The Use of Text Mining in Identifying Means of Enhancing the ESP Teaching Process

Alexandra-Lavinia Istratie-Macarov
"Ovidius" University of Constanta, Faculty of Letters, Romania
lavinia.istratie@gmail.com

Abstract

Nowadays, English is studied at all academic levels. However, this has not always been so. With the advent of lifelong learning and due to the increasingly dynamic labor market, more and more people choose to take undergraduate courses and some of them do not have formal training in English. Thus, ESP teachers are facing the challenge of working with heterogenous classes. This article is analysis of 20 unstructured responses to the task "Agree or disagree with the statement that business and ethics are incompatible", given by 1st year accounting students enrolled in the distance learning program of "Ovidius" University in the academic year 2018-2019. The method used is text mining, by means of a free online tool, with the aim of identifying the degree of readability of the texts and any peculiarities of specificities of the discourse which may be used as anchors for curriculum development.

Key words: text mining, readability, ESP, curriculum development
J.E.L. classification: Z13

1. Introduction

When working with a heterogenous ESP classroom, it is difficult to choose or come up with a set of didactic materials which will fit all levels of competence and will bring added value to all students. In order to do so, there is need for an extensive analysis of the needs, knowledge and competences of the target group. While grammatical structures are the focus of pre-university structured learning, one of the aims of ESP is to build field-specific vocabulary. But neither approach can be exclusive, especially in the case of classes with different or no level of prior formal language instruction. Hence, this micro-level analysis of texts written by 1st year accounting students enrolled in the distance learning program of "Ovidius" University of Constanta, in order to identify commonalities and to develop the curriculum in a value-added manner.

2. Literature review

Text mining is a means to "extract and aggregate numerical data from textual documents" (Schouten, et al., 2019, p. 68). Besides being an increasingly appreciated research and analysis tool, it is being used by social media platforms, since it proved to be very profitable (Schouten, et al., 2019, p. 68).

A simple search in the Web of Science database yielded a staggering 2,098 results for papers which contain the words "text mining" in their title. However, only a handful concentrate on using this method in order to improve the learning experience, with an increasing interest since 2001.

Text mining has been used a computational linguistics tool. For instance, Alina Buzarna-Tihenea (Galbeaza) used this method in view of developing written business communication (Buzarna-Tihenea (Galbeaza), 2019, p. 146). Furthermore, together with Lavinia Nadrag, she made a cross-linguistic analysis of contracts of maritime law texts, in order to contribute to better understanding and translating such texts into Romanian (Nadrag & Buzarna-Tihenea (Galbeaza), 2016, p. 35).

As a support for teaching, Jason West proposed the use of text mining to "empirically analyze the breadth and depth of existing tertiary-level curricula to quantify patterns in curricula through the use of surface and deep cluster analysis" (West, 2017, p. 389), while Harakova and Rydval analyzed 6 storybooks used in class, in order to establish patterns of repetition and recycling for low-frequency words (Harakova & Rydval, 2015, p. 167). Similarly, Joorabchi et al. proposed to text mine specialized forums in order to identify the latest trends in the field of interest (Joorabchi, et al., 2015, p. 1170).

Although several studies analyze learner-generated content, most of them do not aim at improving the learning experience. Kong et al., however, designed an experiment in which keyword analysis is used before and after the teaching activity, in order to identify the frequency of the target vocabulary (Kong, et al., 2018, p. 369).

3. Research methodology

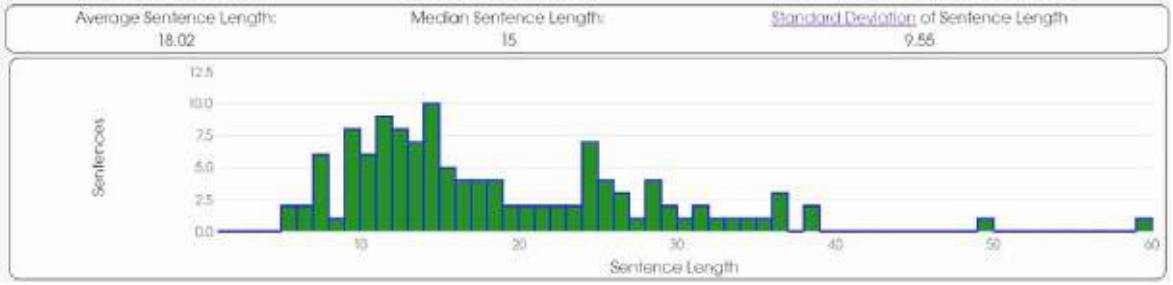
Our endeavor is a micro-level analysis of learner-generated contentment, which may provide useful information and prompt in a certain direction in the further development of the curricula. Twenty accounting students enrolled in the distance learning program of "Ovidius" University of Constanta, Romania, in the first year of study, were asked to agree or disagree in writing with the statement that business and ethics are incompatible. Other than that, no restrictions were imposed regarding the structure of the response or the content and no specific vocabulary or grammatical structures were activated. The students had different levels of competence in English, ranging between A2 and B2 (CEFR).

The obtained texts were analyzed using the free online tool located at the address http://www.analyzemywriting.com/. We focused on the following aspects: word length, sentence length, word frequency, readability and lexical density. The 20 inputs were analyzed separately and together, yielding similar results, with the exception of only one input, which exhibited a higher level of language competency.

4. Findings

The obtained corpus contains short answers, consisting in 1-6 paragraphs. The average sentence length is 18.02 words, with the highest consistency between 9 and 14 words. The sentences containing more than 30 words belong to the same paper. Therefore, a propensity for the simple, classical SPO structure of the sentence is remarked.

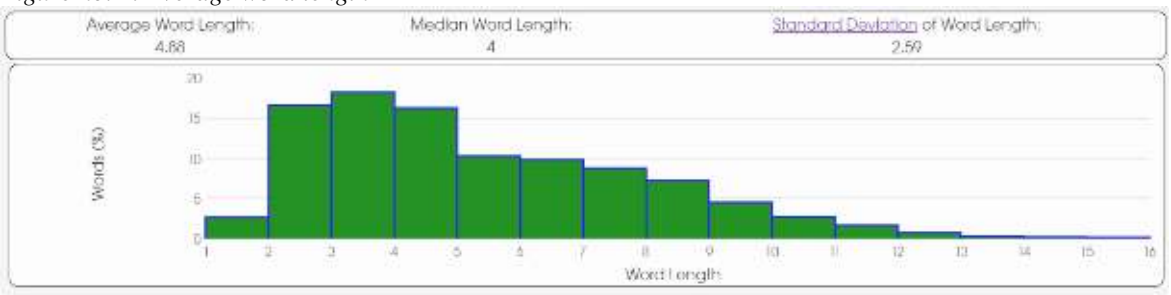
Figure no. 1. Average sentence length



Source: Generated on analyzemywriting.com, based on the corpus (Analyze My Writing, n.d.).

The average word length is 4.88 letters. 51.06% of the words have between 2 and 4 letters. The longest word has 16 letters.

Figure no. 2. Average word length



Source: Generated on analyzemywriting.com, based on the corpus (Analyze My Writing, n.d.).

The most common lexical word which was not provided in the question is "profit". This data shows the use of a limited, poor vocabulary, primarily made up of short words, consistent with a low level of language competence.

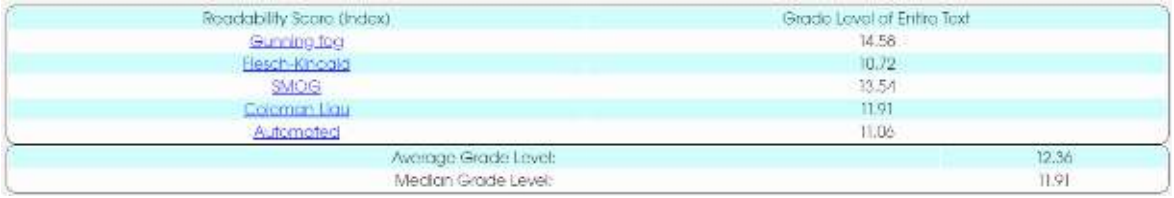
Table no. 1. The most common words

Rank	*Word*	*Number of occurrences*	*Percentage of total words*
1	the	89	4.1%
2	business	71	3.27%
3	to	70	3.22%
4	and	64	2.95%
5	ethics	54	2.49%
6	are	53	2.44%
7	that	49	2.26%
8	is	48	2.21%
9	in	43	1.98%
10	of	41	1.89%
11	a	34	1.57%
12	not	33	1.52%
13	I	24	1.1%
14	for	24	1.1%
15	profit	22	1.01%

Source: Generated on analyzemywriting.com, based on the corpus (Analyze My Writing, n.d.).

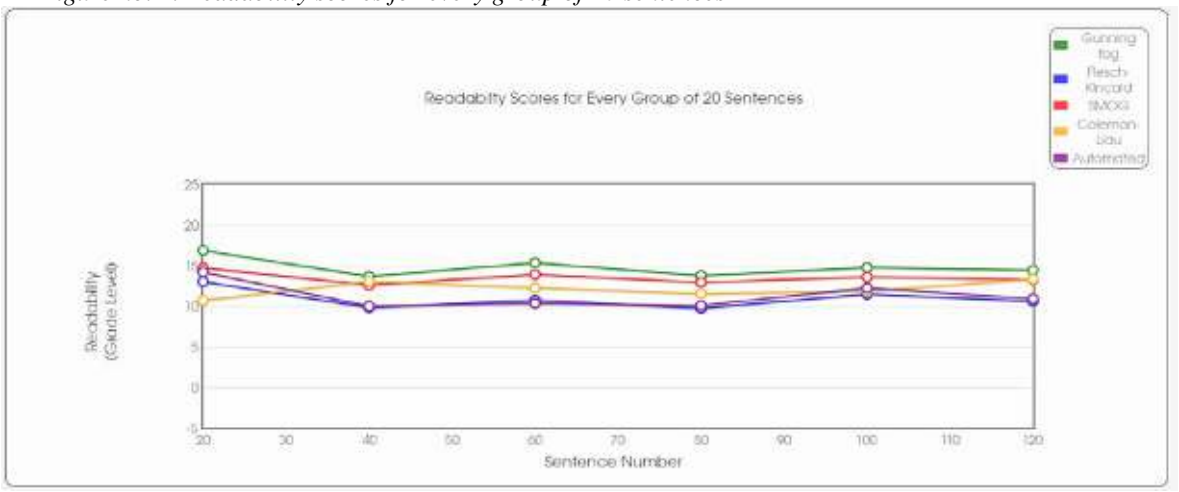
The degree of readability of the corpus was calculated based on 5 methodologies (Gunning fog, Flesch–Kincaid, SMOG, Coleman-Liau and Automated), with an average level of 12. This equates to the reading level of a High school senior in the USA, according to the Gunning fog index and the Coleman-Liau Index or of a college graduate, according to the Flesch–Kincaid index. This information is consistent with the participants' level of study.

Figure no. 3. Readability level of the corpus



Source: Generated on analyzemywriting.com, based on the corpus (Analyze My Writing, n.d.).

Figure no. 4. Readability scores for every group of 20 sentences



Source: Generated on analyzemywriting.com, based on the corpus (Analyze My Writing, n.d.).

The average lexical density of the corpus is 53.13%. The most common parts of speech are nouns (27.39%), prepositions (13.95%), verbs (11.69%), adjectives (8.84%), auxiliary verbs (8.38%), adverbs (5.2%) and pronouns (4.7%). This structure confirms the low level of lexical diversity and the simple sentence structure.

5. Conclusions

This study is limited to the specific target group, consisting of 20 first-year distance-learning accounting students with ages between 22 and 49 and with different levels of competence (mostly between A2 and B1, with one B2-level student). The rather short answers are consistent with the current trend of systematization in social media and electronic communication.

The results of the study are valid for the target group. They confirm the low level of competence in English, but the correct and consistent application of simple grammatical structures common to both languages (Romanian – the mother tongue and English – L2).

The study reveals that the target group requires a communicative approach, with the focus on lexical words both in the general spectrum and in that of accounting.

6. References

- Buzarna-Tihenea (Galbeaza), A., 2019. Written Business Communication. Case Study: Auditor's Report. *"Ovidius" University Annals, Economic Sciences Series,* 19(1), pp. 140-147.
- Harakova, T. & Rydval, J., 2015. Using text mining for the improvement of didactic tools in language acquisition. *Efficiency and responsibility in education 2015,* pp. 167-173.
- Joorabchi, A., English, M. & Mahdi, A. E., 2015. Text mining Q&A websites for supporting course design and curriculum development in higher education. *INTED2015: 9TH INTERNATIONAL TECHNOLOGY, EDUCATION AND DEVELOPMENT CONFERENCE,* pp. 1170-1178.
- Kong, S. C., Li, P. & Song, Y., 2018. Evaluating a Bilingual Text-Mining System With a Taxonomy of Key Words and Hierarchical Visualization for Understanding Learner-Generated Text. *Journal of Educational Computing Research,* 56(3), pp. 369-395.
- Nadrag, L. & Buzarna-Tihenea (Galbeaza), A., 2016. Aspects of Legal Translation in Contracts of Carriage. *Ovidius University Annals, Series Economic Sciences,* 16(1), pp. 35-40.
- Schouten, K., Frasincar, F., Dekker, R. & Riezebos, M., 2019. Heracles: A framework for developing and evaluating text mining algorithms. *Expert Systems with Applications,* Volume 127, pp. 68-84.
- West, J., 2017. Validating curriculum development using text mining. *Curriculum Journal,* 28(3), pp. 389-402.
- * * * Analyze My Writing, n.d. *Analyze My Writing*. [Online]
  Available at: http://www.analyzemywriting.com/ [Accessed 31 October 2019].