Running Head: Lab 1 - CLASH PRODUCT DESCRIPTION

LAB 1 - CLASH Product Description

James A. Ward

CS411

February 9,2015

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

CLASH, or Color Lexical Analysis algorithm and Slash Handler is a software tool which automates two manual processes presently used in English as a Second Language (ESL) courses. The software interface is a web-page incorporating two components. The first component 'COLRS' examines a section of text and denotes each word's part of speech by color coding all nouns as one color, verbs as another color, prepositions as a third color et cetera for all remaining parts of speech. The second component is a slash reading module. The slash reading module examines the structure of a sentence, decomposes it into individual thought groups. In this domain thought groups are referred to as lexical bundles. The web-page then displays those lexical bundles one by one in the original order without the surrounding text. Both of these processes are currently manual operations performed by the ESL course instructor. The present method due to being manually performed, limits the number of examples available to the students. The colorizing process is achieved with a different color pen or font color and cannot be quickly shifted to demonstrate separate subsets of the parts of speech. Other similar methods used include circling or underlining various parts of speech in an example sentence. The slashing process is accomplished by marking a document with slash marks surrounding each lexical bundles within a sentence. Exempli gratia – “The students/ understood incrementally more/ of English grammar/ from day to day.” Old Dominion University ESL instructor Greg Raver-Lampman has stated the purpose of these processes are to respectively, improve the comprehension of English sentence structure and accelerate reading speed.

2 PRODUCT DESCRIPTION

The CLASH software package through automation and preservation of slash reading and

color coded parts of speech examples give ESL instructors time to cover more content or provide individual attention; it concurrently gives students greater opportunity for practice of the subject matter. The primary domain of the CLASH software tool necessitates various design criterion of less importance in other domains. The software being intended to provide expanded corpus of reading examples, requires that the students be able to access the system outside of the classroom. This implies a web based system. Using a web based system allows for access away from the classroom and allows the use of JavaScript eliminating any need for additional software installation. Similarly, as the students are English language learners the interface must be overtly sparse and use intrinsically non-technical terms and phrases such as “enter” or “access your assignments” rather than “login”. The target domain's administrators being subject matter experts in the fields of multilingual education or the English language also benefit from these design considerations. The following subsections of the product description section detail the design criteria of a completed product. Section 4 - Product Prototype Description, describes the functional components and limitations of the proof of concept prototype.

2.1 Key Product Features and Capabilities

The CLASH software package shall be accessed by all users through a web browser. The user interface presented alters by user role. User roles include Administrator, Instructor, and Student. Each user account shall be configured to only permit the use of system functions applicable for their user role or those of lower level accounts. Administrator accounts are permitted to access all features of both lower level account types. Instructor accounts are likewise permitted to access all features of student accounts. The primary purpose of administrator accounts is to create other accounts and maintain the system. Instructors' accounts exist to load reading content to into the

system, review the output for correctness and preserve the output for use in the classroom or reading assignments. Reading content provided to the system shall be text typed or pasted into a form on the web-page or a file uploaded to the system. File formats accepted are limited to Word (.doc or .docx), Portable Document Format (.pdf), Open Document Text (.odt), plain text (.txt). It shall be the instructors responsibility to ensure no copyright infringements or violations of fair use laws occur. Instructors may also input enrollment lists to generate new student accounts. Accounts may be archived en mass at the end of term or individually upon withdrawal from the ESL program. Tracking of a single students' or a defined student group's average reading speed improvement shall be available within the instructors view of the web-page. The instructor screen provides an interface for corrections to automatically generated POS marking and Slash locations. The student interface permits reading of preconfigured assignments in either slashed or color coded mode. In the slashed stream mode the student has the ability to start, stop, or increase or reduce playback of the lexical bundle stream. In the slashed paragraph mode the student would see the reading samples as shown in the Introduction to this paper. The student interface also includes the ability to attribute the parts of speech to uncolored texts for graded comparison with the saved version prepared by the instructors.

THIS SPACE INTENTIONALLY LEFT BLANK

2.2 Major Hardware and Software Components

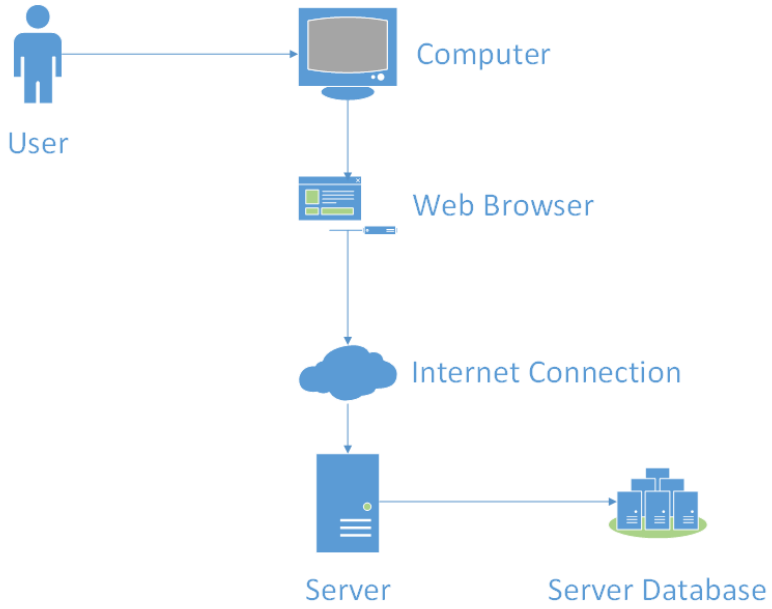

Figure 1: Hardware Requirements

As depicted in Figure 1 there is no special hardware requirement for the CLASH software package. A web server is required, it may be a private server or a rented server such as Amazon.com Incorporated's EC2 cloud based systems. The system is designed as a Single Page Application (SPA) hosted on an Ubuntu Server running MySQL, NGINX, and Node.js. The entire application is written in JavaScript with specific features (delineated below) delegated to other software packages. The NGINX software server is the initial server on the website it redirects web browser requests to the Node.js server. The server.js file is run under Node.js and serves the application to the users web browser. Node.js maintains an open connection to the browser allowing the SPA to render an updated view as required by the interaction with the user. It is not the purview of the development team to perfect the complex task of natural language processing. Decomposition of the input text is passed to the python based software Natural

Language Tool Kit (NLTK). NLTK provides the breakdown by part of speech for the text and returns it to the server.js application. The application then re-parses the NLTK output into a JavaScript Object Notation (JSON) data structure format. This JSON data structure passes through re-parsers to incorporate the color coding values and slash locations. The re-parses do include an exceptions list for phases that the slashing algorithm could misinterpret. This list can be updated by the instructor when additional exceptions are observed to be repeatedly incorrectly handled. The instructor then reviews the automated output for correctness and preserves the sample for future use. Upon selecting “preserve sample” the SPA archives the instructors sample into the MySQL database running on the same server as the SPA. Figure 2 is a simplified diagram of the process flow described above.

THIS SPACE INTENTIONALLY LEFT BLANK

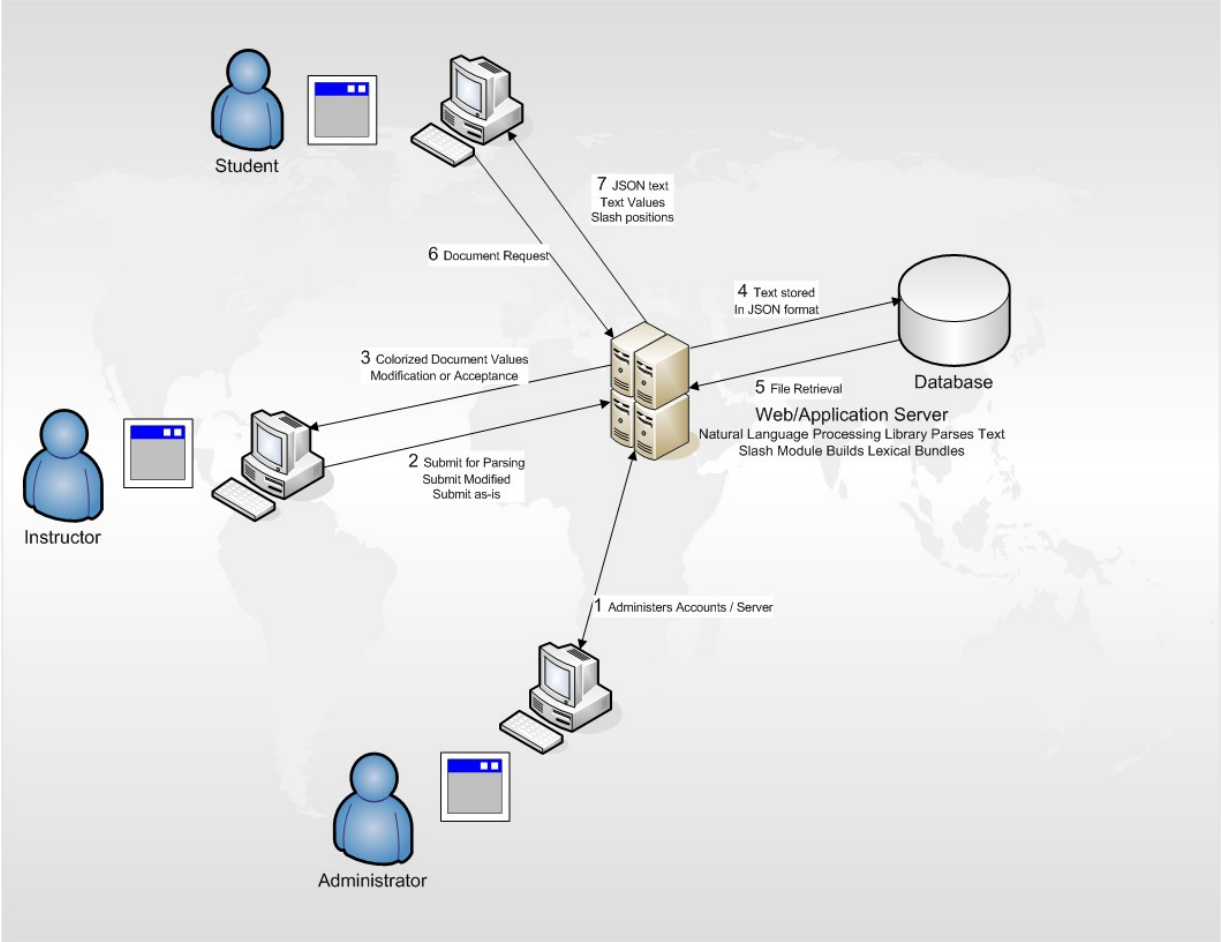

Figure 2: Process Flow

3 IDENTIFICATION OF CASE STUDY

Development of the CLASH software package stemmed from discussions with Mr. Greg Raver-Lampman of the English Language Center (ELC) at Old Dominion University. The ELC shall serve as the test bed for the continued development and enhancement of the software. The software's market potential however is incredibly diverse. According to a study performed in 2005 there were an estimated 5 million ESL students enrolled in American public schools grades K-12. The same study reports drop-out rates among this population to be four times that of their peers. (McKeon) The website of Institute of International Education, Inc. states that for the

2012/2013 school year there were 819,644 (a record high number) of international students enrolled in American universities. The CLASH software package is the only known software designed to improve both reading speed and comprehension within the English as a Second Language domain.

4 CLASH PROTOTYPE PRODUCT DESCRIPTION

Due to time constraints the prototype CLASH software is designed with reduced functionality but in a manner allowing continued development into the full real world product as described in section 2 of this document. The architecture remains the same but excludes elements of the user interface and component modules within the server relating to user statistics generation and reporting or analysis.

4.1 Hardware and Software Prototype Architecture

The server used for prototype development is a virtual machine provided by Old Dominion University. To allow the continued development into the full product it uses the same software packages as the full version namely: Ubuntu Server, NGINX server, Node.js server, MySQL database, the NLTK, and software written to coordinate the interaction from users web browsers with the component modules.

4.2 Prototype Features and Capabilities

Table 1, identifies the limitations of the prototype. The prototype shall parse text entered to a form on the web-page, but will not accept file uploads. In order for the students to access the reading samples they must be stored in the database and this behavior shall not require

modification to implement the full product. The prototype will store the samples in JSON format and interpret them for rendering as a color coded sample or a slashed reading stream in the web browser using the JavaScript programming of the SPA. User authentication is provided to identify particular students and the database includes data structures to handle the futere implementation of progress statistics for the full product version. The only statistic implemented per student in the prototype version is lexical bundles per minute. The homework mode allowing students to select the part of speech from uncolored text is not implemented. Both the slash stream mode and the paragraph mode for slashed text is available to both Instructors and Students.

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 shall 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 or saved documents in the system.	Administrators are able to edit, add, or remove users or saved documents in the system.
SLASH Document Viewing Mode	Ability to view documents with slashes inserted and SLASH Reader.	Ability to view documents with slashes inserted and SLASH Reader.

Table 1: Real Product vs Prototype Feature Comparison

THIS SPACE INTENTIONALLY LEFT BLANK

4.3 Prototype Development Challenges

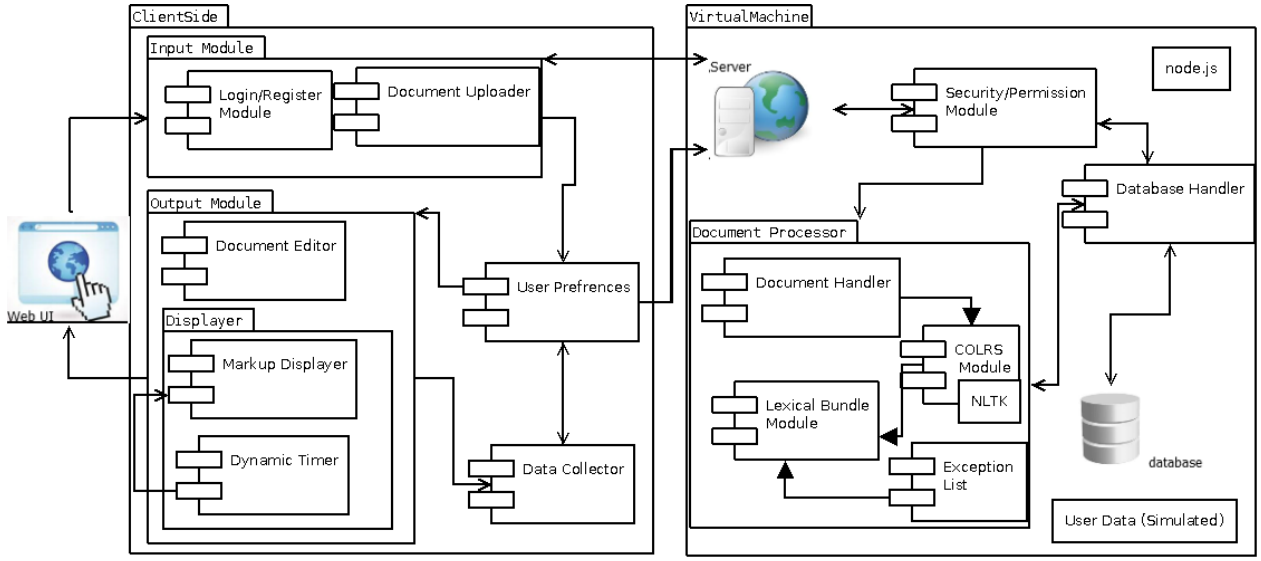

Figure 3: Major Functional Component Diagram

Figure 3 shows the major functional components of the system and highlights the over all complexity of interactions between separate software components. From this complexity the system faces a variety of challenges to its viability. The challenges to develop this software include failures of the natural language processing system, the unfamiliarity of the developers with the required software package components, modules intentionally omitted from the prototype being expected or required by other modules, the building of the exceptions list, and other as yet unforeseen issues. The NLTK software is fairly accurate in determining parts of speech but phrases such as “banana pudding” are “adjective noun” rather than “noun noun” as NLTK would indicate. The design by implementing human corrections to the automatically generated data mitigates this issue. Errors in the slashing procedure will stem from the same situation and again are resolved through the use of human intervention. The developers are rapidly learning the requisite software however mistakes in implementation will require resolution before each module can correctly interact with the others. Modules omitted in the

prototype shall require either manufactured data as place holders during prototyping or completion of components to finalize the development of the real world product.

THIS SPACE INTENTIONALLY LEFT BLANK

GLOSSARY

CLASH - Color Lexical Analysis algorithm and Slash Handler

COLRS – Colored Organized Lexical Recognition Software

ELC – English Learning Center

ESL – English as second language

IBT – International benchmark test

JSON – JavaScript Object Notation

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) for the Python programming language.

Node.js – an open source, cross-platform run-time environment for server-side and networking applications.

POS – Parts of Speech

SPA – single page application, is a highly responsive web application that fits on a single page and does not reload as the web-page changes states.

TOEFL – Test of English as a Foreign Language

Ubuntu – a Debian-based Linux operating system.

VM – Virtual Machine

THIS SPACE INTENTIONALLY LEFT BLANK

REFERENCES

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

Open Doors 2013 Report. (2013, November 11). Retrieved from http://www.iie.org/Who-We-Are/News-and-Events/Press-Center/Press-releases/2013/2013-11-11-Open-Doors-Data

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.

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