| **タイトル**<br>Title | Comparative Lexical Frequency Analysis of East Asian and Middle Eastern Corpora : Motivated by the Extremely Short Story Competition [ESSC] |
| **著者**<br>Author(s) | Hassall, Peter John |
| **掲載誌・巻号・ページ**<br>Citation | Learner Corpus Studies in Asia and the World,1:199-226 |
| **刊行日**<br>Issue date | 2013 |
| **資源タイプ**<br>Resource Type | Departmental Bulletin Paper / 紀要論文 |
| **版区分**<br>Resource Version | publisher |
| **権利**<br>Rights | |
| **DOI** | |
| **URL** | http://www.lib.kobe-u.ac.jp/handle_kernel/81006683 |
Comparative Lexical Frequency Analysis of East Asian and Middle Eastern Corpora: Motivated by the Extremely Short Story Competition [ESSC]

Peter John HASSALL

Zayed University, UAE

Abstract
Consistent with previous papers (Hassall, 1998, 2006a; Hassall & Ganesh, 1996, 1999), this study utilizes Correspondence Analysis (CA) to compare lexical frequency distribution of corpora compiled by the Japanese Association of Asian Englishes (JAFAE) and Zayed University (ZU), United Arab Emirates (UAE), in response to the 50-word Extremely Short Story Competition implemented in formal and informal educational settings in both Japan and the UAE. In 2007, over five hundred ESSC scripts of exactly 50 words in length were assembled by JAFAE and over nine hundred ESSC scripts were assembled by ZU. Both raw, and mediated corpora (edited ready for publication in the UAE) are considered. It is hoped that this study will complement a parallel study examining individual's topic choices using the same corpora undertaken by Matsubara (2013), and also previous studies of ESSC corpora carried out by Fujiwara (2008) and Okaura (2007, 2009).

Keywords
Comparative, Lexical, Frequency, UAE, Japan

I Introduction

Echoing the ‘International Corpus of English, (ICE)’ proposed by Greenbaum (1988, 1991a, 1991b, Ed. 1996) and the ‘International Corpus of Learner English (ICLE)’ initiated by Granger (1998, 2001), the present author proposed an ‘International Corpus of Creative English (ICCE)’ in the journal World Englishes (Hassall, 2006a). It was suggested that this ICCE corpus should be built around creative English similar to the output from the 50-word Extremely Short Story Competition (ESSC) which had been highly successful in formal and informal educational contexts in the UAE. Under the
enlightened leadership of its Founder Professor Emeritus Nobuyuki Honna, the Japanese Association for Asian Englishes (JAFAE), became involved in running its own ESSC, thus establishing the ICCE as a viable international corpus. It is useful to characterize the ICCE using insights gleaned from the compilation of corpora centering on Japan and the UAE. The ICCE may be considered a prospective, purposeful, pedagogic and eclectic corpus. For a detailed description of ICCE see Hassall (2006a).

The ICCE may be considered prospective and purposeful since prior to student-authors submitting stories to the ESSC they agree to their 50-word Extremely Short Stories (ESS) being proofread for publication and their submissions used for intercultural/inter-linguistic research. One of the aims of the ICCE is to encourage participants to develop their individual creative resources in English and to take their creative contributions, i.e. what ‘they want to mean’, out into the wider community. This ensures that the ICCE is not a solely classroom-based activity having instrumental aims closely related to the students’ passing of examinations and their developing English as closely as possible to the norms of native speakers of English. Instead a more holistic and integrative pedagogy is encouraged which allows student-authors, irrespective of their linguistic proficiency, to engage with the English language with the prospect of their receiving palpable appreciation from the community with the likelihood of their ESS appearing in print and the possibility of their linguistic contributions earning them tangible benefits in terms of ESSC awards. Several thousand student-authors have engaged in the ESSC, and hence ICCE, in both Japan and the UAE. This kind of community recognition has been delivered in-house, with national and international consequences including ESSC awards events; daily features in national newspapers, radio presentations, book launches and associated commercial publication of three ESSC anthologies in the UAE (Hassall, Ed. 2006a, Ed. 2007, Ed. 2009) and three ESSC anthologies in Japan (Hassall, Ed., 2006b; Hassall Ed. Trans. Takeshita & Sakai Tanaka, 2010; Takeshita & Sakai Tanaka, 2010). In addition, compilation of an ESSC international traveling exhibition emanating from the UAE and translated into six languages is featured on the Zayed University website: http://www.zu.ac.ae/facets and publication of six years of ‘Outstanding works of the ESSC’ appear on JAFAE website: http://www.jafae.org/essc/en/past_award_work.html

In terms of eclecticism, the ICCE was originally formulated to engage “countries which have tertiary institutions with well-defined populations of student-authors possessing similar cultural and/or linguistic backgrounds” (Hassall, 2006a). This homogeneity reflects the composition of the student body of Emiratis who were readily accessible to the originator via the government tertiary education system in the UAE. A similar accessible homogeneity was also available in Japan to JAFAE; however, the reach of this Japanese association extends to junior high schools, senior high schools, and universities. Therefore JAFAE encouraged a wider age-range to participate in their
ESSC. In the UAE, a website\textsuperscript{7} was constructed to gather ESS of exactly 50 words from throughout the UAE and any text with less or more than 50 words could not be submitted to this website. In Japan, the precise 50-word restriction was stated, rather than being enforced electronically as in the UAE, with the onus being on Japanese student-authors to comply with this directive. Hence stories submitted to JAFAE which did not comply were disqualified for non-compliance with this rule. Naturally, these irregularities produce some inconsistencies as regards what 50 `words\textsuperscript{8}\' should involve.

There are advantages and disadvantages of both the Japanese interpretative implementation and the UAE electronic enforcement of this exact 50-word prescription. The Japanese modus operandi, which puts responsibility for compliance with the exact 50-word prescription on the individual, might allow/disqualify individual ESS depending on an adjudicator `s interpretation; whereas rigid electronic enforcement of this rule in the UAE might produce some elision of spaces in writing, perhaps causing some student-authors to be more concerned about compliance with the electronic word-count regulation than careful attention to meaning. Regardless of which system is deployed, the biggest danger to the reliability of the resultant corpus is the dependence on the integrity of the individual student-authors since the ESS are not collected under examination conditions unlike other corpora such as ICLE. In practice the primary material, collected via both the interpretive and electronic methods, indicates a wealth of `inaccuracies` and very personal observations which suggest that students have generally submitted individual creative compositions with little assistance from others who might have greater proficiency than themselves\textsuperscript{9}.

These initial submissions to the ESSC\textsubscript{, Japan} and ESSC\textsubscript{, UAE} were compiled into `raw` sub-corpora, coded here as \textit{j1raw} and \textit{u1raw}, respectively. In the UAE, as submissions for the \textit{u1raw} corpus were received they were immediately proofread for publication as electronic PowerPoint slideshows and printed-out paper exhibitions and put on display in participating institutions to inspire new student-authors to contribute their own stories to the ESSC. Subsequently these proofread stories were put forward for the judging process and for compilation of an exhibition for an ESSC awards event. Following the end of the competition, this sub-corpus of proofread ESS, coded here as \textit{u2pfd}, was used for publication into an anthology which went on sale commercially\textsuperscript{10}. In 2007, the JAFAE ESSC Steering Committee, emailed JAFAE ESS to the present author for similar proofreading and compilation into a similar publication. This proofread Japanese corpus subsequent to proofreading by Peter Hassall & Jonathan Hassall (Hassall, P. J., & Hassall, J. E., Eds. 2013 forthcoming) is here referred to as \textit{j2pfd}. It is hoped that different countries and regions might, for their own ends, similarly exploit some of the processes and procedures of the ESSC and ICCE, outlined here.
II Literature Review

2.1 Chronology of contributions by ESSC_UAE and ESSC_JAFAE towards the development of a purposeful ICCE

The first anthology of student ESS (Hassall, Ed. 2004) was published in-house at Zayed University (Zayed University, 2007). Two years later a commercial edition of the same anthology (Hassall, Ed. 2006a) was published in the UAE. In response to an exchange of essays between Hassall (2006c) and Takeshita (2006), a Japanese edition of the same anthology was published in Japan (Hassall, Ed. 2006b). Following technical documentation and presentations of the approach (Hassall, 2005a, 2005b, 2005c, 2006b), theoretical background and recommendations for the ICCE were drawn up for publication in Hassall (2006a). The presentation by Hassall (2004a) had inspired the British Council in Seoul, Korea to run an in-house ESSC and this enabled Hassall, shaping the design of the present study, to propose the following directive:

For each national ICCE that is collected using the ESSC, two sub-corpora will be generated: compilation of raw, student text based on the students’ first drafts submitted to the ESSC and another of mediated texts edited for publication. Both national ICCE sub-corpora will be consistent, composed of machine-readable texts of exactly fifty words in length, with the title and authorship being extra. These two sub-corpora may be considered individually, together, or in contrast to each other, in order to determine quantitative and qualitative distinguishing features. (Hassall, 2006a p. 145)

Around this time, papers relating to outcomes from the ICCE detailed pedagogy in the language classroom with Abu-Wardeh (2006) describing how she successfully uses ESS to provoke ‘truly authentic grammar lessons’ whilst teaching English at ZU (p. 80). In Japan, this was complemented by White (2007) outlining classroom procedures and sample ESS to encourage teachers to involve students in writing 50-word stories for the ESSC. When investigating implementation of the ESSC, as part of action research in the classroom, a grammar and writing teacher at a Japanese university explains how students find the ESSC both interesting and stimulating making the grammar lessons more enjoyable (Tsuda, 2007). Similarly La Faye (2007), finds that introducing the ESSC into the writing classroom provides a ready-made goal for the students and invokes a positive response.

In 2007, a number of JAFAE officers, Takeshita, Fujiwara, Miyake and Okaura, delivered a panel presentation in Harbin, China and a number of significant papers have ensued (Takeshita et al., 2007). Takeshita (2008) provides a comprehensive overview of implementation of the ESSC in Japan, and describes JAFAE’s
administration, funding, screening and judging processes, and rates of submission, with samples of some of the award-winning ESS_JAFAE being provided. Miyake (2012) reports a study involving 460 university students over 4 years finding that they all ‘completed’ at least one story. Miyake also observes that introducing students to the creation of ESS, and conforming to the ESSC criterion that insists on students ‘submitting all their own work’ sometimes posed problems for teachers regarding their individual pedagogic role in the classroom and called for further discussion to remedy this and thereby “improve ESSC for English learners” (p. 66).

Since that time, there has been considerable international collaboration between Zayed University and JAFAE, related to the ICCE, including materials development workshops such as Hassall, P. J., Hassall, S. & La Faye, B. E. (2008); panel presentations such as Hassall et al. (2012) and publication of monolingual English anthologies (Hassall, Ed. 2009) and bilingual anthologies, viz. innovative translations by Takeshita and Sakai Tanaka (2010) and Hassall (Ed., Takeshita & Sakae Tanaka Trans. 2010) which have led to further translations posted on http://www.zu.ac.ae/facets. Further documentation of this collaboration is provided in Hassall (2009, 2011), and D’Angelo (2012). Involvement in Europe is detailed in Collins (2011) who explains the ESSC is a project partner with the Evaluation and Accreditation of Quality in Language Services (EAQUALS, 2011) and details educational issues, processes, benefits and particular interest from Greece, Romania, Bulgaria and France (also see Hassall, P. J., & Collins, M. (2009).

2.2 Investigations into text analysis and Correspondence Analysis towards the analysis of ICCE corpora

Being a Japanese national, experiencing day-to-day living in the UAE for a number of years, teaching Japanese to Emirati students and organizing the Japan Club at Zayed University, Matsubara (2011) offers insightful interpretations of ESS text on the topic of ‘Emirati Women’. Most recently (2013) Matsubara uses components of the ICCE to compare the lives and society of Emirati youth, characterized topics for ESS chosen by UAE student-authors which are published in Hassall (Ed. 2007), with similar themes chosen by JAFAE student-authors, appearing in Hassall, P. J. & Hassall, J. E. (Ed, 2007 forthcoming 2013), to explore the potential of the ICCE for comparative sociolinguistic research. It is hoped that the present study, focusing more on language variety complements Matsubara’s perspective.

A more text-oriented approach which considers the linguistic output of the ICCE is taken by Okaura and Fujiwara. Okaura (2007) notes that extensive use of personification, mental process verbs and co-referential expressions appear to characterize ESS submitted to JAFAE and considers a number of ESS samples in terms of relevance theory (Sperber & Wilson, 1986). Additionally, Okaura (2009) examines
characteristic grammatical constructions in 50-word essays of senior high school and university students in Japan with respect to relevance theory. This investigates “ungrammatical but acceptable constructions” influenced by the Japanese language and examines effective usages contributing to the “presumption of optimal relevance in 50-word essays regardless of whether they are prescriptively ‘ungrammatical’ and syntactically ambiguous or not” (p. 226). Okaura concludes from this investigation into the ICCE that Japanese universities and senior high schools should place greater emphasis on international intelligibility than strict grammaticality when teaching writing in English.

Utilizing a methodology closest to that of the present paper, Fujiwara (2008) considers student-authors’ lexical choices in the ICCE Japan to determine creative characteristics of English used by Japanese leaners. Using the ICCE, which Fujiwara notes as being less-prescriptive in terms of topic choice than other corpora, frequency data are examined for three groups of entries by Japanese students: 1) those produced by junior high school students (JHS), 2) those produced by senior high school students (SHS), and 3) those produced by university students (UNI). After subtracting function words, pronouns and demonstratives to produce lists of ‘creative lexis’ these were then submitted to the statistical technique of Correspondence Analysis (CA). Cautioning about the small size of the corpora being considered, Fujiwara concludes tentatively that as students become older their primary concerns, shown by the CA of lexical frequency of the three sub-corpora, indicate thematic development from interest in ‘friendship’ shown by frequent use of the lexical items ENJOY, FINE, INTERESTING, FRIENDS, TREASURE, MUSIC by junior high school students (JHS); through ‘future’ represented by WILL, WISH, DREAM and ‘sorrow’ shown by BAD, SAD, CRY as demonstrated by senior high school students (SHS); to ‘love’ and ‘global issues’ as shown by a more frequent use of JAPAN, EARTH by university students and adults (UNI) (p. 338).

Extending Hassall & Ganesh (1996), Hassall (1998) considers correspondence of lexical items and develops what is termed a relative deviance analysis to underpin the pedagogy of TEIL based around World Englishes (WE) outlined in (Hassall, 1996a, 1996b, 2000, 2004a, 2004b). World Englishes, as proposed by Kachru (1985) includes consideration of the Outer and Expanding Circles of English, in addition to the Inner Circle as represented by the Major Varieties of English (MAVEN) (Svartvik, 1998). Relative deviance is considered a partial measure of variation that results when, for instance, identical language tasks are undertaken by representatives of different language varieties and, as indicated by Mair (1998), in order to determine relative variation between varieties more comprehensive corpora representing the totality of the varieties would need to be compared. Together two descriptors of deviance offered by Crystal (1995, p. 395) epitomize the approach that is to be elucidated in the present
study: a) that deviance may identify various varieties of English and b) that slight degrees of deviance only become apparent through statistical investigation. These are applied to establish the identity of corpora compiled from texts created by different groups of users of WE varieties. Rather than direct contrast with standard Englishes (Standard British English or General American) which would result in a standard relative deviance (relDEV.SBE or relDEV.GA), contrast between the sub-corpora is made by comparing the language of each of these sub-corpora with the aggregate of the corpora involved in the interaction, which may be considered to represent a 'notional EIL'. A later study, Hassall & Ganesh (1999), utilizes Correspondence Analysis to investigate this relative deviance across varieties for five different groups of university student-authors from Japan, Korea, Thailand, Taipei and Ras Al Khaimah (RAK: this being an emirate of the UAE) whilst undertaking a task similar to that on the writing component of the IELTS test. The aim was to contrast between word counts of the most frequent 'etic' units of these sub-corpora with reference to the written language produced rather than their nearness to a standard native English speaking norm.

Correspondence Analysis (CA) (Greenacre, 1984, 1993), was employed to produce a graphical display of the relationship between the language-user groups and the most frequent orthographic units that they each use when compared to the aggregate of these groups. This procedure suggests salient characteristics of text produced by each of the submission groups, when compared to one another and characterizes the international relative deviance analysis (relDEV.EIL) utilized in the present study.

III The Study

3.1 Research Design

Correspondence Analysis is a multidimensional statistical technique which provides an assessment of the interdependence of the rows and columns of a data matrix (primarily, a two-way contingency table). CA facilitates dimensionality reduction and provides graphical displays in low-dimensional spaces. In other words, it converts the rows and columns of a data matrix or contingency table into a series of points on a graph. In its most basic form, CA employs a contingency table of two categorical variables, and portrays the 'correspondence' of categories of variables and forms the basis for developing graphical displays. The graphical display of CA is similar to the perceptual mapping in multidimensional scaling. Proximity of points indicates the level of association among row and/or among column categories.

In statistical terms, CA is a weighted Principal Component Analysis (PCA) of a contingency table, enabling a low dimensional configuration of the association between the rows and the columns. Once the contingency table is presented to CA, the procedure yields a conditional expectation for each row-column combination of categories similar
to that of the Chi-Square test of independence. Once obtained, these values are normalized, and then a process much like PCA defines the lower-dimensional solutions. The total inertia (similar to that of the total variation in PCA) is decomposed to represent the new dimensions. This total inertia is directly proportional to the Chi-Square statistic (for test of independence) and is a measure of total variation of the elements in the table. The number of maximum (new) dimensions obtainable equals \(\min(\text{no. of rows, no. of columns}) \cdot 1\). The low dimensions then simultaneously relate the rows and columns as points in a single plot. The axes of this low-dimensional configuration are called 'principal axes' and are arranged so that the first principal axis accounts for most of the inertia, the second explains the second largest percentage of inertia and so on. It should be noted here that the plot should be thought of as two different overlaid plots, one for each categorical variable (i.e. rows and columns). Distances between category points within a variable (i.e. distances between rows or between columns) have meaning, but distances between category points from different variables (i.e. between a row and a column) do not. In the plots, the points that lie closer to the origin with respect to a principal axis contribute very little to the inertia explained by that axis. Thus, a principal axis can be 'characterized', even given a title, depending on which categories of rows and/or columns contribute the most to this axis.

The present study compares electronic corpora of creative writing produced in English by groups of students undertaking an identical language task. The WE corpora to be compared were assembled from texts gathered in response to the 50-word Extremely Short Story Competition [ESSC] delivered via the internet to educational institutions in Japan, and in the United Arab Emirates. In 2007, Yasuhiro Fujiwara emailed raw ESSC_JAFAE ESS texts from Japan to the current author in the UAE – these were compiled into a corpus j1raw which complemented the u1raw corpus which Hassall had previously collected in the UAE. In 2007, two ESSC anthologies were ‘proofread’ ready for publication: one as a result of proofread submissions to ESSC_UAE (Hassall, Ed. 2007), and another resulting from proofread contributions to ESSC_JAFAE (Hassall, P. J. & Hassall, J. E., Eds. 2007, forthcoming 2013) and it is these two ‘sub-corpora’ which provide additional corpora to include in this study, namely u2pfd & j2pfd. Note that u2pfd and j2pfd refer to the two language user groups outputting corpora as follows:

- **u2pfd** – refers to the sub-corpus u1raw (student-authors in the UAE) with additional input from a small number of native-speaking ‘proofreaders’ (or editors) namely: P. J. Hassall, S. Rice, G. O’Neill, T. Bailey-Sefar and J. Murray Bollard. This group worked under a great deal of time-pressure to output mediated text ready for compilation into an ESSC exhibition displayed at the awards event and publication of an ESSC anthology (Hassall, Ed. 2007).

- **j2pfd** – refers to the sub-corpus j1raw (student-authors in Japan) with input from P. J. Hassall and J. E. Hassall proofreading these texts to output a similar
anthology ready for publication (Hassall, P. J., & Hassall, J. E., Ed. 2007, forthcoming 2013).

3.2 Research Questions

Research questions include the following:

a) Can the sub-corpora *j1raw, u1raw, j2pfd* and *u2pfd* be differentiated solely with respect to frequency of orthographic units of which they are comprised?

b) Is the language of the mediated corpora *j2pfd* and *u2pfd* more similar to each other’s language or to the language of the raw corpora *j1raw* and *u1raw* from which they each derive?

c) Can scrutiny of the frequency differences between the raw language produced by the sub-corpora *j1raw* and *u1raw*, and the ‘proofed language’ produced by the sub-corpora *j2pfd* and *u2pfd* point to necessary modifications which are required to enable raw texts to be published?

d) Would other approaches be useful for investigating raw and proofread texts produced in Japan and the UAE?

3.3 Research Data

In the present study, the word frequency program FREQUENCY (Heatley, Nation, & Coxhead, 2001) which identifies words as being separated by spaces, full stops and apostrophes, was used to produce frequency listings of each word type for each of the sub-corpora. These word types should perhaps more correctly be referred to as orthographic units, since the boundaries between the units are rigorously applied. Thus ‘s‘ referring to the possessive ‘*s* or *s’", or ‘d‘ referring to elision in weak forms such as ‘it’s’ [<= ‘it is’ or ‘it has (go0)’] is identified as a distinct word type *S*. The frequency list therefore not only includes words that are considered well-formed in standard English but also considers all orthographic forms that are bounded by spaces or punctuation. Kenny (1982, p. 66) refers to what here is categorized as ‘orthographic forms’ as 'un-lemmatized words'. He claims that, in a sense, un-lemmatized word counts contain more information since it is always possible, with some effort, to construct lemmatized word counts from un-lemmatized word counts but the converse is not possible. This would appear particularly significant when dealing with comparisons across language varieties, especially when dealing with ‘lesser known’ varieties of English that are inadequately codified or described. In a similar vein, Fujiwara (2008) explains that the components of the ICCE Japan, which he examined, were not grammatically tagged, “... mainly because a numerous number of samples contain what we traditionally call 'errors'. All the automatic tagging programs such as CRAWS or TOSCA Tagger are based on a so-called 'standard English.” In addition, Fujiwara’s caveat also applies, “Please note that due to the difference in the way of counting the number of the words
by means of software used for this analysis, … , the number of tokens does not exactly reflect the number of the words” (2008, p. 336).

In the present study RelDEVEIL methodology is applied to the analysis of sub-corpora whose student-authors have distinct cultures and linguistic backgrounds. These are the \text{j1raw} sub-corpus consisting of 512 submissions to ESSC_JAPAE and the \text{u1raw} sub-corpus comprising 964 responses to ESSC_UAE. Mediated corpora were also considered including the \text{j2pfd} sub-corpus using the same texts as \text{j1raw} but proofread, ready for publication/display, along with the \text{u2pfd} sub-corpus consisting of the same texts as \text{u1raw} but also proofread for publication/display. The consistency of perspective offered by identifying orthographic units and sorting them with respect to the aggregate group, enabled a two way contingency table to be assembled. In total there were 150050 word counts distributed over 7376 levels of orthographic units and four sub-corpora.

IV Results and Discussion

The initial study examines only the first 19 of the most frequent of the 7376 orthographic units. Each of these appeared at least 1000 times when all four groups were 'pooled' together.

Table 1 Contingency table of the 19 most frequent word types against the 4 sub-corpora

<table>
<thead>
<tr>
<th>Word</th>
<th>j1raw</th>
<th>j2pfd</th>
<th>u1raw</th>
<th>u2pfd</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1729</td>
<td>1745</td>
<td>2033</td>
<td>2022</td>
<td>7529</td>
</tr>
<tr>
<td>THE</td>
<td>777</td>
<td>835</td>
<td>2137</td>
<td>2073</td>
<td>5822</td>
</tr>
<tr>
<td>TO</td>
<td>634</td>
<td>649</td>
<td>1280</td>
<td>1271</td>
<td>3834</td>
</tr>
<tr>
<td>A</td>
<td>593</td>
<td>620</td>
<td>1126</td>
<td>1299</td>
<td>3638</td>
</tr>
<tr>
<td>AND</td>
<td>494</td>
<td>483</td>
<td>1326</td>
<td>1249</td>
<td>3552</td>
</tr>
<tr>
<td>MY</td>
<td>463</td>
<td>503</td>
<td>979</td>
<td>963</td>
<td>2908</td>
</tr>
<tr>
<td>IN</td>
<td>373</td>
<td>365</td>
<td>900</td>
<td>790</td>
<td>2428</td>
</tr>
<tr>
<td>YOU</td>
<td>563</td>
<td>565</td>
<td>619</td>
<td>618</td>
<td>2365</td>
</tr>
<tr>
<td>IS</td>
<td>683</td>
<td>540</td>
<td>585</td>
<td>510</td>
<td>2318</td>
</tr>
<tr>
<td>IT</td>
<td>378</td>
<td>373</td>
<td>682</td>
<td>633</td>
<td>2066</td>
</tr>
<tr>
<td>OF</td>
<td>322</td>
<td>306</td>
<td>652</td>
<td>662</td>
<td>1942</td>
</tr>
<tr>
<td>WAS</td>
<td>216</td>
<td>233</td>
<td>717</td>
<td>773</td>
<td>1939</td>
</tr>
<tr>
<td>ME</td>
<td>218</td>
<td>234</td>
<td>447</td>
<td>446</td>
<td>1345</td>
</tr>
<tr>
<td>HE</td>
<td>130</td>
<td>130</td>
<td>535</td>
<td>526</td>
<td>1321</td>
</tr>
<tr>
<td>THAT</td>
<td>169</td>
<td>148</td>
<td>455</td>
<td>428</td>
<td>1200</td>
</tr>
<tr>
<td>T</td>
<td>205</td>
<td>229</td>
<td>320</td>
<td>319</td>
<td>1073</td>
</tr>
<tr>
<td>FOR</td>
<td>200</td>
<td>185</td>
<td>340</td>
<td>331</td>
<td>1056</td>
</tr>
</tbody>
</table>
A major aim of undertaking the correspondence analysis is to determine whether the four groups may be differentiated solely with reference to the relative frequencies of the different word types across the sub-corpora. Once the contingency table is presented to CA, the procedure yields a conditional expectation for each row-column combination of categories. The total inertia is decomposed to represent the new dimensions. The decomposition of the total inertia showed that the first principal axis accounts for 95% of this total inertia followed by 4% accounted for by the second principal axis. In other words, 99% (=95% + 4%) of the information can be accounted for by the first two principal axes; thus the association between the 19 word types considered and the four groups is almost entirely two-dimensional as shown in Figure 1.

Figure 1 Graphical display of the CA for the nineteen most frequent (>1000) word types across the four language user groups

<table>
<thead>
<tr>
<th>WITH</th>
<th>157</th>
<th>167</th>
<th>377</th>
<th>354</th>
<th>1055</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>154</td>
<td>231</td>
<td>283</td>
<td>351</td>
<td>1019</td>
</tr>
</tbody>
</table>

The correspondence between the 19 words and the 4 groups is displayed graphically in Figure 1 for the first two principal axes. Note that, in this display two sets of points are super-imposed, one representing the orthographic types and the other representing the
WE sub-corpora. The order of magnitude between the sub-corpora and amongst the word types is apparent in this display. The variation from left to right along the first principal axis opposes the \textit{u2pf}\textit{d} and \textit{ulraw} groups, against the \textit{j2pf}\textit{d} and \textit{j1raw} groups; whereas the variation from bottom to top essentially opposes \textit{ulraw} and \textit{j1raw}, against \textit{u2pf}\textit{d} and \textit{j2pf}\textit{d}. The CA provides a two-dimensional representation of the relationship between the four sub-corpora and suggests that the greatest contrast is,

\[ \textit{ulraw, u2raw} \Leftrightarrow \textit{j2pf}\textit{d}, \textit{j1raw} \]

A secondary contrast may also be discerned along the second principal axis,

\[ \textit{ulraw, j1raw} \Leftrightarrow \textit{u2pf}\textit{d}, \textit{j2pf}\textit{d} \]

Basing an analysis upon the CA of these nineteen most frequent orthographic forms provides a meaningful description of the synchronic, surface relationship between the sub-corpora and indicates that the sub-corpora \textit{j1raw}, \textit{ulraw}, \textit{j2pf}\textit{d} and \textit{u2pf}\textit{d} may be differentiated solely with respect to frequency of orthographic units of which they are comprised answering research question (a) above. This CA also suggests that when considering the 19 most frequent orthographic forms, the mediated corpora \textit{j2pf}\textit{d} and \textit{u2pf}\textit{d} are more similar to the language of the raw corpora \textit{j1raw} and \textit{ulraw} from which they each derive (see research question b above). However, there is a secondary contrast which opposes the raw corpora against the proofed corpora.

The behavior of the orthographic forms reveals a contrast between words \textit{HE, WAS}, against \textit{IS} along the first principal axis. The words \textit{THAT, AND, THE, YOU, I} also make a moderate contribution to this contrast. The second dimension, however, accounts mainly for the differences between words \textit{IS, THAT, IN} versus the items \textit{S, A, T}. Although, the above behavior of sub-corpora and orthographic forms on their own may be insightful, the interdependence of these two categories is also of interest in this study. The general concept is that, a particular column profile would tend to fall in a position which corresponds to the row categories which are prominent in that column profile. For example, the \textit{j1raw} point lies furthest on the positive side of the first principal axis and any orthographic forms that lie on the positive side of the first axis (i.e. \textit{IS, YOU, I}) may be regarded as ‘influential’ for this group.
Figure 2 Projections of the nineteen most frequent (>1000) orthographic forms onto the secondary axis through the j2pfd sub-corpus.

The relationship between row and column profiles may be examined by considering projections of the ‘orthographic form’ points onto line(s) drawn through the – sub-corpus – point(s) and the origin on the graph. This in turn, enables us to relate, for example, the j2pfd sub-corpus with all 19 orthographic forms as shown in Figure 2. Table 2 elaborates on this information and provides lists of the 19 most frequent orthographic forms in the aggregate corpus ordered so that those exerting the most positive influence (attraction) on a sub-corpus appear at the top of each list whereas those contributing a negative influence (repulsion) appear at the bottom.
An examination of the most frequent orthographic units included in each of the sub-corpora groups permits slight degrees of deviance to be observed. It can be seen from Table 2 that the u2pfd and u1raw groups are drawn towards use of the forms HE, WAS, THE, AND, THAT, in that order. These forms contribute negatively, in reverse order, to the behavior of the j1raw sub-corpus and to a lesser extent to the j2pfd sub-corpus. Hence, in this selection, the word HE contributes most to the behavior of u1raw and u2pfd and least to the behavior of the j1raw and j2pfd sub-corpora. The orthographic forms WAS, THE, AND, THAT also contribute negatively to the behavior of j1raw sub-corpus, in that order and to a lesser extent j2pfd.

In contrast the sub-corpora j1raw and j2pfd may be characterized by the positive influence of a different set of forms. The j1raw and j2pfd sub-corpora share three of their most positively influential orthographic types IS, YOU, I, T but in a different order. These forms IS, YOU, I, T have a negative influence on u1raw and u2pfd but in reverse order with the item YOU contributing the most negative influence on these sub-corpora.

The influence of these two major groupings of forms explain the primary contrast between u1raw and u2pfd as opposed to j2pfd and j1raw. However, the item S is
positively influential in the behavior of j2pfd, and u2pfd but has a negative influence on ulraw and, to a lesser extent, j1raw.

In general it would appear that when considering the 19 most frequent word types only, certain sub-corpora are attracted by particular orthographic types and repelled by others. Some words however contribute in a similar way to each of the sub-corpora, in particular the orthographic forms TO, ME, MY are positioned similarly for all three sub-corpora, lying close to the origin in Figure 1 and hence exert neither a positive nor negative influence. Productive inquiry into relative deviance might best be achieved by examining the outliers YOU, IS, I, S, and HE, WAS, THAT, THE.

Figure 3 Graphical display of the CA for the 45 most frequent (>500) orthographic types across the four sub-corpora

Taking a wider view, Figure 3 provides a representation of CA for the 45 orthographic units that occur more than 500 in the aggregate corpus. When considering the most frequent orthographic units, there is attraction between certain forms and sub-corpora and repulsion between others. Examination of the orthographic units indicates that positive influence on j2pfd is provided by AM, LIKE very, are, have; whereas, positive influence on j1raw is exerted by IS, AM, very, like; whilst, positive influence on ulraw is provided by HIS, her, one, she, he and for u2pfd: HIS, her, she, was.
These orthographic items all warrant further scrutiny as they positively characterize the specific sub-corpora.

Scrutiny of these orthographic items does not immediately generate a distinct pattern comparable to the contrasting relationship detected in Hassall & Ganesh (1999) between the contrasting usage of I, YOU, WE by Middle Eastern and Sino-Pacific groups when undertaking an expository essay on a given topic. This is partly because the creative brief given to the students responding to the ESSC had no restrictions as to theme and therefore produced an extremely open-ended response which caused the production of lexical varietal preferences in response to a specific task considerably less likely than in the focused IELTS-type writing activity for which a relatively closed set of responses was output. In addition, the relDEV.EIL analysis was developed by Hassall (2008) primarily to distinguish between different language-groups, in other words participants who were representatives of different language varieties; whereas, in fact the proofreaders were native speakers of English specifically chosen for their editorial skills to draw out the message of non-native speakers and make this accessible internationally to a general public without transgressing the tenets of the media gatekeepers in the UAE and wider afield. Thus, the proofreaders mediated the language of the groups j1raw and u1raw, and the lexis of the resulting groups j2pfd and u2pfd was not systematically distinct as not all ESS submissions required modification. As a proofreader, the author of this paper became aware that extra orthographic forms had to be inserted to make the message clearer for some of the ESS in both j1raw and u1raw. Because there was a tight 50-word constraint this meant that for proofreaders to insert extra words to ensure coherence some elision was often necessary to prevent exceeding the word limit. Possible mediation by the proofreaders may be seen as the elided word S assumes an appreciably positive significance in both the j2pfd and u2pfd columns of Table 2 when compared to j1raw and u1raw. Conversely in mediation from raw to proofed, IS accepts a less frequent position from most frequent in j1raw to second most frequent item in j2pfd. Likewise, when considering u1raw, the item IS moves from the fourth-least-frequent position exerting negative attraction on the u1raw sub-corpus when considering these 19 items to the 19th (least) frequent position for u2pfd. The increased frequency of S through mediation, which might also include clarification of the possessive S with the addition of an apostrophe, is further elucidated through utilization of the Concordance Plot feature in Figures 4 and 5.
Figure 4. Concordance Plot of IS from bottom upwards in j1raw, j2pfd, ulraw and u2pfd (using AntConc software developed by Anthony, 2011)

Figure 5. Concordance Plot of S from bottom upwards in j1raw, j2pfd, ulraw and u2pfd (using AntConc software developed by Anthony, 2011)

Figure 4, shows that IS has a presence distributed fairly evenly throughout j1raw and ulraw (first and third set from the bottom). It can be seen that in j2pfd and u2pfd the areas of low incidence of IS, represented by areas of white, show they have been mediated and these areas have become less extensive indicating a lower occurrence of the item IS when mediated by the proofreaders. Figure 5 conversely indicates that the lower incidence of S shown in j1raw and ulraw (first and third set from the bottom)
becomes more frequent in the mediated $j2pfd$ and $u2pfd$ as indicated by the darker ‘barcodes’ explaining that $S$ is being used more extensively, thus supporting the suggestion of interrelatedness of these two items through mediation. This supports the idea that mediation is being provided to comply with the task-fulfillment (i.e. to abide by the exact 50-word constraint) whilst at the same time produce coherent text that will conform to the norms of publication and presentation\textsuperscript{12}. An illustrative example follows which shows a text displaying many errors taken from the $ulraw$ sub-corpus which has been extensively proofread and inserted into the $u2pfd$ corpus with one occasion of $IS$ being replace by $S$ as shown in the subsequent paragraph:

\begin{quote}
my brother loes bird and he keeps alot of pigeon on roof of home. he can sale and buy to got same mony;thats goog for a person whose age is 17 only to make trad.I hope for him all the best.In my class I have one student like my brother. \\
\texttt{[ulraw\_171]}
\end{quote}

\begin{quote}
My brother loves birds and he keeps a lot of pigeons on the roof of our house. He buys and sells them to earn some money. It’s good for a person who’s only seventeen to start trading. In my class there is a student who trades birds like my brother. \\
\texttt{[u2pfd\_171]}
\end{quote}

Note that both sub-corpora $j1raw$ and $ulraw$ included texts similar to the paragraphs above and they also contained texts that required no proofreading at all\textsuperscript{13}. The following ESS is a $ulraw$ text which could be considered ‘conceptually creative’, there being two instances similar to this in the $ulraw$ sub-corpus and no texts like this in the present $j1raw$ sub-corpus. Similar to the majority of ESS in both $j1raw$ and $ulraw$, the following $ulraw$ ($& u2pfd$) text has been given an explanatory (extra permissible words) title by the student-author. Here the title is ‘The Stupid Computer’:

\begin{quote}
\begin{center}
\texttt{a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a [ulraw\_887]}
\end{center}
\end{quote}

A $j1raw$ text collected by JAFAE which required little extensive proofreading and mediation now follows together with its proofed equivalent:

\begin{quote}
The morning I woke up there was a sound of pigeons. There was a smell of pancakes, and coffee, which made my stomach growl. There was my sister laughing, and my father staring at the newspaper. Also, there was my dog wagging his tails\textsuperscript{14}. And there, there was my family. \\
\texttt{[j1raw\_284]}
\end{quote}
This morning I woke up there was the sound of pigeons. There was the smell of pancakes, and coffee, which made my stomach growl. There was my sister laughing, and my father staring at the newspaper. Also, there was my dog wagging his tail. And that, that was my family. [j2pfd_234]

These little vignettes of everyday life produced by student-authors using different languages in diverse cultures would be lost without the ESSC and hence ICCE.

V Conclusion

In answer to research question (a) above, the study outlined here suggests that using CA the sub-corpora j1raw, ulraw, j2pfd and u2pfd may be differentiated solely with respect to frequency of orthographic units of which they are comprised. Happily, when considering CA of the 19 most frequent orthographic forms, the greatest contrast is between the corpora compiled in the UAE (ulraw & u2pfd) and the corpora collected in Japan and mediated in the UAE (j1raw and j2pfd). A secondary contrast based upon these 19 most frequent word types is between the raw data compiled into corpora (ulraw and j1raw) and these same corpora mediated by native English-speaking proofreaders in the UAE to output u2pfd and j2pfd. This is very reassuring and provides a negative response to research question (b) which questioned whether the pfd corpora were more similar to each other than to the original sub-corpora from which they originated. Any other arrangement would be inappropriate for the objectives of the ICCE as it would show over-emphasis on the proofreaders’ production detracted from the meaning of the original raw ESS.

In answer to research question (c), using CA to scrutiny frequency differences between the raw language produced by the sub-corpora j1raw and ulraw, and the ‘proofed language’ produced by the sub-corpora j2pfd and u2pfd does not point to the necessary modifications which are required to enable raw texts to be published and made commercially available in the public domain. This is because the CA as introduced here does not provide a delicate enough approach to consider how student-author text may be modified but only suggests a wider perspective examining differences in the frequency of orthographic items related to users of distinct cultures and languages. Proofreaders worked hard to comply with the rigorous 50-word constraint whilst making a wide variety of creative texts coherent as is indicated by the reduced frequency of IS and increased behavior of S in the mediated corpora.

In a similar vein, the number of unique (single) word types produced by the raw corpora compared to the reduced number of unique words in the mediated corpora
provides some indication of intervention by the proofreaders. In the \texttt{j1raw} sub-corpus (26034 total tokens & 3165 types) there are 1614 unique orthographic items appearing which increases to 1649 unique orthographic items appearing in \texttt{j2pdf} (26249 tokens & 3211 types). In contrast, the \texttt{ulraw} sub-corpus (48823 tokens & 5489 types) shows 2923 unique words and this reduces to 2748 single frequency words in \texttt{u2pdf} (48944 tokens & 5326 types). The number of unique words would appear to refer to creativity/diversity of approach and topic, and also idiosyncratic irregularities produced unintentionally. This is an area which warrants further study.

As queried in (d), other more intensive approaches would be useful to investigate the raw and proofread corpora compiled in Japan and the UAE. These might include further investigations of the creative characteristics to explicate varieties as suggested by Fujiwara (2008) and more rigorous examination of grammatical characteristics as outlined by Okaura (2007, 2009) Additional explorations into discourse according to gender/authorship, topic as introduced by Matsubara (2013) and a more delicate consideration of individual texts and their mediation would help serve the objectives of the ESSC and ICCE as would suggestions and involvement from academics throughout the world. In the UAE, presentations are now involving an increasing number of faculty from ZU and other local institutions as well as student-authors/illustrators for ESSC anthologies and also translators (Badariotti et al., 2011; Hassall et al., 2012). To see the most innovative areas of development related to the ESSC and ICCE visit: http://www.jafae.org/essc/ & http://www.zu.ac.ae/facets

References


of Spoken English Interlanguage (LINDSEI) Project


Hassall, P. J. (1996b). Where do we go from here? TEIL: a methodology. World Englishes,


Hassall, P. J. (2004c) World Englishes and English for Global Communications. In Davidson, P. et al. (Eds.). Proceedings of the 9th TESOL Arabia Conference (pp. 224-244). Dubai, UAE.


http://www.zu.ac.ae/elearning/html/schedule.htm


Hassall, P. J. (2006c) International collaboration to promote literature and linguistics via the Extremely Short Story Competition [ESSC]. Asian Englishes, 8(2) 90-95.

Hassall, P. J. (2006d). Building the ICCE in the UAE, South Korea and Japan.
Hassall, P. J. (2006c) International collaboration to promote literature and linguistics
through the compilation of the International Corpus of Creative English [ICCE]
built around the Extremely Short Story Competition [ESSC]. 2nd International
Conference on Language, Education and Diversity (LED 2007). 21-24 November
2007, the University of Waikato, Hamilton, New Zealand.

Hassall, P. J. (2007a) "Emiratia: World English voices of Emirati women". Originally
broadcast on “Talking Books”. Group Interview with I. Abulhoul, L. Park, J. & J.
Macdonald Dubai Eye on 4 November 2006. 20 minutes. 19mb file. Zayed University
Podcast: Download at http://www.zu.ac.ae/podcast/mp3only.html

Hassall, P. J. (2007b) Express Yourself! The Extremely Short Story Competition enters
its third year. Zawaya the official magazine of Zayed University, United Arab
Emirates. <http://www.zu.ac.ae/zawaya/0101/0113.pdf> in English
<http://www.zu.ac.ae/zawaya/0101/0113A.pdf> in Arabic

Hassall, P. J. (2007c) Launch of the Extremely Short Story Competition UAE 2007 on
the UAE Radio Station ‘Dubai Eye’. 31st March 2007 10am-12pm.

Hassall, P. J. (2009). World Englishes as art: the Extremely Short Story Competition
[ESS] in Arabia and the Far East. Presentation for MA students studying at the
Centre for English Language Studies, University of Birmingham, England. 9,
November, 2009. Reported in Birmingham Bite-Size:

Hassall, P. J. (2011). The Extremely Short Story Competition: fostering creativity and
excellence in formal and informal learning contexts in the UAE and internationally.
Learning and Teaching in Higher Education: Gulf Perspectives, 8(2) 1-20.
http://lthe.zu.ac.ae

Hassall, P. J. (2011). Using the Extremely Short Story Competition to enhance creative
language learning. Workshop at the Annual Conference of the European Association
for Quality Language Services (EAQUALS) 8-9 April, National Technical Library,
Prague, Czech Republic.

Center & the Student Literature Club, Zayed University, United Arab Emirates.

University, United Arab Emirates: The Extremely Short Story Development Project.

Hassall, P. J. (Ed.). (2006b) Emiratia: voices of Zayed University. JAFAE’s Official
Guidebook to the Extremely Short Story Competition. The Japanese Association for

Zayed University, United Arab Emirates: The Extremely Short Story Development Project.


http://www.amazon.com/dp/B005S682JM


Traveling Exhibition | 01:00 – 2:30, 15 MARCH Zayed University, Abu Dabi, UAE. 

Women as Global Leaders Conference 2012: Creating a sustainable future for the world.

http://shelter.ae/events/facets-of-emirati-women-a-panel-discussion


Matsubara, N. (2013, forthcoming). The ESSC as a tool for understanding writers’ lives and society: Comparison of young people in the United Arab Emirates and Japan by examining each country’s ESSC stories. Learner corpus studies in Asia and the


http://www.zu.ac.ae/zawaya/0101/0113A.pdf (Arabic).

Notes

1 A form for eliciting a student-author profile and consent was devised, modelled on the profile forms for ICLE (Granger, 1998, 2001) and LINDSEI (De Cock, Granger & Petch-Tyson, 1999).

2 Recent pilot projects (Hassall, Osman, Sellami & Walsh, 2009, 2010) and in-house ESSC at Zayed University have also been carried out bilingually in English and Arabic, hopefully this will be replicated to the advantage of students and other languages in different contexts.

3 Student-authors are permitted to submit stories informally at any time of the day or night and ESS would not necessarily have to be created formally inside the classroom, although this is not proscribed.

4 For three years, from 2005 to 2008, three ESS were published every day in the UAE ‘Khaleej Times’ newspaper.

5 These were particularly significant for Arab parents (mainly Arabic-speaking) to interact with their progenies’ English language medium university. In addition, in an extensive ESSC held in 2008 throughout the Gulf Cooperation Countries (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, UAE, and also Yemen) student-authors were provided special dispensation to cross national borders. For example a number of Yemenis were given special permission to enter the UAE to attend the ESSC Awards Evening at ZU.

6 To date the Facets of Emirati Women International Traveling Exhibition [Facets FITEx] has visited universities in Dundee (Scotland), Birmingham and Preston (England), Paris (France), Milan (Italy), Prague (Czech Republic), Tripolis (Greece), Bucharest (Romania), Kyoto & Nagoya (Japan) and Palmerston North (New Zealand). The Facets FITEx virtual Exhibition is available on the web and can be viewed in English translated into Arabic (soon to be published by Zayed University Press) and English bilingual versions are also available for Greek, French, Italian, Chinese, Burmese: http://www.zu.ac.ae/facets The Facets FITEx Exhibition is also available bilingually as a Japanese-English e-book translated by Yuko Takeshita and Fujimi Sakai Tanaka: http://www.amazon.com/dp/B005S682JM

7 Originally the ESSC_UAE was delivered by email to the two campuses of Zayed University in Dubai and Abu Dhabi (two hours apart by car) which at that time provided tertiary education solely for female students. To extend the reach of the ESSC throughout the UAE (to both male and female tertiary campuses) a website was built which would only accept exactly 50 words. This is documented in Hassall (2005a).

8 For instance, should weak forms such as ‘isn’t’ be considered one or two words? Electronically ‘isn’t’ is considered one word by the Microsoft Word word-count which considers white space rather than punctuation (here apostrophe) for word boundaries. However, the same orthographic form may be considered one or two words if available
for interpretation by adjudicators unless specifically stipulated. Similarly should loose, unattached punctuation (question marks or full stops) be ignored as would be expected interpretively by JAFAE, or should they be considered individual words as they would be if treated electronically in the UAE.

9 In particular, there is a real danger of assistance from family members, friends or teachers who might be intent on supporting success of their student-author associates.

10 Note that without the ‘proofreading’ process it is unlikely that ESSC anthologies authored by Emiratis would have been made available commercially in bookstores in the UAE and hence ‘what the students wanted to mean’ would not have been readily accessible to the general public. This has contributed to development of some cultural understanding in a country where the Emirati population comprises a small enigmatic proportion of the residential population of the UAE. Note also that proceeds from sale of these anthologies go to the Extremely Short Story Competition Development Fund at Zayed University to “encourage the development of literature and literacy throughout the UAE and in needy countries” (Hassall, Ed. 2009: bookjacket).

11 In contrast to an ‘emic’ approach as proposed by Pike (1964) an ‘etic’ approach was adopted as assumed in phonetic and graphitic analysis, where the physical patterns of language are described with a minimum of reference to their functions within the language system.

12 More extensive investigation into this phenomenon will be provided in a later paper.

13 As in note 13 above, more extensive investigation into the extent of mediation of the proofed sub-corpora will be provided in a later paper.

14 Note the plural item ‘tails’ is a non-standard construction produced by a student-author showing considerable proficiency and sensitivity in English. This would appear more likely to occur in the JAFAE sub-corpus, than the UAE sub-corpus, primarily because of the less-frequent use of English conversationally in Japan, compared to the regular use of English both productively and receptively in the UAE, see Hassall (1984, p. 227) which examines student use of English in the UAE, including participants’ use of English as a secret language between their brothers and sisters to keep things secret from their parents and older family members.