ON DESIGNING INTERLANGUAGE CORPUS OF INDONESIAN STUDENTS

WITH ERROR ANALYSIS ANNOTATION

Prihantoro

Universitas Diponegoro

[prihantoro2001@yahoo.com](mailto:prihantoro2001@yahoo.com)

Abstract

This paper proposes a model for the design of interlanguage corpus with error analysis annotation. The data is obtained from ICNALE (Ishikawa, 2013) corpus, a corpus of English learners in Asia. I focus the extraction on the Indonesian data. The sample for this model is obtained from English texts written by proficient Indonesian students. The data is processed by a Local Grammar (LG) based corpus processing software called UNITEX (Paumier, 2008). The experiment begins with error identification. Parts of the texts where errors take place are marked with error tags. The suggested forms to revise the errors are written in Local Grammar Graphs (LGGs) that are applicable to the corpus text. When these LGGs are applied in the corpus, they retrieve parts of the texts that are marked by the error tags and the suggested forms embedded in the LGGs. Users are provided with the options to display concordance in several ways. First, it may identify parts of the texts that are marked by error tags. Second, it may identify the error and display the suggested forms as well. Third, it allows the suggested forms to replace the error tags. Hence, the errors are hidden. This corpus may benefit teachers and linguists in terms of material development and L1 interference-analysis support system.

1. WHY ERORR ANNOTATED CORPUS?

One type of corpora is learner corpus or interlanguage corpus, where the data is obtained from learners of second/foreign language (Mc Enery & Hardie, 2012). In this paper, I want to present my design for future interlanguage corpus where the data is obtained from Indonesian learners. The data is annotated, not only with POS tags, but also error tags so that further natural language processing (Allen, 2006) tasks can be performed.

Annotated corpus has several advantages, where automatic retrieval beyond character match can be performed. With this automatic retrieval, teachers would no longer have to index the errors manually on printed documents. As for this paper, the annotation (Mitkov, 2005) was performed on the errors made by the Indonesian students studying English as second/foreign language. The annotation is useful in terms of error classification and identification, so that teachers may later focus their teaching on the error parts, or design teaching material on that basis. Granger (2003) believed that this semi-automatization of error analysis to the corpus is a promising synergy, as corpus is widely distributed among users.

To some extent, error analysis by computer has been adopted into some text processor programs. For instance, when you use microsoft office word program and you write <I are> or <you is> then the program will give you green underline. This means that it is grammatically wrong, as subject and verb must agree. However, consider this sentence, ‘you is a pronoun’. This sentence is correct, but still, the program detected this as a wrong subject verb agreement.

As for the corpus prototype in this paper, it is made offline. However, when the data is large enough it will be made online with free distribution, so more and more users can access the data or even contribute new data finding.

1. CORPUS DATA

English is a widely used foreign/second language used by In Asia. So far there are many error analysis research concerning the influence of first language (L1) to the target language (English) as the mapping of these errors can be used determine the focus of language teaching or teaching material. However, research where the data is obtained from a specific corpus is rare. One of them is MICASE corpus built by the University of Michigan (Simpson-Vlach & Leicher, 2006) which can be accessed online[[1]](#footnote-2). I started with this corpus, but it seemed that the data in this corpus was obtained from Indonesians studying in US universities’ masters and Ph.D programs, whose competence is most likely to be more proficient than undergraduate Indonesian student or Indonesian students studying in non-english-speaking countries.

I then turned to ICNALE (Ishikawa, 2013), where the data is obtained from ESL learners in Asia, including learners from Indonesia. As learner corpus from Indonesia is not presently available, this is the closest that I could get. Data from Indonesian learners were extracted and further processed by offline corpus processing software called Unitex (Paumier, 2008). Unitex is a multi-language platform corpus processing software and has been used in linguistic or computing field, not only for English, but also some other languages such as French (Laporte, Nakamura, & Voyatzi, 2008), Korean (Aerim, Laporte, & Jeesun, 2010), Arabic (Traboulsi, 2009) and even Indonesian (Prihantoro, 2011).

1. METHODOLOGY

The first part of this section begins with the description of the raw data. ICNALE data is open for download. When completing the download, there are several types of data format. The corpus processing software employed in this research, UNITEX, allows the processing of raw <.txt> format, which is present among the ICNALE format types. As the data is obtained from Asian Learners with L1 background and different level of English proficiency, the data for further processing are selected based on the following criteria. First, the L1 is Indonesian. Second, only data of the most proficient users is used. I managed to obtain a file consisting of 3350 tokens (words) from learners categorized as B2. The categorization follows CEFR (Common European Framework of Reference for Languages) in Alderson (2007). This is the highest level that is present for Indonesian learners’ data in ICNALE.

This second section of methodology deals with my experiment on writing error analysis annotation. The raw data that has undergone filtering process is used as research corpus in UNITEX. I modify the text by marking each error with a unique tag. The same tag format is rewritten in Local Grammar Graphs (LGGs). LGGs are finite state automaton based (Gross, 1993) abstract machine that is useful for automatic retrieval. LGG can both identify input (tag) and create output (transducer) at the same time. Consider the sample as shown by figure 1:

Figure 1. LGGs to retrieve DET + Any Animal Noun with Output

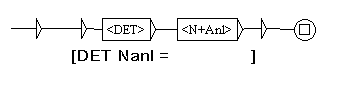
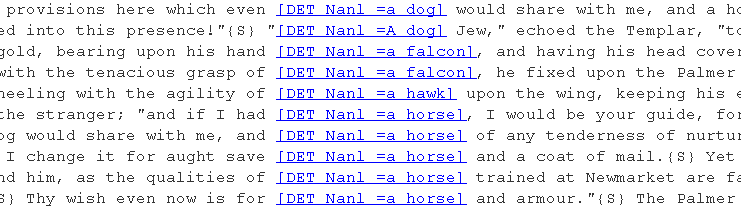


Figure 2. Concordance with Input and Output from LGGs in figure 1

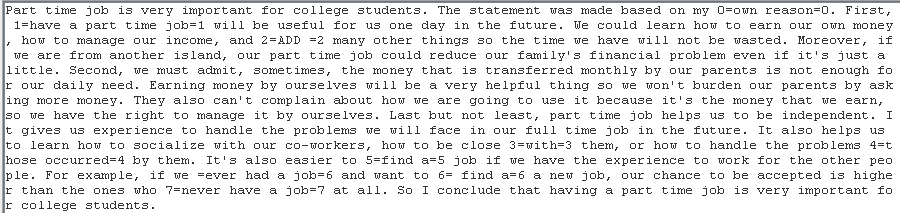


The LGGs are used to identify the input (parts of the text marked with error tags), and creating output (the suggested revision of the error). Each tag rewritten in LGGs also has a unique output (suggestion). The coverage of the first LGGs is around 20% of the data (around 700 tokens). After the testing showed satisfactory result, I improved the coverage to 100% and categorized errors of the same type to specific LGGs. The purpose is to allow users to customize the retrieval for particular error (spelling error, subject-verb agreement error, article use, or etc). Besides it allows the users to customize the retrieval, it also allows the users to manipulate the concordance display.

1. Results and Discussion

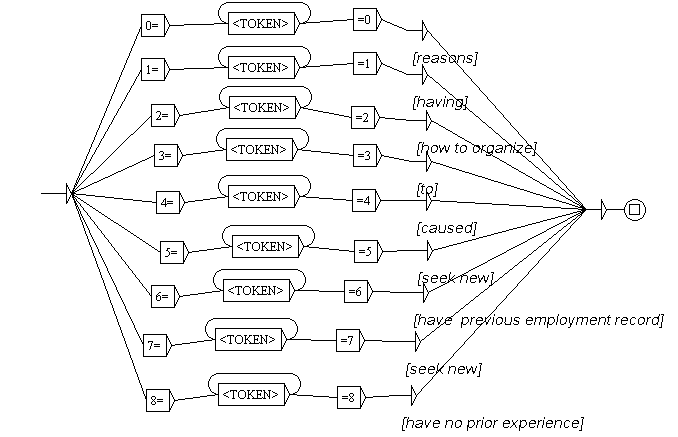
The raw corpus in this study was marked by error tags. The tags are required to let the computer understand how to identify parts of the texts that are error. In this case, I use the following convention. Tags consist of open and close tags for each error. They are composed of numeric digit, plus a metasymbol ‘=’. The numeric digit is a unique identifier that distinguishes one error and another. This is important, as some errors need to be revised in different way although they fall to the same type.

Figure 3. Error Marked Corpus



The second step is to create LGGs that are used to 1) identify the tags, and 2) show the revision. Identifying tags is considered as an input. The LGGs are composed of the tags, where tokens are looped. It means that it retrieves all expressions interfere between the tags. The output is used to show revision or the suggested forms. Please check figure 4:

Figure 4. LGGs with Suggested Forms



The tag format is <number><=><TOKEN><=><number><[suggested forms]>. The opening and closing tag shows the parts of the text that are identified as error. Each tag is unique, but later each tag will be categorized by the type of errors. The parts can be consisting of one or more tokens. For this reason, the token is looped. The suggestion is given as LGGs’ output marked by square brackets. I applied the LGGs to the corpus that is already marked by error tags. The LGGs then retrieved the tags. Figures of coverage and recall are shown, and the retrieval of the result is shown in the form of concordance. At this point, users can choose to display only the input, only the output, or both the input and the output. See figure 5, 6 and 7.

Figure 5. Input (Error) Only

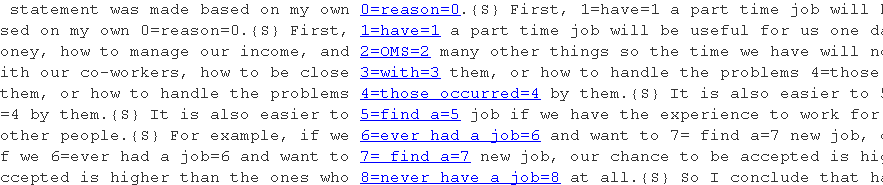


Figure xxx shows how the identification of error tags are shown by concordance. At his point, it does not show any correction of the errors. Teachers, may show this to the students and ask them to revise the errors. Another method is to show both errors and the corrections or revised version as shown by figure xxx:

Figure 6. Input (Error) and Output (Suggestion)

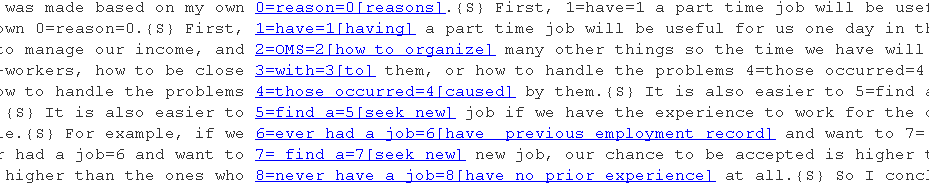
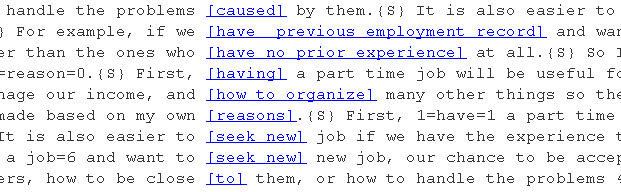


Figure xxx shows both errors and revisions. As for this paper, the tag is simple. However, it can be improved to show the type of errors. For instance, if you see 0=0 tag, it shows the error on plural form. In this case, instead of using plural form, the students used singular form. We can classify the same errors to one LGG, and later retrieve errors specifically on singular/plural form. Another variation of LGGs is to show only the revised version. This might be useful to show the quantity of errors in, for example, one students writing.

Figure 7. Output (Suggestion)



1. Conclusion

In this paper, I have presented my proposal for the creation of interlanguage corpus for English learners in Indonesia. It suggests that this issue must seriously be taken into account. Improvement might be made on the size of the data. Some future works may concern the distribution method and the distribution system to bring more annotators as in this paper, the accuracy relies heavily on the human annotator. I suggest that more and more human annotators are involved in this kind of research so that machine error, as I have commented previously, can be avoided. When meeting the standard of interlanguage corpus, it might be shared online.

**References**

Aerim, A., Laporte, E., & Jeesun, N. (2010). Semantic Polarity of Adjectival Predicates in Online Reviews. Seoul International Conference in Linguistics (pp. 20-27). Seoul: Korea Linguistic Society.

Alderson, J.-C. (2007). The CEFR and the need for more research. The Modern Language Journal, 91(4) , 659-663.

Allen, J. (2006). Natural Language Understanding. California: Benjamin Publishing.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. CALICO, 20 (3) , 465-480.

Gross, M. (1993). Local Grammars and Their Representations by Finite State Automata. In M. Hoey, Data, Description, Discourse: Papers on the English Language in Honour of John Mc.Sinclair (pp. 26-38). London: Harper Collins.

Ishikawa, S. (2013). The Icnale And Sophisticated Contrastive Interlanguageanalysis Of Asian Learners Of English. In S. (. Ishikawa, Learner Corpus Studies In Asia And The World 1 (pp. 91-118.). Kobe: Kobe University Press.

Laporte, E., Nakamura, T., & Voyatzi, S. (2008). A French Annotated Corpus for Multiword Nouns. Language Resources and Evaluation Conference, (pp. 27-31). Marrakech.

Mc Enery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

Mitkov, R. (2005). Handbook of Computational Linguistics. Oxford: Oxford University Press.

Paumier, S. (2008). Unitex Manual. Paris: Universite Paris Est Marne La Valee & LADL.

Prihantoro. (2011b). Local Grammar Based Auto Prefixing Model for Automatic Extraction in Indonesian Corpus (Focus on Prefix MeN-). Proceedings of International Congress of Indonesian Linguists Society (KIMLI) (pp. 32-38). Bandung: Universitas Pendidikan Indonesia Press.

Prihantoro. (2014). Machine Readable Grammar For Optimizing Automatic Retrieval In Text Corpus: A Comparison of Regular Expression and Local Grammar Graph. Proceeding of International Linguistic Congress (held by MLI: Indonesian Linguists Society) (pp. 424-428). Bandar Lampung: MLI Press.

Simpson-Vlach, R.-C., & Leicher, S. (2006). The MICASE handbook: A resource for users of the Michigan Corpus of Academic Spoken English. Michigan: University of Michigan Press.

Traboulsi, H. (2009). Arabic Named Entity Extraction: A Local Grammar Based Approach. Proceedings of the International Multiconference Science and Information Technology (pp. 139-143). Mragowo, Poland: Polish Information Processing Society.

1. http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple [↑](#footnote-ref-2)