# Gender-Specific English Language Use of Malaysian Blog Authors

*Syahrir Mat Ali*
*syahrir.matali@icloud.com*
*UniversitiKebangsaan Malaysia*

*Pramela Krish*
*pramela@ukm.edu.my*
*UniversitiKebangsaan Malaysia*

## ABSTRACT

Gender-based research on the language use in blogs has its roots in the long-standing notion that men and women speak and write differently. This paper reports an empirical study on the use of English in a blog context involving Malaysian blog authors. Specifically, the study aimed to identify gender-specific English use among Malaysian blog authors and determine the differences in the language use. Using an ensemble text analysis approach, Malaysian female blog authors are more inclined towards using more verbs, adverbs and pronouns than their male counterpart, with a significant difference, while the males are inclined towards using more adjectives, nouns, determiners and prepositions/subordinating conjunctions than the females, with a significant difference. There are also differences between females and males in terms of the function words, neologisms/blog words as well as use of tag questions and adverbs initiating sentences. However, there are minimal differences between the females and males in terms of length of sentences and that the use of intensifiers, hedges, empty adjectives and emotions, thus concluding that they are not necessarily gender-specific differences. The findings can serve as useful language markers that can benefit the applied linguistics and particularly gender-based and forensic linguistic research.

**Keywords**: gender differences; English use; blogs; Malaysian blog authors; neologisms

## INTRODUCTION

Since the 1980s, sociolinguists and researchers have looked at the relationship between gender and language use in spoken or oral interactions (e.g., Lakoff, 1975; Cameron, 1998; Holmes & Meyerhoff, 2003). Other studies have also focused their investigation on various aspects of the written use of language such as women's language use (Lakoff, 1973), computational methods of authorship attribution (Koppel, Schler & Argamon, 2009), automatic categorization of written text by author gender (Koppel, Argamon & Shimoni, 2002; Argamon et al., 2009) and the usefulness of function words for authorship attribution (Argamon & Levitan, 2005). While these studies did not specifically focus on blogs, they are highly influential in the field of authorship profiling and most of their findings and methods were heavily cited and replicated in various blog authors' gender research.

Within the increasing use of the Internet world wide, web blogs, which are web sites containing online personal journals where users as authors post, reflect, comment and even use hyperlinks (Merriam-Webster, 2015) have been an integral part of the Internet ever since the late 1990s. Ever since then, language researchers have looked at various angles in order to come up with approaches and techniques that can be used to linguistically profile a blog author based on the most predominant and available information such as the blog text itself. Gender studies in blogs, the focus area of the current paper, have their roots in the long-standing notion that men and women speak and write differently. According to Zaini Amir et al. (2012, p. 105), "in general, society has constructed the belief that men and women act and behave differently to images of masculinity and feminity". This statement is in line with Lakoff's (1973) findings that suggested women speak differently from men. Some empirical studies on blog authorship-profiling focused on age and gender (Schler et al., 2005; Prasath, 2010), while others focused only on gender (Kobayashi, Matsumara & Ishizuka, 2007; Yan & Yan, 2006; Mukherjee & Liu, 2010; Zhang & Zhang, 2010; Zaini Amir et al., 2012).

Although previous gender-based research, in general and particularly on blogs, have enriched our understanding of the different linguistic patterns between males and females, the question that persists now is related to ascertain whether such findings, or rather convictions, will be valid to explain and describe the situation for all English speakers and environments or whether such differences in the language use between males and females identified in previous research will be applicable to ESL learners in other contexts. While bearing in mind that there is an expected distinction between male and female English writers, there is a need for a Malaysian-centric research in order to verify, based on the collected samples, if there is indeed any gender-specific English language use. Therefore, the current study aimed to identify gender-specific English language use among Malaysian blog authors and to determine if such differences are significant and classifiable.

## GENDER-BASED LANGUAGE USE

In the field of sociolinguistics, differences in the language use between males and females represent an interesting research topic of a long history for several studies (Cameron, 1998; Eckert, 1989; Holmes & Meyerhoff, 2003) basing their foundation on the different patterns of oral linguistic behavior. This is not even a new research area in English language and linguistic studies since numerous efforts have been made to establish and understand the differences and similarities between men and women in language use in formal and informal contexts.

As stated by Lakoff (1973, p. 49), "Women's language" shows up in all levels of the grammar of English; and that differences can be found "..in the choice and frequency of lexical items; in the situations in which certain syntactic rules are performed; in intonational and other super segmental pattern". Many previous studies have substantiated this by providing an empirical evidence of such differences. For instance, based on their analysis of a large corpus of 14,324 text files, Newman et al. (2008) found that female English writers/speakers use more pronouns and social words, other psychological process references and verbs compared to male. However, the men apparently exceeded the women in terms of word length, numbers, articles, and prepositions.

Previous studies have also identified differences between males and females in the use of written language. For instance, Koppel, Schler and Argamon (2009) provided a convincing evidence of the male and female differences in writing styles in a corpus of books and articles in English. Argamon et al. (2003) also identified several linguistic properties characterizing gender differences on the use of written language. For instance, the use of more pronouns is an indicative of the written language among females whereas the use of more quantifiers and determiners is an indicative of the males' written language.

## GENDER-BASED LANGUAGE USE IN BLOGS

Online communication via various technological tools including blogs has attracted the interest of many sociolinguists and language researchers to look at whether and how the gender dichotomies in the language use in the traditional or face-to-face context would be identified in the online context. For instance, in examining the written language in instant messaging, Baron (2004) identified differences between males and females that somehow reflect the differences in using the oral language in face-to-face communication. These include the use of longer turns in messages and fewer contractions among females, thus concluding that instant messaging reflects the females' writing style than males' style.

In relation to the language use in blogs, in a study by Schler et al. (2005) on how content and writing style vary between male and female bloggers, based on a corpus of 71,000 English blogs, the researchers found that female blog authors used more pronouns and assent/negation words while males used more articles and prepositions. While females' language emphasized involvedness, males' language emphasized information. In addition, Prasath's (2010) analysis of 95,245 blog.myspace.com blog posts using a rather different linguistic approach than Schler et al. (2005) showed that for blog authors of each gender, there is a clear distinction between usages of a few slangs which seemed to be augmented with other features such as content words. The same researchers also argued for further analysis of non-dictionary words in blogs.

With respect to the extraction of gender-related words, Kobayashi, Matsumara and Ishizuka (2007) found that female authors tend to use family-related words more than male authors, which is similar to the findings by Schler et al. (2005). Their results also showed that there was no significant difference between males and females in using verbs. In using the part-of-speech (POS) analysis approach as well as other features such as f-measure, stylistic features, gender preferential features, factor analysis and word classes, Mukherjee and Liu (2010) indeed showed a significant increase of gender classification accuracy compared to other similar gender estimation/prediction approaches or systems. While that may be true, the accuracy of their findings was uniquely relative to their own corpus. By analyzing the language use in both asynchronous and synchronous discussions in forums, Herring (2003) found that males posted longer messages that strongly assert their opinions, show their use of crude language including insults and manifest their adversarial orientations towards females. However, females tended to post shorter messages showing how they attempted to justify and qualify their assertions, express apologies and social support, thus manifesting their aligned orientations towards males.

In an attempt to predict or identify the gender from blog posts among authors from various backgrounds, Zhang and Zhang (2010) focused on features such as (i) Words, (ii) Average word/sentence length, (iii) Part-of-speech (POS) tags and (iv) Word factor analysis. Their results obtained from analyzing 3226 blog posts showed that it was hard to determine what would constitute a sentence as some of the blog posts do not even have punctuations. The researchers also emphasized the importance of dimension reduction or feature selection since there are lots of noises among the features that may harm the classification process including stop words. In spite of this, Zhang and Zhang (2010) retained the frequency counting of NN (noun) and VB (verb) in their results. This is since NN and VB are perhaps almost just as common as the stop words which appear in any form of utterances, long or short. This indicates that gender analysis of the language use on the Internet, including blogs, is more complicated in the sense that it combines the features of both oral and written forms of the language. In other words, in spite of the characteristics of the written language online, there are features of the language in posts and comments that resemble those in speech or oral language (Baron, 2000; Crystal, 2001).

In the Malaysian ESL context, the work by Zaini Amir et al. (2012) on gender differences in the language use of Malaysian teen bloggers seems to be the only local work available at the time of the current research. Through a qualitative approach to analyzing the language use of 2 female and 2 male Malaysian teenage blog authors, five categories of language features that were notable from the blogs, namely, the intensifiers, hedges, tag questions, empty adjectives and adverbs were identified. Their study showed that that the female blog authors used intensifiers, hedges, tag questions and empty adjectives more than men – and that the use of adverbs is not gender-specific. Although their findings were based on only four sample blogs with an unequal frequency of blog posts, we took note of the language features and later incorporated them into our research, specifically for the data analysis. This calls the need for further analysis of the gender differences in the language use among Malaysian blog authors.

## THE CURRENT STUDY

The current study aimed to identify gender-specific English language use among Malaysian blog authors and to determine whether there are significant differences in language use between the males and females. The study used a mixed approach to data collection which is a combination of both quantitative and qualitative approaches. Although the research was primarily based on the use of POS analysis in order to identify the gender-specific English use among the selected Malaysian blog authors, the analysis over the acquired data was both quantitative and qualitative in nature. Part-of-speech or POS is a traditional class of words distinguished according to the kind of idea denoted and the function performed in a sentence. The POS is represented with notations or tags such as, but not limited to, NN for noun, JJ for adjective, RB for adverb, DT for determiner, UH for interjection and many more. The POS analysis in this research was based on Beatrice Santorini's (1990) POS Tagging Guidelines for The Penn Treebank Project.

The reason behind the selection of POS analysis for this study was due to the fact that, firstly, it is an approach or one of the approaches taken by most researches that have been conducted in this area. In their work, Mukherjee and Liu (2010) combined various methods that also included POS analysis as one of the features. Newman et al. (2008) in their seminal effort also focused on the occurrences of function words, social words, pronouns, verbs, articles and prepositions can be best replicated through a POS analysis. Secondly, there has not been a research on Malaysian blogs, especially on the English language use between the two genders that are primarily led by a POS analysis approach. Although, Zaini Amir et al. (2012), have studied the differences in language use by female and male Malaysian teenage blog authors based on the frequency of words, their work was based on the use of a word counting tool over a pre-defined and limited categories of 8 intensifiers, 27 hedges, tag questions, empty adjectives and adverbs; which is almost similar to the objectives to be gained from a POS analysis – i.e. among other things is to count the frequency and mean averages of occurrences of certain POS tags.

## DATA COLLECTION AND MALAYSIAN BLOGS

The sample of the research population was represented by a total of 60 body of blog text collection written by 60 Malaysian blog authors who primarily write in English; of which the number of female and male authors was divided evenly (i.e. 30 female and 30 male Malaysian blog authors). The blog text sampled from each of the authors amounted to about 2000 words per blog for a total of almost 120,000 words.

Data collection was made via the *Blogged. my* and *Project Petaling Street* websites. Both are blog aggregator websites that list the most recent blog posts by participating Malaysian blog authors. The data collection procedure was carried out following three stages: (i) Blog selection, (ii) Blog text selection and formatting and (iii) Blog text classification. This ended with classifying each of the formatted body of blog text (for each blog author) according to a naming convention based on gender, count and number of words as per depicted in Table 1.

TABLE 1. Examples of classified blog author files

| Female blog classification | Male blog classification |
|---|---|
| f12.2060 | m4.2048 |
| f13.2028 | m5.2176 |
| f14.2133 | m6.2143 |

## DATA ANALYSIS

The analysis was performed on the classified blog author files that were gathered during the data collection phase. It was initiated by blog text processing in which the authors' files were fed to a POS tagging application and developed using the Python programming language. This was followed by converting the file output of the blog text processing into an MS Excel format using the common .xls file format. The third step was concerned with cleaning up the output manually. This involved characters and symbols from the text that are of little use for a POS analysis, such as "-", "(" and "*";numbers in numerical forms such as "4.5", "100" and "5"; and foreign words that are confirmed and verified to be non-English. After this, the final clean text was analyzed using the MS Excel application to take advantage of its wide-ranging arithmetical and value sorting properties. The final step was the advanced text analysis which was performed adopting the following approaches:

i. Newman et al.'s (2008) approach in analyzing the use of function words such as articles, conjunctions, auxiliary verbs and other grammatical words that have little lexical meaning and only serve to express grammatical relationships in a sentence
ii. Schler et al.'s (2005) approach in analyzing the use of blog words or neologisms;
iii. Zhang and Zhang's (2010) analysis of average word length and average sentence length in blog text written by male and female blog authors;
iv. Zaini Amir et al.'s (2012) analysis on the use of selected intensifiers, hedges, tag questions, empty adjectives and adverbs which is based on Lakoff's (1975) theory of language use among women;
v. Words that are most frequently used by either male or female blog authors;
vi. Words that are exclusively used by only either male or female blog authors; and
vii. Yan and Yan's (2006) analysis on the use of emoticons by blog authors.

## FINDINGS

The findings are presented in accordance with the two types of analyses used: first, the results of the POS tags i.e. JJ, NN,VB, RB, PRP, DT and IN obtained through the text analysis, and second, the results of other aspects of language use such as the use of function words; the use of blog words and neologisms, average word length and average sentence length, the use of selected intensifiers, hedges, tag questions, empty adjectives and adverbs, words that are most frequently used by either female or male blog authors; words that are exclusively used by only either female or male blog authors; and the use of emoticons by the blog authors obtained through the advanced text analysis.

### FREQUENCY DISTRIBUTION AND AVERAGE OF POS

In terms of Frequency Distribution, the results showed that there are some distinguishable differences in terms of POS tag occurrences between the females and males. There are also POS tag occurrences with minor variations. The females' text corpus is based on 59,347 POS tag words, while the males' text corpus is based on 61,684 POS tag words. The analysis generated the following summarized results as shown in Table 2 and Table 3. Table 4 sums up which POS tag occurred more in texts written by which gender.

TABLE 2. POS tag analysis: female blog texts

| Occ | JJ | NN | VB | RB | PRP | DT | IN |
|---|---|---|---|---|---|---|---|
| Total | 4056 | 14035 | 11331 | 4424 | 7315 | 4976 | 6551 |
| Avg | 135.2 | 467.8 | 377.7 | 147.46 | 243.83 | 165.86 | 218.36 |
| % | 6.83 | 23.64 | 19.09 | 7.45 | 12.32 | 8.38 | 11.03 |

Table 3: POS tag analysis: male blog texts

| Occ | JJ | NN | VB | RB | PRP | DT | IN |
|---|---|---|---|---|---|---|---|
| Total | 4473 | 14799 | 11635 | 4120 | 5947 | 6230 | 7451 |
| Avg | 149.1 | 493.3 | 387.83 | 137.3 | 198.23 | 207.6 | 248.36 |
| % | 7.25 | 23.9 | 18.86 | 6.67 | 9.64 | 10.09 | 12.07 |

TABLE 4. POS Tag frequency summary

| Gender | JJ | NN | VB | RB | PRP | DT | IN |
|---|---|---|---|---|---|---|---|
| Female |  |  | ✔ | ✔ | ✔ |  |  |
| Male | ✔ | ✔ |  |  |  | ✔ | ✔ |

Based on the results of the frequency distribution and average analysis, it is safe to assume that Malaysian female blog authors are more inclined towards using more verbs, adverbs and pronouns than their male counterpart, with a significant difference especially on the usage of pronouns which is rather noticeable at a 2.67% difference. The difference rates recorded were 2.75% and 1.55%, respectively. On the other hand, the Malaysian male blog authors are inclined towards using more adjectives, nouns, determiners and prepositions/subordinating conjunctions than the females, with a significant difference on the usage of the latter two i.e. determiners and prepositions/subordinating conjunctions.

### USE OF FUNCTION WORDS

The analysis of the use of function words was made over the 60 body of blog text, amounting to a total of about 120,000 words and representing 30 female and 30 male Malaysian blog authors. Table 5 shows the use of function words in terms of their occurrences in both texts by females as well as males.

TABLE 5. Occurrences of Function Words

| Gender | Occurrences | Total Words |
|---|---|---|
| Female Text | 23301 | 59347 |
| Male Text | 24240 | 61684 |

The findings show that the Malaysian male blog authors use more function words than the female blog authors with a difference of 3.95%. Interestingly, this can be clearly seen when the 3.95% difference is compared to the POS tag analysis findings earlier, none of the variance goes beyond 3%.

## USE OF BLOG WORDS/NEOLOGISMS

The analysis of the use of blog words or neologisms in blog text written by the Malaysian female and male blog authors has shown that Malaysian female blog authors use more blog words or neologisms. Words such as *wtf* (what the fuck), *lol* (laughing out loud), *fren* (friend) and *omg* (oh my god) were identified in their blogs and surpassed the male blog authors by 49 occurrences or 54.44% (see Table 6).

TABLE 6. Use of blog words or neologisms

| Blog Words | Meaning | Female | Male |
|---|---|---|---|
| wtf | *what the fuck?* | 2 | 1 |
| btw | *by the way* | 1 | - |
| lol | *laughing out loud* | 1 | 2 |
| haha | | 41 | 24 |
| hehe | | 32 | 7 |
| huhu | | 3 | - |
| muahaha | | - | 3 |
| fren | *friend* | 6 | 3 |
| friend-nemy | *friend-enemy* | 1 | - |
| omg | *oh my god* | 2 | 1 |
| fugly | *fucking ugly* | 1 | - |
| Total Occurrences | | 90 | 41 |

This finding is actually similar to the findings made by Schler et al. (2005) in which they have also found that female blog authors use blog words more than male blog authors, although they recorded a difference of 28.3%.

## AVERAGE LENGTH OF WORDS AND SENTENCES

The analysis of the average number of words and average sentence length depicted in Table 7 that the Malaysian male blog authors write sentences that are, on average, longer than those written by the female blog authors by 1.41 words or 11.5%. On the other hand, the average length of word between the two types of texts is very minimal, and therefore negligible. Given the low variance between the average length of the sentence and word of Malaysian female and male blog texts, it can be deduced that both elements have little contribution in identifying gender-specific English language use.

TABLE 7. Average length of words and sentences

| | Female | Male |
|---|---|---|
| No. of sentences | 5545 | 4522 |
| Average length of word | 4.27 characters | 4.47 characters |
| Average length of sentence | 11.51 words | 12.92 words |

This finding is in accordance with the findings made by Zhang and Zhang (2010) in which they have found that male blog authors write slightly longer sentences; and of almost the same average in terms of length of words. It is also important to highlight here that Zhang and Zhang (2010) used a blog data set that is not confined to any particular nation and nationality of blog authors; while this research is focused on blog texts that have been verified as written by Malaysian blog authors. Yet, both researches have managed to arrive at an almost similar conclusion. This may indicate that, universally, there is not much difference regarding the length of sentences and words written by either gender when it comes to blog writing.

## USE OF SELECTED INTENSIFIERS, HEDGES, TAG QUESTIONS, EMPTY ADJECTIVES AND ADVERBS AT THE BEGINNING OF A SENTENCE

The current study also focused on the occurrences and percentages of the intensifiers, hedges, tag questions, empty adjectives and adverbs used at the beginning of a sentence by the Malaysian females' and males' blog texts. Generally, the analysis indicates that the use of intensifiers, hedges and empty adjectives are not necessarily gender-specific. Meanwhile, the use of tag questions among the Malaysian female blog authors scored 66.67% more than their male counterpart.

Moreover, another difference was found in the use of adverbs at the beginning of a sentence since the number of such adverbs initiating a sentence posted by the males was higher 46.5% than the number of the adverbs posted by the female blog authors. Indeed, this showed a significant variance of occurrences and thus can be regarded as gender-specific and attributed to the gender of the blog authors.

### USE OF SELECTED INTENSIFIERS

The analysis on the use of selected intensifiers has returned the following results as shown in Table 8 below:

TABLE 8. Use of intensifiers

| Gender | No. of Intensifiers | % in Text |
|---|---|---|
| Female | 778 | 1.3% |
| Male | 633 | 1% |

The above results showed that Malaysian female blog authors have been using more intensifier words than their male counterpart slightly by 0.3%. Given the very small amount of variance, it appears that the use of intensifiers is not necessarily gender-specific and therefore its number of occurrences cannot be attributed to the gender of the blog authors.

### USE OF LEXICAL HEDGES

The analysis on the use of lexical hedges has given the following results:

TABLE 9. Use of lexical hedges

| Gender | No. of Hedges | % in Text |
|---|---|---|
| Female | 274 | 0.5% |
| Male | 241 | 0.4% |

The results (Table 9) showed that Malaysian female blog authors have been using more lexical hedges than their male counterpart slightly by 0.1%. Again, given the very small amount of difference between the female and male blog texts, it appears that the use of lexical hedges is not necessarily gender-specific and therefore its number of occurrences cannot be attributed to the gender of the blog authors.

### USE OF TAG QUESTIONS

The analysis on the use of tag questions has generated the following results:

TABLE 10. Use of tag questions

| Tag Questions | Female | Male |
|---|---|---|
| isn't' it? | 2 | - |
| did you? | 1 | - |
| aren't I? | 1 | - |
| aren't you? | 1 | - |
| are they? | - | 1 |
| would you? | 1 | - |
| right? | 18 | 7 |
| huh? | - | 3 |
| okay? | 1 | - |
| ok? | 8 | - |

The results (Table 10) showed that Malaysian female blog authors have been using 66.67% more of the tag questions in their blog writings; or specifically a total of 33 tag questions compared to 11 used by Malaysian male blog authors. Given the significant difference in the use of tag questions between the female and male blog texts, it appears that the use of tag questions is gender-specific and therefore its number of occurrences can be attributed to the gender of the blog authors.

### USE OF EMPTY ADJECTIVES

The analysis on the occurrences of empty adjectives has returned a very low count; and the results are described in Table 11 below.

TABLE 11. Use of empty adjectives

| Gender | No. of Empty Adjectives | Type. of Empty Adjectives |
|---|---|---|
| Female | 2 | *sweet*, *brilliant* |
| Male | 2 | *wonderful*, *nice* |

Apparently, both gender have only used two instances of empty adjectives, each. The researcher has taken careful consideration to go over all the initial results to verify if indeed they are empty adjectives or regular adjectives that were attached to nouns or adverbs; and has only found the above four occurrences of empty adjectives. This would prove that Malaysian blog authors, of either gender, do not commonly use empty adjectives in their blog writings; and therefore its use or occurrences cannot be considered as gender-specific.

Some example uses of the empty adjectives gathered from the blogs are as follows:

Malaysian female blog texts:
a. *Sweet.*
b. *Just brilliant!*

Malaysian male blog texts:
a. *Wonderful, now that I've already bought the box, I can go back to the earlier shop for the present.*
b. *this is what i mean.. foreign style. Nice~*

### USE OF ADVERBS AT THE BEGINNING OF A SENTENCE

The analysis on the occurrences of adverbs at the beginning of a sentenced has returned a low count yet distinguishable between the two genders; and the results are described in Table 12.

TABLE 12. Use of adverbs at the beginning of a sentence

| Gender | No. of Adverbs |
|---|---|
| Female | 31 |
| Male | 58 |

The above results showed that Malaysian male blog authors have been using more adverbs at the beginning of their blog sentences than their female counterpart by 46.5%. And given the significant variance of occurrences, it appears that the use of adverbs at the beginning of a blog sentence can be attributed to the gender of the blog authors.

### MOST FREQUENTLY USED WORDS

The analysis yielded 100 most frequently used words in the blogs written by the Malaysian female and male blog authors and the top 10 most frequent words used by female or male blog authors. The results are aligned with some of the findings of the POS frequency distribution and average analysis earlier. It was found that the Malaysian female blog authors use more pronoun words when compared to the male blog authors. This can seen by occurrences of the pronoun "I" as the second most frequently used word, next only to the ubiquitous "the" – which is well expected. Yet, in terms of variety, it seems that the Malaysian male blog authors tend to use more types of pronouns compared to the females. This would show that although Malaysian female blog authors may be using the most number of pronouns in terms of frequency count, their counterpart, on the other hand, are using more types of pronoun words in the construction of their body of blog texts.

### WORDS UNIQUELY USED BY FEMALE AND MALE MALAYSIAN BLOG AUTHORS

This analysis resulted into identifying two sets of words – each listing unique word occurrences for the Malaysian female and male blog texts amounting to 4670 words and 5142 words respectively, and along with their POS tag classes that are exclusively used by only either Malaysian female or male blog authors. It was found that the differing elements are dominated by nouns (NN), verbs (VB), adjectives (JJ), adverbs (RB) and numbers – in such order of frequency count.

In order to gain a meaningful analysis in distinguishing the differing aspects of the unique words used by either gender, the researchers opted to omit items classified under the noun (NN) POS tag and numbers. This had to be performed since the occurrences of nouns (NN) can give only a minor value to the analysis in that most of the tagged items are made up proper names (which are nouns as well) such as names of individuals, nicknames and institutions while the occurrences of the unique numbers were also ignored for obvious reasons.

TABLE 13. Occurrences of Unique Words by Gender

| Gender | Adjective (JJ) | Adverb (RB) | Verb (VB) |
|---|---|---|---|
| Female Text | 232 | 74 | 470 |
| Male Text | 316 | 120 | 726 |

The results in Table 13 showed that Malaysian male blog authors used more unique adjectives, adverbs and verbs than their female counterpart. This may suggest that Malaysian male blog authors possess a richer vocabulary in blog writing particularly with respect to the use of adjectives, adverbs and verbs. The staggering differences are indeed very useful, and collectively, they can be grouped together into a meaningful unit and possibly used as a language marker that can assist in classifying the author's gender of an unknown blog text.

Some examples of the words (under the classes of adjectives, adverbs and verbs) that were exclusively used by either Malaysian female or male blog authors are depicted in Table 14.

TABLE 14. Sample of unique words by gender and POS tag

| Gender | Adjectives (JJ) | Adverbs (RB) | Verbs (VB) |
|---|---|---|---|
| Female | adorable, affordable, alphabetical, anti-aging, blemish-free, chaotic, chronic, chronological, eccentric, enjoyable, fairytale-like, flower-looking, gorgeous, half-hearted, hysterical, idiotic, ill-willed, laughable, over-the-counter, plenary | awesomely, awfully, blatantly, deliberately, diligently, exponentially, frantically, interestingly, memorably, occasionally, philosophically, sarcastically, sequentially, spontaneously, stylishly, sufficiently, tremendously, truthfully, unconditionally, violently | bragged, captivated, cherished, disgusted, dodged, flabbergasted, kissed, rendered, reproduced, puked, stumbled, tanned, texted, gushing, craving, depressing, dieting, showering, snuggling, hugged |
| Male | accountable, anal, analgesic, apathetic, arduous, asexual, auspicious, colloquial, blue-eyed, consensual, conscientious, diplomatic, distasteful, customizable, fatal, filthy, illogical, immoral, imperative, insidious | abominably, abruptly, consistently, conveniently, defiantly, diagonally, exceptionally, horizontally, metaphorically, invariably, ironically, judicially, ruthlessly, swiftly, terribly, undoubtedly, unexpectedly, unintentionally, ostensibly, painstakingly | alleged, bludgeoned, challenged, colonized, conspired, convicted, contested, destroyed, entrenched, diverted, doctored, flanked, fielded, forfeited, frequented, fucked, installed, mobilized, tendered, engulfing |

### BLOG AUTHORS' USE OF EMOTICONS

The analysis of the use of emoticons by the Malaysian female and male blog authors obtained surprisingly minimal results with noticeable differences in terms of frequency count and the types of emoticons being used. Based on the surprisingly low count of findings, it appears that each gender used only two types of emoticons. In terms of the frequency, the total count of 20 emoticon occurrences is rather negligible when compared to the approximately 120,000 words in the corpus of texts of both Malaysian female and blog authors. However, statistically, the results still show that the Malaysian male blog authors use more emoticons than the females by 60%. What is even more interesting is that 92.3% of emoticons used by Malaysian male blog authors are" :) " which represent or stand for *smiling*. This shows that Malaysian male blog authors express positive emotions more than their female counterpart. Also, it was found that all of the emoticons used by the females are either the " T.T " or " T_T ", which both represent crying.

## DISCUSSION

The results of the present study uncovered several interesting differences in the use of English among Malaysian male and female blog authors. First, concerning the use of POS as identified and tabulated in this study, the results indicated that the use of more verbs, adverbs and pronouns is an indicative of the Malaysian female blog authors' written language. On the other hand, the use of more adjectives, nouns, determiners and prepositions/subordinating conjunctions is an indicative of the Malaysian male blog authors' written language. Such results corroborate the results of some previous studies showing that there were differences in the frequency or occurrences of POS components in the language use between males and females. The Malaysian females' use of more verbs and pronouns in this study supports what Newman et al. (2008) found among the female speakers of English and what Argamon et al. (2003) identified in females' written language that was dominated by the use of pronouns. This is also somehow in line with the results of some studies in terms of the use of more determiners (Argamon et al., 2003; Schler et al., 2005) that characterizes males' written language online or in blogs. This finding is particularly interesting since it also reflects those found by Zhang and Zhang (2010) that recorded a difference rate between females and males. Yet, this contradicts the results obtained by Kobayashi, Matsumara and Ishizuka(2007) that led them to believe that there is no significant difference between males and females in using verbs. This could imply that this variance largely depends on the differences represented by the cultural context as well as the discourse.

It is also interesting to find that there are differences in the use of function words in blogs between females and males in the Malaysian context. This particular result is in accordance to the suggestion made by Newman et al. (2008) who claimed that the largest differences between males' and females' language use would be in the use of function words. In addition, in this study, the Malaysian female blog authors were found to be using more blog words in their written language than the male counterpart, a result that substantiated the findings made by Schler et al. (2005), who found that female blog authors use blog words more than male blog authors by 28.3%.

Concerning the average number of words and average sentence length, the results of the present study showed that such differences were just minimal, thus indicating that the average number of words and average sentence length did not contribute much to our understanding of the gender-based language use in blogs. Such result stands contradictory to the results reported by Newman et al. (2008) concerning the males' higher average number of words than females and those by Baron (2004) regarding the females' longer turns or sentences as well as those by Herring (2003) showing that the males produce longer messages. Based on our findings, such differences were minimal. It is also important to note here that for a similar analysis, Zhang and Zhang (2010) used a blog data set that is not confined to any particular nation and nationality of blog authors; while this research is focused to blog texts that have been verified as written by Malaysians. Yet, both researches arrived at an almost similar conclusion. This may indicate that perhaps, universally, there is not much difference of significant importance regarding the length of sentences and words written by either gender in blog writing.

In this study, the analysis replicated Zaini Amir et al. (2012) analysis of the use of selected intensifiers, hedges, tag questions, empty adjectives and adverbs at the beginning of a sentence (which in turn was based on Lakoff's (1975) theory of language use among women). The results of our study do not completely replicate the results that they have obtained. This is because while our study showed no differences in terms of intensifiers and hedges between the females and males, they on the other hand found that the most noted differences between the females and males were the use of intensifiers and lexical hedges. We believe that the difference could be attributed to the situations in which the language was used. Another difference is that while this study showed that only the males used a higher number of adverbs as starters of sentences, Zaini Amir et al. (2012) found that the females' adverbs initiating sentences outnumbered the males' adverbs with an insignificant difference. The result of this study related to the females' more frequently used tag questions confirms the result obtained by them. In this regards, the higher number of females' tag questions can explain how the female blog authors seemed to be closer to their audience or readers.

The results of the present study also showed the use of written language among the Malaysian male blog authors was distinguished from the language used by the females by unique words in the form of adjectives, adverbs and verbs. However, the use of emotions did not contribute to our understanding of the female-male distinction in terms of their written language in blogs. What is interesting about this is that the males' emotional signs tended to be more positive than those of the females. This finding confirms those obtained by Barrett and Lally (1999) showing that males' online messages were more emotionally oriented than those of females.

## CONCLUSION

This research has discovered a number of significant and distinguishable differences in English language use between female and male Malaysian blog authors that can benefit the applied linguistics field of study, particularly in the area of forensic linguistics. Some of the gender-specific English language uses, such as blog words and neologisms, tag questions, adverbs at the beginning of a sentence and the use of certain adjectives, verbs and adverbs; can manifest further meaningful insights into understanding in what way does a female and male Malaysian write differently when blogging in English.

Another important insight that we have gained from the findings is the fact that male Malaysian blog authors have used more unique adjectives, adverbs and verbs than their female counterpart – which means that these adjectives, adverbs and verbs (inflected form or otherwise) are not being used by the female blog authors in their writings. This may also suggest that they are uniquely tied to the male gender.

While this may also suggest that Malaysian male blog authors possess a richer vocabulary, it is not the most intriguing aspect of the finding. The fact that the analysis has managed to establish the existence of adjectives, adverbs and verbs that seems to be used exclusively by either gender is itself very essential. This proves that there is a possibility to group together a number of adjectives, adverbs and verbs to serve as a language marker and later attributed to represent either gender.

The research has also learned that some of the other gender-specific English language uses, such as the occurrences of Part-of-Speech (POS) tags and the use of function words may not be practical and reliable to primarily distinguish the author's gender of a given known text, let alone of an anonymous one. This is attributed to the fact that the different percentages of their use by both genders are rather small.

All things considered, we believed that our study has gathered various results which facilitated our understanding of the use of written language in blogs from the gender perspective. This was made possible through the application of diverse text analysis approaches and not limiting the analysis scope to any particular angle. The huge size of the data set has also contributed greatly in ensuring that the results of the data analysis are based on a relatively large sample data and therefore, increases the reliability of the knowledge and information gained.

The findings of this study indicate that there are differences in the language use between the Malaysian female and male blog authors, but it has also raises numerous other questions, for example on the practicality of leveraging the differences in real world applications. Similar studies can perhaps be carried out with different focuses and parameters such as by increasing the size of the corpus, increasing the number of respondents (i.e. the text authors), segmenting the age group of the text authors or even on taking the language use differences to the task of predicting the author's gender of a given text. Other studies that are concerned with texts from different domains such as social networking sites, emails and chat applications can also experiment with, re-use or even improve the research methodology of this study.

## REFERENCES

Argamon, S. & Levitan, S. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the 2005 ACH/ALLC Conference*.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts.*Text*. 23, 321-346.

Argamon, S., Koppel, M., Pennebaker, J. W. & Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. *Communication of the ACM*. *52*(2), 119-123.

Baron, N. S. (2004). See you Online Gender Issues in College Student use of Instant Messaging. *Journal of Language and Social Psychology*. *23*(4), 397-423.

Barrett, E., & Lally, V. (1999). Gender Differences in an Online Learning Environment.*Journal of Computer Assisted Learning*. *15*(1), 48-60.

Cameron, D. (1997). Performing Gender Identity: Young Men's Talk and the Construction of Heterosexual Masculinity. In Sally Johnson and Ulrike Meinh of (Eds.). *Language and Masculinity* (pp. 173-187). Oxford, U.K.: Blackwell.

Crystal, D. (2001). *Language and the Internet*. Cambridge, UK: Cambridge University Press.

Eckert, P. (1989). *Jocks and Burnouts: Social Categories and Identities in High School*. New York: Teachers College Press.

Herring, S. (2003). Gender and Power in On-line Communication. In J. Holmes & M. Meyerhoff (Eds.). *The Handbook of Language and Gender* (pp. 202-228). Malden, MA: Blackwell.

Holmes, J., &Meyerhoff, M. (2003). *The Handbook of Language and Gender*. Oxford: Blackwell.

Kobayashi,D., Matsumara, N. & Ishizuka, M. (2007). Automatic Estimation of Bloggers' Gender.International conference on weblogs and social media. Colorado, USA, 27 March.

Koppel M., Argamon S. &Shimoni A.R. (2002). Automatically Catagorizing Written Texts by Author Gender. *Literary and Linguistic Computing*. *17*(4), 401-412.

Koppel, M., Schler, J. &Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology. 60*(1), 9-26.

Lakoff, R. (1973). Language and Woman's Place. *Language in Society.2*(1), 45-80.

Lakoff, T. R. (1975). *Languages and Woman's Place*. New York: Harper & Row.

Merriam-Webster (2015). Internet user. Retrieve July 1st, 2015 from http://www.internetlivestats.com/internet-users/

Mukherjee, A. & Liu, B. (2010). Improving Gender Classification of Blog Authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 207-217.

Newman, M. L., Groom, C. J., Handelman, L. D. &Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples Discourse Processes. *A Multidisciplinary Journal. 45*(3), 211-236.

Prasath, R. R. (2010). Learning Age and Gender Using Co-Occurrence of Non-Dictionary Words from Stylistic Variations. *Rough sets and current trends in computing - 7th International Conference,* pp. 544-550.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). Technical Reports (CIS). Department of Computer &Information Science, University of Pennsylvania.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. (2005). *Effects of Age and Gender on Blogging.* American Association for Artificial Intelligence.

Yan, X. & Yan, L. (2006). Gender Classification of Weblog Authors. AAAI Spring Symposium Series. Stanford University, Stanford, California, March 27-29.

Zhang, C. & Zhang, P. (2010). Predicting Gender from Blog Posts. University of Massachussetts Amherst, USA.

Zaini Amir, Hazirah Abidin, Saadiyah Darus & Kemboja Ismail. (2012). Gender Differences in the Language Use of Malaysian Teen Bloggers. *GEMA Online® Journal of Language Studies. 12*(1), 105-124.

## ABOUT THE AUTHORS

Syahrir Mat Ali is a language consultant specializing in localization and quality assurance and currently providing services for major technology companies. He earned his B.Sc. in Information Studies (Hons) from the Universiti Teknologi MARA and his MA in English Language Studies (Applied Linguistics) from the Universiti Kebangsaan Malaysia.

Pramela Krish (PhD) is an Associate Professor at the School of Language Studies and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia. She specializes in online language learning, qualitative social research, educational technology and language education.