T. Hirashima et al. (Eds.) (2011). Proceedings of the 19th International Conference on Computers in Education. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education

EAGLE: an Error tAGger for Learners of English

Akkharawoot TAKHOM, Kanokorn TRAKULTAWEEKOON, Ananlada CHOTIMONGKOL, Peerachet PORKAEW, Sanooch Segkhoonthod NA-THALANG and Thepchai SUPNITHI

Human Language Technology (HLT) Laboratory,
National Electronics and Computer Technology Center (NECTEC), Thailand
{akkharawoot.takhom, kanokorn.trakultaweekoon, ananlada.chotimongkol, sanooch.segkhoonthod_nathalang, thepchai.supnithi, peerachet.porkaew}@nectec.or.th

Abstract: This paper describes the design and development of EAGLE, an Error tAGger for Learners of English. EAGLE combines all the processes necessary for the analysis of learners' language, such as creating an error tagset, tagging learners' writing, and reporting error tagging results, within the same tool. EAGLE has been developed to allow more flexibility in the creation of tagsets as well as to support many tagsets. These functionalities are achieved with the use of a hierarchical tree tagset and offset annotation. Researchers can develop their own tagsets from different theoretical frameworks and even apply them to the same document. EAGLE also provides multiple ways of viewing and comparing the statistics of tagged errors. All these features allow taggers to compare their ideas and work together to create an error tagset to render the analysis more reliable and accurate. Since EAGLE was not designed specifically for any learner language, it could be applied to tag errors produced by learners of other second languages as well.

Keywords: Written learner corpus, error tagset, error tagging tool

Introduction

To develop tools to help specific second language (L2) learners write in a foreign language effectively, one needs information on how they use the language because L2 learners' language is said to be unique and specific. First, 'errors' produced by learners are categorically different from 'mistakes' produced by native speakers [1]. Second, errors produced by learners from one L1 background are known to be different from errors produced by those from other L1 background, an error type commonly known as interlingual error. As a result, a Thai learner corpus of English and an analysis of errors produced by Thai learners of English are significant to the development of a writing tool to help improve Thai learners' writing skill. Limitations of existing CALL technologies and their applications on a writing tool for Thai learners, and the need of a corpus-based tool were discussed in [2].

So far, the existing tools to collect L2 learners' writing and to tag errors have been developed for specific research purposes, thereby, having specific ways to collect and analyse learners' language, and to design an error tagset. It is therefore not always practical to use these tools in other research projects with different purposes without any modifications. Furthermore, these tools were often developed separately when in fact they serve as processes involved in the analysis of learners' language, ranging from creating a database of language learners, uploading learners' writing into the system, creating an error tagset, tagging learners' writing with the tagset created, reporting error tagging results, to exporting tagged files for further analysis.

To facilitate the process of learner corpus creation, we have developed a tool to collect and analyse learners' language called “Error tAGger for Learners of English” or

EAGLE which combines all of the aforementioned processes. The most important process in analysing learners' language is to tag errors. Most of the available tagging tools were created for a specific tagset. By implication, they may not be able to support tagsets other than the one originally built with the research project, and given the present circumstances, may not be applicable to tag errors produced by Thai learners. Taking that into account, EAGLE has then been developed to allow more flexibility in the creation of tagsets as well as to support many tagsets. These functionalities are achieved with the use of a hierarchical tree tagset and offset annotation. Researchers can develop their own tagsets from different theoretical frameworks and even apply them to the same document. EAGLE also allows tagging researchers to compare their ideas and work together to create an error tagset to render the analysis more reliable and accurate. Another important process is the analysis of error tagged documents. EAGLE provides multiple ways of viewing and comparing the statistics of tagged errors. The analysis results together with tagged documents can also be exported for further processing by other tools.

The rest of this paper is organized as follows: existing error tagging tools are reviewed in Section 1. Section 2 describes the overview architecture of EAGLE. Section 3 explains design principles and key features in detail. System evaluation is provided in Section 4. Finally, Section 5 concludes our work and discusses possible future directions.

1. Related Works

One of the important tasks of analysing learners' language is to tag errors. Many error tagging tools have been developed. Some are developed specifically for tagging learner errors. TagEditor [3] was implemented for transcribing and tagging the Standard Speaking Test (SST) corpus using a tagset designed specifically for Japanese learners of English. It has features including basic functions like text editor, tag validation, and a simple concordancer. The error tagset is represented using XML tag. UCLEE [4] was created by John Hutchinson for use with International Corpus of Learner English (ICLE). It supports a hierarchical error tagset but does not allow any modification on the tagset. EARS [5] was developed based on a multi-level, fine-grained taxonomy of errors from a corpus of Spanish learners of English. The error tagset of EARS is a hierarchical structure that supports a tagset creator to adjust any error tags. EXMARaLDA [6] is a tagging tool that is based on a standoff XML format which allows multi-layer and overlapping annotations to be applied on a learner corpus. The system focused on German learner of corpus Falko. Table 1 summarizes the features of existing tagging tools.

Table 1 : Existing error tagging tools comparison

Tagging Tool	Hierarchy structure	Unlimited level	Overlapping Annotation	Error analysis
TagEditor	×	×	×	✓
UCLEE	✓	×	×	×
EARS	✓	✓	×	×
EXMARaLDA	✓	✓	✓	×

Many tools support a hierarchical tagset structure because it is flexible and makes relationships between levels easy to understand; however, some tools limit the number of levels in the structure. Furthermore, most existing tools do not support overlapping annotation which could also be useful. In addition, the existing tools were developed only for tagging; an error analysis function is usually not included. Taking these limitations into account, EAGLE has then been developed to allow more flexibility in the creation of tagsets as well as to support many tagsets. These functionalities are achieved with the use of a hierarchical tree tagset and offset annotation. EAGLE also combines all processes necessary for the analysis of learners' language within the same tool.

2. System Overview

The analysis of learners' errors generally involves a number of processes, including collecting learners' data, storing the data, sorting the data, designing a tagset, tagging errors, retrieving errors for further analysis and so on. Our EAGLE incorporates all the processes deemed necessary for the analysis of learners' language with the hope that it can offer flexibility to researchers who want to make use of EAGLE with language learners other than Thai learners of English. The details of EAGLE are as follows:

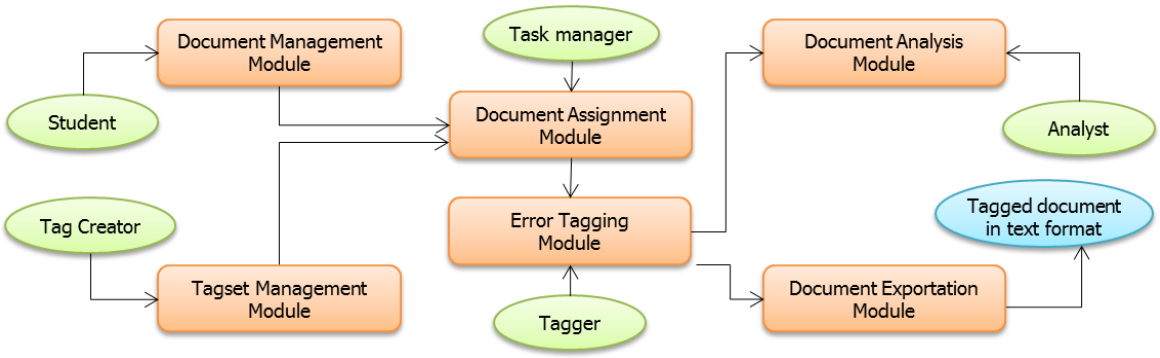

Figure 1 : System Architecture of EAGLE

Based on the system architecture displayed in Figure 1, EAGLE can be used by 5 groups of users. First, students access the tool through a Document Management module in order to provide their profile information and do an English test to identify their levels of English proficiency. They also provide the main input, i.e. the writing tasks, for the research project, which then can be directly uploaded into the system or manually stored (by a typist) in the system in case the traditional paper-based data collection method is used. The design of a tagset is carried out by a tagset creator using a Tagset Management module. Once created, the tagset will be used by an error tagger who will assign appropriate tags to identified errors. A task manager is the one who assigns documents to taggers and, for the case of multiples tagsets, decides which tagset will be used to tag which document. The results of the tagging appearing in the tagged document, along with some statistical results, will be carefully studied by an error analyst with the help of a Document Analysis module. The analysis results together with tagged documents can also be exported for further processing by other tools.

EAGLE is a web-based application which works on a client-server basis. Most data processing is done on a server side while user interactions are done on a client side through a web browser. There are 7 modules in EAGLE.

2.1 Document Assignment Module

The Document Assignment module is used to assign documents to error taggers after the first two modules were used to collect learner documents and create tagsets. Document assignment is done by selecting both the document and the tagset to be used with that document for each tagger. There is a task status panel that shows the progress of each tagger. Given the flexibility of the system, more than one tagset can be assigned to tag the same document, and similarly more than one tagger can work on the same document.

2.2 Error Tagging Module

The Error Tagging module is used to tag documents assigned by the Document Assignment module. When a tagger selects an assigned document, error tagging interface is displayed. The tagger tags an error found in the document by selecting an appropriate error tag from the

assigned tagset. The tagger can also insert an error tag for a word or phrase that is missing from the document. The Error Tagging module is discussed in more detail in Section 3.2.

2.3 Group Management Module

The Group Management module allows a user to select documents based on specific criteria, such as the document topic, learner proficiency, etc., from the entire document collection and then save them under a defined group name. Once created, groups will be used in a Document Analysis module, described in Section 2.6, and a Document Exportation module, described in Section 2.7. Document grouping helps users analyse a subset of learner data that they are interested in. The users can also compare and contrast between groups of documents in the Document Analysis module.

2.4 Document Analysis Module

The Document Analysis module is used to generate statistical results of tagged documents. Analysis results are shown by a bar chart or a plot graph. A user can a subset of documents by utilizing groups created by the Group Management module. He can also compare statistical results between groups or analyse only error tags of interest. This module is discussed in more detail in Section 3.3.

2.5 Document Exportation Module

The Document Exportation module is used to export tagged documents into text files so that they can be used by other data analysis tools such as WordSmith Tools [7]. Error tags are embedded in an exported text file in the following format:

original text [error:tag-level1|tag-level2|…|tag-levelN:correct_form:dependency] original text

3. Design Principles and Key Features

As mentioned in Section 1, most of the available tagging tools were created for a specific tagset. By implication, they may not be able to support tagsets other than the one originally built with the research project. Taking that into account, EAGLE has then been developed to allow more flexibility in the creation of tagsets as well as to support many tagsets. These functionalities are achieved with the use of a hierarchical tree tagset and offset annotation to be discussed in Section 3.1 and 3.2 respectively. Another important feature is error analysis of tagged documents. In most existing tagging tools, this module is not included. In EAGLE, multiple ways of viewing the statistics of tagged errors are provided. This feature is explained in Section 3.3.

3.1 A hierarchical tree tagset

The notion of the hierarchical tree tagset appeals to us since we deem that it can offer a lot of flexibility in creating a tagset and supporting many tagsets created for different purposes and from different theoretical frameworks. Each node in a tree corresponds to an error tag which can be linked to its parent node, showing a clear relationship between the parent and the child. By storing the link between each error tag and its parent node, EAGLE supports a hierarchical tagset with unlimited tag levels.

Our error tagset is designed to consist of 3 parts, namely (1) the overall view of the tree structure, (2) error tags and (3) the tagset management part, the last of which comprises adding, editing or deleting tags. To create a tag, one must decide on the parent node. In our case, the parent nodes are classified into 3 nodes, i.e. Grammar, Lexicon and Writing

Mechanics. Once decided, that parent node must be selected, and then the 'New Child Tag' node must be selected for a tag to be created. For example, “spoke” in “We spoke to her everyday” will be tagged as:

Grammar (Parent) > Verb error (Child node) > Wrong tense

In a similar fashion, “beauty” in “She's a beauty girl” will be tagged as:

Grammar (Parent) > Adjective error (Child node) > Wrong Adjective form

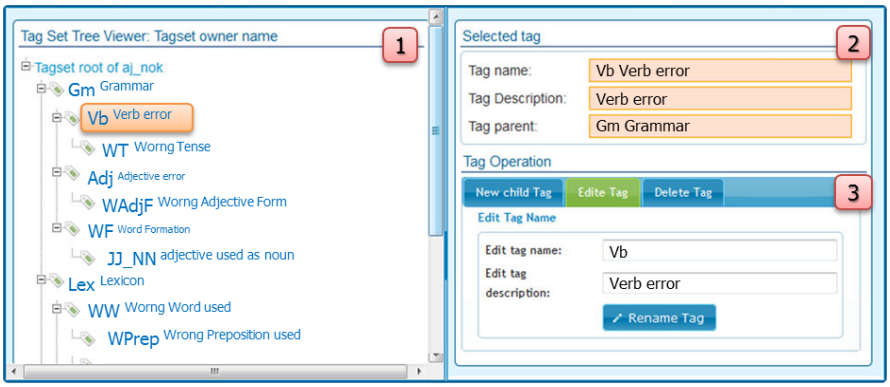

Figure 2 : Tagset Management Interface

3.2 Offset annotation

Offset annotation is an annotation style where tags and their attributes are stored separately from a document. In our case, error tags and related information such as correct form and error dependency are stored in a separate database. This is quite similar in principle to a standoff annotation used in EXMARaLDA [4]. Many tagging tools discussed in Section 1 used inline annotation where tags and their attributes are embedded in a tagged document, for example, typical XML or HTML documents. One shortcoming of the inline annotation is that it does not allow overlapping between tags. In contrast, the offset annotation supports overlapping annotation by keeping indexes of errors in a document and their corresponding error tags in a separated database. To tag the same document with different tag sets, another set of indexes are created. The same technique is applied when more than one tagger works on the same document. This technique allows more flexibility than embedding error tags in the document directly.

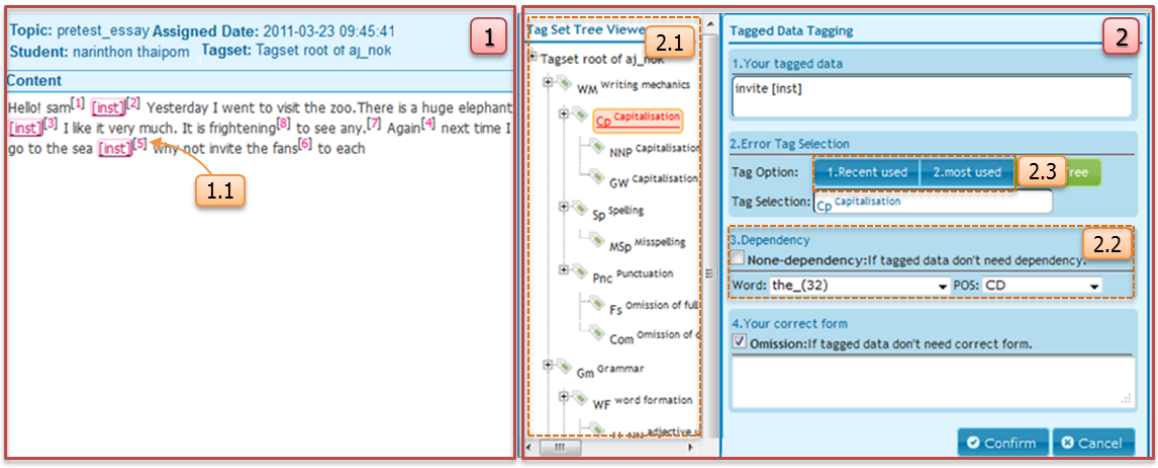

Figure 3 : Error Tagging Interface

Error Tagging, shown in Figure 3, is divided into 2 parts, namely, the part where a document is displayed for tagging (1) and the part of error tagging (2). Once a file is selected and uploaded, the tagger can start tagging errors by selecting tags from the Tagset Tree

Viewer which stores all the tags available (2.1). In our design, we require that the tagger provide the correct form for each error identified along with its dependency if possible (2.2). The part of speech (POS) of the dependency should also be noted. This will be useful for the analysis later on, especially when we develop a writing assistant tool to help learners' writing. To achieve a higher degree of efficiency in tagging, the used tags are stored either as 'recently used' or 'most used' for quick access (2.3). The words or phrases once tagged will be automatically marked with superscript numbers which are displayed in the ascending order as shown in (1.1).

3.3 Error Analysis

The Document Analysis module in EAGLE deals with the results of the tagging of errors done in the Error Tagging module. In general, this module can display results of the analysis in 2 main ways, i.e. an overall view and a comparative view. An overall view displays the overall results of error tagging by frequency counts or percentages of the tags. For a comparative view, researchers can see the results of the analysis grouped by learners' proficiency levels, gender, years of exposure to English, major subjects and so on, depending on research purposes. Examples of error statistics in a comparative view are shown in Figure 4. In addition, a user may choose to analyse only a subset of learner data that he/she is interested in by applying groups created through the Group Management module. Furthermore, researchers can also choose to look at the results of the analysis of any specific tags such as the frequency count of 'Extra Word' and 'Wrong Word Choice'.

That the Analysis module allows the data to be displayed in various ways as described above adds an advantage to the EAGLE system. That is, the researcher can interpret the results of the analysis from many perspectives. For example, in terms of the subject (here referred to as 'group') of the research, there can be many factors that may affect the production of certain errors, ranging from proficiency levels, years of exposure to English, to gender and major subjects. With the Analysis module, the researcher can group the subjects in the way s/he would like to see, so that the results can be interpreted more meaningfully.

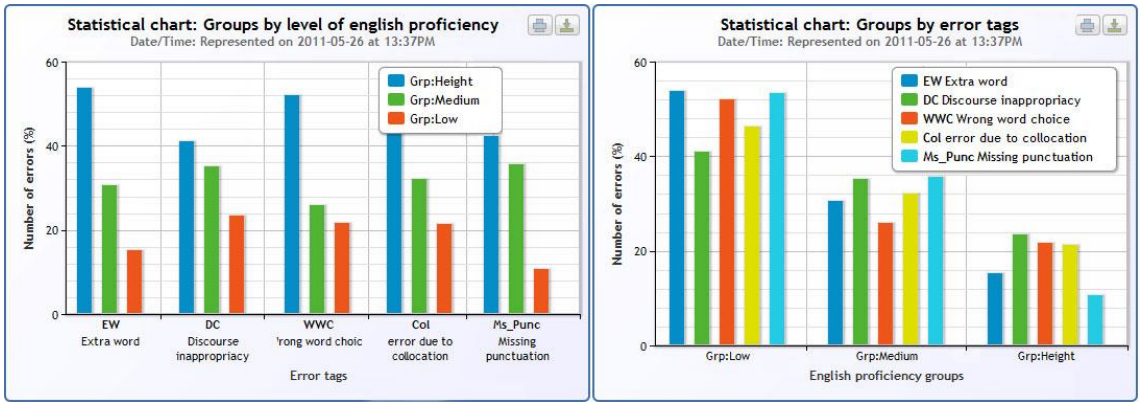

Figure 4 : Graphs of errors grouped by English proficiency and grouped by error tag

4. System Evaluation

To test our system, we used EAGLE to collect profile information and scores on the English pre-test from a total of 400 senior high school students. This data collection process revealed that the system is highly stable even with a great number of users logging onto the system at one time, and allows quick access to the data. 400 paper-based essays were collected from this group of students and entered to the system. A tagset designer and a

tagger have been working on this data set to create an error tagset suitable for errors produced by Thai learners of English. From the flexibility of the system, more than one tagset can be developed depending on the conceptual and/or theoretical frameworks held by the developer. Given this, researchers will have more tagsets at their disposal and they can choose the one suitable for their research purposes.

As shown in Figure 5, two tagsets have been created with two different concepts. The tagset shown in (1) was developed with the 3-level hierarchy, categorising errors according to the parts of speech in English, e.g. Nouns, Verbs, Prepositions, etc. The other tagset illustrated in (2) is non-hierarchical, with the errors being classified in terms of types of errors, e.g. misspelling, omission, redundancy, wrong word use, etc. In Figure 5, (3) and (4) show the results of the tagging using the tagset models (1) and (2) respectively.

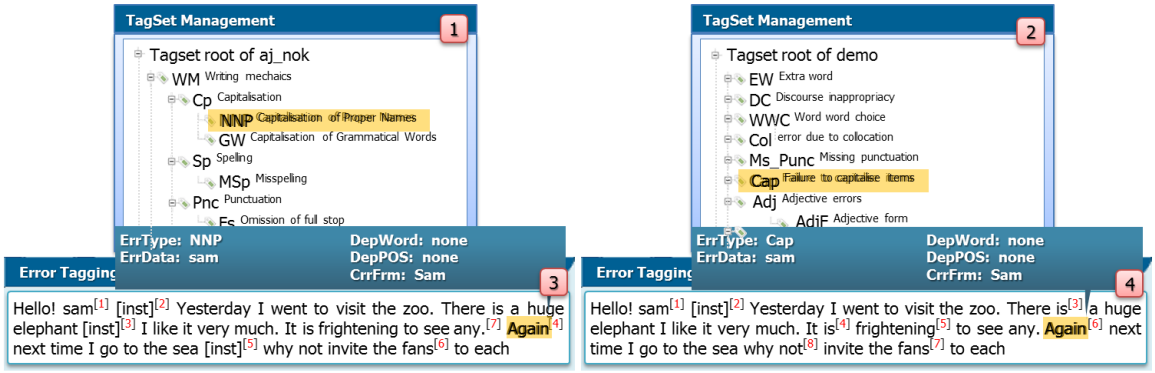

Figure 5 : The various tagsets by different concepts

Figure 6 shows the results of the tagging carried out by two different taggers. One tagger did not consider "frightening" as an error, so it was not tagged as shown in (1), while the other tagger considered it an error of adjective form and provided the word "frightened" as the correct form, as shown in (2).

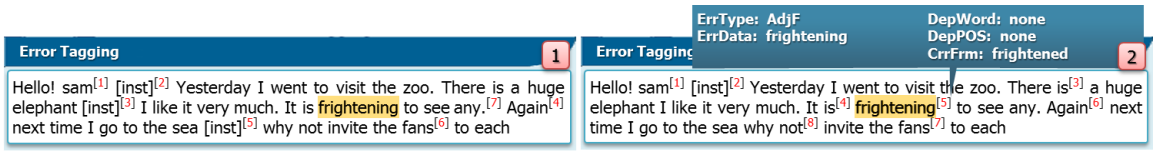

Figure 6 : The document tagging by two different taggers

Figure 7 shows the result of tagging by using two different tagsets. In (1), the word "Again" was tagged as an error of Literal translation from one tagset while it was tagged as Extra word from another (2).

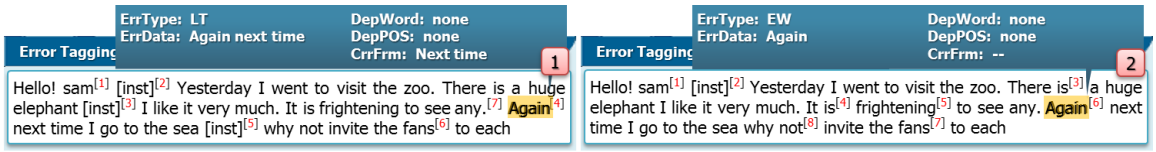

Figure 7 : The document tagged by using two different tagsets

From our preliminary analysis, we found that 63.0% of errors in student essays are grammatical errors while 19.3% and 17.7% are writing mechanics and lexical errors respectively. Figure 8 shows the overall grammatical errors by 11 subcategories. Thai learners often make errors on noun, verb and article. There is no article in Thai so students often omit it. Thai also does not have noun declension and verb conjugation, so students often use noun in generic sense without making a distinction between singular and plural. Incorrect verb forms and subject-verb agreement are other common problems. More detail errors analysis is still underway as we have been working on refining our tagset to better suit the errors produced by Thai learners.

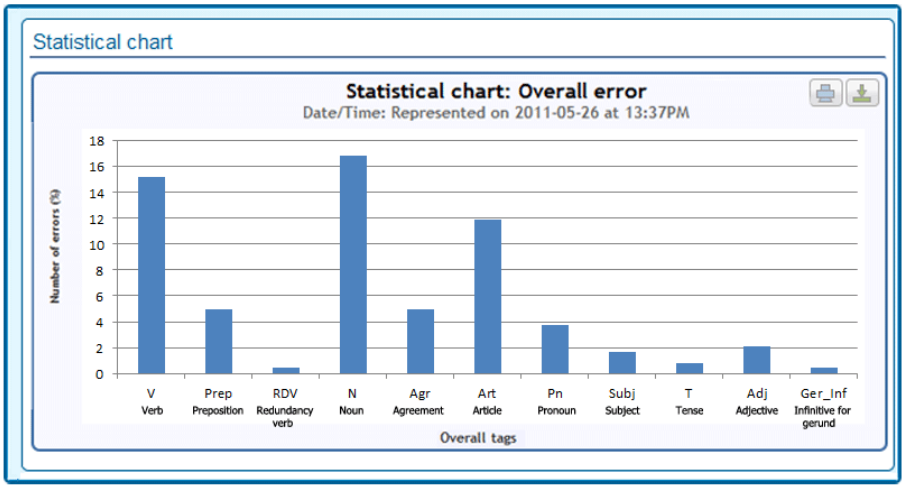

Figure 8 : A graph of overall grammatical errors

5. Conclusion and Future Work

EAGLE facilitates the process of document tagging in many respects. First, EAGLE has incorporated all necessary processes in its system, including uploading the document, developing a tagset, tagging errors, and reporting the results. This offers great convenience to researchers and linguists since it is an all-in-one tool, unlike other tagging tools which typically have some of the processes and not the others. Secondly, EAGLE is also designed to allow more flexibility in the creation of tagsets as well as to support many tagsets. Researchers can develop their own tagsets from different theoretical frameworks and even apply them to the same document. These functionalities are achieved with the use of a hierarchical tagset and offset annotation. Lastly, EAGLE incorporates the Document Analysis module where different groups of data can be displayed to show the results of the analysis. Given this, the researcher can look at the data from many perspectives and interpret them in a more meaningful way.

EAGLE has been presented at an international conference on English language teaching where it attracted a lot of attention from the audience for further research. EAGLE could be applied to tag errors produced by learners of other second languages as well, as it was not designed specifically for any second languages.

References

[1] Corder, S. P. (1981). Error analysis and interlanguage. Oxford: Oxford University Press.
[2] Na-thalang, S., Chotimongkol, A., and Supnithi, T. (2010). "Toward the development of SWAN (Student Writing AssitaNt) for Thai learners", Proceeding of the 7th CULI.
[3] Tono, Y., Kaneko, T., Isahara, H., Saiga, T. and Izumi E. (2001). The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. Lee, S. (ed.) ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary. The Scond Asialex International Congress, 257-262.
[4] Estelle, D., Sharon, D. and Sylviane G. (1998). Computer-aided error analysis. System: An International Journal of Educational Technology and Applied Linguistics 26(2), 163-174.
[5] Díaz-Negrillo, A. and García-Cumbreras, M.A. (2007). A Tagging Tool for Error Analysis on Learner Corpora. ICAME Journal, 31, 197-203.
[6] Lüdeling, A., Walter, M., Kroymann, E. and Adolphs, P. (2005). Multi-level error annotation in learner corpora. Proceedings of the 2005 Corpus Linguistics Conference.
[7] Scott, M. (2008). WordSmith Tools version 5. Liverpool: Lexical Analysis Software.