# Are the aspects of writing for domestic and international students invariant over time?

Clinton Ali

April 2019

## 1 Introduction

In 2016, Zenan Li conducted an analysis on data from The Writing Centre (TWC) located in University of Toronto - Scarborough (UTSC) for the 2014-2016 academic school years. It's purpose was to find patterns within the data from tutorials conducted in TWC. Li established three key findings in his analysis. First, Li established that international students are more likely to focus on sentence level concerns than their domestic cohorts. Secondly, it was also found that 62.5% of students who had multiple appointments do not change their focus to sentence level concerns. Finally, amongst the students with multiple appointments who do change their aspect of writing, international students were more likely to change their aspect of writing to sentence level concerns and domestic students were more likely to change from their aspect of writing that does not include sentence level concern [2].

However, based on 2016-2017 University of Toronto (UofT) annual enrollment report it was established that international students make up of about 20.5% of undergraduate students enrolled in all three campuses and it is slated to grow higher by the 2021-2022 academic school year [3]. This finding brings rise to a question of whether Li's key findings stays consistent with the evolving student body in UTSC.

This report aims to test whether aspects of writing for domestic and international students are invariant over time. This is done by replicating Li's analysis with data from the 2014-2018 academic year. Moreover, additional analysis will be provided by processing tutor comments using text mining methods.

## 2 Data

There are 6349 observations collected from TWC for the 2014 to 2018 academic school year. The data contains a mix of single and multiple visit. The features (Fig. 1) of the data are mainly categorical and textual which reflect the information each tutor gathers after an appointment. These categorical features can be divide into two, either it pertains to aspect of writing or student data. conveys what topics in writing the student tackled in an appointment. While the textual features can be separated into structured and unstructured textual data. The structured textual data were gathered to collect identifying information such as the tutor's name, course code and student status. While the unstructured textual data comes in the form of tutor comments.

*ID, Year, Term, Appt Date, Appt Time, Attended, Analytical Essay, Research Essay, Lab Report, Critical.Review, Literature Review, Annotated Bibliography, Outline, Proposal, Journal Entry, Personal Statement, Pharmacy Exam, Essay Exam, Other Type, Clarify Topic, Sounding Board, Planning.Paper, Formulating thesis, Developing argument, Flow Coherence, Clarity precision, Documentation, Grammar, Identify Errors, Suggest.handbooks, other.aspect, Course, Tutor, Gender, International, Program, Coop, Status, Year__1, Description*

Figure 1 - The entire set of features provided in the data

# 3 Preprocess

The categorical and structured textual data will be used in the in section 4.1 and 4.2, and it will be called model data henceforth. Similarly, the unstructured textual data will be used in section 4.4 and it will be called text data henceforth.

## 3.1 Model Data

The categorical features of the model data that pertains to aspect of writing (e.g. Clarify Topic, Sounding Board) were simplified into the feature "Aspect" that denotes whether the student focused on sentence level concerns (i.e. Include editing) or was not focused on sentence level concerns (i.e. Not include editing). Moreover, the text data that serve as student identifier were cleaned to a standard format. For example some tutor inputted room numbers in the tutor name and the room name needed to be removed.

After cleaning the data, some entries were removed according to the following criteria. First of all, appointments that were not attended were removed. Secondly, appointments conducted with tutors Nancy, Sheryl and Sarah were removed as their frequency was very low compared to the rest of the tutors. Finally, appointments that are not with undergraduate students were removed. The final number of observations are 3897 which is approximately 61% of the original data are leftover after this process. The text data were left untouched from this process as it requires a different method which haven't been covered by Li's previous work. The process of cleaning the text data will be explained in the next section.

Note: More information of how the cleaning was done on the model data can be found Li's report [2].

## 3.2 Text Data

The text data were processed by converting the text into tidy text format, where each row of entry corresponds to one word. Then punctuation and numeric characters were removed as they do not provide meaningful information. Moreover, entries that matched common stop words were removed from the observation. These words provide little meaning. Similarly, negation, modal or adverb words were removed because they added sentiment towards the comment where tutors were supposed to give objective comments. However, a domain specific set of stop words (Fig. 2) were also targeted to be filtered out. These were words that would appear very common amongst all the tutor comment. For example if the observation was about a student who needed help on their analytical essay then *analytical* or *essay* would appear at least once in the corresponding text data and therefore it was pertinent to filter for such words.

*essay, research, bibliography, annotated, analytical, critical, review, journal, entry, lab, report, literature, personal, statement, proposal, research, student, paper, thesis, assignment, draft, appt, 1pm, note, draft, twc, annotated, bibliography, critical, literature, statement, personal, journal, entry, discussed, discussing, review, reviewed*

Figure 2 - Set of self defined user stop words

Besides removing for stop words, stemming were used for the text data. Stemming is a process that reduces derivatives of similar words such as *try* and *tried* into a root word *try*. However, the downside of stemming is the loss of information. For example *universal* and *university*, which have two different meanings, will be stemmed into the same root *univers*.

Note: More information on text cleaning can be found in Julia Silge and David Robinson's literature [1].

# 4 Analysis

The analysis will be divided into three major section. The first section will be an exploration on the student distributions. The second section will be an analysis on modelling aspects of writing on the full data and change of aspect of writing on multiple visit data only. The third section will be devoted towards the text mining analysis.

## 4.1 Student Distributions

### 4.1.1 2016-2018 Model Data

There were five main observations that were found through the student distributions of the 2016-2018 model data.

| Nationality | Aspects | count | prop |
|---|---|---|---|
| Domestic | Include Editing | 414 | 0.3415842 |
| Domestic | Not include Editing | 798 | 0.6584158 |
| International | Include Editing | 341 | 0.5182371 |
| International | Not include Editing | 317 | 0.4817629 |

Figure 3 - Frequencies and proportions based Nationality and Aspect of Writing

First, there were more domestic students (65%) visited writing center than the international students (35%), while at the same time for both nationalities the more junior students, 1st and 2nd year, visit TWC more often than their senior counter parts, 3rd and 4th year. This is consistent to Li's previous findings (Fig 3 & 4). Second, approximately 59% of students who visited TWC are first and second years (Fig 4). Third, more domestic students focus on sentence level editing regardless of the year the student is in (Fig 3). Forth, amongst all international students the more junior students focused less on sentence level editing compared to their more senior peers (Fig 4). Finally, 52% of international students focused on sentence level editing regardless of year of study compared to their domestic counterparts (Fig. 3).

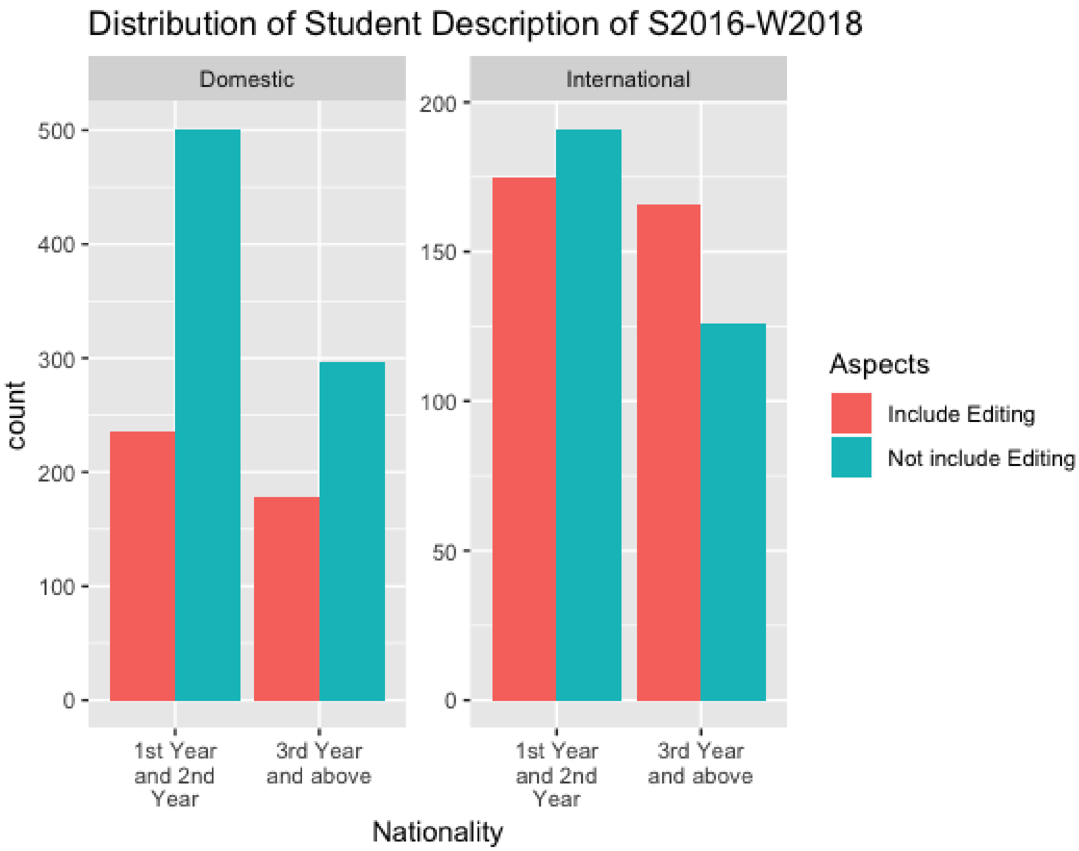


Figure 4 - Student distributions of aspect of writing for both nationalities and years of studies

### 4.1.2 2014-2018 Model Data

There were four main observations that were found through the student distributions of the 2014-2018 model data.

First, domestic students (68%) visited writing center than the international students (32%), and simultaneously the more junior students visited TWC more often than their senior peers (Fig. 5).

| Nationality | Aspects | count | prop |
|---|---|---|---|
| Domestic | Include Editing | 825 | 0.3132118 |
| Domestic | Not include Editing | 1809 | 0.6867882 |
| International | Include Editing | 637 | 0.5043547 |
| International | Not include Editing | 626 | 0.4956453 |

Figure 5 - Frequencies and proportions based Nationality and Aspect of Writing

Second, 50% of international students that visited TWC focused on sentence level editing than their domestic cohorts who only had 30% of appointments that focused on it (Fig. 5). Third, more domestic students focus on not include editing for both underclassmen and upperclassmen categories (Fig. 6). Finally, junior international students did not focus on sentence level editing when compared to their senior counterparts (Fig. 6). This is consistent to Li's findings.

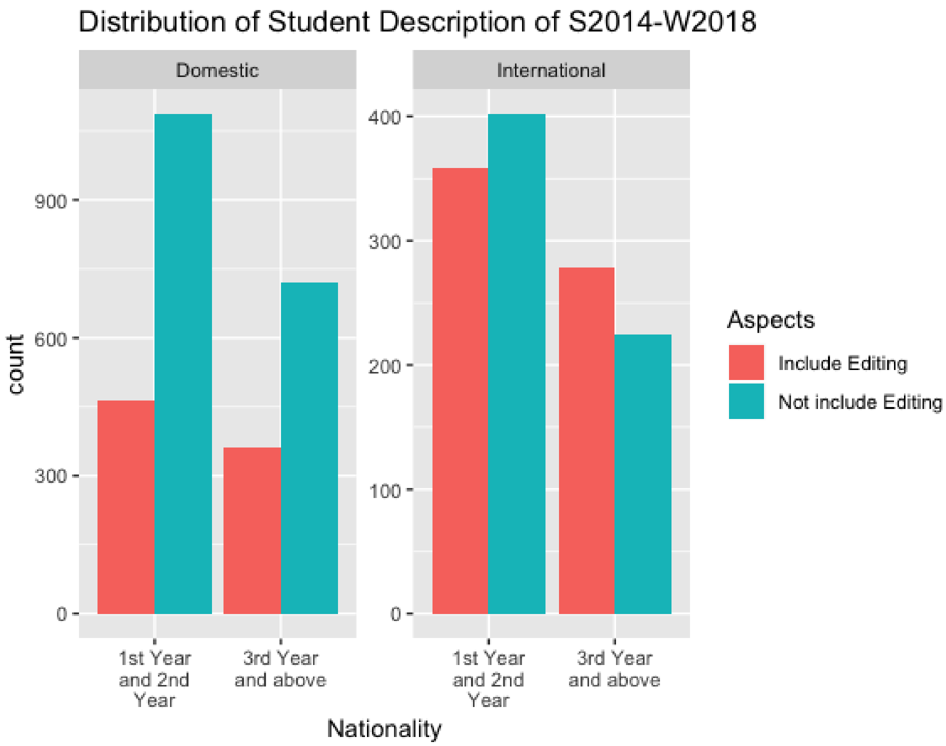


Figure 6 - Student distributions of aspect of writing for both nationalities and years of studies

Besides the distribution of the full data, a key analysis done in Li's work was with the multiple visit data. Out of the 3897 data points, 1912 entries were multiple visit and this accounts of almost half of the data. The purpose of this is to track changes in aspect of writing based on most representative visits of each student by each assignment. Out of the 1912 multiple visits, only 673 entries are left for the most representative visits and it is approximately 35% of the multiple visits. Furthermore, they were further pruned into two groups as follows:

- Type 1 are for students that their 1st visit does not include sentence level editing.
- Type 2 are for students that their 1st visit includes sentence level editing.

Type 1 and type 2 account for approximately 61% and 39%, respectively, of the changes in aspect of writing. Furthermore, for domestic students it was observed that (Fig. 7a):

- Probability that students shifted from not include editing to include editing: 28%
- Probability that students shifted from include editing to not include editing: 59%
- Probability that students not changing from not include editing: 72%

- Probability that students stayed on include editing: 41%

Also, for international students it was observed that (Fig. 7b):

- Probability that students shifted from not include editing to include editing: 51%
- Probability that students shifted from include editing to not include editing: 48%
- Probability that students not change from not include editing: 49%
- Probability that students stayed on include editing: 52%

Finally, for both students it was observed that (Fig. 7c):

- Probability that students shifted from not include editing to include editing: 34%
- Probability that students shifted from include editing to not include editing: 54%
- Probability that students not change from not include editing): 65%
- Probability that students stayed on include editing: 46%

| 7a | Include Editing | Not Include Editing |
|---|---|---|
| Include Editing | 57 | 81 |
| Not Include Editing | 86 | 217 |
| 7b | Include Editing | Not Include Editing |
| Include Editing | 66 | 61 |
| Not Include Editing | 54 | 51 |
| 7c | Include Editing | Not Include Editing |
| Include Editing | 123 | 142 |
| Not Include Editing | 140 | 268 |

Figure 7 Transition Matrix for (a) Distribution for domestic students
(b) Distribution for international students (c) Distribution for both domestic and international students

## 4.2 Logistic Regression

### 4.2.1 Aspect of Writing for 2016-2018

In figure 8, both year of study and nationality of students are statistically significant explanatory variables for aspect of writing. The logarithmic odds of focusing on sentence level editing an international student who are 3rd or 4th years will increase by approximately 1.08.

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | -0.8617 | 0.0865 | -9.96 | 0.0000 |
| Level3rd Year and above | 0.3545 | 0.1008 | 3.52 | 0.0004 |
| NationalityInternational | 0.7342 | 0.1009 | 7.27 | 0.0000 |
| Assignment_typeAnnotated.Bibliography | -0.7985 | 0.5913 | -1.35 | 0.1769 |
| Assignment_typeCritical.Review | -0.3815 | 0.1745 | -2.19 | 0.0288 |
| Assignment_typeJournal.Entry | 0.8121 | 0.3548 | 2.29 | 0.0221 |
| Assignment_typeLab.Report | 1.3182 | 0.2417 | 5.45 | 0.0000 |
| Assignment_typeLiterature.Review | -0.4017 | 0.2980 | -1.35 | 0.1776 |
| Assignment_typePersonal.Statement | 0.2599 | 0.4887 | 0.53 | 0.5949 |
| Assignment_typeProposal | -0.4875 | 0.2642 | -1.85 | 0.0650 |
| Assignment_typeResearch.Essay | 0.1769 | 0.1169 | 1.51 | 0.1301 |

Figure 8 - Logistic regression model for 2016-2018 academic year

### 4.2.2 Aspect of Writing for 2014-2018

In figure 9, both year of study and nationality of students are statistically significant explanatory variables for aspect of writing. The logarithmic odds of focusing on sentence level editing an international student who are 3rd or 4th years will increase by approximately 1.04.

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | -1.0293 | 0.0632 | -16.29 | 0.0000 |
| Level3rd Year and above | 0.2305 | 0.0704 | 3.27 | 0.0011 |
| NationalityInternational | 0.8113 | 0.0712 | 11.39 | 0.0000 |
| Assignment_typeAnnotated.Bibliography | 0.1984 | 0.2038 | 0.97 | 0.3302 |
| Assignment_typeCritical.Review | -0.1616 | 0.1282 | -1.26 | 0.2075 |
| Assignment_typeJournal.Entry | 0.4960 | 0.1913 | 2.59 | 0.0095 |
| Assignment_typeLab.Report | 1.2624 | 0.1804 | 7.00 | 0.0000 |
| Assignment_typeLiterature.Review | -0.1957 | 0.2404 | -0.81 | 0.4156 |
| Assignment_typePersonal.Statement | 0.3356 | 0.2822 | 1.19 | 0.2343 |
| Assignment_typeProposal | 0.0841 | 0.1676 | 0.50 | 0.6159 |
| Assignment_typeResearch.Essay | 0.2843 | 0.0816 | 3.48 | 0.0005 |

Figure 9 - Logistic regression model for 2014-2018 academic year

### 4.2.3 Tracking Changes in Aspect of Writing 2014-2018 for Multiple Visit Appointments

For both types of change (Fig. 10 & 11), it shows that nationality is the statistically significant explanatory variable. For type 2 change, an international student will have a 0.558 decrease in logarithmic odds compared to domestic students to change. For type 1 change, an international student will have a 1.1857 increase in logarithmic odds compared to domestic students to change. These results fall within a 5% confidence level of Li's previous work and thus is consistent.

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | 0.7134 | 0.3339 | 2.14 | 0.0326 |
| NationalityInternational | -0.5580 | 0.2573 | -2.17 | 0.0301 |
| times.visit | -0.1025 | 0.0779 | -1.32 | 0.1882 |
| Assignment_typeAnnotated.Bibliography | 16.0578 | 1385.3778 | 0.01 | 0.9908 |
| Assignment_typeCritical.Review | -0.0779 | 0.5603 | -0.14 | 0.8894 |
| Assignment_typeJournal.Entry | 16.2375 | 895.0346 | 0.02 | 0.9855 |
| Assignment_typeLab.Report | -0.2977 | 0.6185 | -0.48 | 0.6303 |
| Assignment_typeLiterature.Review | 0.8888 | 1.1826 | 0.75 | 0.4523 |
| Assignment_typePersonal.Statement | -0.1952 | 0.8513 | -0.23 | 0.8186 |
| Assignment_typeProposal | -0.1949 | 0.8478 | -0.23 | 0.8182 |
| Assignment_typeResearch.Essay | -0.2089 | 0.2814 | -0.74 | 0.4579 |

Figure 10 - Logistic regression model for type 1 change

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | -1.2108 | 0.2550 | -4.75 | 0.0000 |
| NationalityInternational | 1.1857 | 0.2361 | 5.02 | 0.0000 |
| times.visit | 0.0252 | 0.0609 | 0.41 | 0.6790 |
| Assignment_typeAnnotated.Bibliography | 0.2411 | 0.9663 | 0.25 | 0.8030 |
| Assignment_typeCritical.Review | -0.6789 | 0.4967 | -1.37 | 0.1717 |
| Assignment_typeJournal.Entry | 0.9707 | 0.8510 | 1.14 | 0.2540 |
| Assignment_typeLab.Report | 0.7911 | 0.8157 | 0.97 | 0.3321 |
| Assignment_typeLiterature.Review | 0.7240 | 0.7546 | 0.96 | 0.3373 |
| Assignment_typePersonal.Statement | 0.5676 | 1.0583 | 0.54 | 0.5918 |
| Assignment_typeProposal | -0.3878 | 0.6185 | -0.63 | 0.5306 |
| Assignment_typeResearch.Essay | 0.5193 | 0.2503 | 2.07 | 0.0381 |

Figure 11 - Logistic regression model for type 2 change

## 4.3 Text Mining

The analysis conducted for the text data will be by n-grams as it provides the easiest way to analyze information conveyed from a set of unstructured textual data. Also, the analysis will first explore the distributions of the text data and will move onto unigrams and bigrams. Trigrams and higher were not included for the analysis as the frequencies of valid combinations of words were greatly reduced by the text cleaning done earlier.


Figure 12 - Word clouds for (left) unigrams (right) bigrams

### 4.3.1 Distribution of Text Data

The number of words each tutor comments have follow a non central distribution (Fig. 13). Thus, median will be used to describe average compared to the mean henceforth.

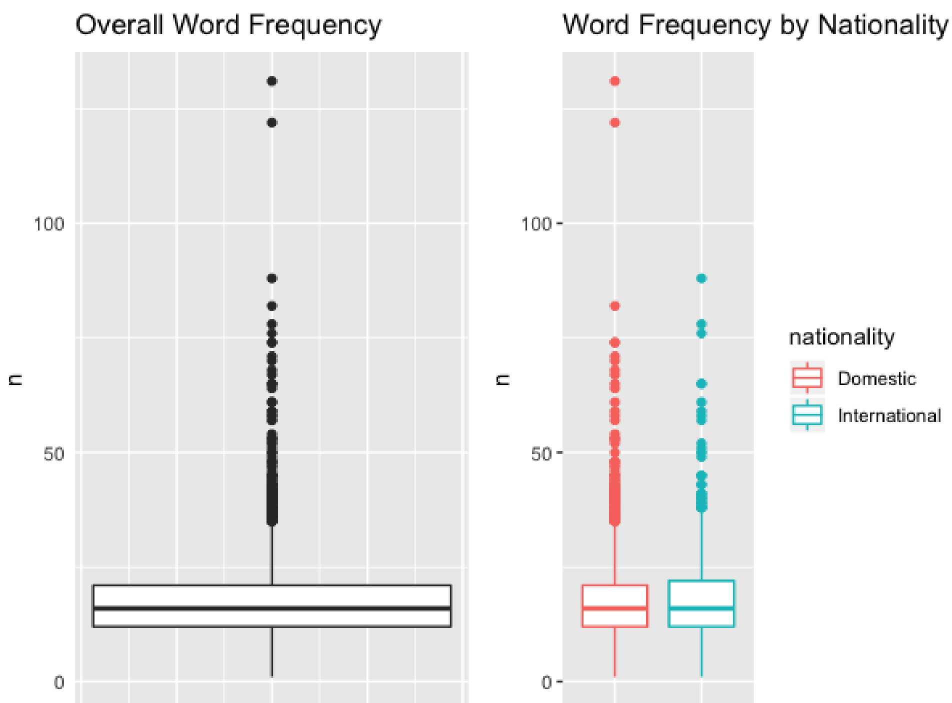


Figure 13 - left) Overall distribution of tutor comments
right) Distribution of tutor comments controlled by nationality

When controlled for tutors (Fig. 14), it can be observed that for the majority of observations each tutor have different average number of words as they all do not fall within the interquartile range of each other. When controlled for assignments, the average amount of words per assignment type vary greatly with each other. On the other hand, when controlled for nationality (Fig. 13) there is evidence that the majority of each groups observations are similar to each other. Therefore, the analysis of n-gram below will focus on domestic,international and both group of students.

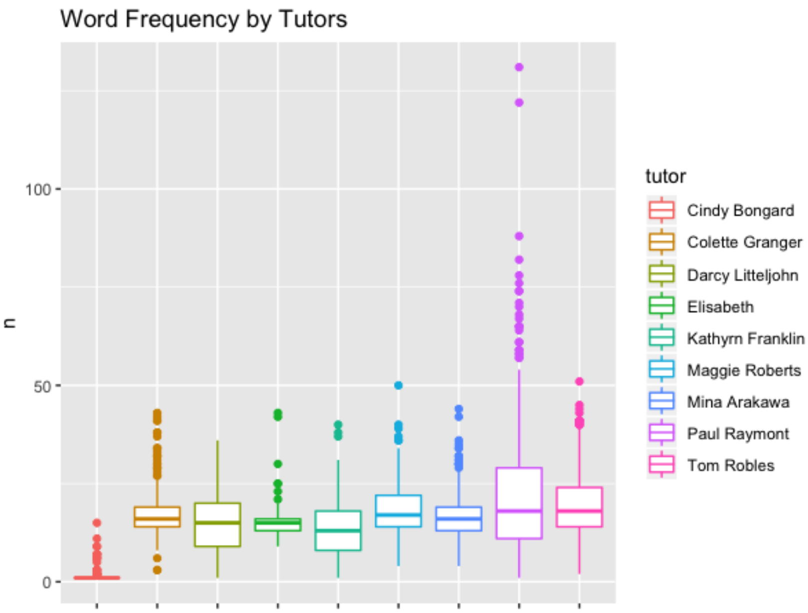


Figure 14 - Distribution of tutor comments controlled by tutors

| Overall | | | Domestic | | | International | | |
|---|---|---|---|---|---|---|---|---|
| word | n | prop | word | n | prop | word | n | prop |
| write | 859 | 0.0175884 | write | 587 | 0.0180094 | write | 272 | 0.0167436 |
| feedback | 655 | 0.0134114 | feedback | 416 | 0.0127631 | feedback | 239 | 0.0147122 |
| goal | 568 | 0.0116300 | goal | 405 | 0.0124256 | grammar | 186 | 0.0114497 |
| discuss | 549 | 0.0112410 | argument | 387 | 0.0118734 | discuss | 180 | 0.0110803 |
| research | 534 | 0.0109339 | discuss | 369 | 0.0113211 | research | 172 | 0.0105879 |
| argument | 518 | 0.0106063 | research | 362 | 0.0111063 | criteria | 167 | 0.0102801 |
| analysi | 500 | 0.0102377 | analysi | 356 | 0.0109223 | word | 164 | 0.0100954 |
| idea | 499 | 0.0102172 | idea | 347 | 0.0106461 | goal | 163 | 0.0100339 |
| question | 492 | 0.0100739 | question | 345 | 0.0105848 | import | 163 | 0.0100339 |
| import | 478 | 0.0097873 | import | 315 | 0.0096644 | articl | 160 | 0.0098492 |

Figure 15 - Top 10 unigrams for (left) all students (middle) domestic students (right) international students

### 4.3.2 Unigrams

Based on the top 10 unigrams (Fig. 15) it can be seen that comments made for domestic students are about on *argument* and *ideas*. These words account for 1.18% and 1.06%, respectively, of the entire pool of unigrams made for domestic students. On the other hand, comments made for international students had a focus on *grammar* and *word*. These words account for 1.14% and 1%, respectively, of the entire pool of unigrams made for international students. This can be interpreted as international students focusing more on sentence level editing than domestic students and also agrees with the analysis done previously.

Furthermore, since there are a greater number of comments made for domestic students influences the overall top 10 unigrams. The unigram *grammar* has been pushed away from the overall top 10 out of the four words that were selected earlier.

#### 4.3.3 Bigrams

Based on the top 10 bigrams (Fig. 16) it can be seen that there are about 37 more occurences of the *word choic* for international students than domestic students. On top of that, the bigrams such as *correct grammar* and *sentenc structur* shows the reoccurring theme that indeed international students focus more on sentence level editing. Similarly, domestic students focus on non sentence level editing as bigrams such as *bodi paragraph* and *topic sentenc* appear. Again this reinforces the notion that domestic students focus less on sentence level editing than international students. Additionally, the bigrams that dominate for both groups are *academ write* and *restrict unifi*. These are bigrams that relate to essay writing. Especially *restrict unifi* which is from a phrase "restricted unified precise" that is correlated with essay writing. Different than the unigrams, the top 10 overall bigrams generate a good mix of aspects of writing.

| Overall | | | Domestic | | | International | | |
|---|---|---|---|---|---|---|---|---|
| bigram | n | prop | bigram | n | prop | bigram | n | prop |
| academ write | 287 | 0.0234152 | academ write | 210 | 0.0262205 | word choic | 87 | 0.0204802 |
| restrict unifi | 184 | 0.0150118 | restrict unifi | 132 | 0.0164815 | academ write | 77 | 0.0181262 |
| unifi precis | 178 | 0.0145223 | unifi precis | 127 | 0.0158572 | restrict unifi | 52 | 0.0122411 |
| word choic | 147 | 0.0119931 | word choic | 60 | 0.0074916 | unifi precis | 51 | 0.0120056 |
| sentenc structur | 76 | 0.0062005 | bodi paragraph | 57 | 0.0071170 | correct grammar | 41 | 0.0096516 |
| bodi paragraph | 75 | 0.0061190 | main bodi | 50 | 0.0062430 | sentenc structur | 41 | 0.0096516 |
| main bodi | 66 | 0.0053847 | health studi | 46 | 0.0057435 | minor grammar | 33 | 0.0077684 |
| semi colon | 62 | 0.0050583 | sentenc structur | 35 | 0.0043701 | semi colon | 32 | 0.0075330 |
| correct grammar | 55 | 0.0044872 | topic sentenc | 31 | 0.0038706 | verb agreement | 30 | 0.0070621 |
| health studi | 54 | 0.0044056 | semi colon | 30 | 0.0037458 | subject verb | 29 | 0.0068267 |

Figure 16 - Top 10 bigrams for (left) all students (middle) domestic students (right) international students

# 5 Conclusion

Based on the student distributions and logistic regression model, international students focus more towards sentence level editing than domestic students (Fig. 8 & 9). This implies that Li's first key finding. Next, based on multiple visits, international students are more likely to switch and stay in sentence level editing than domestic students (Fig. 10 & 11). This agrees with Li's second key finding. Finally, there are about 66% of students who do not change to sentence level concerns and this agrees with Li's third key finding. Thus, it can be concluded that the model data agree with Li's key findings [2] and that it is invariant from the 2014 to 2018 period.

Besides the model data, text data were analyzed with the use of n-grams. Based on both unigrams and bigrams (Fig. 15 & 16) when controlled for nationalities, there are evidence that there are more occurrences of words that pertain to sentence level editing for international students than domestic students. Therefore, it can be concluded that international students do focus more on sentence level editing than domestic editing.

Overall, both analysis done on model and text data agree with the established notion that international students focus more towards sentence level editing than domestic students. Thus, the aspects of writing that domestic and international students are invariant over time.

# 6 Acknowledgements

I would like to say thank you to both Dr. Sohee Kang and Dr. Sarah King for offering guidance in this project. Also, I would like to say thank you to my colleagues Kamilla Kan and Veronica Fu for help in the analysis.

# References

[1] David Robinson Julia Silge. Text mining with r - a tidy approach, 2019.

[2] Zenan Li. Using existing writing centre data to explore how attention to sentence-level issues varies over time and with learner type, 2016.

[3] University of Toronto. Enrollment report 2016-17, 2017.