usable for Arabic. A better selection of the representative noise words may get us better results.

9) Zipf's Law:

The law characterizes the relationship between word frequency and rank in a large enough text corpus. The degree to which the corpus text has the expected distribution is a characteristic of the corpus quality. Zipf's law is an empirical law and is based on the observation that the frequency of occurrence of some events is a function of its rank in the frequency table[14,24]. Zipf's law for a corpus word is given in Fig 7:

$$f = C / r^{\alpha} \quad (7)$$

Where f is the frequency of the word, α is a constant close to 1 and r is the rank of the word and C is a constant. This equation states that multiplying word relative frequency by its rank is a constant and so the most frequent word will occur twice as often as the second most frequent word and so on. This means we can say that quality corpora should obey Zipf's law, thus we can test our corpora frequency tables and compare them with the ideal generated from Zipf's law and see how close is our corpora to the ideal. Closeness of the real to a straight line is an indicator of better quality[6].

Fig. 14, shows the log-log result for the Arabic Wikipedia and AlQuds newspaper corpora.

We also applied Zipf's Law to our Arabic Wikipedia categories. All categories exhibited similar behavior.

To assess the adherence of word frequencies of a corpus to Zipf's law some authors use Kullback-Leibler (KL) distance measure (DKL)[15,17]. We introduced DKL earlier, however here P is the ideal Zipf's distribution (based on rank) and Q is the actual corpus distribution

Table VI, shows the DKL values for Arabic Wikipedia, AlQuds Newspaper and the Arabic Wikipedia categories, for the top common 1000 words.

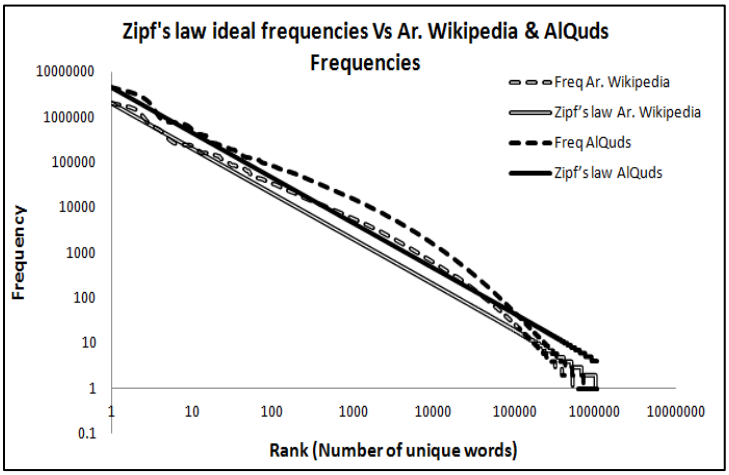

Fig 14. Zipf's law ideal/real for Wikipedia and AlQuds Corpora

TABLE VI. DKL VALUES FOR AR. WIKIPEDIA, ALQUDS AND AR. WIKIPEDIA CATEGORIES (BASED ON ZIPF'S LAW)

Corpus	DKL	Corpus	DKL
AlQuds	-0.084	Medicine Related	-0.136
Ar. Wikipedia	-0.100	Physics Related	-0.139
Computing	-0.143	Politics	-0.128
Economics	-0.134	Religions	-0.139
History	-0.134	Sports	-0.106
Literatures	-0.132		

10) Corpus Hardness Estimates

When a corpus consists of topical sub-corpora and may be used for categorization, one would like to test if the corpus is sufficiently diverse to support the role of a gold standard for categorization. [15] addresses this issue and gives several parameters to measure for that. We discuss the simpler among them here.

a) Domain broadness evaluation measures: basically characterizing how distinct the different categories of the corpus.

Given a corpus C with the constituent categorized corpora C_i for $i \in \{1,2...k\}$ then the vocabulary based broadness measure is given as (8)

$$SVB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|V(C_i)| - |V(C)|}{|C|}\right)^2} \quad (8)$$

Where $|V(C)|$ is the size of vocabulary of C, i.e. the number of distinct words in C. $|C|$is the size of C.
In the absence of categorization one can use (8')

$$UVB(C) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|V(D_i)| - |V(C)|}{|C|}\right)^2} \quad (8')$$

Where $\{Di\}$ for $i \in \{1,2...n\}$ are the constituent documents of C, or in the absence of that we may take Di to be the i^{th} chunk of C when C is divided into n, say 10, equal chunks.

b) Shortness: In order for the classification to work properly, one needs to have sufficiently long documents, with a large enough vocabulary in each. Shortness is meant to characterize this aspect of the corpus.

Here are the formulae to assess that applied to documents.

$$DL(C) = \frac{1}{n}\sum_{i=1}^{n}|D_i| \quad (9)$$

$$VL(C) = \frac{1}{n}\sum_{i=1}^{n}|V(D_i)| \quad (10)$$

c) Class Imbalance: A categorized corpus needs to maintain a balance between the sizes of subcorpora Ci in the various categories. This may be expressed in terms of document length. See (11).

$$CI(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(|C_i| - ENDC(C))^2} \quad (11)$$

Where ENDC(C) is the average number of documents per category = number of documents/number of categories and |Ci| is the number of documents in category C_i.

Table VII shows the result of applying Equations 8-11 to our corpora (where applicable).

IV. Discussion, Conclusions and Future Work

We have presented a collection of corpus assessment measures that we think can be used to evaluate general and specialized corpora. Our main argument can be that the Wikipedia corpus, with the performed cleaning and removing of suspect articles, with small word count and non-real content, can serve as the gold standard and deviations from that can be used as measures of corpus quality. One can take into account the dynamic nature of the Wikipedia corpus and may want to update the figures as the Wikipedia develops.

We performed a much larger suite of tests than reported here to conserve space. We plan to make these result accessible to the scientific community.

One may want to conduct more comparisons with other languages or with dialect material. We performed some experiments on the latter but the results are not complete. One may also want to employ different approaches for certain tasks. For example we experimented with using "well-formedness" of words as judged by a stemmer as a spell checking tool with encouraging results, but more work is needed there. One may also want to consider other, less explanative genres, like romance, and observe differences.

Another interesting aspect of our work is to compare the properties of the discussed corpora with others mentioned in the literature. Also coming up with an aggregate single measure to characterize corpora quality maybe interesting.

ACKNOWLEDGMENTS: The authors acknowledge the support of the Palestinian Ministry of Higher Education through grant 22/1/2013 and thank the 2 referees for their useful comments.

TABLE VII. Hardness Parameters Values

Variable	Arabic Wikipedia	AlQuds
SVB(C)	7.00E-03	-------
UVB (C)	5.33E-5	6.01E-5
DL(C)	335.45	--------
VL(C)	206.26	1281.82
CI(C) = 2,117,726 for Wikipedia Categories		

References

[1] M. Abbas and K. Smaili, "Comparison of Topic Identification Methods forArabic Language". International conference RANLP05 : Recent Advances in Natural Language Processing , 21-23 September 2005.

[2] A. Alarifi, Alghamdi, M, Zarour, M.Alraqibah, A. Alsadhan, K and L. Alkawi. "Estimating the Size of Arabic Indexed Web Content," Scientific Resaerch and Essays. Vol7(28). July 2012.PP. 2472-2483

[3] L. Al-Sulaiti, and S. Atwell. "The design of a corpus of contemporary Arabic," Int. J. Corpus Linguistics 11(2):135–171. 2006. http://www.comp.leeds.ac.uk/eric/alsulaiti06ijcl.pdf

[4] C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski and T. Zesch, "Scalable Construction of High-Quality Web Corpora". JLCL 2013 – Vol. 28 (2). PP 23-59.

[5] E. Darrudi, M. Hejazi and F. Oroumchian, "Assessment of a Modern Farsi Corpus". Proceedings of the 2nd Workshop on Information Technology and its Disciplines WITID2004.

[6] M. de Camp, "Explorations into Unsupervised Corpus Quality Assessment". Masters Thesis. Tilburg University. 2008.

[7] T. Eckart, U. Quasthoff, D. Goldhahn, "The Influence of Corpus Quality on Statistical Measurements on Language Resources". LREC 2012: 2318-2321

[8] A. Goweder, and A. De Roeck, "Assessment of a significant Arabic corpus". Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001. http://www.abdelali.net/ref/ACL-EACL%202001_goweder.pdf

[9] C. Grouin,"Certification and Cleaning up of a Text Corpus: Towards an Evaluation of the Grammatical Quality of a Corpus". LREC'08. 2008.

[10] J. Halskov, D. Hansen, A. Braasch and S. Olsen, "Quality Indicators of LSP Texts — Selection and Measurements Measuring the Terminological Usefulness of Documents for an LSP Corpus". LREC'10, 2010.

[11] A. Kilgariff, "Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity Between Corpora". Proceedings ACL-SIGDAT Workshop on Very Large Corpora, Hong Kong. (1997)

[12] A. Labadié and V. Prince, "The impact of Corpus Quality and Type on Topic Based Text Segmentation Evaluation," Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008, Wisla, Poland, 20-22 October 2008. PP. 313-319.

[13] M. Mahmoud, "Arabic Online Content: Web metric Study (in Arabic)," Veecos Website, April 16, 2011. Available at: http://www.veecos.net/portal/index.php?option=com_content&view=article&id=5997:2011-04-16-08-53-48&catid=42:doctorah&Itemid=180

[14] C. Manning, and H. Schuetze, "Foundations of Statistical Natural Language Processing," MIT Press.Cambridge, MA. 2013

[15] D. Pinto, P. Rosso, and H. Jimenez-Salazar, "On the Assessment of Text Corpora". Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems NLDB'09,Saarbruecken, Germany, PP. 281—290.Springer-Verlag.Berlin, Heidelberg. 2009

[16] U. Quasthoff and C. Biemann. "Measuring Monolinguality," In Proceedings of LREC-06 workshop on Quality assurance and quality measurement for language and speech resources. 2006.

[17] P. Rosso, Y. Benajiba and A. Lyhyaoui, "Towards a Measure for Arabic Corpora Quality," Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV, Damascus, Syria. PP11-14. 2006

[18] M. Saad, "The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification," Master's thesis, Faculty of Engineering, The Islamic University, 2011, Gaza, Palestine. http://library.iugaza.edu.ps/thesis/91986.pdf.

[19] A. Sarkar, A. De Roeck and P. Garthwaite, "Easy Measures for Evaluating non-English Corpora for Language Engineering: Some lessons from Arabic and Bengali". Open University Technical Report 2004/05. 2004 http://www.abdelali.net/ref/Tech_Rep_2004_05.pdf

[20] A.Yahya, "On the Complexity of the Initial Stages of Arabic Text Processing". First Great Lakes Computer Conference. Kalamazoo, MI. 1989.

[21] A. Yahya and A. Salhi, "Arabic Text Correction Using Dynamic Categorized Dictionaries: A Statistical Approach," Proceedings of 4th International Conference on Arabic Language Processing-CITALA 2012; Rabat, Morocco. May 1--2, 2012.

[22] A. Yahya and A. Salhi, "Efficiency Enhancement Tools for Arabic Search Engines: A Statistical Approach". Proceedings of Innovations2011: The Seventh International Conference in Innovation in Information Technology: Special Session on Arabic NLP ; Abu Dhabi, UAE. April 25-27, 2011.

[23] A. Yahya and A. Salhi. "Arabic Text Categorization Based on Arabic Wikipedia," 13, 1, Article 4 (February 2014), 20 pages. DOI=10.1145/2537129 http://doi.acm.org/10.1145/2537129

[24] G. Zipf. "Human Behaviour and the Principle of Least-Effort. Addison-Wesley," Cambridge MA:. (1949)