Machine Learning in the Academic Vocabulary Use in Chinese-English L2 Learners' Academic Writings

Gillian Bausch, Vaasu Taneja, Dhruvkumar Patel, Susan Liu,

1. Project Formulation

Research Problem

Academic success is essential to students. Yet many students experienced academic failure due to the underdeveloped literacy skills. Academic vocabulary acquisition causes widespread challenges to English as second language (ESL) learners (Lesaux, 2014; Schmitt, 1998; Schmitt,Jiang and Grabe, 2011). One of the major reasons is that there is little academic vocabulary teaching pedagogy with appropriate teaching materials in ESL classrooms to help these L2 learners to retain academic words. Consequently, for L2 learners of some languages such as Chinese, academic vocabulary are rather underrepresented in L2 learners' mental lexicon.

Research Question

Given the importance of academic vocabulary and L2 learners' needs on learning academic words, we conducted an exploratory study to seek for a comprehensive linguistic description on how L2 students use academic words in their academic writings. In our study, we focus specifically on L2 learners whose first language is remote from English, such as Chinese. We explore the following research questions:

1) What linguistic patterns can we identify in the academic word choice of Chinese-English L2 students' essays?

2) How well can Chinese-English L2 learners' word choice in their essays predict their language fluency?

Dataset

The present study aims to investigate the linguistic features of academic word uses by college-level Chinese-English L2 learners in their academic writings. The ultimate goal is to explore the Chinese-English 'L2 learners' usage-signature' of academic word uses in order to develop a vocabulary intervention curriculum in the future, which is tailored to the L2 students in developing academic literacy. To this end, this study will apply a machine learning techniques to analyze over a hundred argumentative essays (n=119) from Chinese-English L2 students (shown in Table 1) , whose language proficiency ranges from lower intermediate to advanced level (L1:lower intermediate; L2: higher intermediate; L3: fluent; L4: advanced fluent).

Table 1. Descriptive Statistics of Chinese English L2 students' argumentative writings

	Counts	Means (*M*)	Median (*m*)	Standard Deviation (*SD*)
Total number of words	61937	520.48	511	131.58
Total number of paragraphs	n/a	5.73	6	1.68
Total number of sentences	n/a	24.67	23	7.64

Note. *n = 119*

2. Methods

In this section, we report our methods in the sequence of data analysis processes.

Data preprocessing

First, we took the collection of 119 essays and performed natural language processing (NLP) to obtain a matrix containing features for the essays. Through NLP, we removed numbers and words with length less than five. Our list of features are unique words that appear in the essays, and the value of that feature for each essay will be based on word frequency counts. Eventually, we will end up with a 119 * 4899 features (unique words that appeared in the raw data) dataset.

Dimension Reduction and Feature Selection - PCA

Our data matrix had 4899 features so we used PCA decomposition to reduce the dimensionality of the data while retaining most of the variance. Our goal was to retain 90% of the variance as this was the generally recommended acceptable variance (Zaki and Wagner, pg. 197). By this point, each new principal component was accounting for less than 0.5% of variance too so it was a good place to stop. We ended up with 75 principal components meaning we reduced our dimensionality from 4899 to 75.

Unsupervised Learning to Identify Linguistic Patterns of Academic Word Use

To answer the first research question, we performed three cluster analysis, namely Kmeans, Hierarchical and Density-based (DBscan) Clustering. The processes for conducting the three cluster analyses are similar. For Kmeans and DBscan, we performed k-nearest-neighbour to obtain optimal value of centroid and epsilon. For DBscan, we used agglomerative single-link clustering. Then, we trained our models to gain the final clusters. We compared the results of the three methods and made our decision on the major method we would use to conduct further analysis.

With the overall dataset(119*4889) and cluster analysis results, we performed PCA on overall dataset and in individual clusters. The goal was to obtain and visualize the factor loadings of

each component. The factor loading scores can account for the pattern of the target word use.

For PCA in individual clusters, we also drew our conclusions based on 5 distinctive linguistic features (as shown in table 2). For each linguistic feature, we calculated the mean of the general population and the mean of the features in each cluster. Finally, we compared the overall mean of the population to the mean of the individual cluster.

Supervised Learning to Predict Students' Language Proficiency

1. **Support Vector Machines and Linear Discriminant Analysis**

   Using our data matrix, we used Linear Discriminant Analysis and a Support Vector Machine classifier and to predict students' language fluency based off of their essays. In our dataset, each student essay had a language fluency level associated with it: L1, L2, L3 or L4. For our analysis we group L1 and L2 levels together into lower level student essays and L3 and L4 together into higher level student essays.

2. **K-Fold Cross-Validation**

   To train our models and measure their success, we used 10-Fold cross-validation and calculated the average F-measure for both SVM and LDA. We then compared this average F-measure to the one for our randomized trials.

   The way the randomized trials worked is that instead of using the actual class values for each data point, we randomized their classes. So we would randomize the classes, run 10-fold cross validation on this random trial and get the average F-measures for SVM and LDA. We then repeated this process 100 times to get 100 different average F-measures for the randomized trial which we then averaged together to get our final randomized trial values.

3. Experimental Results and Evaluation

PCA: We reduced the dimensionality of the data from 4899 features to 75 features and retained approximately 90% of the variance.

Q1: Unsupervised Learning

Cluster analysis using three algorithms: Figure 1 shows the scatter plots of cluster analyses using the three algorithms. Table 2 shows the clustering results of each analysis. For DBscan, we use epsilon = 25.3 and the minimum points = 30. DBscan showed 11 outliers. We checked on the original data, and the 11 data points are not real outliers. Therefore, we did not remove them from the results. Kmeans and Hierarchical clustering gave very similar results to each other, except for a few data points clustered in different groups, which are shown in Table 2. We compared the results and decided to use Kmeans as the main clustering methods, which means the further analysis was based on Kmeans cluster results.

Table 2

Comparison of the Results Obtained from Three Cluster Analysis

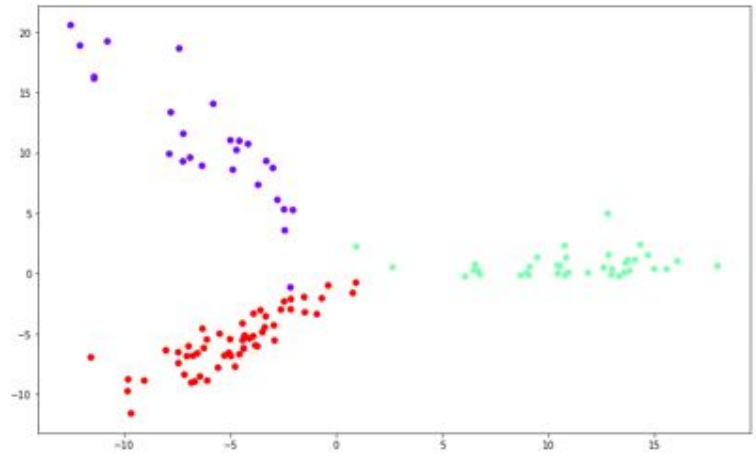
	N_CLUSTER	CLUSTER RESULTS	DIFERENCE
DBSCAN	1 cluster 11 outliers?	Outliers: [14, 20, 29, 39, 40, 46, 49, 64, 70, 101, 114]	
HIERARCHIAL	3 clusters	Cluster1: 26 Cluster2: 38 Cluster3: 55	
KMEANS	3 clusters	Cluster1: 24 Cluster2: 36 Cluster3: 59	No. 47,36 No. 15, 30

Figure 1 *Scatter Plots of Three Cluster Analysis*

DBscan:

Kmeans:

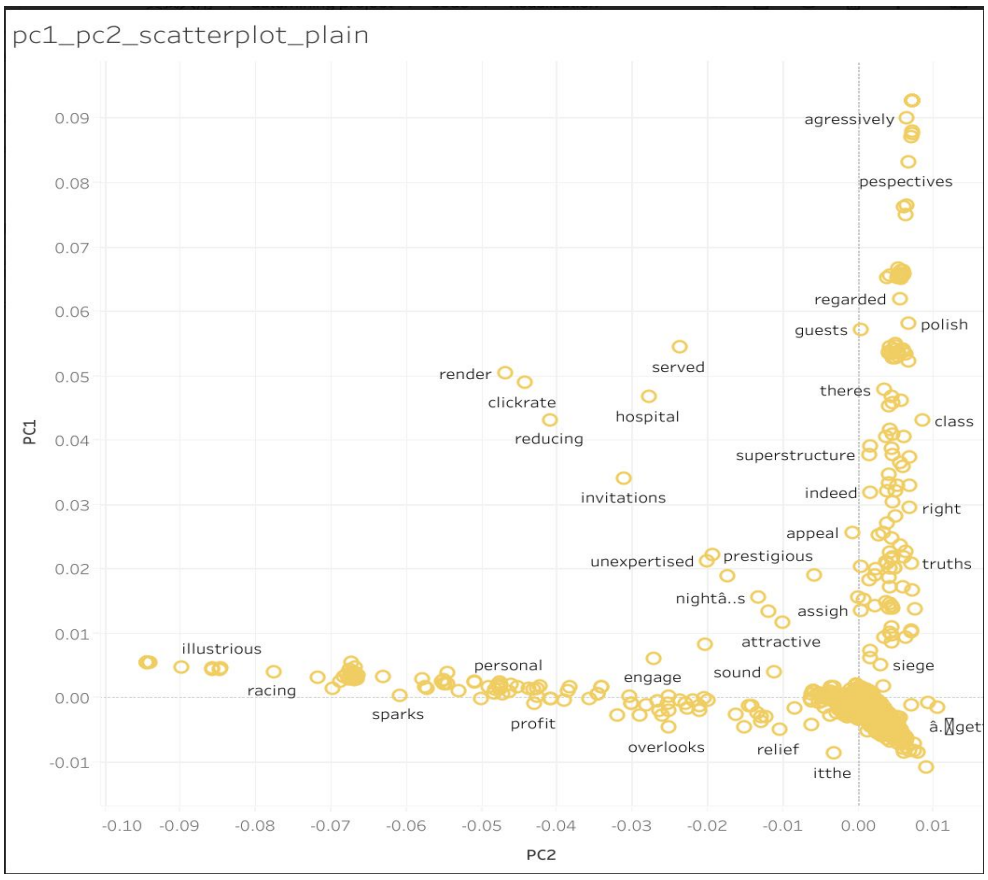
Hierarchical:



PCA analysis in overall dataset and in each cluster: The PCA analysis overall dataset and in each individual cluster shows factor loadings of each component. In this part of analysis, the factors in each component are individual words. As shown in the following Figures, Figure 2 illustrates the factor loadings of the overall dataset; Figure 3 to 5 are the biplots of each cluster. All of the plots have Dimension 1 and Dimension 2 from the PCA results as x-axis and y-axis. For Figure 3 to 5, the total variance accounted by Dimension 1 and Dimension 2 are over 50%. For Figure 2, the total variance is low (around 5%). However, it still can show examples of frequent word use by the L2 students. In Section 4, we will discuss the results in detail.

Figure 2

Plot of Factor Loadings Using Overall Dataset



PCA factor loadings for the individual cluster show one data point with extremely high factor loading (as shown in Figure 3 to 5). A close inspection presents that the high loading (MOOCs, Sleep, and Professors) points are referring to the topic of the prompt. Therefore, Figure 3 contains more students' writings on discussing the benefits of MOOCs; Figure 4 shows writings on talking about Sleep Deprivation; Figure 5 shows writings discussing if professors should be on TV.

Figure 3

Biplot of Cluster 3

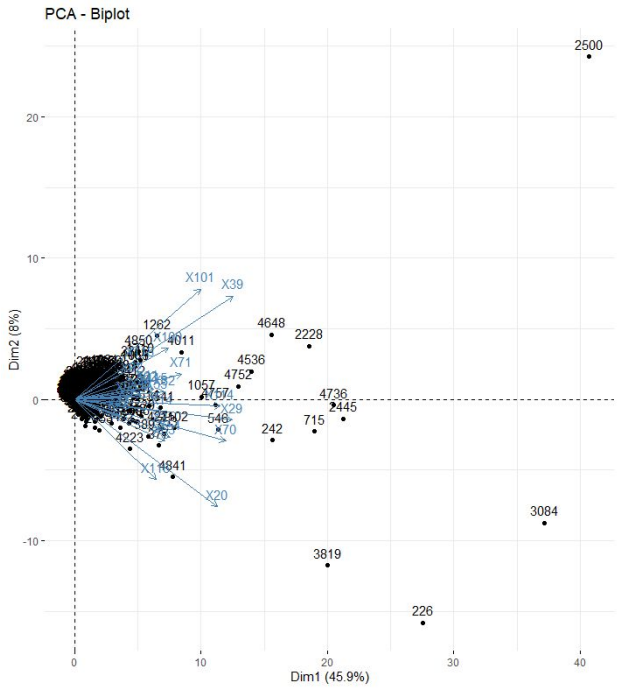


Figure 4

Biplot of Cluster 2

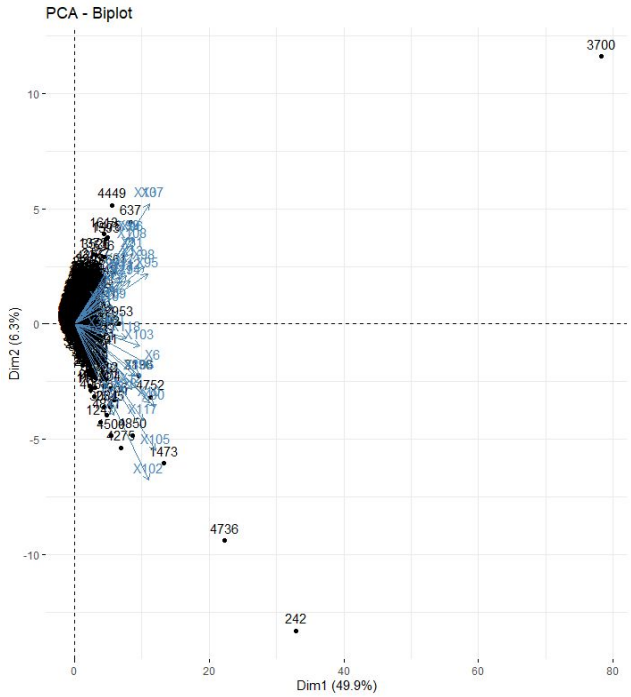
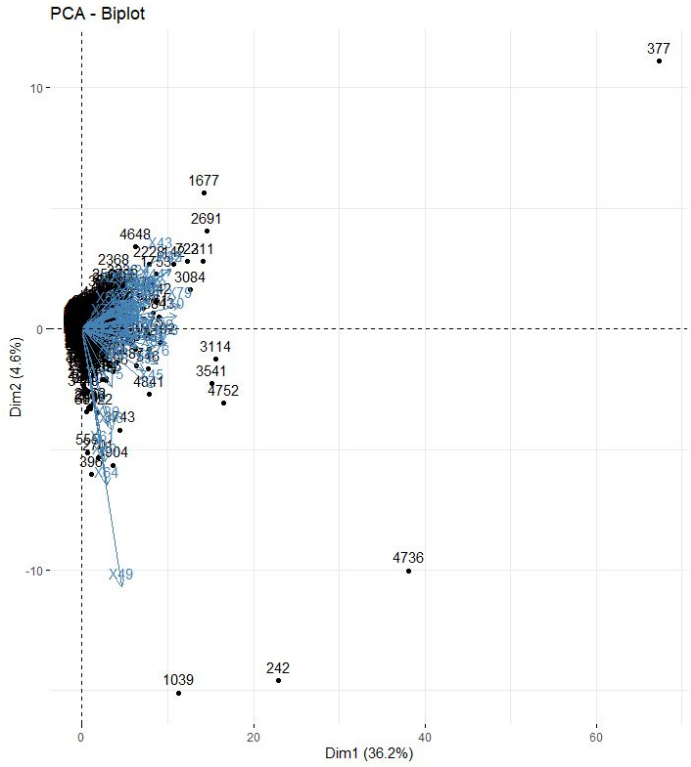


Figure 5

Biplot of Cluster 1



The results of mean comparison are presented in Table 3. The result shows that Cluster 1 has the most (3 out of 5) of the linguistic features above the means, whereas Cluster 3 has the most (4 out of 5) of features below the means. Combining the mean comparison result with topic clustering result, we can get more information on the students' writings and academic word use, which is discussed in Section 4.

Table 3

	μ	CLUSTER 1	CLUSTER 2	CLUSTER 3
PARAGRAPH	5.731	$> \mu, 6.02$	$< \mu, 5.5$	$< \mu, 5.391$
TOTAL WORDS	522.441	$> \mu, 530.278$	$< \mu, 508.139$	$< \mu, 517.217$
MORPH.COMP. WORDS	67.782	$> \mu, 69.759$	$< \mu, 65.333$	$< \mu, 66.304$
NOMINALIZATION	0.620	$< \mu, 0.616$	$> \mu, 0.627$	$< \mu, 0.616$
TTR	0.483	$< \mu, 0.0.477$	$> \mu, 0.487$	$> \mu, 0.487$

Q2: Supervised Learning

To evaluate our success, we consider the F-measures we calculated for our data and compare them to the F-measures we obtained from our randomized trials. If our F-measures are significantly better than the ones with randomized classes, it may be a good predictor. Essentially, the F-Score for the randomized experiment serves as a baseline to compare our actual F-Score to.

F-Score

	SVM	**LDA**
Actual Classes	**0.747**	**0.730**
Randomized Classes	**0.357**	**0.463**

From our results, we can see that SVM and LDA for the actual classes outperformed the randomized classes and were both above 0.7 demonstrating that the classifiers work well for our dataset.

4. Summary of the Results

Question 1:

From PCA factor loading in the overall dataset, we can draw 3 major conclusions:

1)In terms of morphology, Figure 2 shows that words with high loading scores are mostly derivational (word formed by adding prefix and suffix).

2)The words with bounded morphemes (e.g. aggressively, ability) are used more often than free morphemes (e.g. appeal,siege), which indicates that, on one hand,students might find word with bounded morphemes are easy to comprehend, since the meanings are more predictable than free morphemes; on the other hand, the result also indicates students' good mastery of word formation rules.

3)However, good mastery of word formation rules also results in violations in terms of semantics. For example, "unexpertised" has a high loading on Dimension 1, which is a pseudo word. Students acquire the formation rules and apply the rule universally, attempting to create new words which do not actually exist.

From the mean comparison, we found more linguistic patterns of each linguistic feature corresponding to the topic clusters:

1) When students discuss the topic regarding professors, they use more morphologically complex words than with the other two topics. However, their lexical diversity is relatively low. This is a very interesting finding, which needs some qualitative research work to identify the reason.
2) Nominalization indicates how many nouns that students use in their writings. There are studies showing that L2 learners have little knowledge of nominal cases (Bentz & Winter, 2014). The results of our study are to some extent aligned to the findings. Although Cluster 2 has the mean of nominalization above the mean, the statistic may not be significant. And all the other clusterings are all below the mean.

Question 2:

We found that SVM and LDA were successful for our data. This means that students' vocabulary in their essays is a good predictor of one's Chinese-English L2 language fluency. From this, we can conclude that considering one's academic vocabulary alone is sufficient in order to be fairly confident of one's level of fluency. In the future, it would be interesting to see how much better the models will perform if we also consider features relating to students' grammar and syntax.

Also, in the future it would be interesting to see how the results change for different datasets where students spoke different languages. For our dataset, all the students were Chinese-English L2 students, but it's possible that if the students' primary language were not Chinese or if they were not learning English, then classifying on academic vocabulary and word choice may be more successful or less successful. It would be interesting to see how our results change for different languages, and this could yield valuable insights into how certain languages should be taught compared to others..

References

Bentz, C., & Winter, B. (2014). Languages with more second language learners tend to lose nominal case. In *Quantifying Language Dynamics and Change*, 3(2017), 1-27.

Corson, D. (1997). The learning and use of academic English words. *Language Learning*, 47(4), 671

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. American Educational Research Journal, 51(6), 1159-1194.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26-43.

Covington, M. A., & McFall, J. D. (2008). The Moving-Average Type-Token Ratio (MATTR). *Linguistic Society of America*, *35*(8), 16-19.

Zaki, Mohammed J., and Wagner Meira. *Data Mining and Analysis : Fundamental Concepts and Algorithms* . New York: Cambridge University Press, 2014. Digital.

List of Packages

Bisong, E. (2019). Matplotlib and Seaborn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 151-165). Apress, Berkeley, CA.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(3), 90-95.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, *25*(1), 1-18.

McKinney, W. (2010, June). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In *Applied natural language processing: Identification, investigation and resolution.* IGI Global, 188-205.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Datamining Project Breakdown

We all made significant contributions to the project.
We listed the tasks that we worked on that we all did almost 100% for except for some other help we got from our teammates. We all analyzed results and discussed conclusions together so we did 25% of the work each for that. For the project report and powerpoint, Gillian and Vaasu wrote most of it, but Susan and Dhruv both reviewed it and helped with parts too.

Vaasu:
- Coded Python file to Build Word Matrix from directory of text files (filesToMatrix.py)
- Coded Python file for PCA (projectpca.py)
- Coded Python file for SVM and LDA analysis (part2.py)
- Defined research questions
- Wrote sections of Project Report
- Made sections of Project Powerpoint
- Analyzed Results and Discussed Conclusions

Susan:
- Modified filesToMatrix.py to improve the word parsing functionality (datamining.ipynb)
- Created visualizations for initial PCA clustering (pc1_pc2_scatterplot.twbx)
- Reviewed code for SVM and LDA analysis
- Reviewed and Edited sections of Project Report
- Reviewed and Edited sections of Project Powerpoint
- Analyzed Results and Discussed Conclusions

Gillian:
- Coded Python file for DBSCAN clustering analysis (DBSCAN-edited.py)
- Coded Python file for Hierarchical clustering analysis (Hierarchical Clustering.py)
- Code Python file and R for PCA analysis for each cluster obtained from Kmeans(component analysis-kmeans.R)
- Defined research questions and wrote introduction
- Wrote sections of Project Report
- Made sections of Project Powerpoint
- Analyzed Results and Discussed Conclusions

Dhruvkumar:
- Coded Python file for K Means clustering analysis (kmeans.ipynb)
- Reviewed code for DBSCAN clustering analysis (DBSCAN-edited.py)
- Reviewed and Edited sections of Project Report
- Reviewed and Edited sections of Project Powerpoint
- Analyzed Results and Discussed Conclusions