Lab 1 – CLASH Product Description Team Blue

Erich O’Saben

CS 411W

Professor Price and Professor Brunelle

March 4, 2015

Version 2

Table of Contents

[**1.** **INTRODUCTION** 3](#_Toc410998409)

[**2.** **CLASH PRODUCTION DESCRIPTION** 5](#_Toc410998410)

[2.1 Key Product Features and Capabilities 5](#_Toc410998411)

[2.2 Major Components (Hardware/Software) 7](#_Toc410998412)

[**3.** **IDENTIFICATION OF CASE STUDY** 10](#_Toc410998413)

[**4.** **CLASH PRODUCT PROTOTYPE DESCRIPTION** 12](#_Toc410998414)

[4.1 Prototype Architecture (Hardware/Software) 13](#_Toc410998415)

[4.2 Prototype Features and Capabilities 14](#_Toc410998416)

[4.3 Prototype Development Challenges 15](#_Toc410998417)

[GLOSSARY 18](#_Toc410998418)

[REFERENCES 20](#_Toc410998419)

List of Figures

[Figure 1. Major Functional Component Diagram 8](#_Toc413229889)

[Figure 2. CLASH Process Flow 9](#_Toc413229890)

[Figure 3. Prototype Major Functional Component Diagram 14](#_Toc413229891)

List of Tables

[Table 1. Real World Product vs Prototype Continued 13](#_Toc413229944)

1. **INTRODUCTION**

During the 2013-2014 academic year, the United States welcomed nearly 900,000 international students to its colleges and universities around the country. This is the largest quantity of international students ever enrolled in U.S. secondary education in a single school year (Haynie, 2014). While these numbers are record breaking, they are not surprising. Over the past 15 years, enrollment of international students into U.S. institutions has grown 72% and there is no sign of this number decreasing in the near future (Open Doors, 2014). International students are not only important to the diversity of a university, but they are also important to national and local economies. Between 2013 and 2014, international students contributed $27 billion to the U.S. economy (Open Doors, 2014), of which Virginia’s economy received $487.5 million (Institute of International Education, 2014).

The state of Virginia ranked 14th nationally in the number of international students enrolled in their higher education school system. A ranking of the top five Virginia colleges and universities with the highest number of international students found Old Dominion University (ODU) to be lacking in the total number of international students, with ODU failing to make the ranking of the top five universities and colleges (Institute of International Education, 2014). To be competitive in attracting international students, ODU needs to gain a competitive advantage in English as a Second Language (ESL) education.

Most traditional majors and curriculums have found innovative ways to leverage technology in the classroom. Some curriculums, such as ESL, still struggle to find ways to improve the student experience and help students be successful in their courses. Current software solutions designed to help students learn parts of speech (POS) or increase reading speed continually fail English Language Learners (ELLs) because they are not the primary intended audience of the tools currently on the market. For example, many current software tools that assist with learning POS are geared towards children, making the games childish and inappropriate for adults. Other products for learning POS are on the other end of the spectrum, geared towards users that are more technically savvy. The output of such tools are difficult to consume, even for a native English speaker. In addition, tools that are made to help with speed reading such as Spreeder (7-Speed-ReadingTM) break sentences into individual words so ELLs continue to learn to read word-by-word, which does not help in learning to read faster or with improving reading comprehension. Thus, the ELLs are at a disadvantage when they required to read college level material because more time and effort must be spent on required course readings.

Due to the short-comings of the products currently available, the 2015 CS411 Blue group has designed and will implement the Color Lexical Analysis algorithm and Slash Handler, or CLASH for short. CLASH was designed for use in college-level ESL courses with a college-aged end user in mind. It gives course instructors control over the reading materials selected and reports progress and metrics back to the instructor, to gauge the progress of each student.

(This space intentionally left blank.)

1. **CLASH PRODUCTION DESCRIPTION**

CLASH is a web-based, Software as a Service (SaaS), Single Page Application (SPA) that is being developed to accomplish three main goals. The first two goals are to assist ELLs improve their reading speed and reading comprehension, and the final goal assist ELLs in identifying the correct part of speech for all words in a sentence. CLASH will accomplish these goals by providing ELLs with two different ways to consume a document that has been assigned by their ESL instructor. The first way that ELLs will be able to view a document is through the SLASH Handler. The SLASH Handler will increase students’ reading speed and reading comprehension by teaching them to read text in Lexical Bundles. The second way that ELLs will be able view a document is through the COLRS module. The COLRS module will assist students in identifying the parts of speech by colorizing words within the document. Each part of speech will have a designated color so that the part of speech of a word will be easily identified without the use of a separate lookup table. A simple example to highlight this point is that all nouns could be colored blue and therefore whenever the user sees a blue word, they will know it is a noun.

(This space intentionally left blank.)

* 1. **Key Product Features and Capabilities**

The first module, called SLASH Handler, is aimed to improve the reading speed and reading comprehension of students by displaying text on the screen as individual Lexical Bundles. A Lexical Bundle is a group of words that repeatedly occur together within the same register (Hanel, 2011). Some common Lexical Bundles include “I don’t want to”, “in the United States”, and “at the end of.” Teaching students to read in Lexical Bundles reduces the number of fixations and regressions by the reader, allowing them to understand sentences by stringing together thought groups rather than by stringing together each individual word of a sentence.

The SLASH Handler provides a playback environment that displays a selected document in Lexical Bundles one at a time. The display is similar to the scrolling text on a teleprompter used by television hosts. The user will have the ability to start, stop, fast-forward, and rewind playback. The user will also be able to speed up and slow down playback to allow the student to practice at a speed that they are comfortable with.

The second module, called COLRS, will colorize the eight essential POS within the document to assist the student in learning the uses of those words. These eight POS (nouns, verbs, adjectives, adverbs, pronouns, conjunctions, and prepositions) will be displayed with a separate color for each POS and the user will be able to select which POS will be highlighted at any given time. A primary goal of COLRS is to ensure that the user is able view the output text in the original sentence and paragraph structure, making it easy for the user to consume.

CLASH will be built using a modern web-based architecture called a Single Page Application (SPA). A SPA is a highly responsive web application that fits on a single page and does not reload as the web page changes states. While SPAs can be built in many different programming languages, building a SPA completely in JavaScript is the most logical choice. JavaScript has become a very mature open source programming language and it rivals the computational speeds of traditionally compiled programming languages (Powell, 2014). There are now servers (Node.js), databases (MongoDB), and data formatting standards (JSON) that are all built-in and native to JavaScript. By utilizing these tools, an entire web application can now be built by using one programming language.

(This space intentionally left blank.)

* 1. **Major Components (Hardware/Software)**

Figure 1 is the Major Functional Component Diagram (MFCD) of CLASH, which highlights the major software and hardware components. CLASH is structured in two major areas, the Client Side and the Server Side. The main goal of the Client Side is to provide the user with a seamless experience to SLASH Handler and COLRs by providing the interface used to interact with the software. It will be written completely in JavaScript, while using some HTML and CSS as needed for additional text formatting.

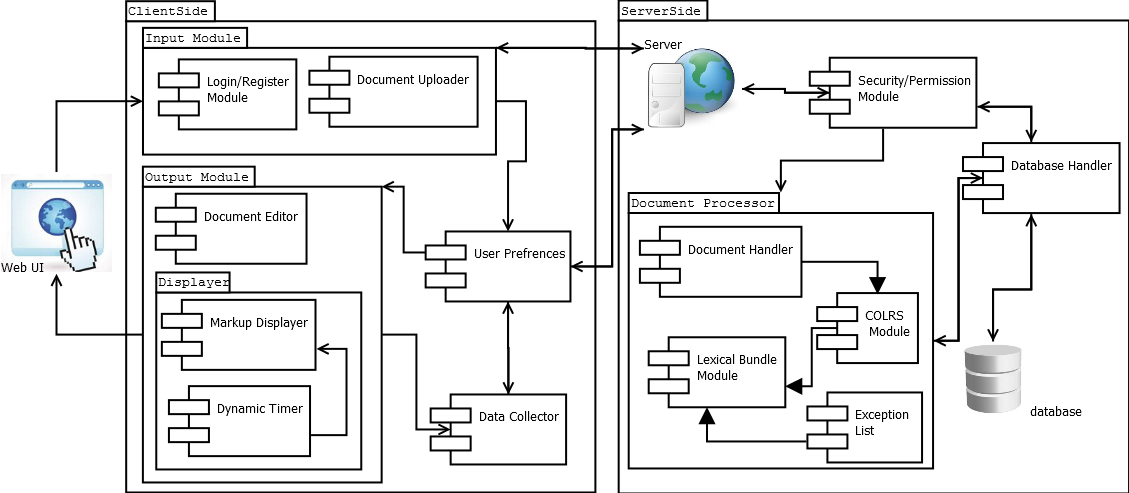


Figure . Major Functional Component Diagram

The Server Side is composed of three major elements. The first component is Node.js, which is the web and application server. Node.js is a prime choice for the application because it is also written completely in JavaScript and allows for synchronous communication with the Client Side. Node.js provides the functionality to allow the Client Side to change states without refreshing the webpage.

The next component on the Server Side is the database server. The database server that CLASH will utilize is MySQL. MySQL is an open source, production ready database server that will allow CLASH to scale both horizontally and vertically as more users and universities adopt the system. The database’s main function is to store user accounts, previously processed text documents, and user SLASH Handler performance metrics that can be reported to ESL instructors.

The final component of the Server Side is the Document Processor (DP). The DP is responsible for processing the text that will be viewed in the COLRs module and SLASH module. The DP takes the text entered by the instructor in the Client Side and parses the text into individual words, called tokens. Each token is then tagged with a POS using the open source Natural Language Processing (NLP) library, Natural Language Tool Kit (NLTK). Once all the words in the text are tagged with a POS, the text is sent through the SLASH module to build Lexical Bundles. Once the DP is complete, the document is saved to the database server and the final result is sent back to the client to be rendered in the SLASH Handler or the COLRs module. Figure 2 shows the process flow and lifecycle of how a document is entered, processed, and returned back to the user.

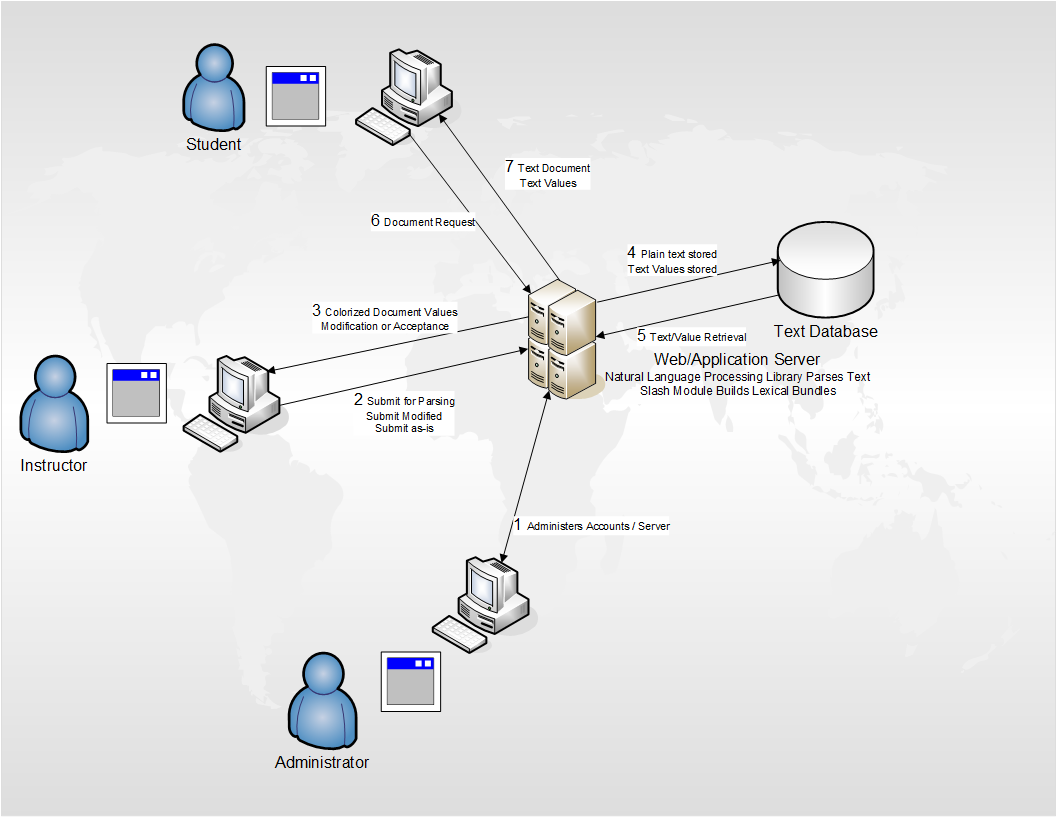


Figure 2. CLASH Process Flow

1. **IDENTIFICATION OF CASE STUDY**

Most colleges and universities require incoming international students to take the TOEFL test. The TOEFL test measures a student’s understanding of the English language at the university level (Educational Testing Service, 2015). While some students score high enough on the test to be granted full enrollment to a university, other students have the necessary academic record and a strong foundation in the English language; however, they require additional language training in order to be successful in an English speaking university. Students in the latter category can still be accepted for enrollment on a conditional basis pending they complete an Intensive English Program (IEP). According to a 2013 report, students who qualify for an IEP required on average 14.8 weeks of English language training in order to be successful in an English speaking college program (Institute for International Education, 2014). Based on available information posted on college and university websites, the typical IEP lasts seven to nine weeks meaning the average international student cannot successfully complete an IEP on the first try.

CLASH is being developed for Professor Raver-Lampman, who teaches ESL in the English Language Center at ODU. Professor Raver-Lampman found that the traditional method for teaching ESL students to read faster could be improved by teaching them to read in Lexical Bundles, which is how native English speakers read, usually without realizing it. At ODU, where the IEP is seven weeks, international student’s English skills are nowhere near their native English speaking student counterparts, but they are required to complete the same courses within the same time frame. What complicates matters more is that student visas are only valid for four years. Since many international students must first complete an IEC before they can begin a full-time degree, they must increase their course load to ensure they finish within four years. If an international student does fall behind in their studies and their visa expires, they must return to their country of origin to get the visa extended and then return to the U.S. to finish their studies on a full-time basis. Therefore, finishing the curriculum on time is paramount to an international student, and using CLASH to learn to read in Lexical Bundles will help them accomplish their goal by ensuring that the student can successfully finish the IEP on their first time through, and by providing the student the ability to read faster and better comprehend what they have read.

(This space intentionally left blank.)

1. **CLASH PRODUCT PROTOTYPE DESCRIPTION**

The CLASH prototype was designed to function as closely as possible to the real world product. However, given the time-frame provided to build CLASH, there were some concessions that had to be made. Although the prototype will not have all the features of the real world product, the prototype will be a highly functional proof of concept that can have additional functionality added in subsequent CS411W sessions. Table 1 on the following page provides an overview of the primary features of the prototype and real world product.

The biggest difference between the real word product and the prototype is that the prototype will not have the Homework Mode. The Homework Mode would allow instructors to assign readings to students to test their ability to identify POS within the document. Once the student completes the assignment, the instructor would receive a status report indicating how the student performed on the assignment. Given the time and complexity of building the Homework Mode user interface and backend document management, the Homework Mode will not be included in the prototype.

Another difference between the prototype and real word product is that the prototype will only be able to process plain text documents. The real world product would have the functionality to process a wide selection of document types such as plain text, MS Word documents, and PDFs. However, MS Word and PDFs have their own proprietary encodings which make it difficult for the NLP library and SLASH Algorithm to process, therefore making the complexity and time to implement too much for a single semester.

The final major difference between the prototype and the real world product is that the prototype will have limited reporting capabilities. The real world product would have a robust reporting suite of pre-built reports that the Instructor could use to track progress of their ELLs. However, this functionality is not essential to show a proof of concept, and therefore only limited reporting will be included.

|  |  |  |
| --- | --- | --- |
| **Features** | **Real World Product** | **Prototype** |
| Parsing Capabilities | Ability to Parse different kinds of documents | Ability to parse text copy and pasted in text block |
| Text Modification | Ability to modify and store previously parsed documents | Ability to modify and store previously parsed documents |
| Color Capabilities | Ability to color chosen parts of speech using a JSON format and JavaScript functions. | Ability to color chosen parts of speech using a JSON format and JavaScript functions. |
| Slashing Capabilities | Ability to identify Lexical Bundles through the inserting of slashes. | Ability to identify Lexical Bundles through the inserting of slashes. |
| Displaying Lexical Bundles in a single bundle form | Ability to speed up, slow down and pause Lexical Bundles being displayed. | Ability to speed up, slow down and pause Lexical Bundles being displayed. |
| Exception list | Lists of commonly used expressions that would otherwise be incorrectly handled by the SLASH Algorithm. | Lists of commonly used expressions that would otherwise be incorrectly handled by the SLASH Algorithm. |
| Login interface | User Authentication in a stand-alone environment | User Authentication in a stand-alone environment |
| Student Data Reporting | Tracks individual and collective student progress. To include words per minute, total time and total Lexical Bundles. Data to be stored in database. Displayed in graphs and statistics. | Limited basic student metrics will be available such as Lexical Bundles per Minute. |
| Homework Mode | Instructors have the ability to remove coloring of words and have students correctly identify the part of speech. | Not Included. |
| Administrative Privileges | Administrators are able to edit, add, or remove users and saved documents in the system. | Administrators are able to edit, add, or remove users and saved documents in the system. |
| SLASH Document Viewing Mode | Ability to view documents with slashes inserted and SLASH Handler. | Ability to view documents with slashes inserted and SLASH Handler. |

Table 1. Real World Product vs Prototype Continued

(This space intentionally left blank.)

* 1. **Prototype Architecture (Hardware/Software)**

The CLASH prototype will be built as a modified version of the SPA architecture. CLASH will differ from the traditional pure JavaScript SPA stack by using a relational database rather a NoSQL database. However, the user-interface will still be built in JavaScript and Node.js will still be used as the web and application server.

Figure 3 is the MFCD for the CLASH prototype. This MFCD is the same as the MFCD of the real world product in Figure 1 in section 2.2, however, the User Data within the database will be simulated. Since this is a prototype, CLASH will not be able to integrate with any ODU systems that are used to store student records. Therefore, the prototype will need to simulate this data to provide adequate student data to test with and to demonstrate the functionality of the basic reporting features.

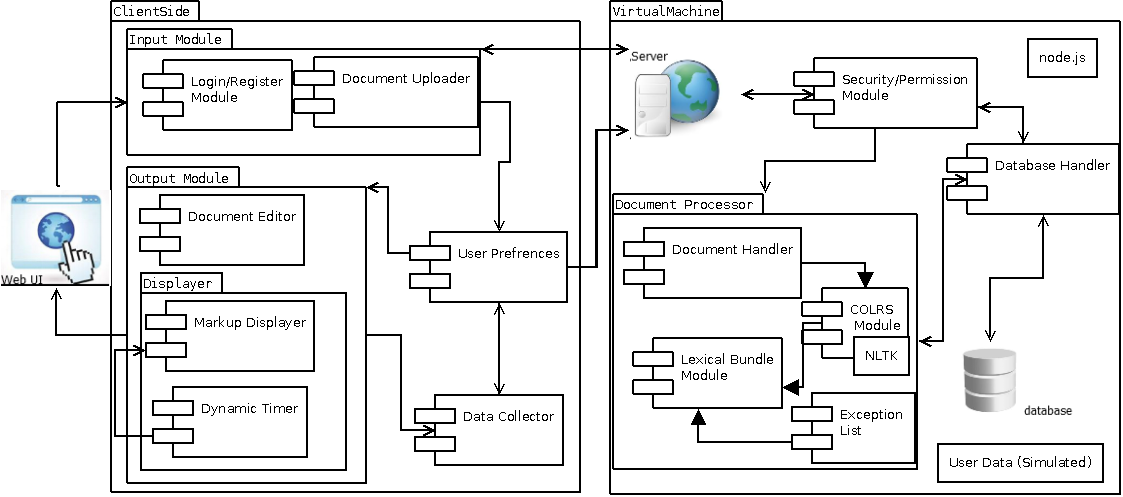
****

Figure 3. Prototype Major Functional Component Diagram

* 1. **Prototype Features and Capabilities**

The primary features of the CLASH prototype will be the SLASH Handler, the COLRS module, and the ability to allow an instructor to edit the text that was returned by the Document Processor. In order to successfully demonstrate these features, the CLASH prototype will also include a User module that will have three different user types Administrator, Student, and Instructor. Since the CLASH prototype will not integrate with any of the ODU systems, student user accounts will be manually entered by an Administrator user type or will be simulated in the database. These four features are the core functionality of the CLASH real world product and by successfully implementing these features it will demonstrate that CLASH is a viable product for the ESL department at ODU.

The SLASH Handler and COLRS module presented similar development risks. First, both features must be able to accurately identify the Lexical Bundles and parts of speech within the document. Not only do the Lexical Bundles and parts of speech need to be accurate, the Document Processor must be able to process the document efficiently so when the instructor enters the document, there will be minimal waiting time until the Document Processor finishes and the instructor can review the final version of the document.

Because of the complexity of the English language and the limitations of the NLP Library being utilized in the Document Processor, the CLASH prototype cannot guarantee that the Lexical Bundles will always be correctly formed or that each word will be properly assigned the correct part of speech. Since this is essential to the success of the prototype, CLASH will provide the Instructor user type with the functionality to edit the Lexical Bundles and the parts of speech of the words in the document before assigning the document to the student. This will ensure that the student will always have a document that is to the instructor’s standards of accuracy.

The time to process a long document by the Document Processor could take longer than what the average user is willing to wait. Not only is this unpleasant to the user, but it could also reduce user adoption of CLASH. Therefore, in order to mitigate this risk, CLASH was designed with the SPA architecture which provides synchronous communication between the Client Side and Server Side of the application. Therefore, this decreases the response time of the Document Processor to the Client Side. Additionally, to continue to mitigate this risk, CLASH will limit the size of the document to five pages, which is roughly 3,000 words.

(This space intentionally left blank.)

* 1. **Prototype Development Challenges**

The key development challenge that will be faced in the development of CLASH is the accuracy of NLTK and the SLASH algorithm. These two entities are key to the success of CLASH because any inaccuracy would result in students learning improper Lexical Bundles and POS. In order to mitigate the potential for inaccuracy, CLASH will have an edit mode for instructors to change a POS for a word and the content of a Lexical Bundle before publishing the document to their students.

Another development challenge is learning new technologies. The SPA architecture that CLASH will use requires the developers to learn and use state of the art web development technologies that are not taught in the ODU Computer Science undergraduate program. Therefore, team members must master the skills necessary to implement CLASH outside of their regular studies.

The last development challenge is project management. The 2015 CS411W Blue team is a merger of two CS410 teams resulting in 12 total members. Bringing together code developed by 12 people independently will be impossible without a clear and concise development plan. Additionally, extra time will need to be allocated to ensure team members stay on track and are developing their assigned piece of the project with the end result in mind.

(This space intentionally left blank.)

## GLOSSARY

**CLASH**: Color Lexical Analysis algorithm and Slash Handler.

**Client Side**: The user-interface of CLASH.

**COLRS**: Colored Organized Lexical Recognition Software.

**Document Processor:** A Server Side component responsible for processing the text entered by an Instructor user type.

**ELC**: English Learning Center at Old Dominion University.

**ESL**: English as second language.

**ELL**: English Language Learner.

**JSON**: JavaScript Object Notation. A nested data structure commonly used to pass data between a server and a client.

**Intensive English Program**:A short and intensive English language training program offered by US colleges and universities to improve the English language skills of international students who did meet the minimum TOEFL scores for typical enrollment.

**Lexical** **Bundle**: A group of words that occur repeatedly together within the same register

**MFCD:** Major Functional Component Diagram.

**NLTK**: A suite of libraries and programs for symbolic and statistical natural language

processing (NLP).

**Node.js**: Open source, cross-platform run-time environment for server-side and networking

applications.

**POS**: Part-of-Speech such as noun, adjective, verb, etc….

**Server Side**: The back-end of the CLASH system responsible text processing, the database, user-authentication, and web-hosting.

**SPA**: Single page application. A highly responsive web application that fits on a single page

and does not reload as the web page changes states.

**TOEFL**: English language proficiency test required by universities for enrollment for internationally based students.

**Software as a Service (SaaS)**: Software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet.

**Token**: Text that has been processed into individual words by the Document Processor

**VM**: Virtual Machine.

## REFERENCES

Educational Testing Service. (2015). *About the TOEFL iBT® Test*. Retrieved February 7, 2015, from ETS.org: http://www.ets.org/toefl/ibt/about?WT.ac=toeflhome\_ibtabout2\_121127

Hanel, S. D. (2011, March 5). *Lexical Bundles*. Retrieved February 7, 2015, from Communicating in English - An Internet-based, English Language Resource : http://sdhanel.com/corpuslinguitics/lexicalbundles.html

Haynie, D. (2014, Novemeber 14). *Number of International College Students Continues to Climb*. Retrieved February 6, 2015, from US News: http://www.usnews.com/education/best-colleges/articles/2014/11/17/number-of-international-college-students-continues-to-climb

Institute for International Education. (2014). *Intensive English Programs: Leading Places of Origin*. Retrieved February 7, 2015, from Institute for International Education: http://www.iie.org/Research-and-Publications/Open-Doors/Data/Intensive-English-Programs/Leading-Places-of-Origin/2012-13

Institute of International Education. (2014). *Open Doors Fact Sheet: Virginia.* IIE with support from the U.S. Department of State's Bureau of Educational and Cultural Affairs.

Open Doors. (2014). *Open Doors 2014: A 15-Year Snapshot.* Washington DC: Institute of International Education.

Powell, M. &. (2014). *Single Page Web Applications.* Shelter Island: Manning Publications Co.