DOI: 10.15439/2017F433
ISSN 2300-5963 ACSIS, Vol. 11

Unsupervised tool for quantification of progress in L2 English phraseological

Krzysztof Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kwolk@pja.edu.pl

Agnieszka Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: awolk@pja.edu.pl

Krzysztof Marasek
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kmarasek@pja.edu.pl

Abstract - **This study aimed to aid the enormous effort required to analyze phraseological writing competence by developing an automatic evaluation tool for texts. We attempted to measure both second language (L2) writing proficiency and text quality. In our research, we adapted the CollGram technique that searches a reference corpus to determine the frequency of each pair of tokens (bi-grams) and calculates the t-score and related information. We used the Level 3 Corpus of Contemporary American English as a reference corpus. Our solution performed well in writing evaluation and is freely available as a web service or as source for other researchers.**

I. Introduction

A person's second language, or L2, is a language that is not the native language of the speaker but is used in the locale of that person. In contrast, a foreign language is a language that is learned in an area where that language is not generally spoken. Some languages, often called auxiliary languages, are used primarily as second languages, or lingua francas. More informally, a second language can be said to be any language learned in addition to one's native language, especially in the context of second language acquisition, (that is, learning a new foreign language) [1]. A person's first language is not necessarily their dominant language, the one they use most or with which they are most comfortable. For example, the Canadian census defines first language for its purposes as "the first language learned in childhood and still spoken," recognizing that for some, the earliest language may be lost, a process known as language attrition. This can happen when young children move, with or without their family (because of immigration or international adoption), to a new language environment [2].

In the process of language development, lexical indices are not as popular as the utilization of syntactic procedures. In the area of foreign linguistics, there has been a constant lexicalization of the teaching curriculum, which has a phraseological basis. Moreover, it is also recognized that the process of language production is affected by the pre-pattern segments described by [3]. Corpus language methods have highlighted the broad range of word combinations that were previously analyzed.

It is important to analyze the role of corpus linguistic studies in the grading of L2 writing. In such grading, it is essential to analyze the writing based on functional skills and the independent construction of written text to communicate in a purposeful context. A human writer cannot be used to demonstrate the requirements of the standards, as this does not meet the requirement for independence. In writing assessment, we should consider whether or not information and ideas were presented concisely, logically, and persuasively. It is also important to determine whether or not a writer clearly presented information on complex subjects, used a range of writing styles for different purposes, and employed a range of sentence structures, including complex sentences and paragraphs, to effectively organize their written communication. We should also evaluate the accuracy of punctuation in written text using commas, apostrophes, and quotation marks. Lastly, written work should fit the purpose and audience, with accurate spelling and grammar that support clear meaning [4].

Corpus analysis is both qualitative and quantitative in nature. One of the biggest advantages of using corpus language is that we can easily provide quantitative data to assess concerns for which intuition cannot be considered reliable. In other words, much more than just counting bi-grams is involved [5]. Prior research highlights the variety of questions that need to be addressed on the vital role played by L2 writing [6].

II. Evaluation Methods Using N-grams

An n-gram is a contiguous sequence of n items from a given sequence of text. Depending on the application, the items can be phonemes, syllables, letters, words, or base pairs. N-grams are typically collected from a text or speech corpus. When the items are words, n-grams may also be called shingles. An n-gram of size 1 is known as a unigram (1-gram), size 2 is a bigram (2-gram), size 3 is a trigram, and so on. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of an (n-1)-order Markov model [7]. The n-gram models are widely used in probability, communication theory, computational linguistics (e.g., statistical natural language processing), computational biology (e.g., biological sequence analysis), and data compression. Two benefits of n-gram models (and algorithms that use them) are simplicity and scalability [7].

The n-gram based evaluation method consists of removing the arrangement of n-words for bigrams, learner corpus data, and data that contain native combinations of tokens [5]. The results of different n-gram models are not directly comparable, as they utilize different criteria to identify relevant units. However, they can indicate some general trends in L2 writing that rely on the most restricted repertoire of lexical bundles as compared to that of native writers [8]. L1 writers utilize more phrases that are familiar with poor sequences and fewer native-like phrases. They also report having difficulty while introducing speech-like phrases into their official language. These studies highlight various features of L2 phrasing because of the lack of huge longitudinal corpora of L2 writing and the effort required to collect them.

Complex lexical phrases are very rarely used by the lowest skilled writers. Learning traditional strings of words at an elementary level has been found to be productive at advanced and secondary levels. However, this finding relied only on the frequency of multi-word units and paid no attention to the degree of association within units. Very common words stand a much greater chance of frequent arrangements than uncommon words of different varieties.

Mutual information (MI) and t-score are used in this research to calculate the comparative frequency of occurrence of word sequences in a reference corpus. They also indicate the probability of a word sequence appearing due to the frequency of the words of which it is composed. MI will highlight word sequences that are developed from a small rate of word reoccurrence, for example, the term "tectonic plates" is a very low-occurring word sequence. Similarly, t-score works with word sequences of highly recurrent sets of words. However, a study by Durrant and Schmitt [9] focuses on one type of sequence, an adjoined pair of words used as a modifier. The studies show that, unlike native writers, L2 writers of English use collocations with the highest MI ratings at a very low frequency. This means that the usage of MI with high frequency is not very popular among them, whereas collocations with t-scores are frequently used. The same pattern can be observed with transitional and sophisticated learners. Learners in their transitional phase are more inclined to very often use frequently-occurring collocations and make minimal use of lower frequency collocations. The present study has utilized the same methodology but is unique in two aspects. First, it uses a preset system to obtain word sequences from a tagged part of speech. Second, it simulates longitudinal corpora. We evaluated L2 writers who had multiple levels of proficiency. Therefore, it was very important to assess the phraseological index of the longitudinal data in question; our study strongly considered this aspect. This study has incorporated both longitudinal and pseudo-longitudinal approaches that assist in recognizing the given input of all the research designs in the analysis of L2 writing [10].

III. Data and Methodology

Our writing evaluation application consists of three main sub-tools. First, a user interface, implemented in ASP.NET and shared as a web service, handles user inputs, manages them, and requests solutions from the other software components on behalf of the user. The website is responsible for loading the user input files and generating the final download link of the results as a ZIP file for the user. The results are output in Excel-compatible CSV files. Each separate file contains different analysis results for each bi-gram, such as frequency in the L2 text, frequency in the reference corpus, mean frequency in the reference corpus, MI score, and t-score. For multi-file analysis, the tool calculates the number of unique 1-grams and 2-grams, the number of 2-gram types, the number of 2-grams collocated in the reference corpus, the percentage value of L2 coverage in the reference corpus, and a summary that includes how many 2-grams were not found, MI, and t-scores.

Second, we employ the CLAWS part-of-speech (POS) Tagger[1] for better text tokenization and identification of the proper parts for speech recognition and comparison with n-grams in the reference corpus in their correct form. We also use it for recognition of Germanic genitive markers. In our web service, we used the web crawler and demo version of CLAWS. For the full version, a CLAWS license must be purchased.

As a reference corpus, we used an n-gram model based on the largest publicly-available, genre-balanced English corpus - the 520 million word Corpus of Contemporary American English (COCA)[2]. With this n-gram[3] data (2, 3, 4, and 5-word sequences, with their frequency), we conduct queries. The main advantages of using this corpus are that it is already genre balanced and includes part-of-speech tags. In addition, it includes the lemmatized forms of words and pre-calculated word and phrase frequencies. For faster processing, we converted the n-gram COCA corpus into an SQL database and pre-calculated all required 1-gram and 2-gram dependencies.

Our solution relies heavily on an automatic procedure. First, each part of the learner's text is tokenized and tagged with POS. This step aids the recognition of proper names and punctuation marks. In this context, CLAWS [11] was used, due to its high degree of accuracy. When we are comparing corpora of diverse sizes, it is important to normalize the frequencies of occurrence to a common base, such as per million tokens. Next, bigrams are extracted from each L2 text. Association scores are then computed. In this step, each bigram is searched in the corpus and is assigned its corresponding MI and t-scores, which are calculated by the formulas reported in Evert [13].

[1] http://ucrel.lancs.ac.uk/cgi-bin/claws72.pl

[2] http://corpus.byu.edu/coca/

[3] http://www.ngrams.info/intro.asp

The last step is the computation of our tool profiles. Our profiles of L2 texts are designed to use three major indices: MI, mean t-score, and proportion of absent bigrams. They are estimated by a combination of tokens and types in the texts. The MI score lets us count the association between two words depending on the independent relative frequency of the given two words. It does not depend on the size of the corpus. Even if the given corpora are of different sizes, it can be calculated. It outputs detailed information about lexical behavior.

The calculations are made in accordance with the following equations:

- Expected Frequency

$$E(\mathrm{w}1, w2) = \frac{f(w1)f(w2)}{N}$$

- MI

$$MI(w1, w2) = \log\frac{O(w1, w2)}{E(w1, w2)}$$

- T-score

$$t\,(w1, w2) = \frac{O(w1, w2) - E(w1, w2)}{\sqrt{O(w1, w2)}}$$

The solution topology, shown in Fig. 1, illustrates the time sequence and user actions during the lifetime of the solution:

- The vertical dashed line represents the lifetime of each component of the application, the time that component is active and running.
- The arrow represents an action triggered by one object to another (or to itself if the arrow is curved to start and end to the same object).
- The rectangle represents an object.
- An orange rectangle represents an object inside our solution.
- A blue rectangle represents an object outside our solution.

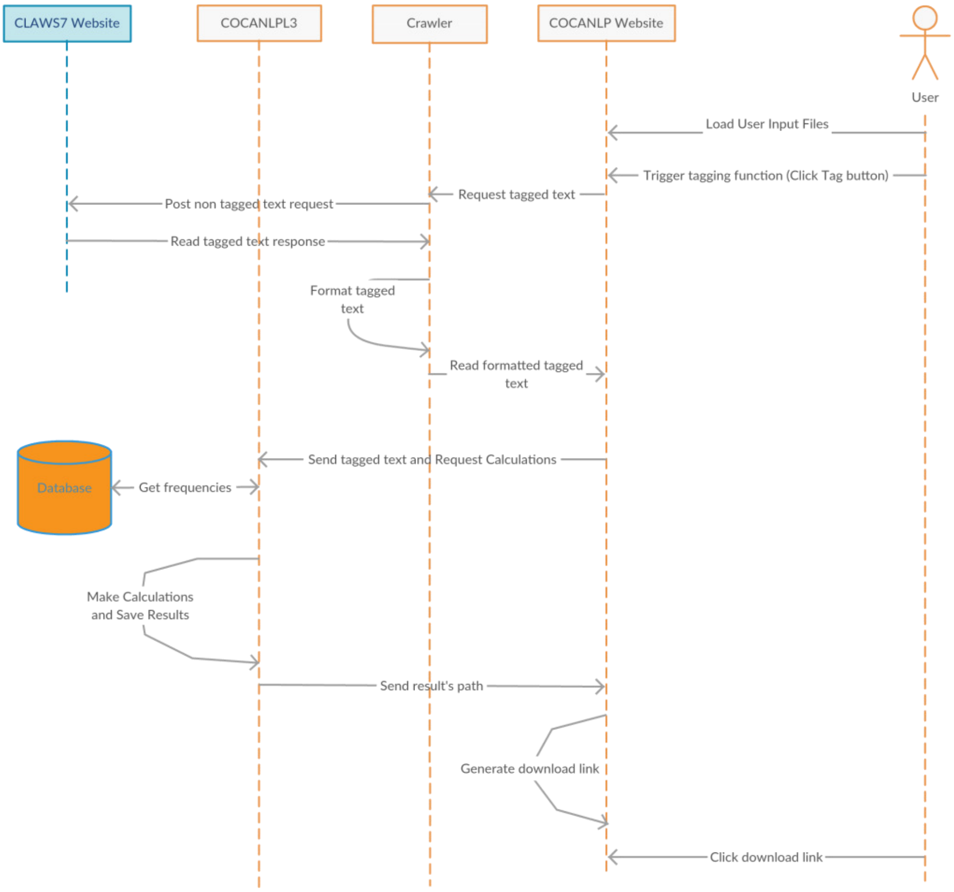
CLAWS7 Website
COCANLPL3
Crawler
COCANLP Website
User
Load User Input Files
Trigger tagging function (Click Tag button)
Request tagged text
Post non tagged text request
Read tagged text response
Format tagged text
Read formatted tagged text
Send tagged text and Request Calculations
Database
Get frequencies
Make Calculations and Save Results
Send result's path
Generate download link
Click download link

Fig. 1: Application topology

Time-sequence:

- Each point represents an arrow on the diagram.
- The application starts with a user action on the website (http://localhost/default) to select the input files (browsing the system's files and selecting the desired input text files).
- The user then triggers another action.
- The web application uploads the input files, reads them, and then sends a request to the crawler tool with the non-tagged text of each input file, requesting the corresponding tag text.
- The Crawler formats the non-tagged text appropriately.
- The Crawler waits for the response and reads it.
- The Crawler then extracts the output from the response of the HTML page.
- Then the Crawler sends the formatted tagged text to the website.
- The website redirects the formatted tagged text to our web application.
- Web application takes the tagged text, generates the bigrams, then communicates with the database to get the unigram and bigram frequencies using SQL stored procedures.
- Finally, our application makes all the calculations described above and saves a local (on the server) copy of the results as a zip file.
- Then sends the website a link to the saved copy of the results.
- Finally, the website generates a link of the zip file and displays it to the user.

IV. Experimental Evaluation

To evaluate our tool, 50 participants were asked to write an article on the same topic ("Memories from the best trip of their life."). Those stories were supposed to be between 1,000 and 1,200 words and were evaluated by our tool and by 10 random native English speaking teachers (having certified high proficiency in English). Instead of giving grades, they were supposed to mark all the corrections that needed to be done and to calculate the translation error rate (TER) metric.

TER was designed to provide a very intuitive machine translation evaluation metric, which requires less data compared with the other techniques while avoiding the labor intensity of human evaluation. It calculates the number of edits required to make a translated text exactly match the closest reference translation in fluency and semantics [13]. The TER metric calculation is defined in [14].

$$TER = \frac{E}{w_R}$$

where E represents the minimum number of edits required for an exact match. The average length of the reference text is given by w_R. Edits may include the deletion of words, word insertion, word substitutions, and changes in the word or phrase order [13]. In our research, this metric was used to measure the difference between students work and corrections made by teachers. It provided us a much more accurate evaluation than just grading. The TER result were compared with MI and t-scores, as presented in Table 1. All TER, t-score, and MI metrics were normalized to fit between 1 and 100 scale, where 100 means that the writing was perfect.

Table I.

Results of mining after progress

Sample No.	**TER**	**MI**	**t-score**
1	72.98	67.65	79.45
2	86.56	69.23	83.44
3	76.62	68.34	83.19
4	71.98	67.12	78.52
5	87.29	70.34	84.47
6	82.36	68.79	81.97
7	79.20	67.86	80.28
8	75.47	64.13	78.45
9	83.20	71.43	81.44
10	89.23	73.57	84.22
11	75.69	68.89	80.43
12	79.28	67.91	80.42
13	82.12	71.91	82.04
14	76.53	65.78	79.67
15	86.79	72.86	83.65
16	85.23	72.73	83.25
17	70.98	66.36	77.39
18	76.58	65.12	78.24
19	71.29	63.28	77.42
20	84.28	72.37	82.07
21	82.19	72.01	82.13
22	89.14	75.12	85.91
23	87.48	74.27	84.24
24	78.95	67.89	89.12
25	77.23	61.23	65.29
26	81.49	71.24	81.86
27	85.57	73.03	83.49
28	75.78	64.28	77.11
29	72.20	64.34	76.38
30	73.16	65.87	77.91
31	83.35	72.95	82.49
32	87.69	74.34	84.37
33	86.29	73.48	83.29

34	74.82	66.29	76.89
35	76.46	67.15	78.11
36	86.18	74.58	86.12
37	75.12	67.29	63.29
38	87.24	74.59	84.52
39	85.34	73.29	84.11
40	89.28	75.82	85.49
41	86.34	73.39	84.52
42	85.26	72.79	84.12
43	73.29	66.89	75.31
44	71.39	65.79	75.21
45	76.28	68.12	72.12
46	79.68	69.13	79.14
47	73.37	67.21	78.72
48	87.78	74.61	87.12
49	78.75	67.79	79.28
50	88.24	74.87	85.69

The results showed in Table I reveal a positive correlation between TER and MI scores, which means our tool is well suited for automatic student evaluation.

V. Discussion and Conclusions

In summary, our tool is capable of tracking the development of phraseological competency in L2 writing [12]. It can be easily adapted to support other languages. Only a language model change is required, along with use of a language-specific POS tagger and tokenizer. However, languages like Mandarin will require an additional segmenting step in the data pre-processing phase.

Our tool can identify collocations that are frequently used by learners, particularly native speakers of the language. Such information can help in writing L2 instruction.

Our technique evaluates the associated scores of every bigram, which are calculated on the basis of a reference corpus. A bigram is described by the study as any adjacent pair of words in the L2 text. This technique is also known as the unsupervised CollGram technique, on which there has been extensive research [10]. Previously mentioned research was also used to quantify the collocation power of each of three measures:

1. The mean MI score indicates the number of collocations that are produced from uncommon words.

2. The mean t-score measures the number of collocations produced from the collection of common words.

3. We also calculated the proportion of bigrams that are not present in the reference corpus and, therefore, will not be a part of any associated rating.

In the future to further improve the tool, we envision using multiple parameters to obtain the best analysis of the learner texts. For instance, we can remove spelling errors from identical pairs of words. Similarly, instances of adding or reducing one or two letters can also be discovered. POS tagging can be very useful in achieving the goal.

From the dataset, we empirically observed that the MI value relates to the bigrams. Such bigrams can contain a flawed combination of words or even a slightly creative combination. However, we have also observed that if there are punctuation marks in the text, then it will eventually interfere with the bigrams. This is because punctuations marks will not let the system record the readings and scores, and hence proper calculation will not be taken into account.

We can categorize the highest and lowest rated bigrams in the learner corpus. They can be categorized in diminishing order of the unqualified value of the MI and t-scores. The lowest rated bigrams in the category are the ones that exist in the reference corpus and will occur at a very small frequency.

Bigrams in the learner corpus that are not present in the reference corpus should have a prominent place in the analysis of the categories. On the basis of the theoretical framework, we can say that bigrams are of two types. First, one is the creative combination, which will most probably be used by advanced learners. Second, erroneous combinations will be produced in a very small quantity by advanced learners.

Statistical correlation is observed between the quality of text that was already scored, the MI score, the fraction of bi-grams not present in the system, and a combination of the two indices in question. This result enhances the quality of the prediction. [10]

Lastly, in the future we plan to extend the tool so that it can also calculate MI and t-score using trigrams and quadragrams. This is expected to improve the accuracy and analytic scope for linguists. We also plan to conduct an evaluation of domain-adapted language models [15].

VI. References

[1] J. Sinclair, John. „Corpus, concordance, collocation." Oxford University Press, 1991.

[2] R. Ellis. „Understanding second language acquisition." Oxford, UK: Oxford University Press, 1985

[3] M. Lewis. „The lexical approach: The state of ELT and a way forward." Hove, UK: Language Teaching Publications, 1993

[4] Office of Qualifications and Examinations Regulation, „Functional Skills Criteria for English Entry 1, Entry 2, Entry 3, Level 1 and Level 2", 2011

[5] R. Garside, N. Smith. „A hybrid grammatical tagger: CLAWS4", in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121, 1997

[6] N. Storch. „The impact of studying in a second language (L2) medium university on the development of L2 writing." Journal of Second Language Writing, 18, 103-118, 2009, DOI: 10.1016/j.jslw.2009.02.003

[7] N. Ellis. „Construction, chunking, and connectionism: The emergence of second language structure." In C. J. Doughty & M. H. Long (Eds.), The handbook of second language

acquisition (pp. 63-103). Malden, MA: Blackwell, 2003, DOI: 10.1002/9780470756492.ch4

[8] Y. Bestgen, S. Granger. „Quantifying the development of phraseological competence in L2 English writing: An automated approach". Journal of Second Language Writing, 2014, 26: 28-41, DOI: 10.1016/j.jslw.2014.09.004

[9] P. Durrant, N. Schmitt. „To what extent do native and non-native writers make use of collocations?" IRAL: International Review of Applied Linguistics in Language Teaching, 47, 157-177, 2009, DOI: 10.1515/iral.2009.007

[10] J. Billiet, B. Maddens, R. Beerten. „National identity and attitude toward foreigners." in a multinational state: A replication. Political Psychology, 2003, 24.2: 241-257, DOI: 10.1111/0162-895X.00327

[11] S. Granger, Y. Bestgen. „The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study." International Review of Applied Linguistics in Language Teaching, 2014, 52.3: 229-252, DOI: 10.1515/iral-2014-0011

[12] K. Wołk, K. Marasek. "Polish – English Speech Statistical Machine Translation Systems for the IWSLT 2014.", Proceedings of the 11th International Workshop on Spoken Language Translation, Tahoe Lake, USA, 2014, p. 143-149, DOI: 10.13140/RG.2.1.1128.9204

[13] S. Evert, "Corpora and collocations." Corpus linguistics. An international handbook 2, 2008, p. 1212-1248, DOI: 10.1515/9783110213881.2.1212

[14] Zhang, Y., Vogel, S., & Waibel, A. (May 2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system?. In LREC.

[15] Ma, W. Y., Ju, Y. C., He, X., & Deng, L. (2014). Language Model Adaptation through Shared Linear Transformations.