Research on Humanities and Social Sciences
ISSN (Paper)2224-5766 ISSN (Online)2225-0484 (Online)
Vol.5, No.4, 2015

Measuring Lexical Richness through Type-Token Curve: a Corpus-Based Analysis of Arabic and English Texts

Khalid Shakir Hussein
English Department, Thi-Qar University, Iraq
E-mail of the corresponding author: khalidshakir74@gmail.com

Abstract
WordSmith Tools (5.0) is used to analyze samples from texts of different genres written by eight different authors. These texts are grouped into two corpora: Arabic and English. The Arabic corpus includes textual samples from the Qur'an, Al-Sahifa al-Sajjadiyya (a prayer manual), Modern Standard Arabic (Mistaghanmy's novel Chaos of Sensations) and Imam Ali's Nahjul-Balagah (Peak of Eloquence). The English Corpus comprises The New Testament, Conrad's Heart of Darkness, Dickens' David Copperfield, and Eliot's Adam Bede. Each textual sample is statistically analyzed to find about its lexical richness or vocabulary size. The number of tokens (total number of words) and the number of types (distinct vocabulary words) are counted for each sample. Then both numbers are plotted against each other using Microsoft Office Excel diagrams. The resulted curves in both corpora give a vivid idea about the lexical richness of each textual sample. They open an active avenue to compare between the different authors in terms of their vocabulary size and the range at which they begin to exhaust their linguistic repertoire by repetition. The curves for Imam Ali's Nahjul-Balagah (Arabic corpus) and Conrad's Heart of Darkness (English corpus) rise up high reaching the maximum. By contrast, Qur'anic Verses and The New Testament have the lowest curve for the ritualistic quality of their texts.
Keywords: corpus stylistics, type-token curve, lexical richness

1. Introduction

As a fast-developing area of investigation, corpus linguistics has criss-cross interdisciplinary relations and influences on contemporary stylistics. It is beyond doubt that language description has been changed forever after the dynamic impact corpus linguistics has had on the scale of the surveyed data. Sizeable collections of computer-readable texts became available and easily downloadable. The corpus –based methodologies make it possible to explore the basic statistics of tremendous amounts of textual data targeting the nature of their different recursive patterns.

One of the central concerns in corpus linguistics is to bring out a rigorous description of various uses of language (Biber, 2011:16). Literature is most definitely an evident example of keen uses of language. Corpus stylistic investigations come into play as the increase in the size of targeted data becomes the rule rather than the exception. It would be more useful and informative to identify a particular stylistic feature coming across not a single poem, a novel, or a short story but a series of novels, an anthology of poems, dozens of short stories.
The basic question that the researcher addresses in this paper is how useful corpus stylistics is to the study of lexical richness, as a stylistic feature, across various literary genres (text-types) in English and Arabic.

2. The Circle of Corpus Linguistic Description and Literary Appreciation

The use of corpus linguistics (the electronic analysis of certain computerized language data) to consider the distributional characteristics of electronically stored literary texts is the major analysis carried out under the umbrella of corpus stylistics (ibid.). Like in traditional literary stylistics, the goal is still intact within corpus stylistics that is related to exploring the relationship that holds between language use and artistic function (Leech, 2007: 11). This relational goal is quite essential in any stylistic analysis whether it was traditional or empirical. What matters most is how to relate the findings of linguistic description with the critic's overdominating concern of literary appreciation. Leech's (ibid:12) traditional cyclic motion figures out a relational bounce whereby linguistic description contributes into the verification of certain literary insights and whereby literary appreciation seeks a sort of linguistic evidence or validity. Figure 1 below represents the cyclic relation between traditional linguistic description and literary appreciation.

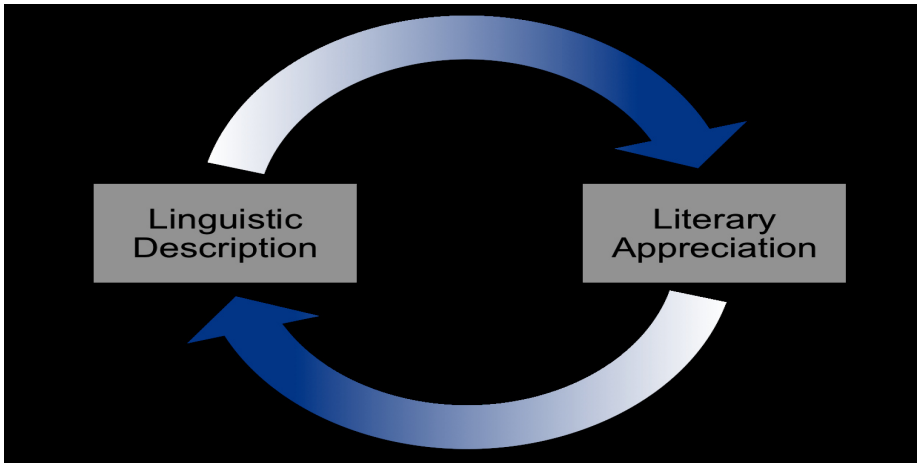

Figure 1. The Circle of Explanation (Taken from Mahlberg, 2013:15)

Within the domain of corpus stylistics, this figure goes into a further modification that backs up a more rigorous description of language based on employing corpus-based techniques to explore the quantitative behavior of words in literary texts.

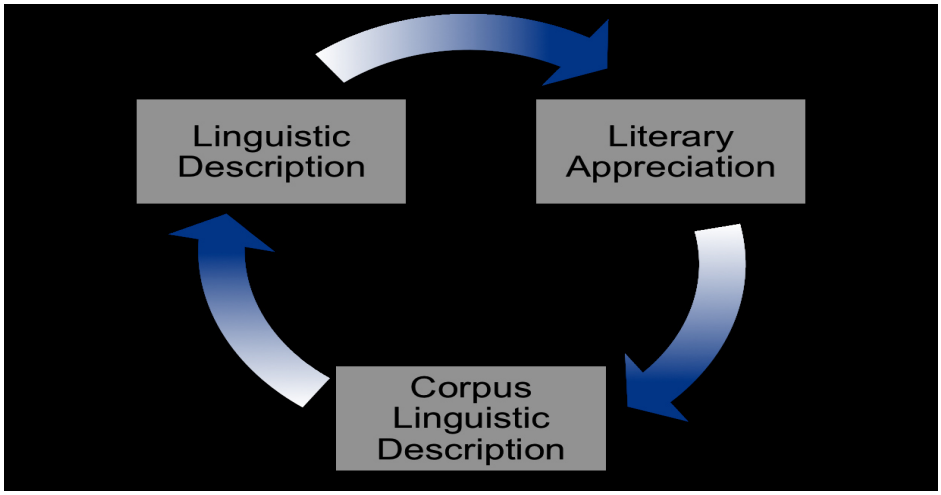

Figure 2. Corpus Linguistic Description slotted within the Circle of Explanation (Taken from Mahlberg, 2013:17)

As Figure 2 above shows, *corpus linguistic description* has a serving role in the circle of explanation. Its explanatory role appears to be of two folds: on the one hand it can bring a sort of confirmation to the intuitions critics might have about style. On the other it attracts attention to a more foregrounded set of features that might otherwise be overlooked by traditional linguistic description, i.e. features that contribute to the empirical and objective measurement of style. However, this empirical type of description is still clinging more to the domain of linguistic description than to the artistic function sought in literary considerations (Mahlberg, 2013:19). It might be put in an equal relation to both linguistic description and literary appreciation, but the numerical data it provides are always adapted to the needs of the objective linguistic description of texts.

3. Lexical Richness and Type-Token Curve

The relation between types and tokens in a text used to be the focus of many quantitative investigations of vocabulary (see Francis and Kucera 1982, Wimmer 2005, Milička, 2009). However, Peirce in 1906 was the first to recognize the distinctive power of this relationship (Wybraniec-Skardowska, 2007:164). This relation is usually reported in a form of ratio between types and tokens for a given sample of text. Type-Token ratio (henceforward TTR) is basically one particular statistical strategy of *tokenization* that encompasses a given set of basic statistical measures used in looking at texts with computerized methods. Whatever was the strategy used in tokenizing a digital text, it always utilizes the same statistical concepts of *tokens* and *types*. A token is any single linguistic unit, most often a word, in a text (Baker et al, 2006: 159). While the number of tokens in a computerized database refers to the total number of words. As for the number of types, it refers to the total number of the unique distinct type of words (ibid: 162). Therefore, a token is any linguistic item that occurs in a

text regardless of its type, whereas a type is a statistical concept that targets only the token-types involved in a surveyed corpus. Comparing the number of *tokens* in the data or corpus to the number of *types of tokens* can tell us much about how large a range of vocabulary is used in the corpus under consideration. For example, the sentence, (The man put the book on the table), contains (eight) tokens (*the, man, put, the, book, on, the,* and *table*) but only (six) types (*the, man, put, book, on,* and *table*). So its TTR is simply (6 ÷ 8 = 0.75) types per token.

As a text gets longer, the number of types that have already been encountered increases, and the likelihood of any given token representing a new type goes down (Hardie & McEnery, 2006:139). For this reason, TTR as a statistic measure is very sensitive to the length of the text under investigation. Therefore, it is extremely crucial when comparing the TTRs of different texts to make sure that the texts concerned are of equal sizes (Baker et al, 2006:138). Generally speaking, a high TTR indicates a large amount of *lexical variation* and a low TTR indicates relatively little *lexical variation* (ibid.).

As with *lexical density* or *richness*, referring to a measure calculated by counting the number of lexical words in a text that occur only once, the TTR can also be used to monitor changes in the use of vocabulary items throughout one text or throughout a group of texts produced by one person (Butler, 1995: 135).

Youmans (1990) introduces this statistical relationship in a quite interesting and straightforward way that visualizes what he calls *type-token vocabulary curves* (ibid:584) or simply *type-token curves*. The statistical rationale behind the visual plotting of these curves is carefully expressed by Youmans (ibid: 586):

> They begin as a straight line, with types = tokens, until the first repeated word. Thereafter, the number of tokens exceeds the number of types, and this margin grows larger with every additional repetition. Consequently, type-token curves rise rapidly at first, then begin to lose momentum as repetitions become more frequent and the author's vocabulary is used up. The number of types reaches its maximum when the author's vocabulary is completely exhausted. Hence, as the number of tokens approaches infinity, the number of types approaches the total vocabulary of the author.

Then, when the curve loses its *momentum*, it stops rising and takes a relatively stable straight line that declares the exhaustion of its lexical size. Statistically speaking, there would be no more *types* to be added and all what would be there is only a series of *tokens* repeated over and over again. The size of the vocabulary used by the writer starts to shrink and this can be very helpful in estimating the lexical richness of a given text.

3.1 Morphological Challenges

Any corpus-based analysis of the lexical richness of digital or computerized texts must involve a clear cut concept of what counts as *word* (token) and *distinct word* (type). Arabic derivational morphology, in particular, is quite complex and demanding. In all layers of Arabic, the bulk of the vocabulary is built on the principle of root and pattern (Watson, 2007:124). To express certain semantic terms (i.e., words), a purely consonantal root carrying the basic semantic information is combined with a limited set of patterns using a fixed sequence of consonants, vowels, and optional prefixes and suffixes (Prochazka, 2006:426). Accordingly, a complete quantitative analysis of a text would most definitely involve dividing words into their component morphemes and affixes. Nevertheless, the researcher borrowed Francis and Kucera's concept (2000) of *graphic word* to conduct this quantitative study: "Graphic word: a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks." (ibid:4). Therefore, the Arabic words such as *أَنُلْزِمُكُمُوهَا, ذهبوا, ذهبا, ذهبت, ذهب, منها, منه, منا, من* would count as single words, as for the apostrophe or hyphens, they count only in English and are unfamiliar in Arabic. It follows then that words such as *can't, cannot, and twentieth-century* count as single words in English texts. Moreover, any identical *alphanumeric strings* would be recognized as being the same graphic word or *type*.

4. Text Corpora

The corpora used in this study consist of texts downloaded from the World Wide Web. Two digital corpora have been compiled throughout the study: one in Arabic and one in English. Each corpus comprises various *genres* (text types). The Arabic includes four genres with five samples of each: *Quranic Verses*, *prayer manual* (Al-Sahifa al-Sajjadiyya), *sermons* (Nahjul-Balagah), and *modern standard Arabic* (Mistaghanmy's novel *Chaos of Sensations*).

The English corpus comprises fifteen digital samples taken out of three English novels, Conrad's *Heart of Darkness*, Dicken's *David Copperfield*, Eliot's *Adam Bede*, and five more samples of *The New Testament*.

5. Analysis Procedures

Five methodological procedures will be followed in analyzing the Arabic and English corpora:

1. Taking random five samples from each of the eight texts involved in this study with approximately 1000-token intervals: 1000, 2000, 3000, . . .etc

2. transcribing every individual sample into *plain text format*
3. the total number of words (tokens) and the number of distinct vocabulary words (types) are computed for each sample via *WordSmith Tools (5.0)*
4. importing *WordSmith Tools (5.0)* outputs into an excel spreadsheet in a form of matrix
5. plotting the number of types against the number of tokens using Microsoft Office Excel diagrams to produce type-token curves.

6. Type-Token Curves Analysis

6.1 Arabic Corpus

A summary of the *type-token statistics* for the Arabic corpus is listed in Table 1 below.

Table 1. Arabic Corpus Statistics

Texts	**Samples Numbers**									
	1		**2**		**3**		**4**		**5**	
	Tokens	Types	Tokens	Types	Tokens	Types	Tokens	Types	Tokens	Types
Quranic Verses	1,159	610	2,299	998	3,481	1,408	4,626	1,771	5,223	1,911
Nahjul- Balagah	1,106	717	2,084	1,316	3,144	1,777	4,161	2,426	5,096	2,892
Al-Sahifa al-Sajjadiyya	1,010	515	2,002	903	3,016	1,313	4,010	1,565	5,018	2,159
modern standard Arabic	1,020	634	2,016	1,149	3,014	1,596	4,025	1,822	5,013	2,199

Plotting the statistics in Table 1 above, Figure 3 below shows interesting type-token curves. One of the most evident features of this Figure is that the type-token curves for *Al-Sahifa* and the *Quranic Verses* are virtually identical for the first 4000 words, after which they gradually diverge. However, this divergence is still a graphical evidence that *Al-Sahifa's* vocabulary is larger than that of the *Quranic* samples.

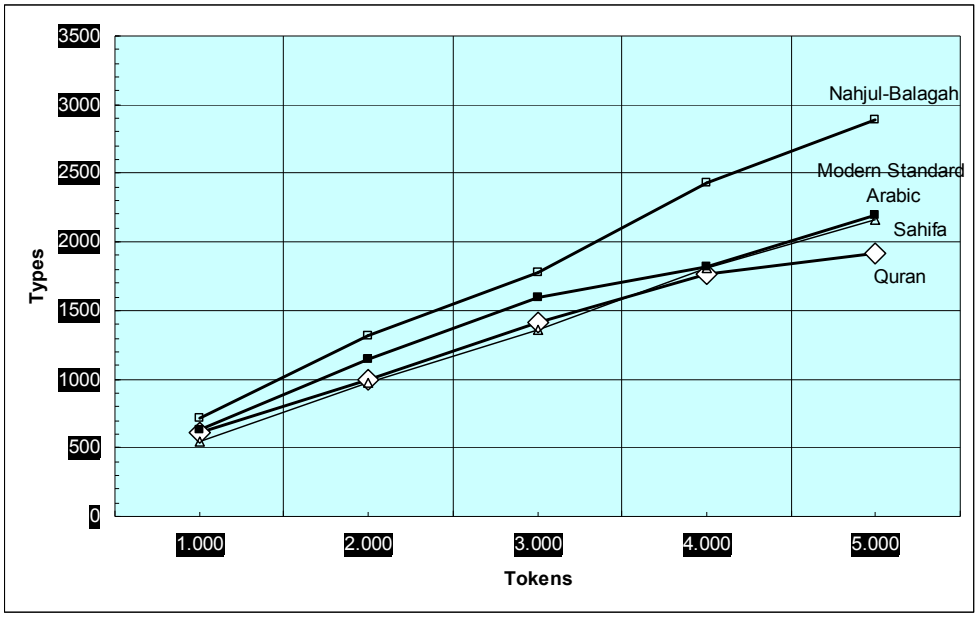

Figure 3. Type-Token Curves for *Quranic Verses*, *Al-Sahifa*, *Modern Standard Arabic* and *Nahjul-Balagah*

Statistically speaking, what brings the *Quranic Verses* and *Al-Sahifa* together is the very ritualistic nature of both. This nature is reinforced by certain types of *liturgical repetitions* that enhance the religious ceremony expected in these two sacred books. For example, the *Quranic Verse* (فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبَانِ - Then which of the favours of your Lord will ye deny?) has been occurred (31) times in the samples of the study, the word (دُعاء - supplication) is scored in different Quranic contexts with its various derivations:

> **Supplicate** your Lord humbly and secretly; He loves not transgressors. (7:55)
> **Supplicate** Allah or **supplicate** the All-merciful. Whichever you **supplicate** - to Him belong the most beautiful names. (17:110)
> **Supplicate** God, making your religion His sincerely, though the unbelievers be averse. (40:14)
> Your Lord has said: `**Supplicate** Me and I will respond to you. Surely those who wax too proud to

worship Me shall enter Gehenna utterly abject.' (40:60)
And when My servants question thee concerning Me - I am near to respond to the **supplication** of the **supplicator** when he **supplicates** Me. (2:186)

The same *liturgical repetitions* are evidently observed in *Al-Sahifa* samples. For example (محمد– Muahmmed) occurs 245 times, (الله– Allah) 395, (صل الله على محمد و آله - May Allah pray on Muhammed and his household) 34, (اللهم - Oh Allah) 41, etc. Besides, every prayer in *Al-Sahifa* has the same overall structure beginning with (praying on Muhammed and his household), then moving to praising *Allah*. As for the final part of the prayer, it is always concluded with (praying on Muhammed and his household) as well.

These types of repetition appear evidently as the major source of depression in the type-token curves for *Quranic Verses* and *Al-Sahifa*. The remarkable convergence of the two curves is an obvious evidence of the similar vocabulary size of both. This might sustain the claims that *Al-Sahifa* is known as "Sister of the Qur'an", "Gospel of the Folk of the House" or "Psalms of the Household of Muhammed" (see http://www.al-islam.org/sahifa-al-kamilah-sajjadiyya-imam-zain-ul-abideen). It is said to be the oldest prayer manual in Islamic sources and one of the seminal works of Islamic spirituality (ibid.).

The next upper range of lexical richness is occupied by a curve representing the samples of *modern standard Arabic* (Mistaghanmy's novel *Chaos of Sensations*). Mistaghanmy's vocabulary looks larger than that of *Quranic Verses* and *Al-Sahifa* especially in the first three samples. The curve diverges upwards in these three samples and then converges into the curve of *Al-Sahifa* in the fourth and fifth samples. This might mean that *Modern standard Arabic* in samples of Mistaghanmy's *Chaos of Sensations* drew upon a larger potential vocabulary than the samples of *Quranic Verses* and *Al-Sahifa*.

Moreover, the highest type-token curve is extraordinarily reserved for *Nahjul-Balagah*. The 2892-type vocabulary introduced in the 5096 tokens of *Nahjul-Balagah* is quite larger than that of its nearest curve (*Chaos of Sensations*) that scored 2199-types vocabulary counted in 5013 tokens. This is why *Nahjul-Balagah* (Peak of Eloquence) is well-known for its highly distinctive *genre* and *style*. The distinctive figurative style of the book and the various genres it involves crucially contribute to its statistical uniqueness within the range of the type-token curves involved in this study. It comprises a collection of *sermons*, *letters*, and *sayings* attributed to Imam *Ali Ibn Abi Talib*, cousin and son-in-law of Prophet *Muhammed*. This richness of *genres* and the different topics involved in the samples contributed most essentially in pushing the type-token curve of *Nahjul-Balagah* into upper limits within the range of the curves in Figure 3 above.

Figure 4 below gives a more vivid diagram of the wide variations for type-token curves and sets the limits between the highest and the lowest possibilities.

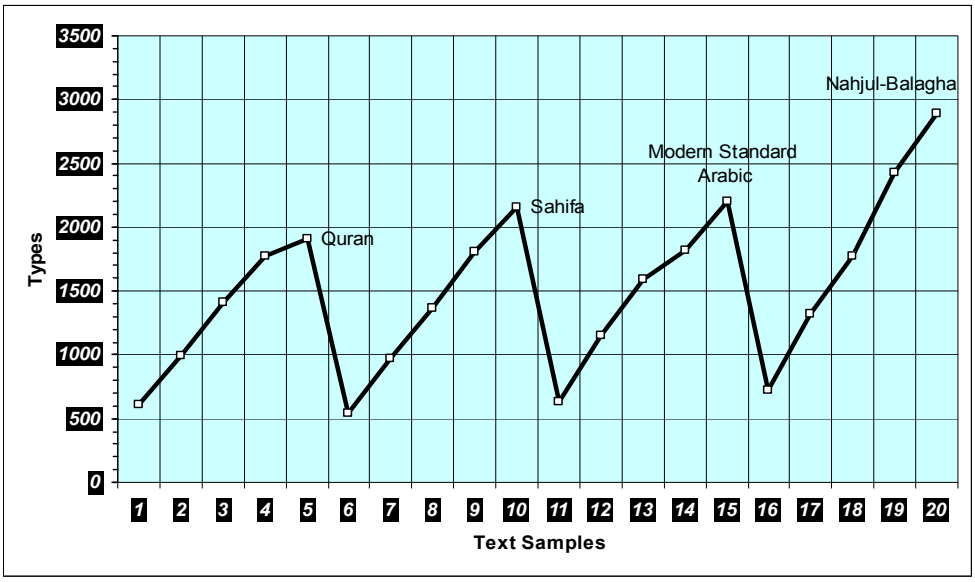

Figure 4. The Order of the Type-Token Curves for *Quranic Verses*, *Al-Sahifa*, *Modern Standard Arabic* and *Nahjul-Balagah*

6.2 English Corpus

Below is a summary of the type-token statistics observed in the English corpus.

Table 2. English Corpus Statistics

Texts	**Samples Numbers**									
	1		**2**		**3**		**4**		**5**	
	Tokens	**Types**	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**
The New Testament	1, 103	369	2, 066	539	3, 025	709	4, 036	824	5, 023	925
Joseph Conrad	1, 102	510	2, 005	766	3, 079	1, 126	4, 079	1, 332	5, 69	1, 414
Charles Dickens	1, 095	464	2, 077	670	3, 064	929	4, 086	1, 014	5, 090	1, 305
George Eliot	1, 028	510	2, 020	783	3, 062	1, 116	4, 059	1, 095	5, 068	1, 401

The diagram in Figure 5 below illustrates a wide range of variations for type-token curves in the English corpus. Conrad's *Heart of Darkness* is the highest in its lexical richness and *The New Testament* is near the minimum.

It is quite evident that the curves of Conrad's *Heart of Darkness* and Eliot's *Adam Bede* converge in an interesting way, though they diverge in sample 4 for each. The types observed in Conrad's and Eliot's samples are approximately identical: 510/510, 766/783, 1,126/ 1,116, 1,414/1,401. The types scored in sample 4 for each novelist are widely different: 1,332/1,095. This is a graphic evidence that Conrad's linguistic repertoire is statistically similar to Eliot's.

However, Dickens' *David Copperfield* comes in the next lower part of the graph. Dickens' curve approaches Eliot's only in samples 4 and 5 but without any convergence between the two.

The New Testament has the lowest type-token curve in Figure 5. It's fifth sample, for example, is 5,023 words long, but it uses a vocabulary of just 925 words (types)- considerably less than the 1,305-word vocabulary in the first 5,090 words of Dickens' *David Copperfield*.

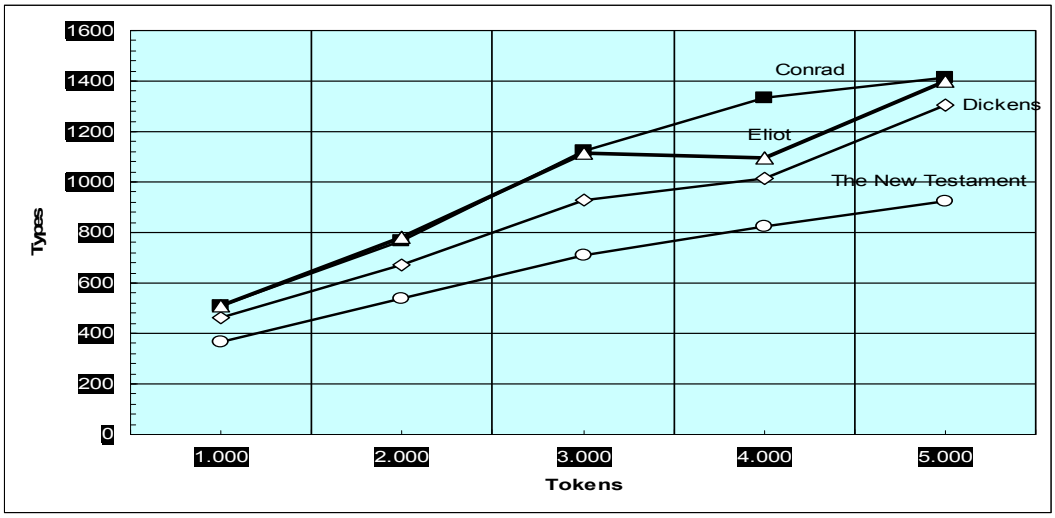

Figure 5. The Order of the Type-Token Curves for *The New Testament*, Conrad's *Heart of Darkness*, Dickens' *David Copperfield*, Eliot's *Adam Bede*

The order of the type-token curves in terms of their own graphical height is subtly shown in Figure 6 below.

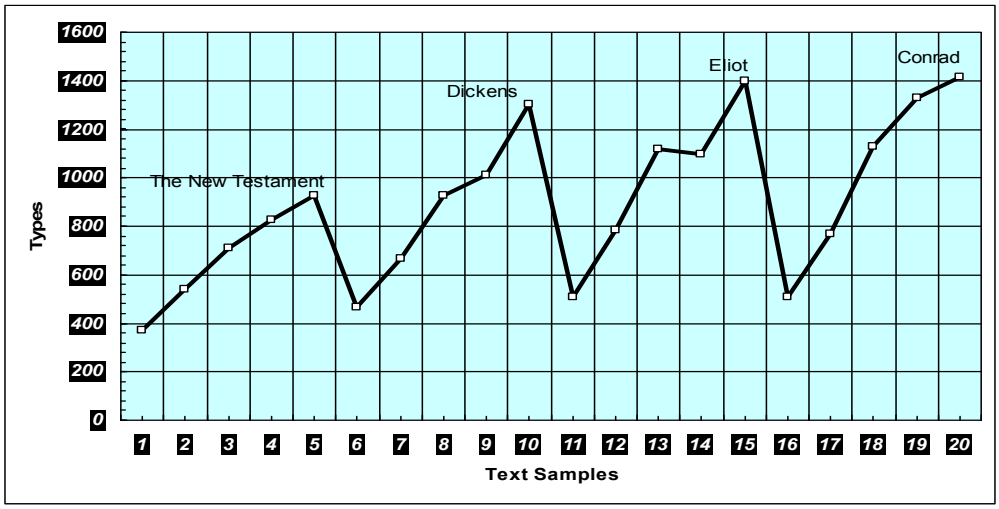

Figure 6. The order of the Type-Token Curves for *The New Testament*, Conrad's *Heart of Darkness*, Dickens' *David Copperfield*, Eliot's *Adam Bede*

7. Conclusions

Rounding off our study, we get back to the question the study started with: how useful corpus stylistics is to the study of lexical richness across various literary text-types in English and Arabic.

Corpus stylistics simply proved useful in targeting the statistical behavior of types and tokens within texts of different languages and genres. It is only by the large- scale data produced by corpus stylistics that the differences in the lexical richness can be quantitatively perceptible.

It became evident that the type-token curve, as a statistical measure, performs quite efficiently and informatively in terms of the vocabulary size used by writers. Each writer, whether in Arabic or English, has his own distinctive type-token curve that provides researchers with a lot of statistical information about the richness of the writer's linguistic repertoire. This curve tells researchers when the writer exhausts his language and measures up the vocabulary size used in a particular text.

Besides, it seems plausible to conclude that the statistical range at which the shifts of type-token curve become more perceptible and radical appears different in English and Arabic corpora: within the Arabic corpus the shifts of the curve are more perceptible within a range of (2000-2500 tokens) and (1000-1500 types) whereas the range in the English corpus reads: (2500-3000 tokens) and (900-1000 types).

Moreover, the type-token curve appears to be highly sensitive to the genre of *sacred texts* (the Qur'an and the New Testament). It seems that the curve of this type of texts is commonly depressed and occurs at the minimum range of lexical richness when compared with other types of texts (novels and sermons). This is why the *Qura'nic Verses* and *the New Testament* samples have the lowest type-token curves in this study. This is not due to the limited vocabulary size of such a genre, rather it is their directive nature and instructional context that require a kind of liturgical repetition to enhance the impact of the divine instructions on the believers' minds. As for the exceptional lexical richness of *Nahjul-Balagah*, it could be ascribed to the diversity of its topics and genres involved. It does not represent one specific genre, rather it comprises various text types: *sermons*, *sayings*, and *letters*: hence its rising type-token curve and amazing lexical richness.

References

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Biber, D. (2011). "Corpus linguistics and the study of literature", *Scientific Study of Literature*, 1(1): 15-23.

Butler, C. (1995). Statistics in Linguistics. Oxford: Blackwell.

Conrad, J. (2014). *Heart of Darkness*. EBook. http://www.gutenberg.net/dirs/etext96/agnsg10h.htm (Accessed 7 September 2014).

Dickens, C. (2014). *David Copperfield*. EBook. http://www.gutenberg.net/dirs/etext96/agnsg10h.htm (Accessed 12 September 2014).

Eliot, G. (2014). *Adam Bede*. EBook. http://www.gutenberg.net/dirs/etext96/agnsg10h.htm (Accessed 24 September 2014).

Francis, W., Nelson, and Kucera, H. (2000). *Frequency analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Hardie, A. & McEnery, T. (2006). "statistics". In Sarah, G. & William, J. (eds.). *Encyclopedia of Language and*

Linguistics. Retrieved from http://www.sciencedirect.com (08 May 2014).
Ibn Abi Talib, A. (1998). *Nahjul-Balagha*. EBook. http://www.google.iq/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved (Accessed 23 September 2014).
Ibn Al-Hussein, A. (2014). *Al-Sahifa al-Sajjadiyya*. EBook. http://www.al-islam.org/sahifa-al-kamilah-sajjadiyya-imam-zain-ul-abideen (Accessed 07 October 2014).
Leech, G and Short, M. (2007). *Style in Fiction*. London: Longman.
Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. London: Routledge.
Milička, J. (2009). "Type-token & hapax-token relation: A combinatorial model". *Glottotheory* 2 (1): 99 – 110. Trnava.
Mistaghanmy, A. (2014). *Chaos of Sensations*. EBook http://www.google.iq/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved (Accessed 22 October 2014).
Procha´ zka, S. (2006). "Arabic". In Sarah, G. & William, J. (eds.). *Encyclopedia of Language and Linguistics*. Retrieved from http://www.sciencedirect.com (02 May 2014).
The Holy Qur'an. (2014). EBook http://www.pdfquran.com/
The New Testament. (2014). EBook https://ebible.org/web/webnt.pdf
Wimmer G. (2005). "The type-token relation". In: R. Köhler (ed.): *Quantitative Linguistics*: *An International Handbook* : 361 – 368.
Watson, J. (2007). *The Phonology and Morphology of Arabic*. Oxford: Oxford University Press.
Wybraniec-Skardowska, U. (2007). "On the Type-Token Relationships". *Bulletin of the Section of Logic*, 15 (4): 164–168.
Youmans, G. (1990). "Measuring lexical style and competence: The type-token vocabulary curve". *Style*, 24 (4): 584-596.

Online References

http://www.al-islam.org/sahifa-al-kamilah-sajjadiyya-imam-zain-ul-abideen

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar