1) Introduction

It is believed that people converse similarly to those whom they interact with. They speak like how they are spoken to. In this report, we investigate this concept by utilising R to generate graphs, tables etc. to further confirm the theory. We were given the dataset of a real online forum and we are to find out if members communicate with one another using similar patterns of language. We must also investigate the change in language used over time.

Throughout the report, we will be considering three main objectives that will support the accuracy of the theory. First, we will observe if members' use of language change over time. Next, we will analyse members' behaviours in different types of threads.

2) Preliminary Analysis

a) Dataset Description:

We have thoroughly studied the metadata and linguistic summary of the online forum which consists of 20,000 posts. Among the 20,000 posts, there are upwards of 2,300 active users whereby all anonymous users are considered as one user. The AuthorID of every single anonymous user is the same: -1. Between January 16, 2002 and December 31, 2011, there are 260 threads created with a wide range of topics.

b) Data Cleaning Process:

Before getting into the actual data cleaning, we first observe the values in the dataset and identify the ones that seem abnormal. In this case, it would be the anonymous users (AuthorID's that are -1) and posts with a 0 word count (WC).

We decided to remove all posts with a word count of 0. The reason there exists posts with a word count of 0 is because what the users posted were pictures, videos or emojis; those do not contribute to the word count. Since the word count is 0, all other data fields will also be 0. Therefore, we do not include such data in our analysis because we are more concern on the language and words being used. This will require information on the types of words used, thus, posts without words will not be of use in our analysis.

As for anonymous users, we will not be excluding them from our analysis. Even though they are anonymous, their posts still consist of words which will be contributing in values of the variables of our interest. Their posts might be the identifier of the of language used in a particular thread. If we were to remove those anonymous entries, there might be a bias in our data selection. The users that chose to remain anonymous have their reasons of doing so. They might be introverts that want to express themselves or just people who value their privacies. We believe that our dataset should include all types of users to better represent the true population.

Columns that could be omitted:

i. WC Column

A multiple linear regression model was fitted with the lm() function, using the 4 summary language variables (analytical thinking, clout, authenticity and emotional tone) to model the word count. The output is shown in Figure 1.1. By looking at the multiple R-squared value, ***it appears that approximately only 3% of the variation in word count can be explained by the variation of the 4 summary language variables***. This suggest that there might be no relationship between word count and all those 4 variables. This fact can be supported by the low correlation values (all less than 0.2) of WC with the 4 variables shown in Figure 1.3.

A multiple linear regression model was fitted again using all variables from column 7 to 23 to model the word count. The output is shown in Figure 1.2. ***It appears that approximately only 5% of the variation in word count can be explained by the variations in all of the variables from column 7 to 23 including the 4 summary variable.*** This indicates that the WC variable is independent and is not affected by any other predictor.

Before carrying out the multiple linear regression models, we hypothesised that posts with a higher number of word count will have a higher proportion of authenticity. We thought that the longer the post, the higher the proportion of analytical thinking there will be. After carrying out the tests, we concluded that the word count variable does not play an important role in our analysis. Hence, we will not be taking the word count variable into much consideration throughout our analysis.

ii. Time Column

We will also be omitting the time variable because we have decided to aggregate our data by months. We concluded that aggregating our data by time would be too strenuous and unnecessary, it would be an overkill. It is not necessary because the volume of posts is not sufficient.

```
> fitWC = lm(webWC$WC ~ ., data = webWC[7:10])
> summary(fitWC)

Call:
lm(formula = webWC$WC ~ ., data = webWC[7:10])

Residuals:
   Min     1Q Median     3Q    Max
-163.0  -71.9  -33.1   23.8 6457.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.54268    4.48525  17.065  < 2e-16 ***
Analytic     0.26042    0.03965   6.568 5.23e-11 ***
Clout        0.65455    0.04558  14.362  < 2e-16 ***
Authentic   -0.08511    0.03739  -2.276   0.0229 *
Tone        -0.44276    0.03038 -14.576  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 150.3 on 19917 degrees of freedom
Multiple R-squared:  0.02924,   Adjusted R-squared:  0.02904
F-statistic:   150 on 4 and 19917 DF,  p-value: < 2.2e-16
```

Figure 1.1 R output from fitting a multiple linear model to Word Count (WC)

```
Call:
lm(formula = webWC$WC ~ ., data = webWC[7:32])

Residuals:
   Min     1Q Median     3Q    Max
-373.6  -67.3  -30.7   21.6 6434.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.50620    5.68385  14.164  < 2e-16 ***
Analytic      0.34874    0.04898   7.121 1.11e-12 ***
Clout         0.38502    0.07909   4.868 1.14e-06 ***
Authentic    -0.07294    0.04046  -1.803  0.07144 .
Tone         -0.33348    0.04341  -7.682 1.64e-14 ***
ppron      -464.94362  191.27357  -2.431  0.01508 *
i           464.49461  191.27478   2.428  0.01517 *
we          468.02868  191.27211   2.447  0.01442 *
you         464.13298  191.27476   2.427  0.01525 *
shehe       466.85132  191.27153   2.441  0.01466 *
they        468.99070  191.27494   2.452  0.01422 *
number       -1.42555    0.24676  -5.777 7.72e-09 ***
affect        5.69892    3.01266   1.892  0.05855 .
posemo       -7.21072    3.02303  -2.385  0.01708 *
negemo       -7.88056    3.06750  -2.569  0.01021 *
anx           1.03864    1.06175   0.978  0.32797
anger         1.46216    0.75356   1.940  0.05235 .
social        0.55101    0.31031   1.776  0.07581 .
family       -3.15204    0.63832  -4.938 7.96e-07 ***
friend       -3.64253    0.79227  -4.598 4.30e-06 ***
work          4.06414    0.45395   8.953  < 2e-16 ***
leisure      -2.64098    0.36494  -7.237 4.76e-13 ***
home          1.96923    1.18030   1.668  0.09525 .
money         2.01729    0.77253   2.611  0.00903 **
relig         1.61389    0.50078   3.223  0.00127 **
swear        -2.15646    0.73758  -2.924  0.00346 **
QMark        -1.35221    0.20159  -6.708 2.03e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148.4 on 19895 degrees of freedom
Multiple R-squared:  0.05507,   Adjusted R-squared:  0.05384
F-statistic:  44.6 on 26 and 19895 DF,  p-value: < 2.2e-16
```

Figure 1.2 R output from fitting a multiple linear model to WC

```
> cor(webWC[,6:10])
                   WC    Analytic       Clout   Authentic        Tone
WC          1.00000000  0.06597761  0.12737942 -0.07263959 -0.10573116
Analytic    0.06597761  1.00000000  0.14218663 -0.12510991 -0.01529677
Clout       0.12737942  0.14218663  1.00000000 -0.41829408 -0.02443453
Authentic -0.07263959 -0.12510991 -0.41829408  1.00000000  0.02824641
Tone      -0.10573116 -0.01529677 -0.02443453  0.02824641  1.00000000
```

Figure 1.3 Correlation table of the 5 variables stated

3) Report Analysis

3.1) Variables analysis and selection

We have chosen a few variables to focus on. We decided to only emphasise on the 4 summary language variables: Analytic, Clout, Authentic and Tone. This is because when those variables are fitted with a multiple linear regression model using all the variables in column 11 to 23, we realised that the adjusted R-squared value is quite high. This suggest that a higher percentage of variation in those 4 summary language variables can be explained by all the variables. Thus, it is sufficient to only use those 4 variables to analyse the change in language used across threads over time. This will be further explained with the aid of the 4 R outputs shown in Figure 2.2.

```
> fitTone = lm(Tone ~ ., data = webWC[11:32])
> summary(fitTone)

Call:
lm(formula = Tone ~ ., data = webWC[11:32])

Residuals:
    Min      1Q  Median      3Q     Max
-400.60  -17.65   -7.60   20.92  409.16

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.94146    0.45728  85.159  < 2e-16 ***
ppron        26.61027   31.23808   0.852  0.39431
i           -26.14754   31.23797  -0.837  0.40258
we          -26.14130   31.23821  -0.837  0.40269
you         -26.33104   31.23856  -0.843  0.39929
shehe       -26.34963   31.23765  -0.844  0.39895
they        -26.99209   31.23829  -0.864  0.38756
number       -0.01400    0.04008  -0.349  0.72689
affect        0.44924    0.49203   0.913  0.36124
posemo        3.62324    0.49309   7.348 2.09e-13 ***
negemo       -5.28214    0.49962 -10.572  < 2e-16 ***
anx           1.20778    0.17318   6.974 3.17e-12 ***
anger         0.36186    0.12290   2.944  0.00324 **
social       -0.12251    0.04281  -2.861  0.00422 **
family       -0.19423    0.10421  -1.864  0.06235 .
friend       -0.53570    0.12931  -4.143 3.45e-05 ***
work         -0.03868    0.07374  -0.524  0.59996
leisure       0.55778    0.05935   9.399  < 2e-16 ***
home         -0.74956    0.19247  -3.894 9.88e-05 ***
money         1.29982    0.12575  10.337  < 2e-16 ***
relig        -0.71617    0.08156  -8.781  < 2e-16 ***
swear         0.53411    0.11926   4.479 7.56e-06 ***
QMark        -0.05860    0.03281  -1.786  0.07410 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.24 on 19899 degrees of freedom
Multiple R-squared:  0.5231,    Adjusted R-squared:  0.5226
F-statistic: 992.3 on 22 and 19899 DF,  p-value: < 2.2e-16
```

Figure 2.1

Figure 2.1 shows the output of the multiple regression model fitted using the all the variables from column 11 to 23 (the list of variables stated in the figure below the word “(Intercept)”) to model the LIWC summary for Tone.

From the output, we can see that the p-values for a few variables is large for which they are greater than 0.1 (i.e. number, work, I, we etc.). Therefore, we can deduce that those variables are not associated with the “Tone” variable. However, the problem with this simple p-value approach is that sometimes predictors can appear to be more associated if we remove other unimportant predictors that are “diluting the signal”. Thus, we used stepwise selection to try to prune out unimportant predictors, using the step() function in R. The first run was done using the Akaike Information Criterion (AIC), which identified the full model including all 22 predictors as optimum. This is an example of a greedy algorithm. At every step, it tries to make changes to our model that results in the biggest improvement to the smallest information criterion. Such algorithm is not guaranteed to find the best model, but it can do reasonably well in many settings. This algorithm starts with the full model and sequentially try to remove predictors until we cannot improve out model (in terms of information criterion score) by removing any more predictors.

- This approach is similar to how we were to add a variable one by one to the model and see if that variable is associated with the dependent variable by looking at the adjusted R-squared value. If the adjusted R-squared value of the model decreases after a certain variable is added, that indicates that the variable is highly unlikely to be associated with our dependent variable. Thus, we would be eliminating that variable from our model.

- So instead of adding all the variables to the model manually one by one, we use the step() function which considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according to the information criterion. These criteria assign scores to each model and allow us to choose the model with the best score.

The output for the Tone variable is shown in Figure 2.2.1, which removed 5 of the variables ('i', 'we', 'number', 'affect' and 'work') as unimportant in explaining the tone variable. This indicates that threads or post with emotional tone is less likely to use words referring to first person point of view and people are less likely to sound emotional when it comes to work.

```
#MLR to model the 4 summary
fitTone = lm(Tone ~ ., data = webWC[11:32])
summary(fitTone)
swTone = step(fitTone)
summary(swTone)
```

```
Console C:/Users/user/Desktop/fit3152/
lm(formula = Tone ~ ppron + you + shehe + they + posemo + negemo +
    anx + anger + social + family + friend + leisure + home +
    money + relig + swear + QMark, data = webWC[11:32])

Residuals:
    Min      1Q  Median      3Q     Max
-400.48  -17.67   -7.68   20.92  408.94

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.84106    0.41048  94.622  < 2e-16 ***
ppron        0.46732    0.04171  11.205  < 2e-16 ***
you         -0.18733    0.08568  -2.186  0.02879 *
shehe       -0.20812    0.10941  -1.902  0.05716 .
they        -0.84766    0.09805  -8.645  < 2e-16 ***
posemo       4.07269    0.03722 109.431  < 2e-16 ***
negemo      -4.83003    0.08668 -55.720  < 2e-16 ***
anx          1.20624    0.17316   6.966 3.36e-12 ***
anger        0.36220    0.12288   2.948  0.00321 **
social      -0.11995    0.03926  -3.056  0.00225 **
family      -0.19559    0.10238  -1.910  0.05610 .
friend      -0.53484    0.12765  -4.190 2.80e-05 ***
leisure      0.55742    0.05920   9.415  < 2e-16 ***
home        -0.74952    0.19243  -3.895 9.85e-05 ***
money        1.28630    0.12261  10.491  < 2e-16 ***
relig       -0.71414    0.08144  -8.769  < 2e-16 ***
swear        0.53368    0.11923   4.476 7.65e-06 ***
QMark       -0.05822    0.03278  -1.776  0.07570 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.24 on 19904 degrees of freedom
Multiple R-squared:  0.5231,    Adjusted R-squared:  0.5227
F-statistic:  1284 on 17 and 19904 DF,  p-value: < 2.2e-16
```

2.2.1: R output from fitting a multiple linear model to Tone

```
Call:
lm(formula = Authentic ~ ppron + we + you + shehe + they + number +
    affect + posemo + social + family + work + leisure + home +
    money + relig + swear, data = webWC[11:32])

Residuals:
     Min       1Q   Median       3Q      Max
-107.453  -20.192   -4.807   18.368  113.294

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.71492    0.50148  77.201  < 2e-16 ***
ppron        3.35475    0.04814  69.690  < 2e-16 ***
we          -3.37994    0.12341 -27.388  < 2e-16 ***
you         -3.10522    0.09908 -31.341  < 2e-16 ***
shehe       -6.32739    0.12537 -50.471  < 2e-16 ***
they        -3.29567    0.11060 -29.798  < 2e-16 ***
number      -0.19307    0.04427  -4.361 1.30e-05 ***
affect      -0.76100    0.06305 -12.070  < 2e-16 ***
posemo       0.39479    0.07121   5.544 2.99e-08 ***
social      -0.56655    0.04437 -12.768  < 2e-16 ***
family       0.34984    0.11392   3.071  0.00214 **
work        -0.69397    0.08131  -8.535  < 2e-16 ***
leisure     -0.34710    0.06553  -5.297 1.19e-07 ***
home         1.11695    0.21254   5.255 1.49e-07 ***
money       -0.34093    0.13825  -2.466  0.01367 *
relig       -0.48687    0.09010  -5.404 6.61e-08 ***
swear        0.19067    0.12342   1.545  0.12238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.78 on 19905 degrees of freedom
Multiple R-squared:  0.2751,    Adjusted R-squared:  0.2745
F-statistic: 472.1 on 16 and 19905 DF,  p-value: < 2.2e-16
```

2.2.2: R output from fitting a multiple linear model to Authentic

```
Call:
lm(formula = Clout ~ ppron + i + we + you + shehe + they + number +
    negemo + anx + anger + social + friend + work + leisure +
    relig + swear, data = webWC[11:32])

Residuals:
     Min       1Q   Median       3Q      Max
-157.935   -7.733    2.612    9.356   94.861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.61826    0.25787 161.392  < 2e-16 ***
ppron        33.63026   18.52749   1.815 0.069515 .
i           -36.33985   18.52744  -1.961 0.049846 *
we          -31.24270   18.52754  -1.686 0.091757 .
you         -31.80463   18.52772  -1.717 0.086068 .
shehe       -33.06597   18.52722  -1.785 0.074321 .
they        -33.39969   18.52763  -1.803 0.071451 .
number        0.29860    0.02374  12.580  < 2e-16 ***
negemo       -0.25639    0.05121  -5.007 5.57e-07 ***
anx           0.22628    0.10254   2.207 0.027346 *
anger         0.27278    0.07261   3.757 0.000173 ***
social        2.15317    0.02330  92.409  < 2e-16 ***
friend       -0.18150    0.07563  -2.400 0.016412 *
work          0.40159    0.04260   9.427  < 2e-16 ***
leisure       0.10338    0.03504   2.950 0.003180 **
relig         0.08230    0.04823   1.706 0.087953 .
swear        -1.31327    0.07025 -18.695  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.38 on 19905 degrees of freedom
Multiple R-squared:  0.6911,    Adjusted R-squared:  0.6909
F-statistic:  2783 on 16 and 19905 DF,  p-value: < 2.2e-16
```

2.2.3: R output from fitting a multiple linear model to Clout

```
Call:
lm(formula = Analytic ~ i + we + you + shehe + they + number +
    posemo + negemo + anx + anger + family + friend + work +
    leisure + home + money + relig + QMark, data = webWC[11:32])

Residuals:
    Min      1Q  Median      3Q     Max
-95.239 -14.365   3.209  16.153 160.341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.74678    0.40027 196.734  < 2e-16 ***
i           -2.81223    0.04074 -69.023  < 2e-16 ***
we          -2.55883    0.09423 -27.156  < 2e-16 ***
you         -2.66927    0.06491 -41.125  < 2e-16 ***
shehe       -2.28260    0.09188 -24.844  < 2e-16 ***
they        -2.87451    0.07915 -36.317  < 2e-16 ***
number       0.42880    0.03734  11.485  < 2e-16 ***
posemo      -0.10089    0.03446  -2.927 0.003424 **
negemo      -0.52329    0.08107  -6.455 1.11e-10 ***
anx          0.30899    0.16166   1.911 0.055971 .
anger        0.79865    0.10878   7.342 2.18e-13 ***
family       0.31551    0.08979   3.514 0.000442 ***
friend       0.25151    0.11350   2.216 0.026704 *
work         0.78341    0.06886  11.376  < 2e-16 ***
leisure      0.42028    0.05542   7.584 3.51e-14 ***
home         0.91009    0.17979   5.062 4.19e-07 ***
money        0.55530    0.11741   4.730 2.27e-06 ***
relig        0.23557    0.07613   3.094 0.001974 **
QMark       -0.33758    0.03065 -11.014  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.65 on 19903 degrees of freedom
Multiple R-squared:  0.308,     Adjusted R-squared:  0.3074
F-statistic: 492.2 on 18 and 19903 DF,  p-value: < 2.2e-16
```

2.2.4: R output from fitting a multiple linear model to Analytic

Figure 2.2

Figures 2.2 show final multiple linear regression models (after eliminating all potentially unimportant variables using the step() function) to model all 4 summary language variables respectively using all other variables from column 11 o 23. From the analysis above, we can infer that:

- In Figure 2.2.1, the R output from fitting a multiple linear model to Tone shows that most variables have very low p-values. The QMark variable has the highest p-value in the model, which means there is little association between being emotional and asking questions. Reason being that when one is emotional, he or she will hardly ever ask questions. He or she would more likely address or confess something, expressing his or her emotions.
- In Figure 2.2.2, almost all the variables have very low p-values. This means that users are authentic and trustworthy regardless of the things that they are talking about. Only the swear and money variables have high p-values. The swear variable, which has the highest p-value, shows that when users are being authentic, they very rarely use any swear words. As for the money variable, which has the second highest p-value, shows that users are seldom authentic and genuine when talking about money.
- In Figure 2.2.3, we can see the p-values of swear words, anger and negative emotions (negemo) are low. This means that there is a strong correlation between Clout summary and the 3 variables: swear words, anger and negative emotions (negemo). The 3 variables are very significant when it comes to measuring the Clout summary of a post. Furthermore, the p-values of all the pronouns (i.e. I, we, they etc.) are very high. Since the p-values are very high, it is more likely that pronouns will be absent from Clout posts. Therefore, we can say that users often do not structure their sentences properly when they are more forceful or aggressive (attributes of Clout summary). We also noticed that the 2 variables, family and positive emotions (posemo), are excluded from the model. This suggests that users seldom associate Clout with family and positive emotions, they hardly react aggressively when it comes to discussions about family and positive emotions. However, this is not the case when discussing about work matters as the p-value for the work variable is very low. This means that users tend to be more assertive because influence and power are very important in the working environment.
- In Figure 2.2.4, almost all the variables have very low p-values, which means that most of them will appear frequently in a post that has a high proportion of Analytic summary. This is because most posts will require analytical thinking regardless of the topic discussed. Posts with high proportion of Analytic summary tend to be associated with question marks. This tells us that there are always questions going back and forth in analytical posts. We also noticed that the p-value of the anx variable is high. It shows that in analytical discussions, anxiety among users are barely present. Everyone is so busy learning and sharing information that there is hardly any sign of anxiety.

3.2) Author-Thread relationship

The average (mean) proportion of the 4 summary variables is calculated after the dataset is aggregated by ThreadID and month. We would be analysing the top 3 most active threads. We determine an active thread by looking at the number of posts per month and the number of months the thread has been ongoing. We found that the top 3 most active threadID are 127115, 472752, and 532649.

First, we look at the largest thread in the dataset. Figure 2.3 is the network graph showing authors who have posted at least one entry in Thread 127115. The thickness of the edge represents how frequently the author has posted on that thread (i.e. it indicates the total number of post contributed by that author). From the network graphs below, we can see that the graph in Figure 2.3 is very hard to read and interpret. However, the one thing we can notice is that there exist many authors that have only posted once in that thread. After removing authors that only contributed one post in that thread, the network graph becomes clearer as shown in Figure 2.4 and we can start to see the difference in thickness of the edges. It is obvious that author 47875 contributed the most in this thread. Thus, we will be studying the author's usage of language in this thread and compare it with the overall change in language used by this thread over time. We will also be comparing how different is the author's usage of language in this thread as compared to other thread he/she is involved in.

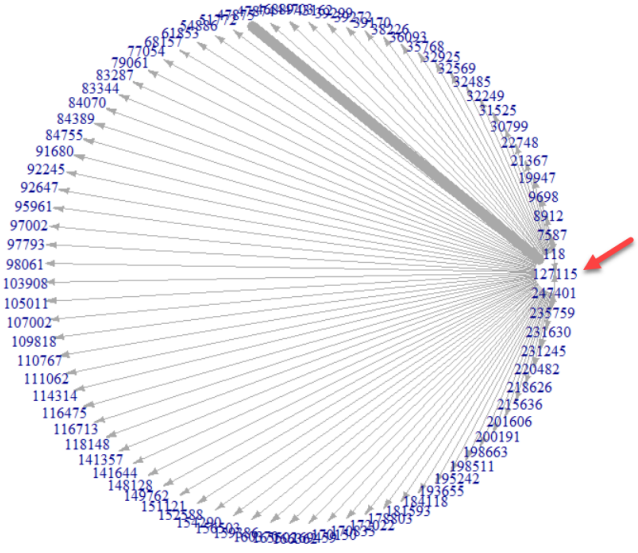


Figure 2.3

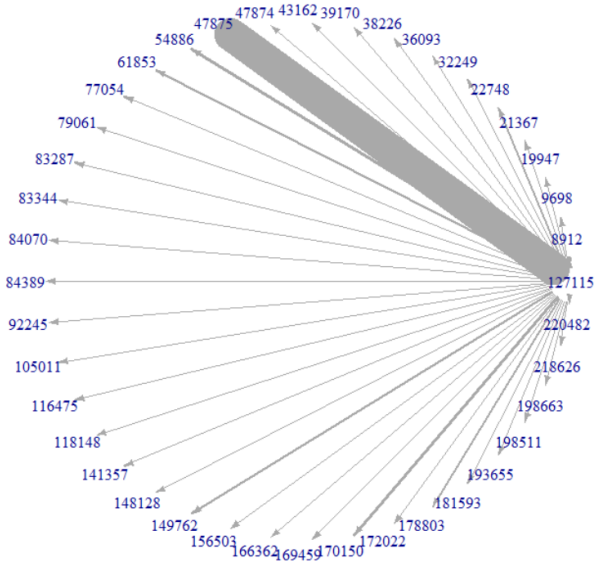


Figure 2.4

- This thread lasted from 2004/4/1 until 2011/11/1
- It has 84 participants and 433 posts.
- The red arrow indicates the threadID, the rest represent the authorID.

ThreadID: 127115

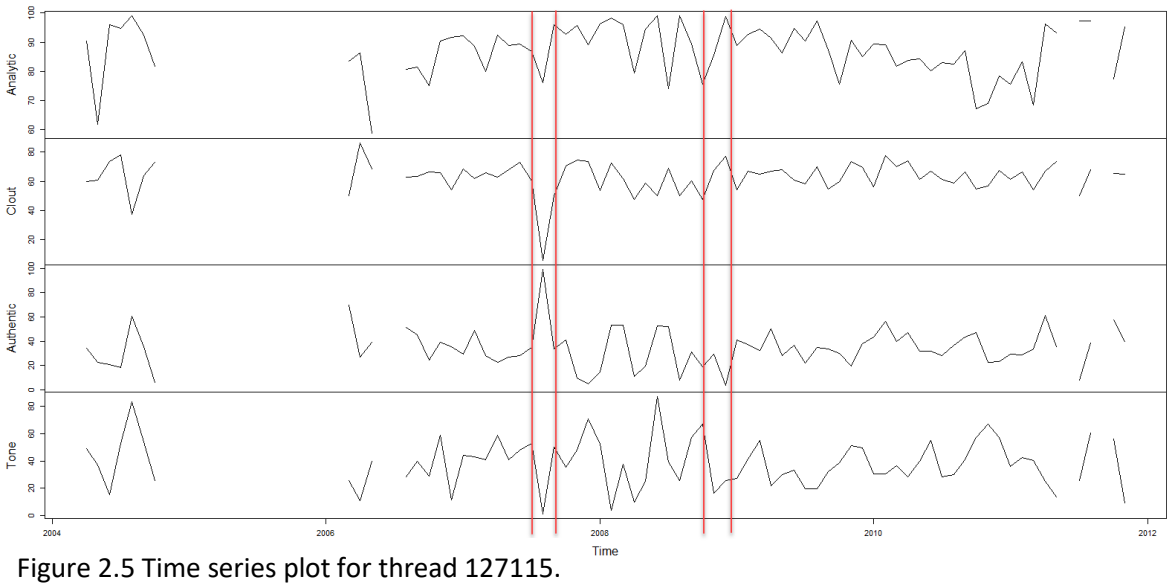


Figure 2.5 Time series plot for thread 127115.

Figure 2.5 plots the changes of proportion of all 4 summary language variables across the time period of that thread being active. The correlation table in Figure 1.3 under the first section already suggest that Analytic is positively correlated with Clout and negatively correlated with both Tone and Authentic. This piece of information can be supported by the time series plot above. Take a look at the lines within the red line, notice how when analytic and clout decreases, authentic and tone increases, vice versa.

There is no trend nor seasonal pattern observed in the time series graph in Figure 2.5. Even after decomposing the time series for all 4 summary language variables (Figure 2.6), we still cannot obverse any pattern in all of the time series components. From the time series analysis (Figure 2.6), it appears that there is seasonality in our time series, however this pieces of information is forged because before decomposing, we fill up all the missing values using a seasonal Kalman filter by employing na.StructTS() from the zoo package (refer to line 133 in our code). This method of dealing with missing values (as there are some months whereby no one posted in the thread) will enhance the seasonal component of the time series. Due to the large amount of missing values, the seasonality of the time series appears to be stronger. This is why for all of the 4 time series analysis on the 4 summary language variable, there appear to be seasonality.

This suggest that the change in language of thread 127115 is random and not consistent. There is no constant increase nor decrease, in any of the 4 summary language variables. However, from the time series we observed changes in language over time. ***This indicates that the language used in a thread does changes over time.*** The language used is not stagnant and will change over time, suggesting that the topic of the thread might divert or the topic of the thread is related to leisure (or any other variables that is significant in modelling all 4 summary language variables) resulting in a random change in the 4 summary variables of a thread over time. We cannot yet conclude our reasoning based on only one thread sample. Thus, we will be analysing on the second most active thread in our dataset which is thread 472752 in the next part of the report.

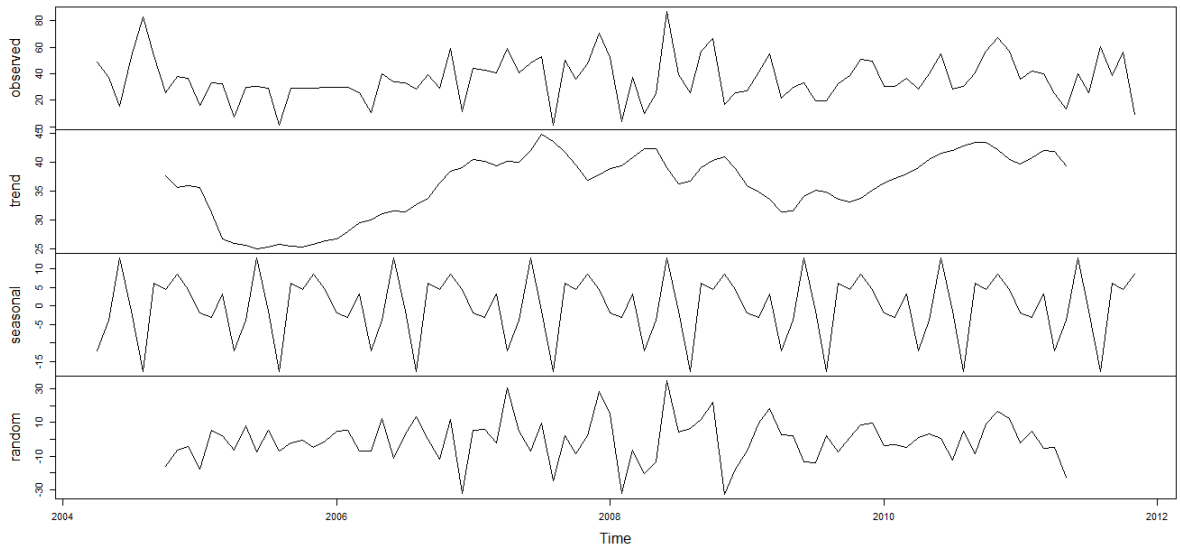


Figure 2.6.1 Time series analysis for the analytic summary language variable for thread 127115

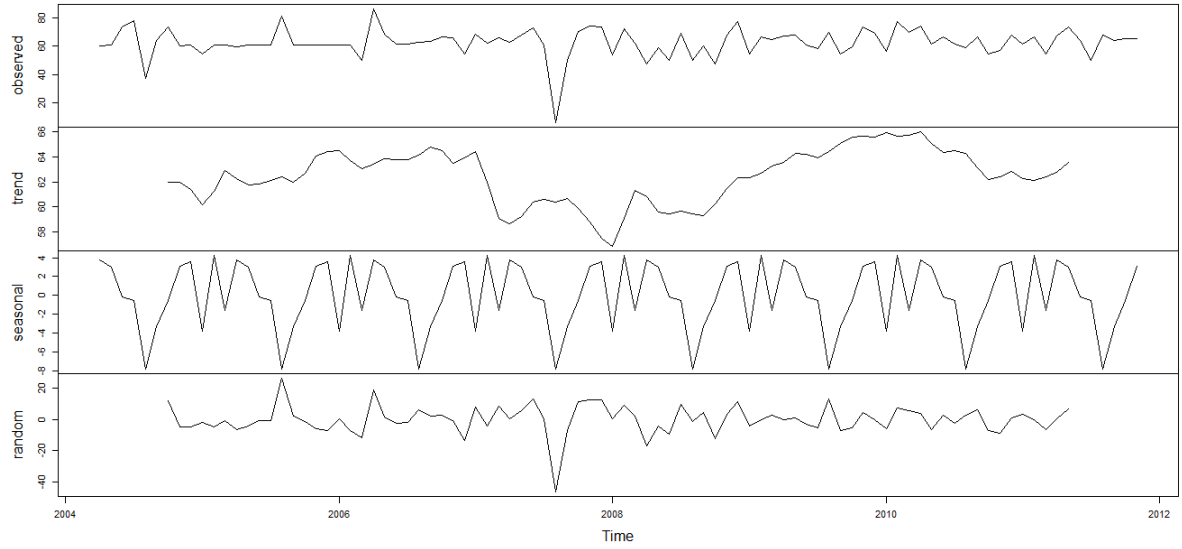


Figure 2.6.2 Time series analysis for clout summary language variable for thread 127115

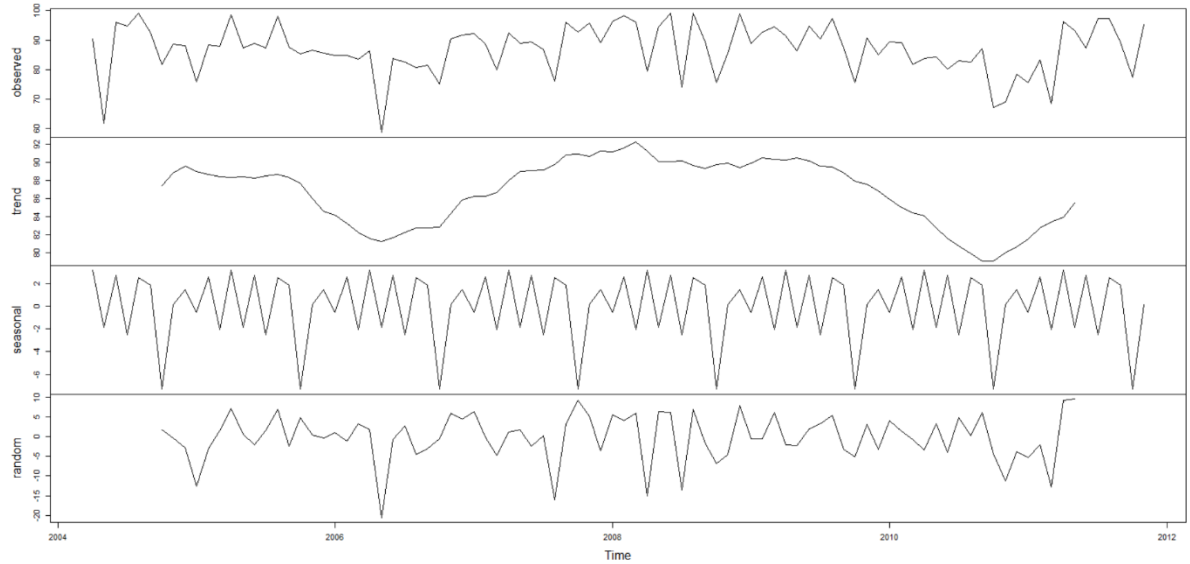


Figure 2.6.3 Time series analysis for tone summary language variable for thread 127115

The top 3 author that contributed in thread 127115 is author 47875 (171 posts), author 8912 (34 post) and author 54886 (19 post). Now we concentrate on the most active author in thread 127115. We will be analysing author 47875's usage of language across different threads that he or she is involved in. We want to see if the author's language changes or remains the same across different threads. We want to find out if people will change the way they speak if they are spoken differently. Author 47875 is involved in 87 different threads (i.e. posted at least once) over different a timespan. He or she is active since 2004/08/27 and the most recent post posted was on 2011/10/31.

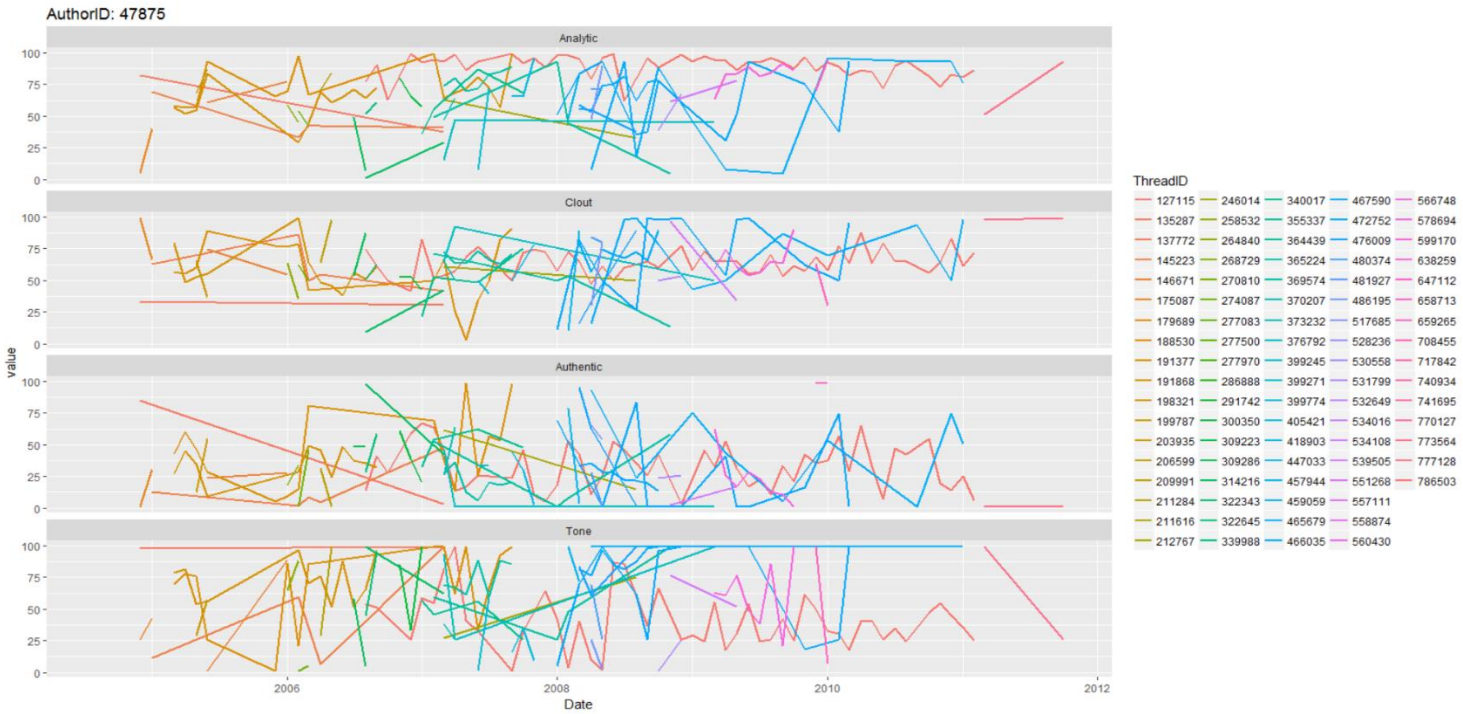


Figure 2.7 Time series plot: the proportion values of the 4 summary language variables for all threads author 47875 is involved in over time

There appears to be no consistent pattern in the changes in language for author 47875. ***The pattern of language used across different threads is random and varies from one thread to another. Author 47875 does not have its own pattern of language***, instead the value of 4 summary language variables changes according to the thread the author contributed in. The time series plot in Figure 2.7 might contain too much information and that might be the reason we cannot see any pattern. Thus, we plotted 2 more time series plot (Figure 2.8 and Figure 2.9) with only information on the top 3 and top 5 threads author 47875 is most active in respectively. Based on both plots, we still cannot observe any consistent pattern in the change of proportion of the 4 summary language variables over time.

The same result is observed when we analyse the changes in language used for 3 other different authors which is author 8912 and author 54886 (as mentioned above they are in the top 3 most active users for thread 127115). The third author with authorID of 61853 is randomly selected (Figure 2.10.2). The point data plots indicate that the author only posted once in that particular thread. We also randomly selected authorID 8912 and plotted bar charts of the 4 summary language variables for different threads he/she is involved in. (Figure 2.10.1)

Figure 2.8 Time series plot: the proportion values of the 4 summary language variables for top 3 threads author 47875 is most active in over time

Figure 2.9 Time series plot: the proportion values of the 4 summary language variables for top 5 threads author 47875 is most active in over time

AuthorID: 8912

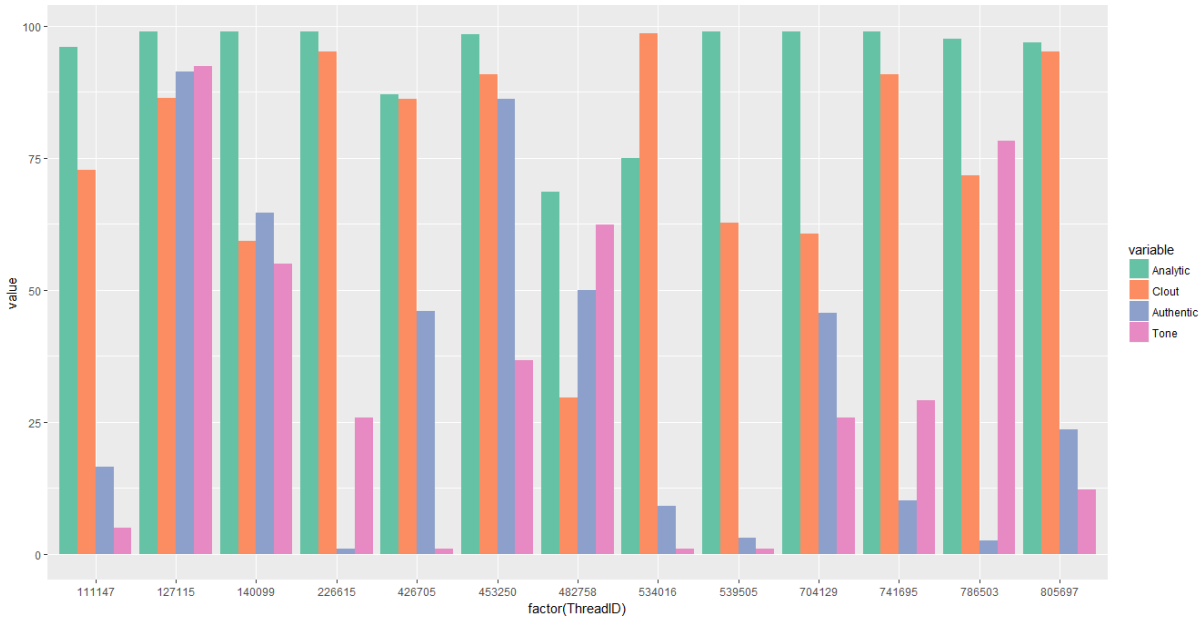


Figure 2.10.1 Analysing the language usage of authors in all of the threads they are involved in

The values for all 4 variables varies across different threads. Expect for the analytic variable which is consistently high is all threads, other the values for the other 3 variables fluctuates randomly across different threads. This suggest that this author is more of an analytical person whereby his/her post always contain more analytical words. But this characteristic varies from author to author.

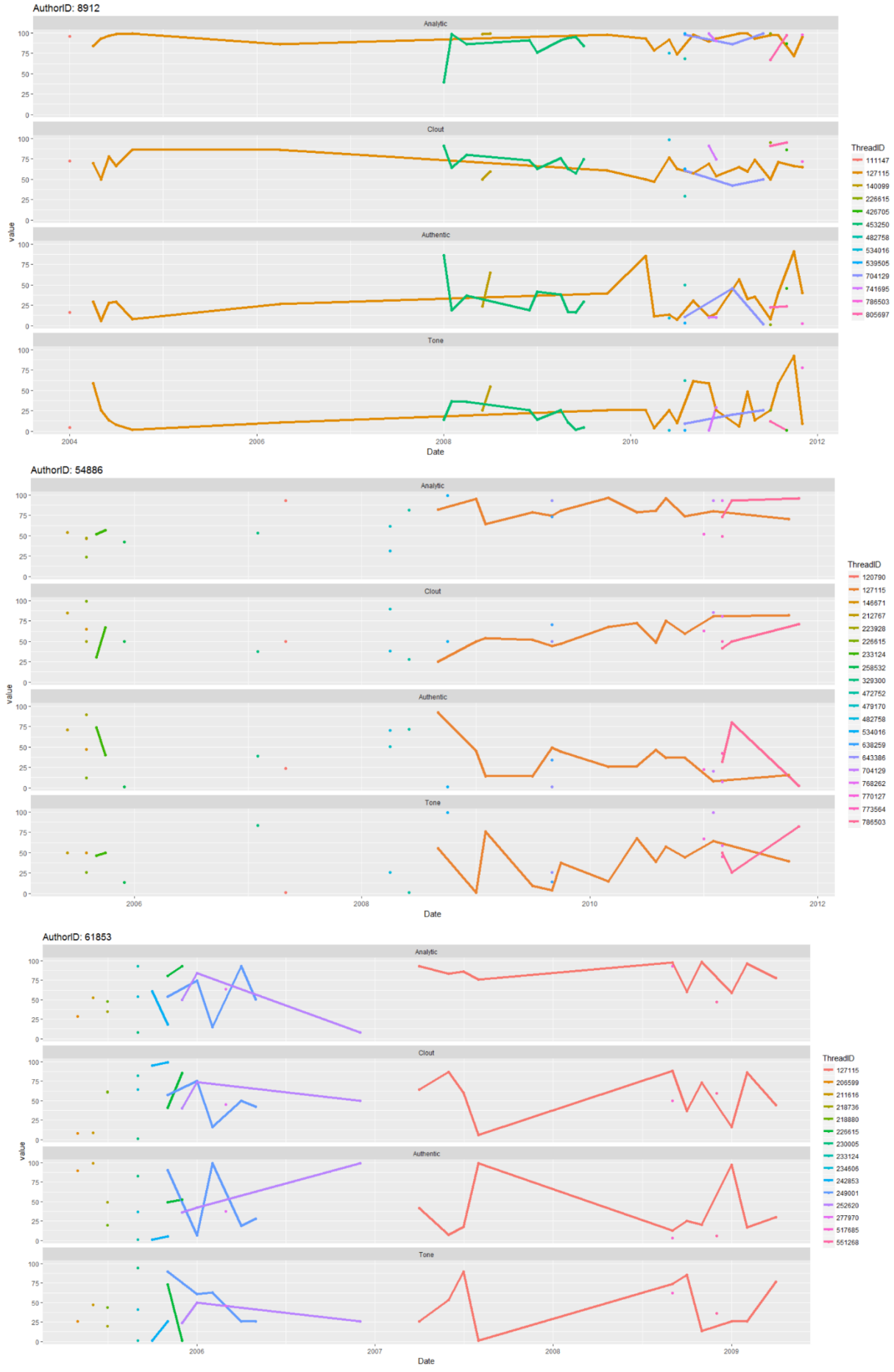


Figure 2.10.2 Analysing the language usage of authors in all of the threads they are involved in

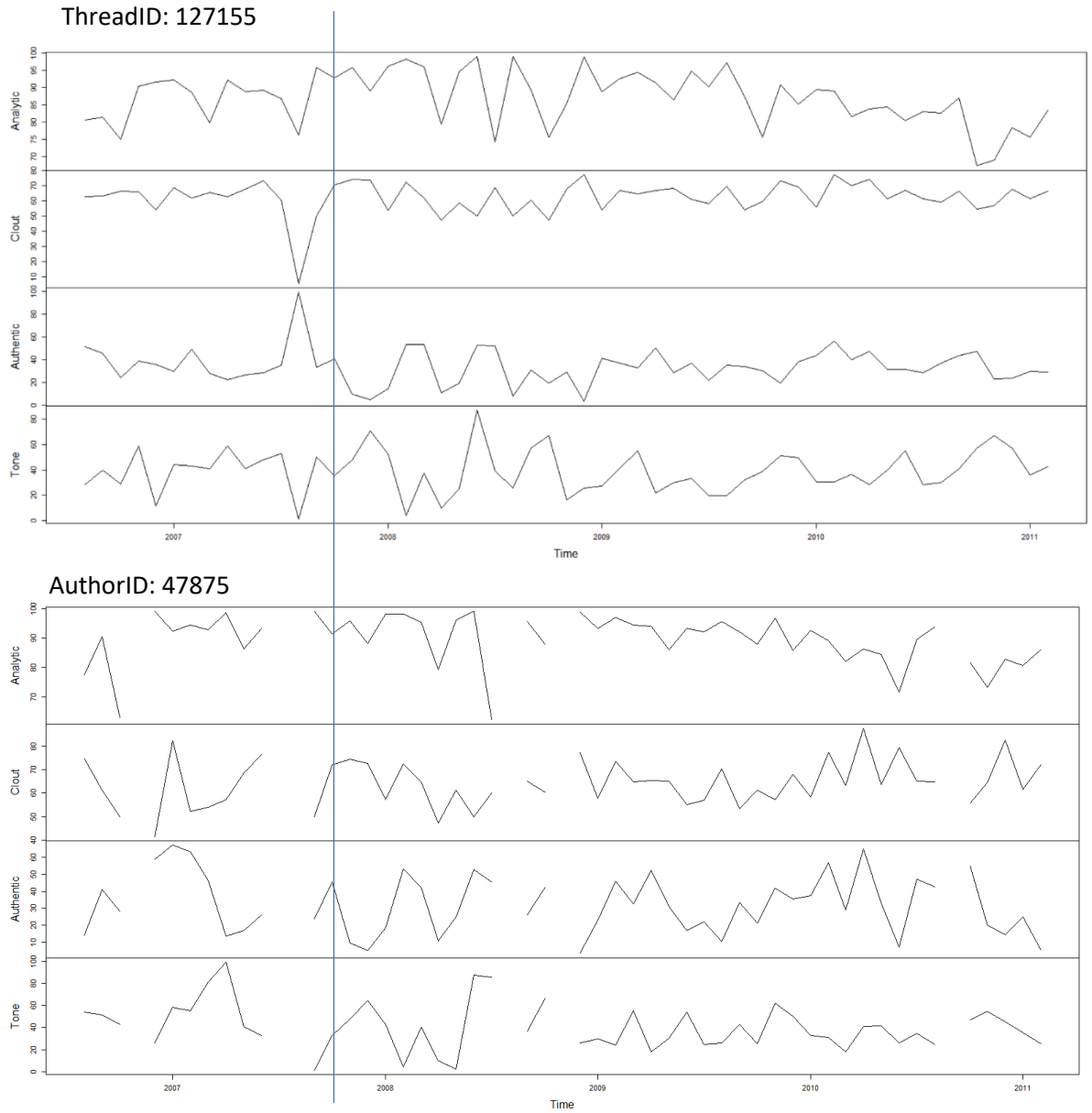


Figure 2.11: Comparison of the changes in language between the thread and its most active

However, when the proportion values of the 4 summary language variables of the author are compared with the values of the thread he or she is active in, we noticed that the change in language that the author uses resembles the change in language of that particular thread over time. Using the blue line in Figure 2.11 as a guideline, we noticed that when either of the language summary variables value increases or decreases, the values for the author follows as well. Similar analysis had been done for the second largest thread (Figure 2.14) and another randomly selected thread and similar results are produced. This phenomenon can only be observed given that the author has been continuously active in that thread long enough. The only thing we are unsure about is how long should the "long enough" be, it varies from author to author, thread to thread, it also depends on the people who are involved in the thread, whether they are influential and depending on the topic the thread is discussing. Just like the second most active author in thread 127155, author 8912 was active in that thread

over a long period of time but he or she is not consistently active (as shown in Figure 2.12). Author 8912 is active from 2004/04/01 till 2004/09/01 then he or she became inactive for 6 years (only posted 2 posts within that time period). Author 8912 then only return to be active in the month of March 2010 and has been active till November 2011. Therefore, we notice that the changes in author 8912's language only resembles the thread at the last few months of author 8912's active participation in the thread. Additionally, the language used by the author does not resemble the change of language of the thread maybe because the author has yet to engage in a conversation in the thread. Thus, the author is adapting languages from other threads that he or she is involved in and applying it in thread 127155. Another reason could be that his/her post contribution is too few until we cannot observe any changes.

Maybe taking the top few most active authors as sample isn't the best approach because the changes in language of the thread is mainly affected by the authors that posted most frequently. So, if there exist one author that post very frequently as compared to other active authors, then the values of the thread's summary language variables will reflect that author's usage of language. Therefore, the thread's changes of language will most definitely reflect the most active author's usage of language. However, if we were to randomly choose any author, the chances of choosing an author that is not active (only contributed one or two posts in that thread) is quite high and such sample provides insufficient data for us to analyse. Thus, we still have to be selective when it comes to nominating sample author. To overcome this problem, we randomly selected a few authors in the top 10 active users and analyse their usage of language in that thread. The result was as predicted whereby the usage of language by the author does reflect the change of language in the thread (Figure 2.13).

3.3) Thread-thread relationship

Figure 2.15 shows the time series plot for the change in language in thread 127155 and thread 472752. From there, we noticed that as time passes, the values of the 4 summary variables will eventually average out. The fluctuation in the values of the 4 variables decreases over time. However, this observation only appears in threads that are active for more than 3 years. This suggests that regardless of any topic the thread is about, the enthusiasm level of the author in that thread will ultimately become inactive as time passes.

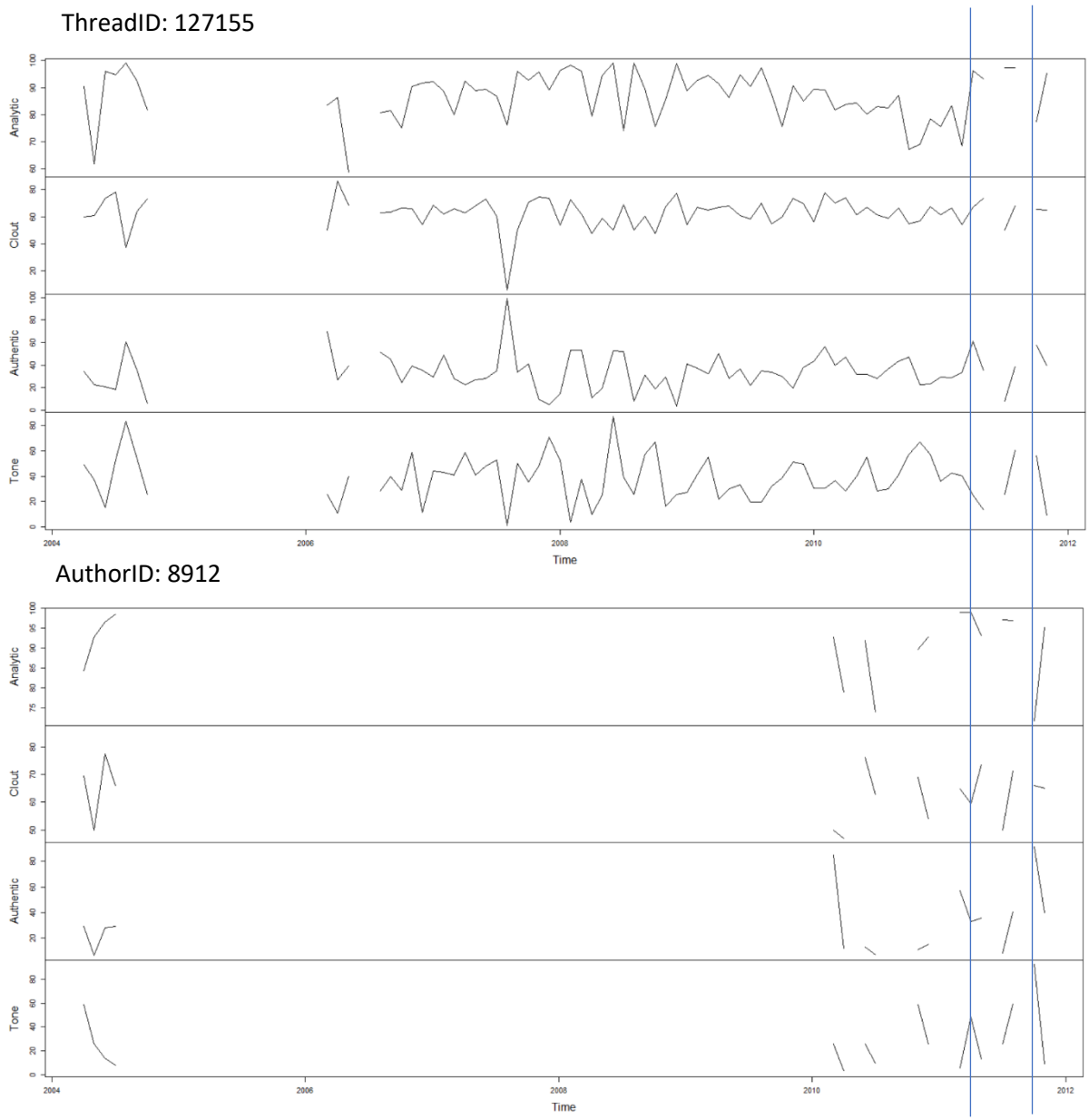


Figure 2.12: Comparing the most active thread with the second most active author in the thread.

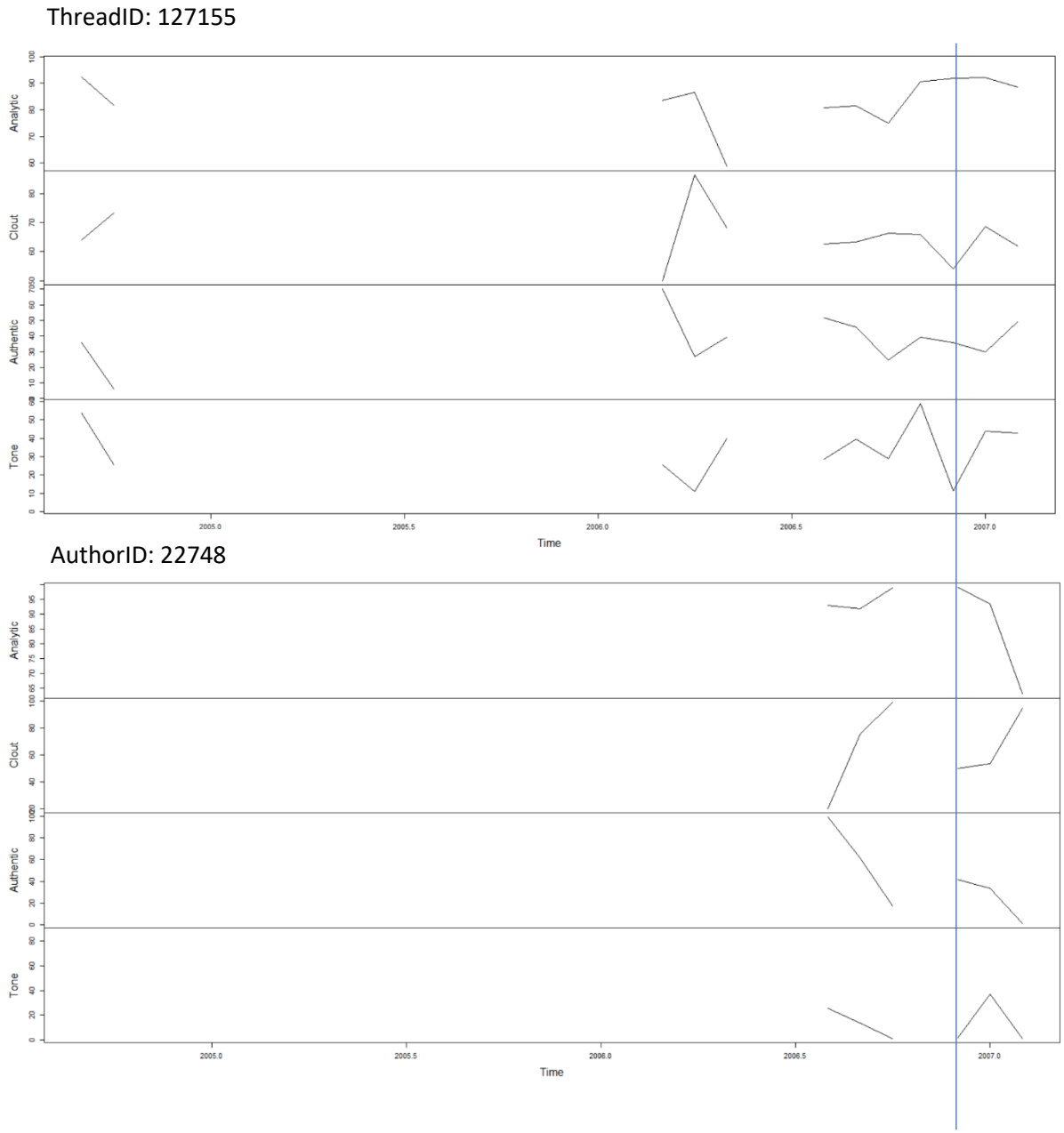
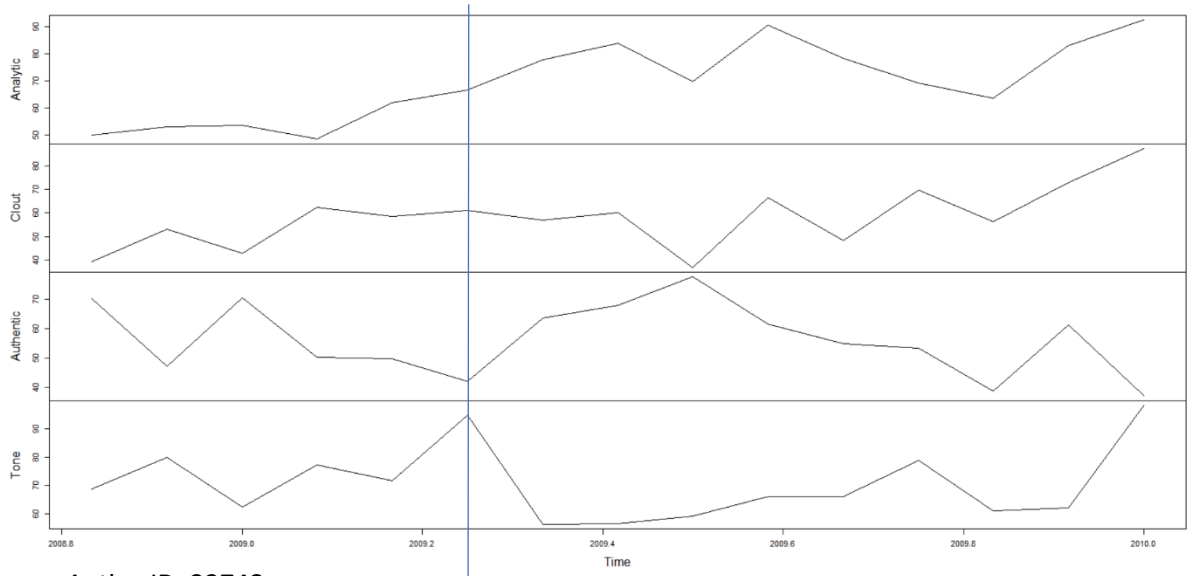


Figure 2.13: Comparing the most active thread with a randomly selected author from the top 10 active authors in the thread

ThreadID: 472752



AuthorID: 22748

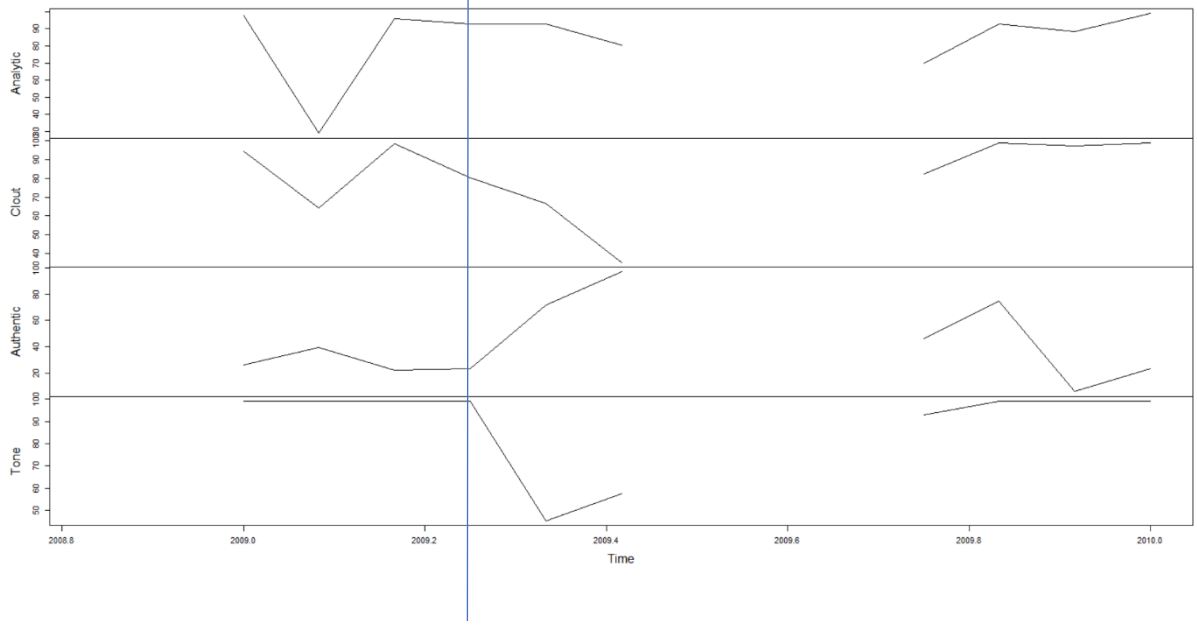
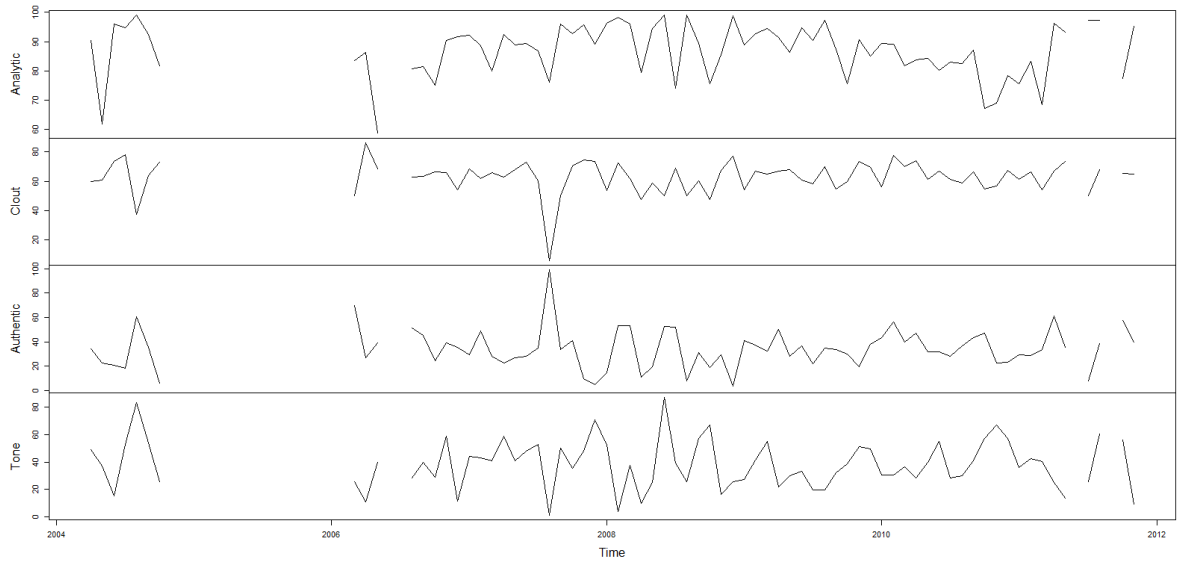


Figure 2.14: Comparing the second most active thread with a randomly selected author from the top 10 active authors in the thread

ThreadID: 127155



ThreadID: 472752

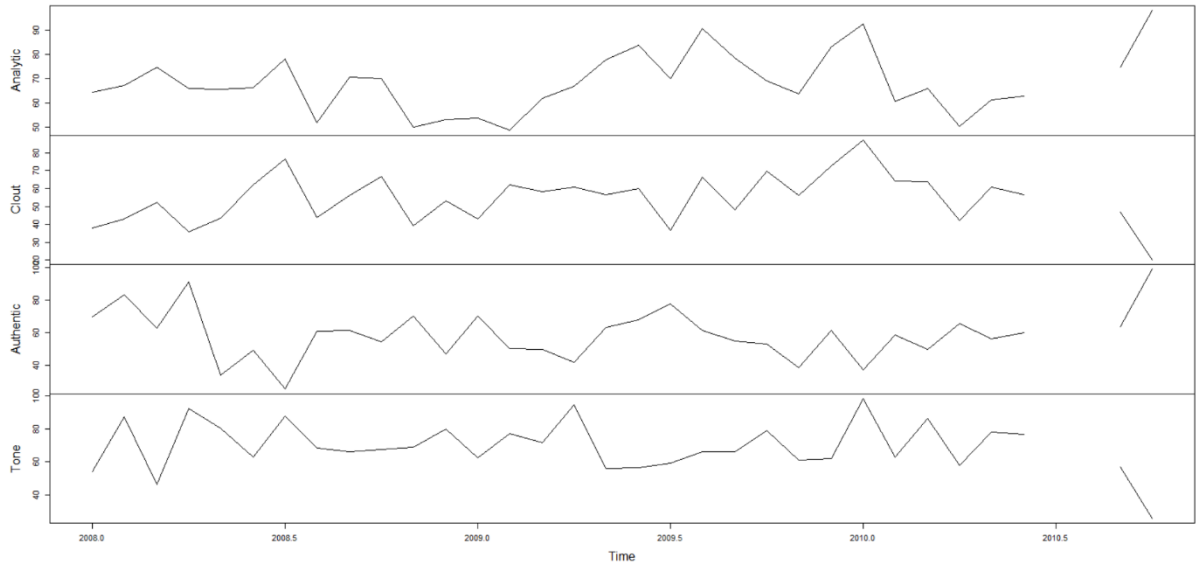


Figure 2.15: Comparing the changes in language for the top 2 active thread

4) Conclusion

After the thorough analysis, we conclude that the language being used in a thread do not remain constant throughout, it changes over time. However, there is not a fixed pattern in the language change, it changes in a random manner. Users will change the way they speak as time progresses. The longer they have been active in a thread the more apparent the change will be. As for our second objective, our analysis tells us that there is a difference in behaviour, but there is not a consistent pattern in the changes in language used over different threads. It is completely random and it differs for every thread. At first, we thought that there will be a change in language used in a thread over time, and that a certain pattern would lead to that change. But we were proven wrong through the analysis that there is no pattern but the language in the thread does change over time. Users' behaviours were also changing randomly according to the threads that they are involved in.

5) Appendix

```
rm(list=ls())

web = read.csv("webforum.csv")

##taking a look at the features of the dataset
# calculating the number of Authors involved, threads exist and the number of posts.
length(unique(web$AuthorID))
length(unique(web$ThreadID))
length(unique(web$PostID))

#remove thread with word count 0. Because we cannot compare forum without any contents
webWC = web[web$WC != 0,]

#convert the data type of column "Date" from factor to Date.
webWC$Date = as.Date(as.character(webWC$Date), format = "%Y-%m-%d")

######Multiple Linear Regression(MLR) to model word count
##code to produce Figure 1.1
fitWC = lm(webWC$WC ~ ., data = webWC[7:10])
summary(fitWC)

##code to produce Figure 1.2
fitWC = lm(webWC$WC ~ ., data = webWC[7:32])
summary(fitWC)

#finding the correlation between the 4 summary language variable with word count
#Figure 1.3
cor(webWC[,6:10])

#MLR to model the 4 summary language variable (figure 2.2)
fitTone = lm(Tone ~ ., data = webWC[11:32])
summary(fitTone)
swTone = step(fitTone)
summary(swTone)

fitAna = lm(Analytic ~ ., data = webWC[11:32])
summary(fitAna)
swAna = step(fitAna)
summary(swAna)

fitClout = lm(Clout ~ ., data = webWC[11:32])
summary(fitClout)
swClout = step(fitClout)
summary(swClout)

fitAuth = lm(Authentic ~ ., data = webWC[11:32])
summary(fitAuth)
swAuth = step(fitAuth)
summary(swAuth)

##Create a new "Month" column because we would like to aggregate our data at months level
```

```
#Basically we would find the mean of the
values of the 4 summary language variables of
all post within a month

webWC$Month = as.Date(cut(webWC$Date,
breaks = "month"))

byThread = aggregate(cbind(webWC$Analytic,
webWC$Clout, webWC$Authentic,
webWC$Tone), by=list(webWC$ThreadID,
webWC$Month), FUN=mean)

colnames(byThread) = c("ThreadID",
"Month","Analytic","Clout","Authentic","Tone
")

#then we would like to calculate the freq of
thread that is active for most number of
month.

ThreadSize =
as.data.frame(table(byThread$ThreadID))

ThreadSize = ThreadSize[order(-
ThreadSize$Freq),]

ThreadSize[which.max(ThreadSize[,2]),]     #
this would give us the thread that has been
active for the most number of months

### Ploting the time series of the change in
language for a particular thread.

## Figure 2.5 is all plotted using this part of R
code.

# we just have to manually change the Thread
ID and the "start" vector values then we can
analyse time series of different threads

plotData = byThread[byThread$ThreadID ==
127115,]

full = as.data.frame(seq(from =
min(plotData$Month), to =
max(plotData$Month), by='1 month'))
colnames(full) = "Months"

full$Analytic =
plotData$Analytic[match(full$Months,
plotData$Month)]

full$Clout =
plotData$Clout[match(full$Months,
plotData$Month)]

full$Authentic =
plotData$Authentic[match(full$Months,
plotData$Month)]

full$Tone =
plotData$Tone[match(full$Months,
plotData$Month)]

new = ts(full, frequency = 12, start =
c(2004,4))

plot(new[,2:5])

#####

##Please remember to detach the 'igraph'
library before runnign the decompose
function using "detach("package:igraph",
unload=TRUE)" command

#Because in the igraph package there is a
function with the same name.

#Figure 2.6

# the function used from the zoo package is
further explained in the report at page 8

library(zoo)

ana = ts(full$Analytic, frequency = 12, start =
c(2004,4))

decomAna = decompose(na.StructTS(ana))

plot(decomAna)

cl = ts(full$Clout, frequency = 12, start =
c(2004,4))
```

```
decomcl = decompose(na.StructTS(cl))

plot(decomcl)

tn = ts(full$Tone, frequency = 12, start =
c(2004,4))

decomTon = decompose(na.StructTS(tn))

plot(decomTon)



####### the thickness of the graph is how
frequent the author post on that thread.

#plot network graph of thread with author
(Figure 2.3 and 2.4)

library(igraph)

library(igraphdata)


fullVec = webWC[webWC$ThreadID ==
127115 & webWC$AuthorID!=-1, 2:3]

thickness =
as.data.frame(table(fullVec$AuthorID))


v = rep(127115, length(thickness$Freq))


nodes
=cbind.data.frame(thickness,ThreadID=v)

#removing authors that only post one time
(figure 2.4)

nodes = nodes[nodes$Freq>=2,]


# to nodes

Tnodes = as.character(nodes[nodes$Freq>=2,
1])

# from nodes

Fnodes = as.character(nodes[nodes$Freq>=2,
3])


graphdata = data.frame(from = Fnodes, to =
Tnodes)

g = graph.data.frame(graphdata, directed =
TRUE)

E(g)$width = nodes$Freq/5

E(g)$arrow.size = .2

plot(g, vertex.shape="none",layout =
layout.circle)


##sort highest to lowest frequency

#to see which author contributed the most in
that particular thread.

sorted = nodes[order(-nodes$Freq),]


#then we take the top n authors who
contributed the most post in the thread

#figure 2.7, 2.8 ,2.9 and 2.10 is produce using
this part of the code

#just change the index value for sortFirst[1:x,]
and change the AuthorID

first = webWC[webWC$AuthorID == 8912,]

byfirst = aggregate(cbind(first$Analytic,
first$Clout, first$Authentic, first$Tone),
by=list(first$ThreadID, first$Month),
FUN=mean)

colnames(byfirst) = c("ThreadID",
"Date","Analytic","Clout","Authentic","Tone")


frequent =
as.data.frame(table(byfirst$ThreadID))

colnames(frequent) = c("ThreadID","freq")


byf = merge(byfirst, frequent, by =
"ThreadID")

```

```
sortFirst = byf[order(-byf$freq), ]

sortFirst$ThreadID =
as.factor(sortFirst$ThreadID)

row.names(sortFirst) = 1:nrow(sortFirst)

#data row 1 to 78 is the top 3 most active
author in the thread

pp = sortFirst[1:78,]

# "melt" data with the melt() function so that
each row is a unique id-variable combination.

#eg:

#mydata

'id        time     x1        x2

 1         1         5         6

 1         2         3         5

 2         1         6         1

 2         2         2         4'

#melt(mydata, id=c("id","time"))

'id        time     variablevalue

 1         1          x1        5

 1         2          x1        3

 2         1          x1        6

 2         2          x1        2

 1         1          x2        6

 1         2          x2        5

 2         1          x2        1

 2         2          x2        4'

# this is to transform our data so that we can
plot 4 plots together as shown in the figures

pdata = melt(pp,
id=c("ThreadID","Date","freq"))

ggplot(pdata,aes(x =
Date,y=value,colour=ThreadID,group=ThreadI
D)) + geom_line(size = 1.5) + geom_point() +
ggtitle('AuthorID: 8912') +
facet_wrap(~variable, nrow = 4)

ggplot(pdata, aes(factor(ThreadID), value, fill
= variable)) +

  geom_bar(stat="identity", position =
"dodge") +

  scale_fill_brewer(palette = "Set2")

########################################
####################################

###Ploting one thread one author

##we will be manually changing the AuthorID
and rerun the whole chunk of code to get the
graph desire

#remember to change the start varibale of the
time serires as well

#Figure 2.11 to 2.15 is produced using this
part of R code

compa = webWC[webWC$ThreadID ==
472752 & webWC$AuthorID == 166362,]

bycompa = aggregate(cbind(compa$Analytic,
compa$Clout, compa$Authentic,
compa$Tone), by=list(compa$AuthorID,
compa$Month), FUN=mean)

colnames(bycompa) = c("AuthorID",
"Date","Analytic","Clout","Authentic","Tone")
```

```
test = as.data.frame(seq(from =
min(bycompa$Date), to =
max(bycompa$Date), by='1 month'))

colnames(test) = "Months"

test$Analytic =
bycompa$Analytic[match(test$Months,
bycompa$Date)]

test$Clout =
bycompa$Clout[match(test$Months,
bycompa$Date)]

test$Authentic =
bycompa$Authentic[match(test$Months,
bycompa$Date)]

test$Tone =
bycompa$Tone[match(test$Months,
bycompa$Date)]

testTS = ts(test, frequency = 12, start =
c(2008,11))

plot(testTS[,2:5])

# only plot the time zone whereby the author
is active in that thread

timeC = full[full$Months %in% test$Months,]

timeCTS = ts(timeC, frequency = 12, start =
c(2008,11))

plot(timeCTS[,2:5])

########################################
#################
```