Lab – 1 CLASH. Product Description

Blue/Purple Teams

Cory Morewitz

CS 411W

Janet Brunelle, Hill Price

Old Dominion University

March 23rd, 2015

Version 3

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

Based on state-reported data, in 2004, it is estimated that 4,999,481 English as a second language (ESL) students were enrolled in public schools. Despite those numbers 15% of these ESL students had no special resources or programs to help them learn the language. This problem is systemic. In 2001, in the states that tested ESL students in reading comprehension, only 18.7 percent of ESL students were assessed as being at or above the norm. That same year, 10% of ESL students in grades 7-12 were retained. In February 2001, it was reported that ESL students had dropout rates up to four times that of their native English-speaking peers. Here at ODU, there is an entire department focused on addressing this problem, but while they work diligently to help ESL students, the processes of teaching reading and grammar are outdated. (McKeon) ESL students go through a year and a half long Intensive English Program and must pass the TOEFL (an English language proficiency test) at the end of that time. (Raver Lampman) If better tools that increased comprehension and retention could be brought to market, those numbers will shrink.

CLASH, or Color Lexical Analysis algorithm and Slash Handler is a web interface that includes three modules, COLRS Slash and Slash Playback. The COLRS module colorizes each part of speech (POS) in a text document with a particular color to help increase comprehension of sentence structure and grammar. The Slash module parses text into lexical bundles (a group of words that occur repeatedly together, or present one single thought), or thought groups to help increase reading speed and comprehension. A lexical bundle is a group of words that occur repeatedly together within the same paragraph. The Slash Playback module will take the parsed text and play it back in a speed reader fashion. Though the program could be useful in a number of settings, CLASH is primarily concerned with ESL students.

Currently, the process is simple, for grammar, the professor writes a sentence on the board, then circles, or marks in some way, each part of speech. This is time consuming, and limits the amount, and size of the examples that the professor can give. Psychologically, it has been proven that color impacts learning, by relieving eye fatigue, increasing information retention, increasing productivity and accuracy, and supporting developmental processes. (Engelbrecht) Extrapolating from that information, the COLRS module should help students identify POS and increase their potential for learning grammar. For increasing reading speed and comprehension, reading assignments are given, or students are directed to sites like Spreeder. These sites only increase the reader's ability to read faster. Do to this focus on speed; they also teach students to read word for word.

2 PRODUCT DESCRIPTION

CLASH is a web application that will help professors teach ESL students POS and parse text into lexical bundles. The graphical user interface will be simple, easy to use, and contain three modules. The website will give the user the ability to highlight different POS to help delineate syntax as well as break text into lexical bundles. Additionally, the site will provide a reader application (Slash Playback) that will act as a speed reader using the lexical bundles provided as opposed to simply reading word for word.

The Slash module will parse documents into lexical bundles. This will help non-native speakers/readers transition from reading word for word to reading in thought groups. The slash module will also feed into a playback function that will allow for students to slowly increase the speed of their reading. The Slash Playback module will accept the input from the Slash module for display. Each lexical bundle will be displayed in order for variable lengths of time based on

words per minute set by the user. This will continue until the entire document has been played back to the user. It has been shown that those who learn to read in lexical bundles read faster and perform better in word and sentence recall experiments. (Tremblay, Derwing, Libben, Westbury)

The COLRS module will parse text, identifying POS and highlighting words that are assigned certain POS their respective colors (i.e. nouns will be red, verbs green, etc.). This will allow for instructors to easily display and therefore teach grammar. By beginning to associate certain words with their POS based on color, ESL students will then begin to recognize syntax more quickly and therefore increase their reading comprehension. It is a similar concept as highlighting, or underlining key information in a textbook. (Hoffman)

2.1 Key Product Features and Capabilities

CLASH will be a web-based application that will provide an easy to use interface. The CLASH site will be accessible through any standard browser with access to the Internet. Using the intuitive user interface, users will be able to display parsed text that is colorized, Slashed, or both. The Slash module will feed into a Slash Playback module that will display the lexical bundles. The COLR module will allow users to single out particular part(s) of speech to display so they can easily study one, or multiple concepts. For example, if they would like to see only the nouns, all other colors and POS will be hidden.

CLASH will have three different user roles, delineated by login: administrator, instructor, and student. The student will be able to view parsed documents, and interact with the Slash Playback module. The student will not be able to enter text, because CLASH is not infallible. There will be times when the part of speech is marked incorrectly in a particular scenario. So as not to lead to confusion, the students will only be able to view documents that have been

reviewed, edited and approved by instructors. The instructor will be able to do everything the student can as well as: input text, edit the parsed output in case of error, and save and delete documents from the server. The administrator will have all the previously mentioned functionality plus the ability to add or remove instructor permissions to particular users, and add or remove student access.

2.2 Major Components (Hardware/Software)

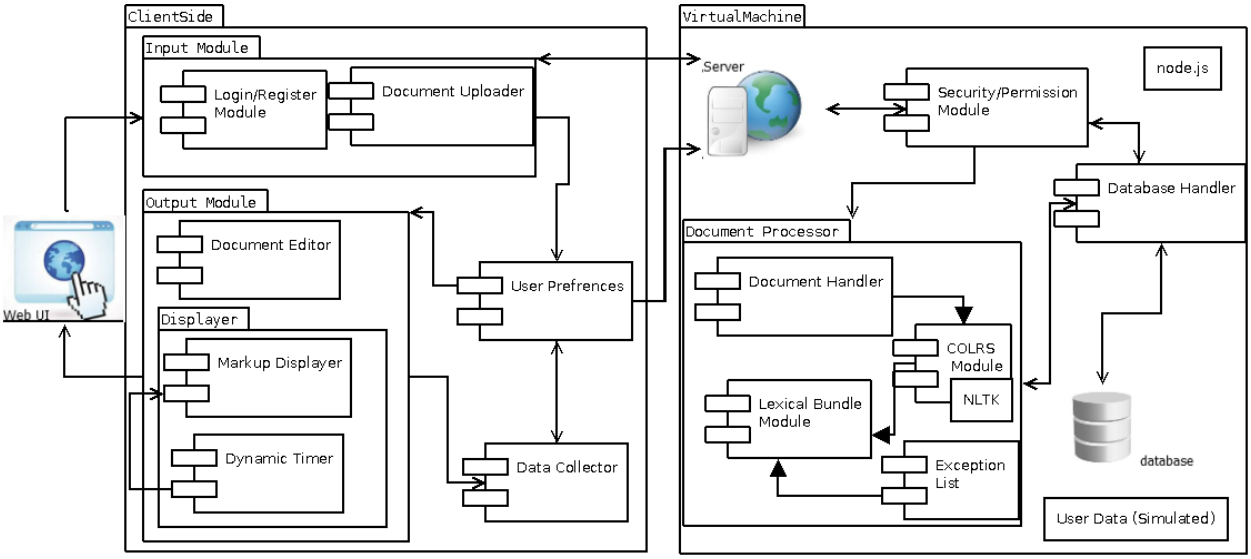

Figure 1. Major Functional Component Diagram Proof of Concept

Figure 1 illustrates the major functional components of the CLASH proof of concept, which is a web application accessed through a browser. There are no special hardware requirements, as CLASH is mostly a software product. The only hardware necessary is the machine the virtual environment hosting the CLASH site and backend.

The COLRS module will make use of a natural language processor (NLP) via python to parse text. This text will then be inserted into a .json file. The .json file will then be parsed by attribute, one of which is the part of speech. A JavaScript will run against the .json file and color each word with the same part of speech attribute tag and color it a matching color.

The SLASH module will take the parsed COLR .json file, and using the same attributes, will parse the text into lexical bundles. It will do this by following a few basic rules. First, the module will compare the document looking for "exceptions" set by the professor. This will ensure those lexical bundles on the "exception" list will not be split up. Then, it starts at the beginning of the document and adds a slash after each period, coma, semicolon, colon, or question mark. Next, the module will add a slash before each proposition and each conjunction.

On the client side, there will be a website which receives text from the user, and then parses it according to what the user requests. In reality, the entire thing is parsed completely, but what is shown to the user varies depending on what the user selects. The user will select from the options: COLR, SLASH, and SLASH PLAYBACK. This will show them colored text delineating parts of speech, the text parsed into lexical bundles, or a playback module. The playback module will be similar to a speed reader.

.

(This space intentionally left blank.)

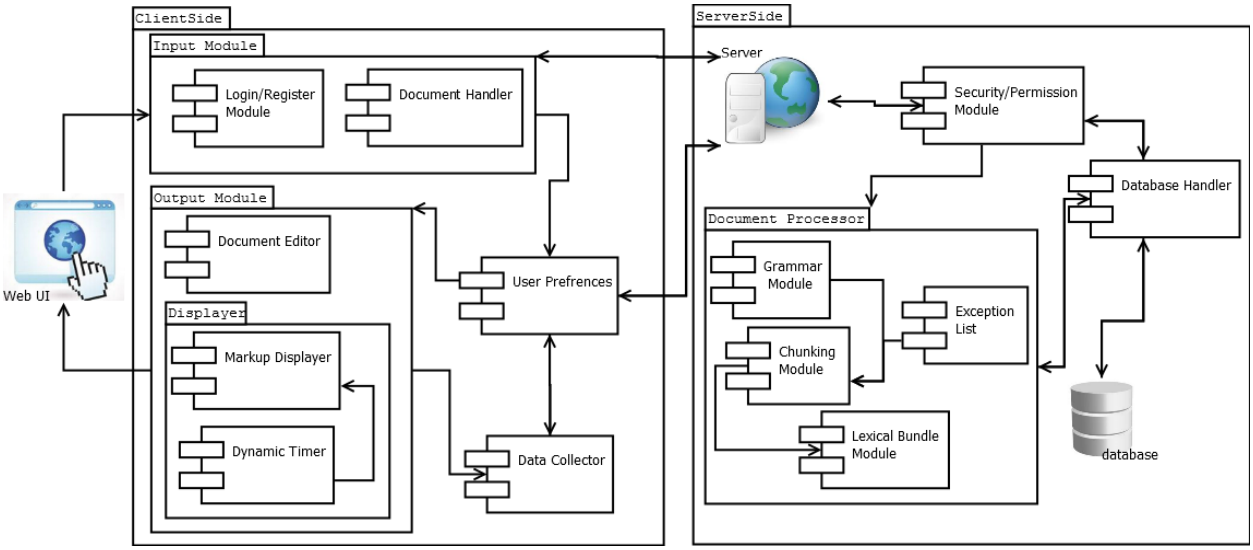

Figure 2. Major Functional Component Diagram Final

Figure 2 illustrates how the MFCD would be different in the real world application. There are very minor differences. For example, the database data will not be simulated. This will contain actual student and Professor data. Also, the real world product may not make use of node.js as it may make use of a native NLP, as opposed to NLTK.

3 IDENTIFICATION OF CASE STUDY

Professor Greg Raver-Lampman; an instructor at the Old Dominion University English Language Center hypothesized that coloring parts of speech and defining and displaying lexical bundles would increase the reading speed and comprehension of ESL students. Therefore, he tasked the CLASH team with making a product to prove his hypothesis. While this is simply a hypothesis, it is an informed one. Currently, he colors parts of speech and displays lexical bundles manually. This process, while effective is tedious, time consuming and prone to error.

(This space intentionally left blank.)

The CLASH web application will expedite this process, as illustrated in Figure 3. It will also allow the instructor to save his documents and reuse them from year to year. Eventually, reporting and tracking algorithms could be added to allow the instructor to track his student's progress and increases in reading speed. It will also better allow students to study on their own time via the website.

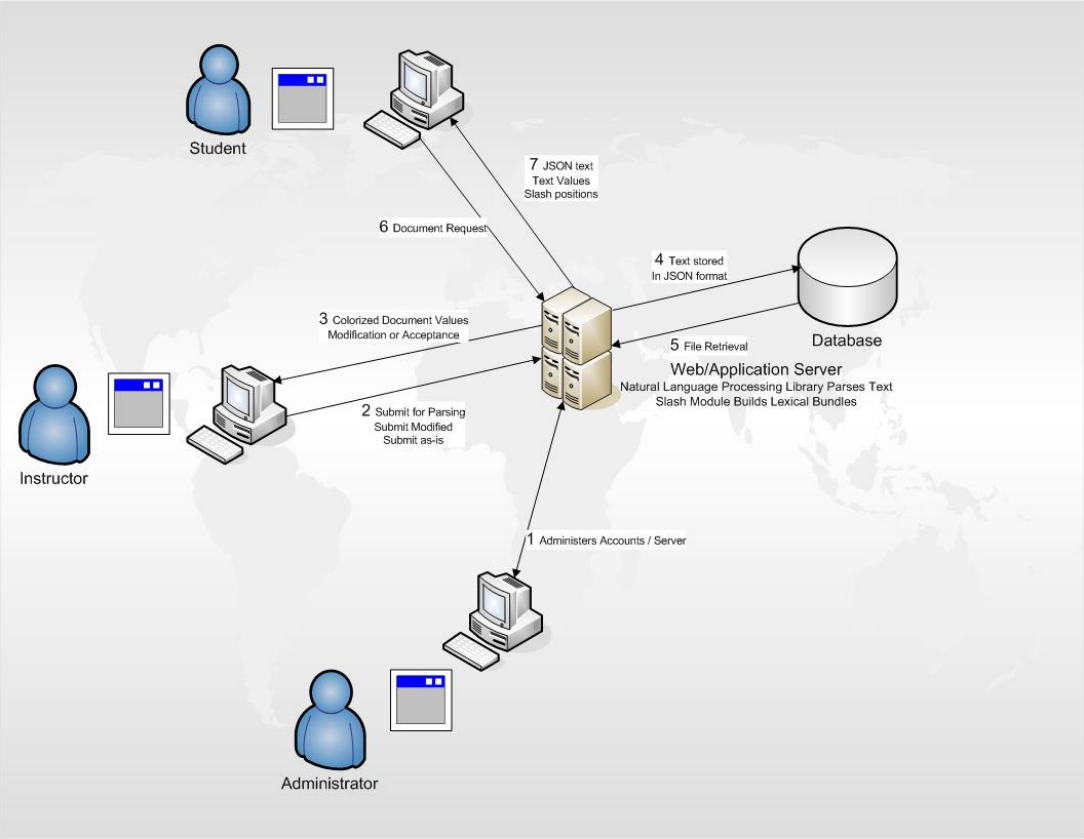

Figure 3. Future Process

4 CLASH PROTOTYPE DESCRIPTION

The CLASH prototype will be built as a single page application (SPA). This form of web application first on a single page and does not reload due to state changes. It will rely heavily on JavaScript as the primary programming language, as will most of the modules interacting with it.

It will be using Node.js, .json file formats, and a JavaScript interface. There will also be a python shell that interacts with the natural language processor and the JavaScript. Where CLASH will differ from traditional SPAs is it will use a relational database, as opposed to a NoSQL database.

There are many benefits to using a SPA. It is easy for the end user to access and use, no matter how limited their technical background may be. It is also beneficial because there is no software for the user to install. As there is no software to install, it also means the user can access CLASH anywhere he has an Internet connection. Finally, being a single page means that the user will not get 'lost' in a path of web pages. All functionality is accessible through the menus on the main page. There will however be a steep learning curve to implement SPA functionality, which should be assessed in the risk matrix.

(This space intentionally left blank.)

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 Playback.	Ability to view documents with slashes inserted and Slash Playback.
Text input	Ability to accept various text inputs (.docx, .pdf, etc.)	Will accept copy and pasted text and .docx

Table 1. Real World vs. Prototype

The prototype will share features with the real world product, as illustrated in Table 1. However, the prototype is merely a proof of concept. As such, some functionality may be

simulated, faked, or non-existent. For example, the COLRS module will not have homework modes for the student. Also the POS tagging will rely on a pre-built library as opposed to learning from the instructor editing the mistaken tagging. The prototype will also only be built for Old Dominion University. It will not provide functionality to expand the user base beyond campus. The real world product will be provided as software as a service (SaaS).

4.1 Hardware and Software Prototype Architecture

From a hardware perspective, as illustrated in Figure 4, there will only be a machine which will run a virtual machine to hold the server which will contain the database, and website files. The software used will be a collection of open source, free to use, and programs the CLASH team has written. The virtual machine will run Ubuntu 14.04 LTS, and the web and application servers will be Node.js. The Node.js server will interact with the MySQL database and the Natural Language Toolkit.

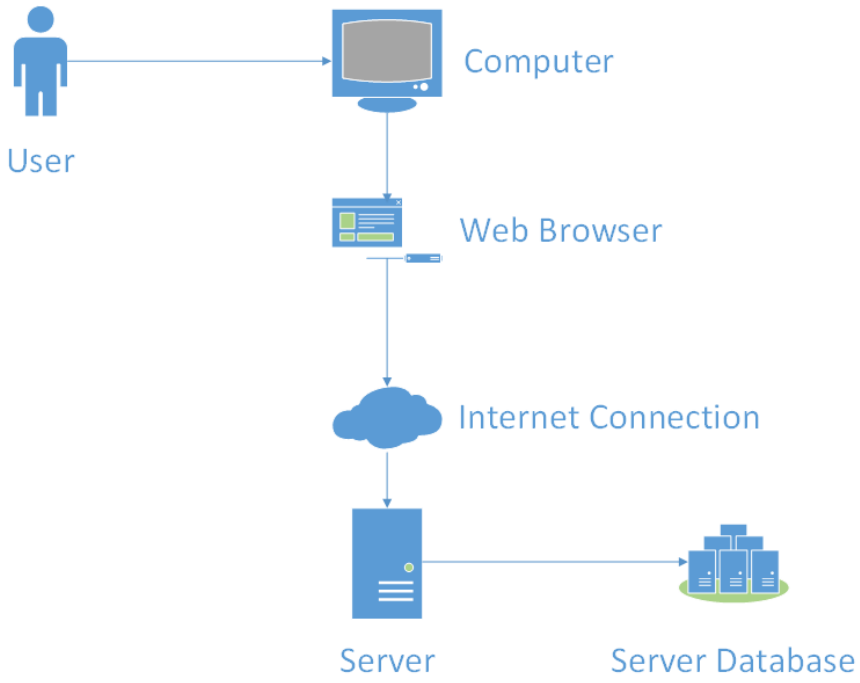

Figure 4. Prototype Hardware/Software Architecture

4.2 Prototype Features and Capabilities

The CLASH prototype will demonstrate the ability to identify specific parts of speech by passing text thru the natural language toolkit (NLTK) which will tokenize each word. Then, the COLR algorithm will color the text while the Slash algorithm identifies lexical bundles using slashes. The prototype will also allow users to display lexical bundles in a 'speed reader' type fashion. In addition the prototype will allow users to display bundles one at a time, increasing, decreasing the speed at which the lexical bundles are displayed, or pausing the playback altogether. Success of the prototype will be determined by the accuracy of the tagged parts of speech and lexical bundles.

Determined risks are to be mitigated through agile development, the involvement of the mentor throughout development of the product, and early and constant beta testing by potential users. Upson completion of the working prototype containing core functionality will result in the customer being able to test the product in a real world situation against a control group, further assisting the development of the product. By completing the proof of concept the program prototype we will be able to demonstrate the feasibility of such a system and together with the prototype testing group, the utility of using the product in an academic environment thus giving the customer the desired outcomes.

4.3 Prototype Development Challenges

There are many challenges being faced during the development of CLASH. The main challenges are scope creep, project management, and learning new software. Professor Greg Raver-Lampman has very high aspirations and great ideas for what CLASH could and should do. As such, we have a very definite problem of managing scope creep. Beyond scope creep, the team is comprised of many people. Attempting to manage a large group with little to no authority

is almost impossible. Keeping everyone busy and on task is a major undertaking. The team also faces the challenge of learning new programming languages and concepts such as .json and SPA. The program itself faces all the challenges inherent with anything concerning the English language. It is a very complex language, and we are merely scratching the surface with CLASH.

(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. Module of CLASH that colorizes part of speech based on what NLTK tokenizes and returns.

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.

ELL: English Language Learner.

ESL: English as second language.

GUI: Graphic User Interface

HTML: Hyper Text Markup Language

IBT: International Benchmark Test

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.

JS: JavaScript

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

Lexical Bundle: A group of words that occur repeatedly together, or represent a single thought group.

MFCD: Major Functional Component Diagram.

NLP: Natural Language Processing

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.

NoSQL: (often interpreted as Not only SQL) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

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

Python: a widely used general-purpose, high-level programming language.

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

Slash: Module of CLASH that Slashes text into lexical bundles and displays them

Slash Playback: Module of CLASH that displays a text stream showing one lexical bundle, of three to five words, at a time with the feature of speed control for display time. Speed reader that plays pre-slashed lexical bundles in the fashion of Spreeder.

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.

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.

Spreeder: Speed reading tool www.spreeder.com

TOEFL: English language proficiency test required by universities for enrollment for

internationally based students.

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

Ubuntu: a Debian-based Linux operating system

VM: Virtual Machine.

(This space intentionally left blank.)

REFERENCES

Anderson, N. (1999, April 1). Improving Reading Speed - Activities for the Classroom. Retrieved February 1, 2015, from http://dosfan.lib.uic.edu/usia/E-USIA/forum/vols/vol37/no2/p2.htm

Engelbrecht, K. (2003, June 18). The Impact of Color on Learning. Retrieved February 25, 2015, from http://sdpl.coe.uga.edu/HTML/W305.pdf

English Proficiency. (2015, February 2). Retrieved February 6, 2015, from https://www.odu.edu/content/odu/admission/proficiency.html

Hoffman, D. (n.d.). Academictips.org - Reading and Highlighting Tips. Retrieved February 25, 2015, from http://www.academictips.org/acad/literature/readingandhighlighting.html

Improve Reading Speed and Comprehension. (2006, January 1). Retrieved February 25, 2015, from http://spreeder.com/

Lab 1 CS 410 Team Blue

McKeon, D. (n.d.). Research Talking Points on English Language Learners. Retrieved December 11, 2014.

Mikowski, M., & Powell, J. Single Page Applications. Manning Publications 2014.

Monarch Diversity. (n.d.). Retrieved February 6, 2015, from https://www.odu.edu/admission/international/global

Nishida, H. (2013). The Influence of Chunking on Reading Comprehension: Investigating the Acquisition of Chunking Skill. *THE JOURNAL OF ASIA TEFL, Vol.10*(No. 4), Pp. 163-183.

Raver-Lampman, Greg. (2014. August). Personal Interview.

The Condition of Education 2014. (2014, January 1). Retrieved February 6, 2015, from http://nces.ed.gov/fastfacts/display.asp?id=96

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011, January 15). Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks. Retrieved December 10, 2014.

(This space intentionally left blank.)