

Construction and Interpretation Of Corpus-Based English Poetry Vocabulary Profile

Omar A. Sheikh Al-Shabab

King Abdullah Institute for Translation and Arabization, Imam University (IMISIU), 225, Anas St, Al-Malqa, Riyadh 13524, Saudi Arabia

E-mail: osalshabab@gmail.com

Received: 04-02-2017

Accepted: 28-03-2017

Advance Access Published: July 2017

Published: 01-09-2017

doi:10.7575/aiac.ijalel.v.6n.5p.51

URL: <http://dx.doi.org/10.7575/aiac.ijalel.v.6n.5p.51>

Abstract

Vocabulary Profilers (VPrs) are deeply rooted in pedagogical purposes. The current investigation, however, uses the Classic and Compleat VPrs to: 1) determine the distribution and content of vocabulary in an English poetry corpus 2) explain differences in the constituents of the vocabulary profile (VP), 3) explore the role of language users in constructing the VP. The corpus includes Extended Corpus (EC: 1.363.225 words), Micro Corpus (MC: 43.200 words) from thirty-six poets, and two poems translated into Arabic. The main results show that Types, Offlist words, Academic and Anglo-Saxon words outline the VP, and that the number of Types and the size of the Individual Mental Lexicon constitute the main features of the translator's VP. The paper concludes that the poet's construction of the poetry VP undergoes multilayer interpretation by the reader/analyst and the translator, who utilize their socio-environmental context to pin down the semantic potential of the VP anew.

Keywords: Vocabulary profiles, Translator Mental Lexicon, Type/Token Ratio, Offlist words

1. Introduction

1.1 Vocabulary and the Mental Lexicon (ML)

Essentially, text is prior to corpus, and vocabulary contextualization in a text predates the text itself; therefore the poet's choices from existing vocabulary, mold new lexical and derivational forms, and necessarily result in an interesting design that embodies the poetic lexicon and message. Hence, the poetic Vocabulary Profile (VP) stands to reveal the poet's environment, experience and identity. Lexical choices establish the fundamental layer upon which all matters of concepts, including terminology, propositions, and sense and reference rest. It is not surprising, therefore, that lexical items, as concepts or technical terms in scholarly works, especially science, have drawn the attention of philosophers, literary critics, mathematicians and scientists alike.

Outside literary criticism and natural language processing, the study of vocabulary in English has been approached from three main traditions that differ in their fields and concerns but unite in their focus on learners of English. The first approach is from lexicography, the second is American readability studies and the third vocabulary counts and vocabulary lists. Lexicography as a profession is a main producer of learners' dictionaries (Hornby, 2005/1948 & West 1953). American research and works on vocabulary is primarily addressed to serve and control school graded textbooks (DuBay, 2006). In British English, the main focus of teaching English was driven by the situation in the colonies, especially India in the twentieth century. Each of Ogden and West had work experience in India and shared these pedagogical concerns. In the twentieth century, the need to teach English to non-English speakers in British colonies and immigrants in the USA surged, especially during World War II and after. But vocabulary and lexis have not been used to establish the distinctive outline of the poet's Vocabulary Profile (VP), or to investigate the relationship between the Communal Mental Lexicon (CML) and Individual Mental Lexicon (IML) in English poetry and its translation into Arabic.

1.2 The Current Study

It is known that poets refine their language including the choice of words, an observation which needs to be investigated and verified. Much has been said about Shakespeare's rich diction and Milton's use of Latin word, but there is a need to test these "hypotheses" and to draw an outline of VP for each poet in the corpus in order to establish the reliability of using such data in author attribution and author identification (Zhao & Zobel, 2006). The size of the corpus needed and the accuracy of the computer programs and their capacity are among the side issues touched on in investigating poetical language and poetical ML in a language and in individual poets. A larger corpus and a more powerful computer program will give a panoramic, more accurate results (Cobb, 2013), though sampling may offset much of the costs incurred by large corpora (Patty & Painter 1931).

But, what are the main constructs of a poet's VP? How and why does a poetic VP emerge? Furthermore, who are the makers of the poetic VP? What are the linguistic processes involved in making it? To answer the *how* and *why* question

one can trace the poet's concern with refining his art through creating a language, which merges experience in its widest meaning, with language and talent. Producing a creative text, especially a poem, is both consuming and risky while the result is not guaranteed. This heavy tapping of experience can be studied through the Interpretive Frame, which includes environment, experience, and identity, and their formalization through the First Person Domain (Author, 2008). The question about the *what* – the constructs – is addressed by specifying the exact distribution of vocabulary in terms of frequency of occurrence and lexical features of the VP in each poet in a representative corpus of poets and English poetry at large. Finally, the linguistic processes will be investigated here as various stages of interpretation. Hence, the construction of the VP is originally the work of the speaker, text maker, i.e. the poet, while the metalinguistic construction of the primary text(s) embodying the VP is the work of the reader, i.e. the researcher here. It must be accepted that the same set of texts by one or more than one text maker can yield a variety of readings and hence, a variety of VPs. The instruments utilized in constructing the VP by the researcher differ from those employed by the maker of the poem, the poet, since the first resorts to all possible devices and techniques and works in an environment and frame of real – physical – time, removed from those of the original user of the Vocabulary being studied. The main task of the current investigation is to reveal the properties of both layers embedded in the same corpus.

1.3 Research Objectives

To pursue establishing a linguistic profile for individual poets or language communities is not to deny the presence of general features of language varieties. Rather it is an effort in the direction of empirically describing the actual role of the poet and that of the analyst in processing the vocabulary choices to reach an outline of the text maker's construction of the VP and that of the reader's, the researcher's in this case..

The objectives of the study can be summarized in four main points:

1. To outline the constructs of the VP and ML through describing the distribution of word frequency and the main lexical features in the VP of individual poets, periods of English poetry and translated poems.
2. To identify and explain the emergence of difference in poetic VPs.
3. To examine the role of the poet and analyst in constructing the VP.
4. To study the ML in various Arabic translations of two English poems from the current corpus.

2. Related Literature

2.1 Vocabulary Lists and VPs in Teaching English Context

The implicit rationale for vocabulary counts and lists stems from the simple observation that learners of English face difficulties in accessing standard English texts, and pedagogical needs require an explicit method for grading and controlling vocabulary input. Thorndike, an American psychologist by specialization, and Ogden, a language specialist, developed “word counts” (Thorndike & Lorge, 1944) and “word lists” (Ogden, 1930) to assist teachers and simplified books writers in controlling their products. To the same end, West (1953) devised the *General Service List* and pioneered limited-vocabulary dictionaries. The spread of teaching English, vocabulary as a topic and language component, has received due attention in curricula and textbooks (Cobb, 2010 & 2013), in teaching methods (Burkett, 2015) and language assessment (Szymańska, 2015), in translation (Sin-wan 2004), corpus studies (Browne, Culligan & Phillips, 2013; & Brenzina & Gablasova, 2015), and text linguistics (Author, 1996). The concern with vocabulary frequency and lists is also present in the works of Béjoint and Cahill, who find good arguments for utilizing the notion of frequency in teaching French (Béjoint, 1989) and Russian (Cahill 1989).

2.2 Vocabulary and Lexis outside Pedagogy

Outside language pedagogy, computational linguistics and text linguistics have approached the question of vocabulary from completely different theoretical and practical perspectives seen in machine translation, corpus-studies, computerized lexicology and author attribution and identification. A number of branches and concerns in language studies and linguistics have resorted to vocabulary to present and resolve standing problems, like Kay who studies words as grammatical units (Kay, 1997). At a different level, Sapir's argument in support of “linguistic determinism” rests on the lexical level since “In the sense that the vocabulary of a language more or less faithfully reflects the culture whose purposes it serves it is perfectly true that the history of language and the history of culture move along parallel lines.” (Sapir, 1921, p. 104). Similarly, the debate about the incommensurability of scientific terms, even within the same language, hinges on the descriptors which determine both the meaning of the term and the requirements of a new scientific theory (Kuhn, 1962 & 1982). In translation, the introduction and generation of new lexical items via translation is attested in lexical borrowing from the SL, lexication and creative contextualization of words (Author, 2017). However, from Ogden and Thorndike to Nation and the NBC (Nation and Waring, 2004), the history and application of vocabulary lists and profiles, has been dogged by the debate about frequency of occurrence in a “representative corpus”.

The present paper examines vocabulary through the working of the VP in the first person domain, a construct which explains the speaker's internalization of language including interpretation, ML and linguistic processing. The study of VP in a poetic corpus should help understand the ML, the frequency of vocabulary in individual poets and poetry, and ultimately leads to constructing a poetic lexicon of the language concerned. Moving to interpretation in translation, the study of the VP in various translators will help understand both the ML and the translator's creative potential, because each translator operates from his/her own experience, identity and assertions.

3. Corpus and Methodology

3.1 Corpus: poems, Poets, Periods and Sources

The present corpus is originally part of a larger project which studies the translation of English poetry into Arabic in the context of investigating the Interpretive Frame and First Person Domain as hypothesized by the author (Author, 2008, chapter 7). The basic assumption is that the language user is the cardinal interpreter of his/her experience and environment, and hence the translator's task is to construct the poet's output creating his/her own text and VP. To investigate the primary corpus and the translated corpus, the analyst will need to build up a sample of poems, from selected poets from various periods of English poetry and a sample of translated poems.

Corpus design and size: The current corpus starts with Thomas Wyatt (died 1542), a nobleman, diplomat and poet, whose output is relatively small. It includes thirty-six men and women poets from England, Scotland, Ireland, United States, and one Lebanese-Syrian poet who lived in the United States during the second and third decades of the twentieth century (see Table 1). Burkett (2015) stresses that "representativeness is relevant" and that "the source and genre of the texts, the age of the texts, and the country of origin" are relevant as well (Burkett, 2015, p. 74). To achieve an acceptable degree of validity, the researcher was guided by three criteria: 1) authenticity of the texts, 2) representativeness of periods and countries, 3) accessibility and manageability through Internet resources (Table 1).

The six periods, covering about five centuries of English poetry, roughly coincide with the traditional chronological divisions of English Literature. The selection of poems and the individual poets is governed by the following considerations:

- 1- Age and country of poet: guaranteeing representation of every period and country.
- 2- Circulation and prominence: passing the time test.
- 3- Gender and ethnicity: relatively fair representation of women poets.
- 4- Availability of digital copyright-free copies on the Internet

Poets are included regardless of their conviction or political views; Donne, Herbert, Milton, Pope and Burns are good cases in point, while the number of poets in each period reflects the state of poetry in the age, since some movements like Romanticism marked great changes and thrived in the nineteenth century. The criterion of time-test is relevant since it is tightly related to circulation and thus availability and prominence today. Geographical representation is seen in including eleven American poets, seven from the twentieth century alone. Geographical distribution also includes Scotland and Ireland. The researcher is also restricted to what is available in public domain on the Internet, which explains the small number of words from certain poets like Eliot and Cummings.

Table 1. English poetry corpus: periods, poets, poems and words

No	Periods & poets			Micro corpus			Extended corpus		
	Period	Poet &	Year of death	Words in text	No of poems	Average words In a poem	Words/tokens In text	Number of poems	Average words In a poem
1	Tudor & Elizabethan	T Wyatt	1543	1200	10	120	5700	34	167
2		C Marlow	1593	1200	6	200	55082	11	5007
3		E Spenser	1955	1200	1*		55062	8+	6882
4		W Shakespeare	1616	1200	11	109	55090	153+	360
5		B Johnson	1637	1200	7	171	55018	107	514
6	17 th century 1631-1674	J Donne	1631	1200	7	171	47140	161	292
7		G Herbert	1633	1200	6+	200	21320	151	141
8		J Milton	1674	1200	3+	400	55000	54+	1018
9	18 th century 1731-99	A Pope	1744	1200	1-	long	54646	61	896
10		R Burns	1796	1200	7+	171	49342	228	216
11	19 th century Romantics	J Keats	1821	1200	5+	240	50408	51	988
12		P B Shelley	1822	1200	10+	120	52260	168	211
13		G G Byron	1824	1200	7+	171	52225	35	1492
14		W Blake	1827	1200	8+	150	32781	143	229
15		S T Coleridge	1834	1200	1+	long	55325	136	406
16		E A Poe	1849	1200	7+	171	19688	78	252
17		W Wordsworth	1850	1200	2+	600	49738	129	385
18		H D Thoreau	1862	1200	7+	171	5831	41	142
19	W Whitman	1892	1200	2+	600	50012	204	245	
20	C G Rossetti	1894	1200	1+	long	50007	282	177	

21	19 th century	R W Emerson	1882	1200	4+	300	41502	118	351
22		E Dickinson	1886	1200	16+	75	36000	414	87
23		M Arnold	1888	1200	2+	600	39194	61	642
24		R Browning	1889	1200	3+	400	49960	44	1135
25		L A Tennyson	1892	1200	2+	600	56066	46	1218
26	20 th century	G K Gibran	1931	1200	4	300	33067	81	408
27		E E Cummings	1962	1200	11	109	14693	134	109
28		R Frost	1963	1200	7+	171	41460	161	257
29		S Plath	1963	1200	5+	240	26387	144	183
30		T S Eliot	1965	1200	2+	600	24164	52	464
31		E Pound	1972	1200	6	200	46659	246	189
32		W H Auden	1973	1200	3+	400	21506	79	272
33		R Lowell	1977	1200	1+	long	7444	27	275
34		X Kennedy	Alive	1200	14+	85	1331	18	74
35		W B Yeats	1939	1200	5+	240	48033	163	294
36		R Dove	Alive	1200	8+	150	4084	31	131
				43200			1363225	3375	403. 91

The initial decision about the size of the Extended Corpus (EC), the main corpus was to include fifty thousand words from each poet, but soon this aim proved to be too ambitious, due to various factors including the size of the poets' original corpus, and limitation of availability of accessible digital texts. Because of various constraints, the EC came to (1.363.225) words from (3.375) poems and extracts written by (36) poets (Table 1). To facilitate computer processing and control of classification and results from the corpus, the researcher assigned a Word file for each poet, and separate file(s) for the results from each poet.

The first (1.200) words from each poet make a Micro Corpus (MC), which accommodates for the smaller capacity of the Classic VPr, and which allows direct comparisons with data from the Compleat VPr. This means that while the EC is processed only once using *Compleat VPr*, the MC the smaller corpus of (43.200 words), is processed four times. The total number of words being processed for various purposes stands at (1.536.025) words (Table 2).

Table 2. Size of EC and MC in terms of processed words

Time span covered by EPC	1543?-to-date (AD)	
Poets included in EC	36	Poets
Size of Eng. Poetry EC used with Compleat VP (ILP)	1.363.225	Words
Size of Eng. Poetry MC used with Classic VP (IML)	43.200	Words
Size of Eng. Poetry MC used with Compleat VP (ILP)	43.200	Words
Size Eng. Poetry MC used with Compleat VP (CLP: periods of English poetry)	43.200	Words
Size of Eng. Poetry MC used with Compleat VP (CLP: English poetry)	43.200	Words
Size of EPC put for processing	1.536.025	Words

The number of words from different periods varies due to the number of poets included in each period (Table 3); while the number of words from each poet consistently remains the same in the MC. Hence, internal consistency in the number of words is maintained in the sample from each poet, and not in the sample from the period. The total number for all six periods adds up to the number in the MC (Table 3).

Table 3. Size of MC from each period

Periods	Years (approximate)	poets	Number of words
1 Tudor & Elizabethan	1542-1637	5	6000
2 17 th century	1631-1674	3	3600
3 18 th century	1731-99	2	2400
4 19 th century Romantics	1821-1894	10	12000
5 19 th century	1882-1892	5	6000
6 20 th century	1932-to-date	11	13200
7 All periods (language)	1542-to-date	36	43200

Most of the corpus is collected from one Web site, www.pomhunter, which is an open domain source and which states copyright restrictions when a poem is removed due to copyright laws. Other sources, especially www.Poetryacademy, have been used in cases where the sample found in the poem hunter site is too small.

The preparation of the texts took considerable time and effort, and covered a number of operations, including selection from websites, control of size and form, editing extracts from dramatic poetry, as in the cases of Shakespeare's Hamlet, and Marlowe's Tamerlane, and Johnson's Volpone. Mainly poetry stretches were taken, while proper names and casual prose sentences are left out.

3.2 Computer Programs and Methodological Observations

Computer analysis of vocabulary and lexical features dates back to the nineteen eighties with programs like Clock, which was capable of achieving vocabulary counts and collocation. The Oxford Concordance Program can calculate Type/Token ratio and collocation (Author, 1996). However, computer processing memory, storage and speed were limited and awkward, to say the least, compared with what is available today. Nation developed his vocabulary program in the early nineties, but the limit of the number of words in the text to be processed was and still is the main handicap (see Author, 2012). *Compleat VPr* is more recent and it draws on a number of large databases first of which are the NBC and COCA. Additionally, it can process up to (60.000) words, which is well beyond its predecessors.

The present study, therefore, makes use of both instruments, which achieves four desirable ends: 1) obtaining a comprehensive description of vocabulary behavior in the corpus, 2) identifying and comparing the VPs of various poets, periods and translation, 3) gaining insight into the construction of the VP in original and translated texts, 4) testing the accuracy and lexical features of the two programs and identifying their limits.

3.3 Arabic Translations: Source and Method of Analysis

Some of the Arabic translations of two poems "Kubla Khan" by Coleridge and "In Memory of W. H. Yeats" by Auden were taken from Internet sources (Table 4). The basic objective is to examine the translators' translational ML and compare it with poets' VP.

Table 4. ST, poets and translators, and source and date of poem and translations

Poem	Poet/translator	source	words	date
1 Kubla Khan	Coleridge	www.Poemhunter	351	1816
2 Kubla Khan	S. Al-Naser	Personal communication	284	1999
3 Kubla Khan	A. Al-Masiri & M Zaid	<i>Selection from Romantic Poetry</i>	286	
4 Kubla Khan	M. N. Al-Neimi	Commissioned	297	2002
5 Kubla Khan	B. Abbas	Internet	286	nd
6 Kubla Khan	Al-Shabab	Author	273	2013
7 Kubla Khan	A.S. Al-Zubeidi	Internet	307	nd
6 In Memory of W. H. Yeats	W. H. Auden	www.Poemhunter		1940
7 In Memory of W. H. Yeats	S. Al-Naser	Personal communication	298	1999
8 In Memory of W. H. Yeats	M. N. Al-Neimi	Personal communication	348	2007
9 In Memory of W. H. Yeats	K. Al-Hirz	Personal communication	329	2007
			2718	

Among the translators, only Al-Naser is a poet, while Al-Masiri & Ziad, Neimi, Al-Shabab and Hirz are academics, the researcher does not know the rest. The poetry translator achieves his work with care and complete devotion, taking the

time needed and consulting dictionaries and literary works. The results, therefore, should be taken as reliable and illuminating, though only the lexical level is being investigated and only a limited number of descriptive features can be obtained from manual analysis, since there are no VPr developed to handle Arabic VPs. The main method employed is the alphabetical ordering of words in a column and colour code to mix and compare the similarities and differences in employing the lexical items, which will help classify the Communal ML (CML) and the Individual ML (IML) of each translator for comparison among translators and with the lexical indicator of the Source Text (ST) later. Two main points will be investigated: 1) the number of words produced in a translation in relation to ST word number and in relation to other translations into the same TL, i.e. word number and directionality, 2) the VP of translators in terms of the CML and IML. Translation from Arabic into English is later used for comparing word number in relation to directionality.

4. Results and Discussion

4.1 Constructs of the VP

The constructs of the VP are based on two types of data: 1) frequency of occurrence and 2) lexical features and linguistic indicators. The poets' works originally provide the first construct of the VP at a level which embodies the poet's own works, while the second construct rests on the analysis and interpretation of the corpus presented by the researcher.

The two basic components of constructing the VP are frequency of occurrence and lexical features, and both will be examined in the current study. The corpus is analysed from various aspects, and the results help explain the rationale and the processing of the text maker or reader. In other words, the explanation should address the profile, the circumstances and processes that create a distinctive VP for each poet, and ultimately for each language user.

4.2 English Poetry VP

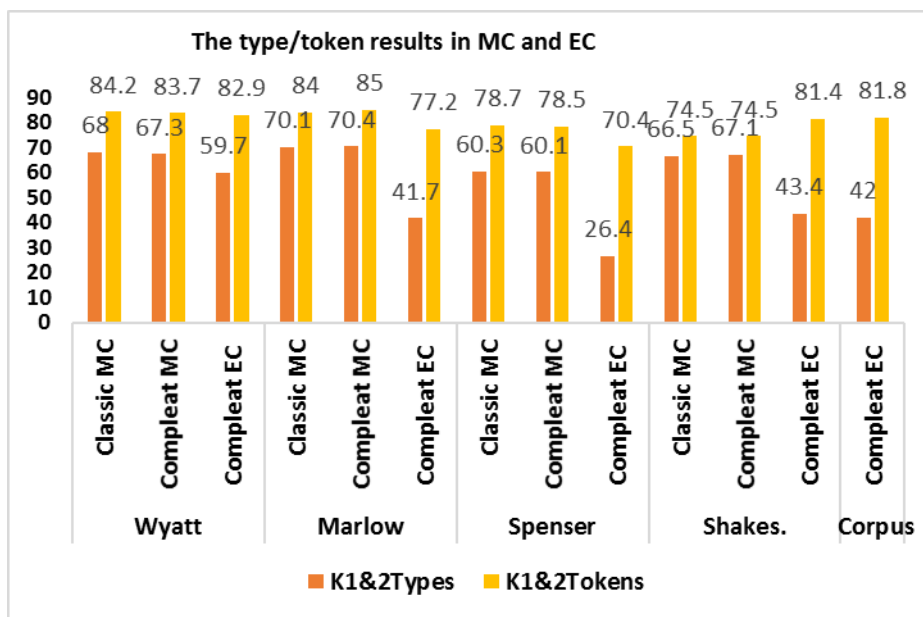
The Classic and Compleat VPrs, overlap and diverge in their functions. The Classic VPr operates two frequency indicators (K1 & K2) working mainly from the General Service List (GSL), in addition to lexical indicators such as the Academic Word List (AWL) and Anglo-Saxon, word counts; while the Compleat VPr operates twenty five vocabulary frequencies (K1, ... K25), but no special count indicator for AWL or Anglo-Saxon words counts. Significantly, the database for each of the two programs is not the same. Neither is the capacity, since the Classic VPr is limited to about (5.000) words in one operation. Therefore, the results from both offer comparisons of K1 & K2 occurrences, and a chance to investigate internal relationships within the sample from one poet or external relationships across samples from different poets, even indicator from periods of English poetry or even relationships between the indicators from all English poetry in the MC on the one hand and each poet or period on the other.

Examining the Distribution of vocabulary in terms of frequency of occurrence, one can easily observe the large percentage of K1 words in the EC and MC, which is natural in light of the fact that K1 words are designated to this category by virtue of being the most frequent in the major lists, corpora and counts such as the GSL, British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). The results from processing the MC and the EC show the special importance of K1 and K2 frequencies when added together. Table (5) shows that K1 & K2 Tokens in the EC in Wyatt, Marlow, Spenser and Shakespeare stand at: in Wyatt 84.2%, Marlow 77.2%, Spenser 70.4% and Shakespeare 81.4%, which can be compared with 81.8% in the EC as a whole.

Table 5. Example of Types and those of Token in MC & EC in the corpus

Poet/ Corpus	VP Program	VP program	Words in text	Types	Tokens	K1+2 Type	K1+2 Tokens
Wyatt	MC	Classic	1200	451	1200	68.0	84.2
		Compleat	1200	450	1113	67.3	83.7
	EC	Compleat	5700	1355	5437	59.7	82.9
Marlow	MC	Classic	1200	506	1229	70.1	84.0
		Compleat	1200	504	1169	70.4	85.0
	EC	Compleat	55092	6413	53985	41.7	77.2
Spenser	MC	Classic	1200	540	1210	60.3	78.7
		Compleat	1200	544	1193	60.1	78.5
	EC	Compleat	55062	7995	54451	26.4	70.4
Shakespeare	MC	Classic	1200	520	1224	66.5	74.5
		Compleat	1200	521	1217	67.1	74.5
	EC	Compleat	55090	6722	53877	43.4	81.4
Total Corpus	EC	Compleat	43.200	7563	41994	42.0	81.8

The percentage of Tokens, which is used in most pedagogical discussion to demonstrate the amount of text being covered, is by no means the best indicator of difference in comparing the frequency of occurrence, and hence the VP. Leaving Wyatt's results which are obtained from a sample of only one tenth of the other three poets, one can notice that there is less than (10%) difference in the number of Tokens in the MC and EC; meanwhile there is (28%) difference of types between MC and EC in Marlow, (23%) in Shakespeare, and about (34%) in Spenser. This means that occurrence of *types* is more sensitive and accurate in reflecting differences in the VP. A second reason in favour of comparing *Types* and not *Tokens* is the fact that *Types* represent the size of the VP, whereas *Tokens* represent the size of the text, and the focus of the current investigation is on the distinctive features of the ML and not the corpus size, which itself may not be important if effective sampling methods could be devised. Moreover, a more interesting observation is the fact that comparing *Tokens* would put Spenser and Shakespeare not far from each other (Spenser 70.4% and Shakespeare 67.1%); while comparing *Types* tells a different story: (Spenser 26% compared with Shakespeare (43.4). If we remember that the average for the sample representing the whole corpus stands at (42.0), (16%) less from Spenser's result and only (3%) less from Shakespeare's, then it becomes clear that *the size of types, the ML, is the more significant indicator of difference as Histogram (1) shows.*



Histogram 1. Types and Tokens from K1&K2 in four poets from the MC and EC

The aim was to collect (50.000) words from each poet, but that proved to be impossible due to limitations of output, availability in electronic copies, or copyright, but despite these restrictions the sample from twenty poets amounts to more than (45.000) words. Poems from only five poets amount to less than (10.000) words each. A second basic fact is that the processed words (Tokens) differ in number from the actual words in text (corpus), due to electronic processing of the two VPr programs, a fact which calls for an investigation of computing programs. A third fact has to do with the capacity of the two VPs, maximum of (5.000) words for Classic and (60.000) for Compleat.

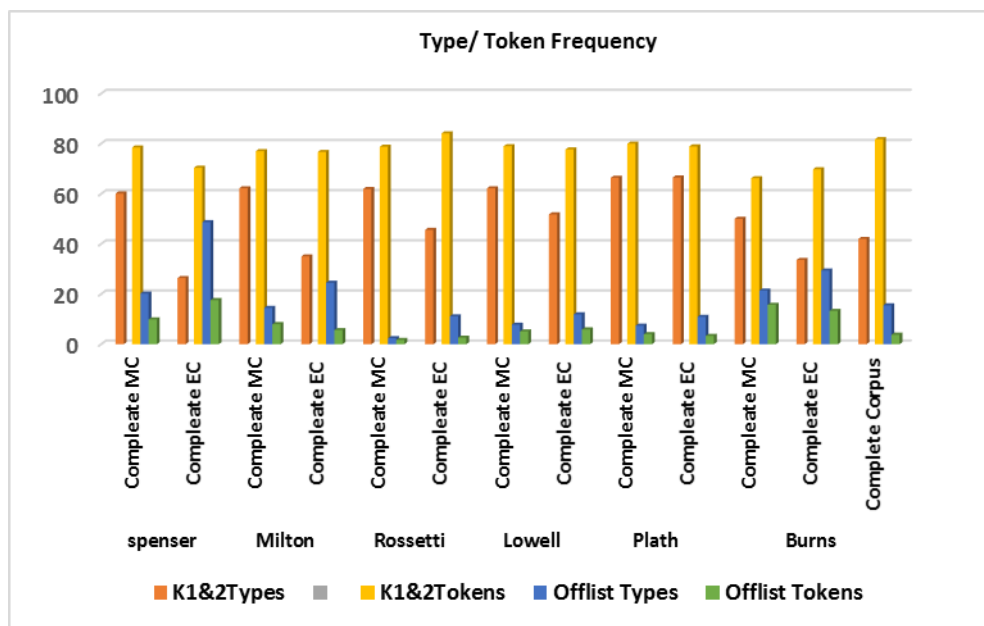
First in the basic results is the relationship between the number of Tokens and the number of Types in the reported percentages; this is neutralized in restricting the words from each poet in the MC to exactly (1.200) words, which are treated by the Classic and Compleat VPrs for comparing results and for assessing the programs on different occasions and input. The second observation about the results is seen in the gaps where two programs do not include the same features or lexical indicators. The third basic result has to do with the internal coherence of the results of the MC, since the average for K1&K2 Types from MC results is about (70%) and the difference between Classic and Compleat VPrs among poets is less than (2%), except in few interesting cases: Spenser (60.3%) and (60.1%); Milton (62.3% and 62.2%); Rossetti (62.5% and 61.9%), Lowell (61.3% and 62.2%), all of which show about (10%) deviation from the average (70%). Plath's MC results sample (65.2% and 66.4%) shows about (5%) compared to the total average, but the real big difference is in Burns' results (49.7% and 50.0%), which is (20%) drift from the average (70%). The internal consistency of the poet's MC is seen in the similarity of the results from Classic and Compleat VPr in each case, which validates the accuracy of the two programs and which points to the cases which merit more attention when external comparisons with other poets are carried out.

However, the low percentage of Type in Spenser and few other poets in the MC and EC calls for further examination to see whether it represents an accidental case or a significant phenomenon. Table (6) below reports selected results, demonstrating the differences among poets in the occurrence of K1&K2 Types in the EC.

Table 6. The results of occurrence of Types, Offlist words and Anglo-Saxon words.MC & EC in the corpus

Poet/ Corpus	VP Program	VP program	K1+2 Type	K1+2Tokens	Offlist types	Offlist Tokens
Spenser	MC	Classic	60.3	78.7	37.4	20.0
		Compleat	60.1	78.5	20.2	9.9
	EC	Compleat	26.4	70.4	48.7	17.6
Milton	MC	Classic	62.2	76.8	35.9	22.2
		Compleat	62.2	77.0	14.5	8.0
	EC	Compleat	35.0	76.7	24.5	5.6
Rossetti	MC	Classic	62.5	80.2	35.8	18.8
		Compleat	61.9	78.8	2.4	1.7
	EC	Compleat	45.6	84.1	11.1	2.6
Lowell	MC	Classic	61.3	78.4	37.8	21.8
		Compleat	62.2	79.0	7.8	5.1
	EC	Compleat	51.8	77.7	11.9	5.9
Plath-	MC	Classic	65.2	80.9	32.4	17.9
		Compleat	66.4	80.0	7.4	4.0
	EC	Compleat	66.5	78.9	10.9	3.3
Burns	MC	Classic	49.7	66.1	48.3	32.9
		Compleat	50.0	66.2	21.4	15.7
	EC	Compleat	33.6	69.8	29.4	13.3
Total Corpus	EC	Compleat	42.0	81.8	15.5	3.8

The consistency of the MC results from both VPr programs, is clear in the two columns representing K1&K2 results in the MC, and in the sensitivity of the Types, ML, is clear in the Compleat EC when comparing results from different poets. The K1&K2 Types and the Offlist Types, are conversely related: more Offlist Types than K1&K2 Types in Spenser; another interesting case is Burns’ balanced K1&K2 Types and Offlist. In Histogram (2), the orange and blue bars represent Types and Offlist respectively. The low Types and high Offlist are clear in Spenser and Burns, especially when compared with the total results of the corpus in the last set of bars (Histogram 2).

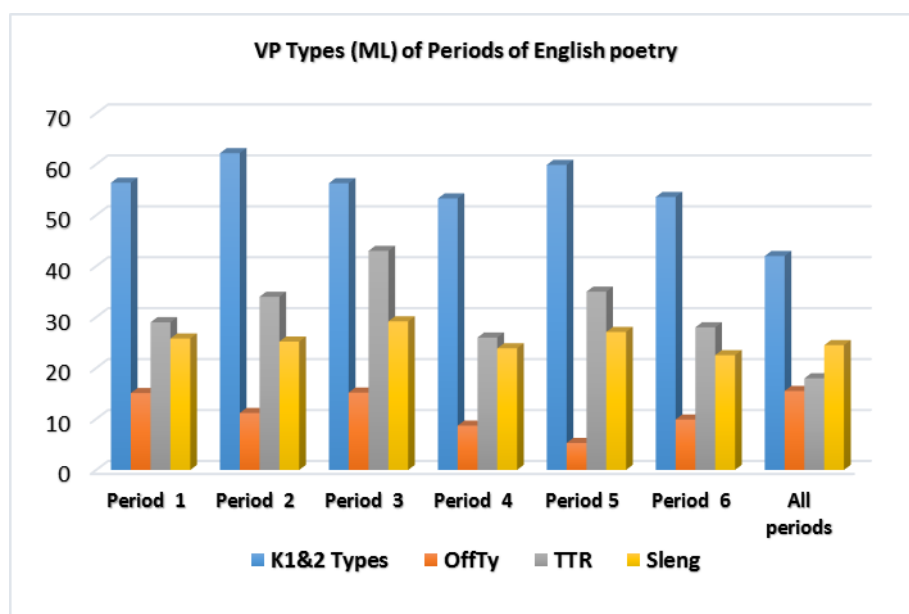


Histogram 2. Types, Tokens and Offlist in the MC & EC from the Compleat VPr.

The lexical indicators show that there is a number of significant relationships between the indicator and frequency in the word counts. First is the Type-Token (T/T) Ratio which is directly influenced by the size of the corpus and the

percentage of Types in relation to Token, as in Spenser's low frequency of Types in the EC which yields a higher T/T Ratio: (K1&K2: 26.4% and T/T Ratio: 0.15) compared with Marlow is (K1&K2: 41.7% and T/T Ratio: 0.12%) and Shakespeare (K1&K2: 43.4% and T/T Ratio: 0.12%). The T/T Ratio takes into consideration the Offlist as Types, of course; but the K1&K2 are from within the lists of the first (25.000) words in Compleat VP. This means that the T/T Ratio does not fully represent the difference in profile, like the Offlist words, especially if Offlist words are unique to an individual Poet. Another significant lexical indicator is the Anglo-Saxon words and here Burns is unique in this respect, using the lowest percentage of Anglo-Saxon words, which is important for the specification of the individual VP. A final indicator is sentence length, which reflects a textual feature where syntax and editing by publishers who apply their standards to manuscripts of certain poets, but not so much to others, a fact which is clear in the standard, up to date style, a result affected by work of repeated editing by scholars and publishers. The vocabulary and lexical profiles represent a complex network of scholarship, interpretation and socio textual culture.

The MC using the Classic VPr and the EC using the Compleat VPrs, show that Tokens are sensitive discriminators, because they show similar percentage of K1 and K2 vocabulary in the various periods (Histogram 3). The Types representing the ML, however, show reasonable differences. The last set of columns (Histogram 3) represents the percentage of all six periods, an average which presents a reference for comparison of various periods. The difference between average of all periods and the second period amounts to (1.8%) of Tokens but (20.2%) of Types, i.e. the ML is very different in this respect, but the distribution of frequency of K1&K2 is not. A most significant result is the occurrence of Offlist words. Internally, the Tokens are less in terms of percentage; but external comparisons show the high gray (Offlist) column in the first three periods, which reflects the selection of word lists and corpora in the first place. This result needs more attention in the discussion.



Histogram 3. Compleat VPr, Periods: Frequency of K1&2Ty, OffType, Types/Token Ratio, and Sentence Length

In Histogram (3), Tokens are not reported, and differences in Types in the first bar, in blue, and Offlist in the second bar, in orange, become clearer, especially when compared with the all Periods results. The Type/Token Ratio (the third bar in gray) is calculated out of (1), but it is multiplied by a hundred here in order to make the column clear. Finally, sentence length, the last bar in yellow, shows about (6 %) variation among periods.

4.3 VP and the ML in Arabic Translations of "Kubla Khan" and "In Memory of W. B. Yeats"

Translation creates new texts in a new language, which raises interesting questions about the VP and the ML of the Translated Text (TT) and consequently the translator. Manual calculations have produced four aspects of the VP of the translators: 1) Type/Token analysis, 2) content/function words, 3) content and size of the ML, 4) TT Arabic/ST English ratio. The TTs are relatively short, (2.609) words, and manually calculated. In three Arabic translations of Auden's "In Memory of W. B. Yeats", the number Arabic words in the TT is less than the ST English words (Table 7, line 1), and so is the number of Tokens (line 4). Types (line 3) and Tokens (Line 4) are different, and although Al-Neimi uses more Types, his large number of Tokens makes his T/T ratio the smallest. The function words used by Al-Naser is the smallest (line 8), while the ML is influenced by the large number of Tokens employed by Al-Neimi (line 5, Table 7). Finally the TT Arabic / ST English ratio demonstrates this fact clearly (Al-Neimi 0.75, Al-Hirz 0.71 and Al-Naser 0.64), which means that the results are overshadowed by the unexpected difference in the number of words used by Al-Neimi in the translation of the same ST, compared with the other two translators.

Table 7. VP of three translators of “In Memory of W. B. Yeats”

Feature	ST*	S. Al-Naser	N. Al-Neimi	K. Al-Hirz
1 Words in text	462	298	348	329
2 Lines in poem	65	74	66	71
3 Types	251	235	261	258
4 Tokens	451	298	348	329
5 TT Ratio	0.56	0.78	0.75	0.78
6 Content (Types)	-	217 (92.3%)	229 (87.4%)	228 (88.3%)
7 Function (Types) %	-	18 (7.6%)	33 (12.5%)	30 (11.6%)
8 Cont./ function Ratio	-	0.08	0.14	0.13
9 ML (total Types)	-	235	261	258
10 IML Types %	-	105 (44.6%)	119 (45.5%)	109 (42.2%)
11 CML Types %	-	130 (55.3%)	142 (54.4%)	149 (57.7%)
12 IML-CML Ratio	-	0.80	0.85	0.73
13 ST (Eng.) Tokens	-	462	462	462
14 TT (Arabic) Tokens	-	298	348	329
15 TT Arabic-ST English Ratio	-	0.64	0.75	0.71

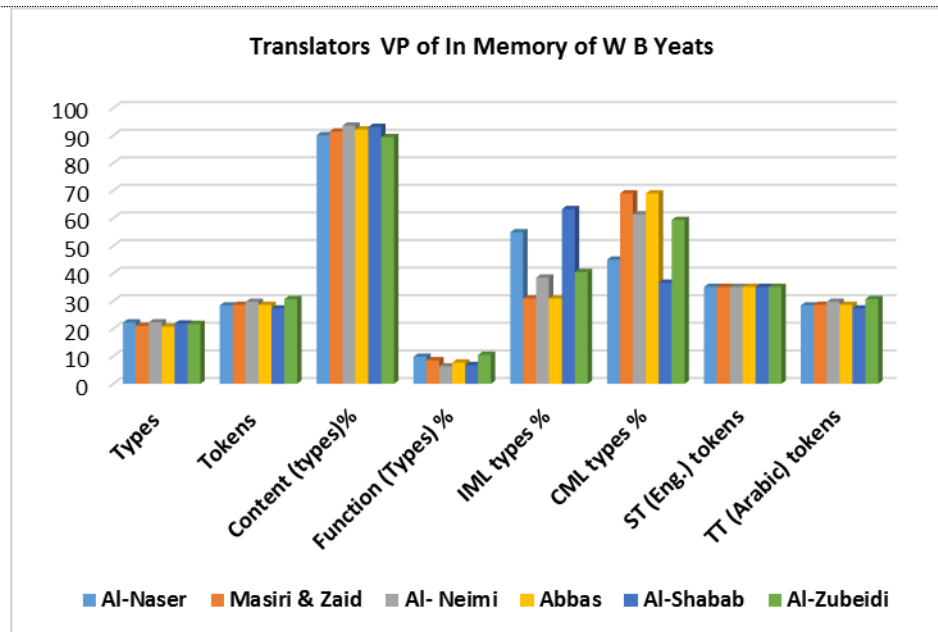
*(Compleat VPr is used for English ST)

The results from the translation of Coleridge’s “Kubla Khan” by six translators provide interesting results concerning the VP.

Table 8. VP of six translators of “Kubla Khan”

Feature	ST	Al-Naser	Masiri & Zaid	Al- Neimi	Abbas	Al-Shabab	Al-Zubeidi
1 Words in text	351	284	286	297	286	273	307
2 Lines in poem	54	55	53	54	52	55	54
3 Types	214	222	210	223	207	218	217
4 Tokens	339	284	286	297	286	273	307
5 TT Ratio	0.63	0.78	0.73	0.74	0.72	0.79	0.70
6 Content (Types) %	-	202 (90.1%)	192 (91.4%)	208 (93.6%)	191 (92.2%)	203 (93.1%)	194 (89.4%)
7 Function (Types) %	-	22 (9.8%)	18 (8.5%)	14 (6.3%)	16 (7.7%)	15 (6.8%)	23 (10.5%)
8 Cont./ function Ratio	-	0.10	0.09	0.06	0.08	0.07	0.11
9 ML (total Types)	-	222	210	223	207	218	217
10 IML Types %	-	120 (54.0%)	65 (30.9%)	86 (38.5%)	64 (30.9%)	138 (63.3%)	88 (40.5%)
11 CML Types %	-	102 (45.9%)	145 (69.0%)	137 (61.4%)	143 (69.0%)	80 (36.6%)	129 (59.4%)
12 IML-CML Ratio	-	1.17	0.44	0.62	0.44	1.72	0.68
13 ST (Eng.) Tokens	-	351	351	351	351	351	351
14 TT (Arabic) Tokens	-	284	286	297	286	273	307
15 TT Arabic-ST English Ratio	-	0.80	0.81	0.84	0.81	0.77	0.87

There are similarities in the number of Types and Tokens (Table 8 and Histogram 4), but Al-Zubeidi’s translation has more Tokens. Al-Neimi uses less function words than other translators, followed by Al-Shabab. Crucial differences appear in the size of the IML, which contains the words used uniquely by one translator. The highest percentage of IML words is found in Al-Shabab’s translation followed by Al-Naser’s. This represents individual creativity and deviation from the common expected occurrence, and hence it defies frequency and prediction. Here, as in the case of the high percentage of Offlist words, an explanation is required. Still, this finding needs to be mined from the large amount of Data and summarized.



Histogram 4. VP of Three Translators of "Kubla Khan"

4.4 Findings

The results of the current study are rich and varied. Below are the main findings:

- 1- The current study of VPs and vocabulary frequency, illustrates that Types are more sensitive than Tokens as indicators of VP, since Types make the ML, which is more significant for poets' distinctiveness and author identification (Author & Baka 2015).
- 2- Offlist words make a significant component of the VP and ML; they also raise questions about the validity of the frequency lists.
- 3- The Classic and Compleat VPrs show coherent main trends in the occurrence of K1 and K2 words.
- 4- Individual poets resort to different reservoirs and configurations of words to construct their VPs. Poets show predictable vocabulary features and lexical indicators in their VPs. Spenser has more Offlist words; Milton utilizes more academic (Latin) words; and Burns has less Anglo-Saxon words.
- 5- Samples from the six periods of English differ in the frequency of K1 and K2 words, percentage of Offlist words, and T/T ratio.
- 6- Periods of English literature manifest different trends of the VP.
- 7- The Offlist glossaries obtained from the EC of English poetry can be extended and used in author identification.
- 8- Translators' VPs show significant differences in the size of the IML, Offlist Types, and TT/ST Ratio (when English is the ST).

The above findings cannot be final, but they offer an evidence for the above facts and the capability of the two programs being used (Classic VP and Compleat VP). The discussion below attempts to place the main findings in a theoretical context.

4.5 Discussion of the Construction of VPs

The discussion covers three topics, the findings revealed in the description of the corpus, the interpretation of the evolution of the VP & ML, the VP in translation context. First, the results obtained from the MC (same 1.200 words) point to the efficiency of the Classic and Compleat VPrs, since the difference between the input corpus and the actually processed words rarely exceeds (4%) and never reaches (5%). Discrepancy in frequency of Offlist words seems to be due to the lists used by each program and to processing capacity. Other problems such as treating homoforms or "multiword units" involves different dimensions such as context and collocation, topics left for future generation of VP Programs (Cobb, 2013).

Naturally, word lists and word counts originate in a corpus i.e. from a definite number of Tokens; but they aim at extracting and describing a set of types, in a list or a glossary or a dictionary. The Types are more significant. Additionally, the percentage of occurrence of Types from K1 and K2 is by and large comparable, but it is different in the case of Spenser's sample. The textual editing of Spenser's texts is relevant in this respect, since the text culture of Spenser's manuscripts has not received the level of attention given to the Shakespearean corpus. For reasons beyond the current paper, Spenser's *Fairy Queen* has not received the repeated editing and updating which Shakespeare's works have known, a factor which has increased the Offlist words in Spenser's sample. However, Burns' employment of less Anglo-Saxon words evokes a different line of thinking, because his geographical dialectal VP is not a matter of editing, but rather a part and parcel of the message he intended to transmit, and hence the VP goes beyond the poet's corpus and

evokes hi/her environment and cultural background. Milton's VP is part of a different narrative. His VP reflects erudition and Latin background, influences that set to create not only texts, but also epics and a *legacy*. The frequency of unique Types helps portray the individual poet as a user of language, a creator of masterpieces and an interpreter of reality. The poetic construction rests on several pillars of which vocabulary is a basic one, which tells a story at every level. Statistical facts, therefore, make but one aspect of that narrative, but initiation, selection and contextual association would take the study of words beyond lexis (Baka, 2014 & Author 2012).

The study of VP of the periods of English poetry covered in the current corpus provides a testimony of the age, a testimony that reflects social and individual creativity and historical trends. Like the dictionary which documents the history of stability and dynamics of social and intellectual movements, the VP is in many ways informative about the vitality of the community of users, their sensitivities, taboos and level of tolerance and acceptability. The social dynamics of words, even their currency and disappearance, provides a registrar of the social and intellectual life of the VP community. The glossary, which represents a community (of poets in this study) or an individual poet, will help draw the boundary of concerns and imagination of the person(s) concerned. The statements about the content and distribution of English poetic lexicon opens the way to wider possibilities of investigating the authenticity of works, the life and history of text culture and of the English poetic potential compared with that of other languages such as French, German or Arabic. Words on the margin, outside the general frequency of (K1, K2 etc.) academic learned words, and Offlist words, offer a field of study for the out of favour, anti-canon, and the hidden aspects of words and their makers.

4.6 Discussion of the Interpretation of VPs

Vocabulary frequency and lexical features function as indicators within a complex network whose main components can be simply shown (figure 1). Human consciousness and environment, the axes of human and matter, are the basic reference for analyzing linguistic realization and human knowledge. Vocabulary is central to text, corpus and the function of the linguistic system realized in the function of words in linguistic structure and propositional content.

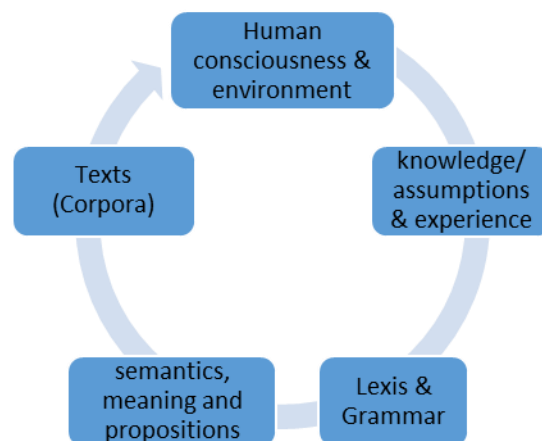


Figure 1. The interpretive & functional cycle of words

Informativity as a component of text hinges on the vocabulary content; and the structure of language does not operate independently of its lexicon. Word meaning does not operate in isolation; nor does it aptly yield to discrete item analysis, even in a dictionary. Lexis is at the heart of the linguistic cycle in Chart (1), and can be interpreted as such. The claim that the lexicon is an appendix to the grammar is based on an overgeneralization that takes lexis and the lexicon for granted. This claim has been rightly set aside previously by linguists like Sinclair and Halliday and more recently by advocates of the mental lexicon (see Mirêlis, 2004 & Ullman, 2001). Even Wittgenstein's "private language" is proposed or rejected through an argument about the possibility of existence of a private lexicon (Wittgenstein 1951, sections 243 & 202 & Kripke 1982, pp. 4-6). Kuhn (1962) has shown that to use a set of terminology is to subscribe to the theoretical frame that gives rise to the terminology, and that a technical term is locked in its semantic spread to that particular theory allowing no commensurability among terms (Words) from a different theory in the same discipline and same language (Kuhn, 1962, pp. 148-151).

The poet's construction of his/her VP relates to his/her topics which are derived from history in Shakespeare, myth and allegory in Spenser, oriental history in Marlow, Modernism, Chinese and Latin in Pound, Modern Europe and mythology in Eliot. Hence, experience equipped with topics and harnessed by imagination charge words in statements with limitless semantic potential as the following examples show

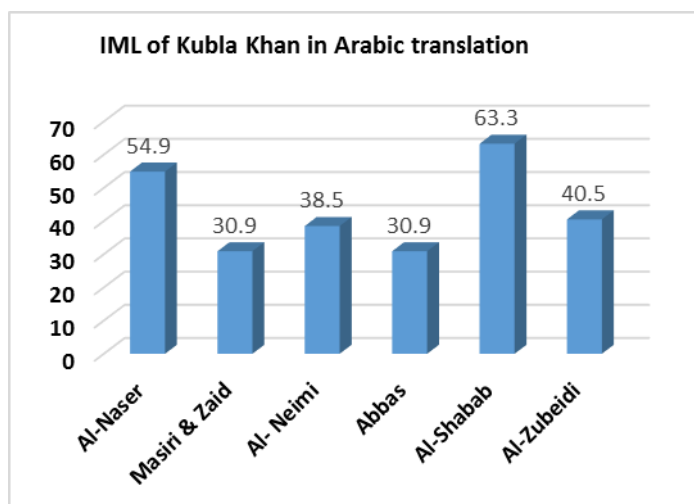
- a) I am not Lazarus (T. S. Eliot, *The Love Story of Alfred Profroke*)
- b) And drank the milk of paradise. (Coleridge, *Kubla Khan*)
- c) She walks in beauty like the night. (Lord Byron)
- d) I never writ, nor no man ever loved. (Shakespeare, *Sonnet 51*)
- e) They flee from me that sometime did me seek, (T. Wyatt, *They flee...*)
- f) Your children are not your children,
They are the children of life. (Gibran K. Gibran)

The VP creates in a unique compilation of words, grammar and context, all of which result from the poet’s creative skill and interpretation attested in the contextualization (Baka, 2014 & Author, 2012). The construction itself originates in interpreting experience and identity in a specific environment, the poet’s own environment. Richards and Ogden connects the linguistic sign in relation to a context which is made of “a set of entities (things or events) related in a certain way” (Richards and Ogden, 1923, p. 58), and that stands “behind all interpretation we have the fact that when part of an external context recurs in experience this part is, ... sometimes a sign of the rest of the external context” (Richards and Ogden, 1923, p. 57). The processing of words in one individual is a matter of interpretation of the context through the “sign”, the word. The poet’s words in the VP reflect the interpretation of the context and the poet’s experience, in a socio-environmental context.

This construct of the VP comes into life through the poet. The VP realized in the textual output, is open to various readers after its introduction to public domain. Here the reader, or the analyst, constructs a new VP out of the original poets depending on the reader’s environment and experience. The reader’s interpretation of the poet’s VP adds a new construct determined by the reader’s environment and experience. The VP, therefore, has one construct in relation to the poet’s own interpretive frame, and a set of “external” constructs by readers using their own interpretation of the original construct. In this sense, the reader makes sense of the poet’s VP in relation to his/her experience and environment as a reader.

4.7 Discussion of The Translator’s VP

The multiple constructs of the VP feed on the processing of interpretation and the ML in First Person Domain (Author, forthcoming), a process which is based on the inherent prerogative of the language user, the creator of the VP and text. The main evidence to the various constructs of the poet’s VP resides in the variety of VPs utilized by translators who render “Kubla Khan” and “In Memory of W. B. Yeats” into Arabic. Hence, in this case, the IML reflects the individual translator’s interpretation of VP (Histogram 5).



Histogram 5. IML representing “Kubla Khan” in Arabic translation

The *types* and *IML* obtained by Baka in her study of Arabic translations of Auden’s “In Memory of W. B. Yeats” show less than (5%) difference from the present results, which may be due to using a different VPr (Baka, 2014, p. 194, 197).

The tagging of the translator’s VP to the original VP transcends time and place, and interprets the poet’s VP in a new socio-environmental context, resulting in a new VP and new poem each time the poem is translated. The words in the VP have their own dynamics “the object which is referred to by a given symbol or word is not static, but relative to each language user”. (Tonkin, 2008, P. 64). It would be interesting to see the results of translating the complete works of a given poet, to examine the amount of agreement or divergence from the original VP.

The difference in the number of Types, i.e., ML, evokes a question about the value of the number of Tokens and the directionality, i.e. changing the SL and the TL. When English is the SL, Arabic as a TL has fewer words, due to morphographology, the interaction between morphology and graphology. The opposite is the case when Arabic is the SL and English is the TL. When English is the TL, more words are found in the English TT, with interesting variation due to the translators’ strategy and phraseology. In a twenty-two thousand words text wide variation is found:

Table 9. TT English -ST Arabic ratio

Feature	Translator				Total text
	A	B	C	D	
ST (Arabic) words	4316	4966	5566	5387	20.235
TT (English) words	4743	6961	9831	7780	29.315
TT English -ST Arabic Ratio	0.90	0.71	0.56	0.69	0.69

The Arabic text is a legal document (a sentence by a Shari'a court) which was translated in four parts by four translators for speed of delivery. Translator D coincided with the general average for the whole document, but the difference between translator A and translator C is (34%), which suggests that translator A has produced a summarized form in some part, while translator C has adopted a paraphrase strategy.

4.8 Concerning the Limits of VPrs

Morris and Cobb among others are aware of the scope and limits of the Classic and Compleat VPrs, even their limitations in pedagogical applications (Morris & Cobb, 2004). The present paper extends the utilization of these VPrs, but it also exerts pressure on them. Four shortcomings of the programs are observed in the course of the present study: 1) limitations of the capacity of processing large number of words, 2) the observed mismatch between the number of words of input and actually processed words, 3) the variation in coverage in terms of frequency and lexical indicators, 4) the inherent circular relation to the ML. The first weakness is well known, since the two programs do not claim processing texts beyond (6.000) words for the Classic VP and (60.000) words for the Compleat VPr. This constitutes a limitation on comparing large corpora, existing trends and lexical features. The second shortcoming is the observed discrepancy in the number of input words and the number of processed words. The third limitation is clear in that the number of word frequency is limited to two thousand (K1 and K2) in the Classic VPr, and twenty-five thousand words in the Compleat VPr. Additionally, lexical indicators like Lexical density, frequency of Anglo-Saxon words, function/content words and AWL, are found only in the Classic VP, while sentence length and cognateness are found only in the Compleat VPr. The fourth limitation is circularity, which is methodological and hence shared with other vocabulary research. This is mainly because if the complete corpus of a poet is described and all the possible features of his VP are known, then one would be able to successfully predict, or attribute, a new anonymous piece to the poet in question. This prediction, however, is based on an earlier judgement about the authenticity of the corpus, but more importantly, cases which require text attribution or author identification are far from the ideal situation in which a stable corpus is already known and only an automatic prediction is needed (Zhao & Zobel, 2006).

5. Conclusion

The poet's Types, frequency and the translator's ML, come together in the VP. Poets and translators enjoy a distinctive VP whose distinctiveness and content call for explanation. The sample, especially from Arabic translation, is rather small, but it provides indicators to be developed later. Two theoretical junctions are crucial for explaining the VP in primary and translated texts. First, words, lexical and functional, operate in networks of sense relations and collocations specific to the VP holder, because they result from the holder's choice of words. This first junction has implications for the uniqueness of the configurations of the VP. The second theoretical junction has to do with the fundamental link between words and the specific environment which originates them as carriers of content, of embodiment of experience and identity and of the interpretation of wider socio-environmental conditions of the VP holder. Briefly, the VP construct rests on the human consciousness realizing the compilations of lexis, and on the interpretation of the socio-environment that generates the specific VP or IML of a particular poet or translator. Consequently, the study of VP in corpora transcends the question of vocabulary and their frequency to evoke a host of questions touching on all aspects of language, including propositional content, the relation between the linguistic sign and its reality, and ultimately linguistic interpretation and the hermeneutic circle.

More research is needed to verify, refine and complement the present results. Computer programs need to be improved and empowered. English poetry corpora need to be enlarged and thoroughly investigated. A larger Arabic translation corpus can shed light on translation processing and strategies in general and the translational ML in particular. Still, the VP presents a streamlined silhouette that hides great deal of emotional content and relational dynamics.

References

- Abbas, B. (Translator). (nd). *qublāya khaṅ (Kubla Khaṅ)*. <http://www.adab.com/world/modules.php?name=Sh3er&doWhat=shqas&qid=80830>
- Al-Zubeidi, A. S. (Translator). (nd). *qublāya khaṅ 'aw vu 'yā fī manām (Kubla Khaṅ: or A Vision in a Dream)*. <http://www.alnoor.se/article.asp?id=30961#sthash.ohOXBr24.dpuf>.
- Author, (1996). *Interpretation and the language of translation: Creativity and Convention in Translation*. London, England: Janus.
- Author (2008). *From Necessity to Infinity: Interpretation in language and translation*. London, England: Janus.
- Author (2012). First person domain: threshold mental lexicon and Arab learners of English, *Proceedings of the second symposium on English language teaching in KSA: realities and challenges: research papers*. Riyadh, Saudi Arabia (9-11 April, 2012), pp. 141-208.
- Author (2017) *Linguistic interpretation: the interpretive frame and first person domain*. Al-Ahsa, Saudi Arabia: King Faisal University Press.
- Author & Baka, F. (2015). Text variants and first person domain in Author identification: hermeneutic versus computerized methods. *International Journal of Applied Linguistics and English Literature*, Australian International Academic Centre, 4(5), pp. 170-181.
- Baka, F. H. (2014). The mental lexicon in poetry translation: examples from translating Auden's In Memory of W. B. Yeats. *Faculty of Languages and Translation Journal*, Al Azhar University, 7(2), pp. 188-216.

- Béjoint, H. (1989). Codeness and lexicography. In G. James (Ed.) *Lexicographers and their works* (pp. 1-4). Exeter, Exeter University Press.
- Brenzina, V. & Gablasova, D. (2015). Is there a core general vocabulary? *Introducing the New General Service List*. <https://doi.org/10.1093/applin/amt018>
- Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List: A Core Vocabulary for EFL Students and Teachers, from <https://www.google.com.sa/search?biw=1366&bih=667&q=The+new+general+service+list%3A+a+core+vocabulary+for+EFL+students+and+teachers&oq=The+new+general+service+list%3A+a+core+vocabulary+for+EFL+students+and+teachers&...>
- Burkett, T. (2015). An investigation of the use of frequency vocabulary lists in university intensive English programs. *International Journal of Bilingual & Multilingual Teachers of English*. 3(2), pp. 71-79.
- Cahill, A. (1989). Word-frequency dictionaries of Russian: the user of perspective, In G. James (Ed.) *Lexicographers and their works* (pp. 18-28). Exeter, Exeter University Press.
- Cobb, T. (2013). Frequency 2.0: Incorporating homofoms and multiword units in pedagogical frequency lists. In C. Bardel, C. Linduist, & B. Laufer (Eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and corpus Analysis* (pp. 79-108). Eurosla Monograph.
- Cobb, T. (2010). Learning about language and learners from computer programs, *Reading in a Foreign Language*. 22(1), pp. 181-200. <http://nflrc.hawaii.edu/rfl/April2010/articles/cobb.pdf>
- DuBay, W. H. (Ed.) (2006). *The classic readability studies*. Costa Mesa, impact Information, pp. 32-43. <http://www.ecy.wa.gov/quality/plaintalk/resources/classics.pdf>.
- English Poetry Corpus. WWW.pomhunter & www.Poetryacademy
- Hornby, A. S. (2005). *Oxford advanced learners' dictionary* (7th ed.). Oxford, Oxford University Press.
- Kay, P. (1997). *Words and the grammar of context*. Stanford, California, CSLI Publishers.
- Kripke, S. A. (1982). *Wittgenstein on rules and private language*. Cambridge, Massachusetts, Harvard University Press.
- Kuhn, T. (1982). Commensurability, comparability, communicability. *PSA: Proceeding of the biennial meeting of the philosophy of science association*. 2, PP. 669-688.
- Kuhn, T. (1962). *The Structure of Scientific Evolutions*. Chicago, the University of Chicago Press.
- Morris, L. & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a second language trainees. *System*, 32, pp. 75-87. www.elsevier.com/locate/system.
- Mirêlis, M. T-M (2004). *Exploring the adaptive structure of the mental lexicon*. PhD thesis, Edinburgh University, Scotland, UK.
- Ogden, Charles K. (1930). *Basic English: A general introduction with rules and grammar*. <http://ogden.basic-english.org>.
- Nation, P. & Waring, R. (2004). *Vocabulary size, text coverage and word lists*. <http://www.wordhacker.com>
- Patty, W. W. & Painter, W. I. (2006 originally 1931). A technique for measuring the vocabulary burden of textbooks. In W. H. DuBay (Ed.), *The classic readability studies* (pp. 32-43). Costa Mesa, impact Information. <http://www.ecy.wa.gov/quality/plaintalk/resources/classics.pdf>
- Richards, I. A. & Ogden, C. K. (1989/1923). *The meaning of meaning*. USA, Harcourt Jovanovich Inc.
- Sapir, E. (1021). *Language: An introduction to the study of speech*. <http://www.ugr.es/~fmanjon/Sapir,%20Edward%20-%20Language,%20An%20Introduction%20to%20the%20Study%20of%20Speech.pdf>
- Sin-Wan, C. (2004). *A Dictionary of translation technology*. Hong Kong, The Chinese University. Press.
- Szymańska, A. L. (2015). The English vocabulary Profile as a benchmark for assigning levels to learners corpus Data. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment*. USA, John Benjamins Company. <https://www.researchgate.net/publication/275021429>.
- Thorndike, E. L. & Lorge, I. (1944). *The teacher's word book of 30.000 words*. New York, NY: Colombia University press. <http://facstaff.bloomu.edu/triley/The%20Complete%20Collection/from%20The>.
- Tonkin, E. (2008). Orthography, structure and lexical choice as identity markers in social tagging environments. *IADIS International Conference on Web Based Communities*, pp.63-69. www.academia.edu/2545427/Orthography_Structure_and_lexical_choiceas_identity_markers_in_social_tagging_environments?auto=download
- Ullman, M. T. (2001). The neutral basis of Lexicon and grammar in the first and second language: The declarative/procedural model. *Bilingualism, Language, and the Cognition*, 4(2), pp. 105-122, Cambridge, Cambridge University Press.
- West, M. (1953). *A General service list of English words*. www.cambridge.org/downloadfile/849064/0/
- Wittgenstein, L. (translator: P. M. Hacker & J. Schulte) (2009/ originally 1551). *Philosophical investigation*. London, Wiley-Blackwell.
- Zhao, Y. and Zobel, J. (2006). Searching with style: Authorship attribution in classic literature. *Twenty-ninth Australian computer science conference (ACSC)*. <http://goanna.cs.rmit.edu.au/~jz/fulltext/acsc07yz.pdf>