First Steps Towards NLP-based Formative Feedback to Improve Scientific Writing in Hebrew

Moriah Ariely
Weizmann Institute of Science
Rehovot, Israel
moriah.ariely@weizmann.ac.il

Tanya Nazaretsky
Weizmann Institute of Science
Rehovot, Israel
tanya.nazaretsky@
weizmann.ac.il

Giora Alexandron
Weizmann Institute of Science
Rehovot, Israel
giora.alexandron@
weizmann.ac.il

ABSTRACT

As scientific writing is an important 21st century skill, its development is a major goal in high school science education. Research shows that developing scientific writing skills requires frequent and tailored feedback, which teachers, who face large classes and limited time for personalized instruction, struggle to give. Natural Language Processing (NLP) technologies offer great promise to assist teachers in this process by automating some of the analysis. However, in Hebrew, the use of NLP in computer-supported writing instruction was until recently hindered by the lack of publicly available resources. In this paper, we present initial results from a study that aims to develop NLP-based techniques to assist teachers in providing personalized feedback in scientific writing in Hebrew, which might be applicable to other languages as well. We focus on writing inquiry reports in Biology, and specifically, on the task of automatically identifying whether the report contains a properly defined research question. This serves as a proof-of-concept of whether we can build a pipeline that identifies major components of the report and match them to a predefined grading rubric. To achieve this, we collected several hundreds of reports, annotated them according to a grading rubric to create a supervised data set, and built a machine-learning algorithm that uses NLP-based features. The results show that our model can accurately identify the research question or its absence. To the best of our knowledge, this is the first paper to report on the application of Hebrew NLP for formative assessment in K-12 science education.

Keywords

Scientific writing, Formative assessment, Natural Language Processing

1. INTRODUCTION

Writing is a critical 21st century skill, and a high level of writing proficiency is required to succeed in academia and workplaces [1]. In science, writing is one of the primary means of communication in the scientific community and a crucial aspect of scientific literacy. Thus, developing writing skills has become a major educational goal in high school science education [11].

Numerous studies have shown that developing scientific writing skills among high school students poses considerable difficulties for both students and teachers [9, 15, 23]. A lot of this may be due to the lack of formative feedback, which is known to be essential for the development of these skills [17, 10]. Formative feedback aims to guide and improve students' learning by providing them with information about the gap between their current and the desired performance. In the context of formative feedback on scientific writing, it has been shown that in order to support students in improving the quality of their writing, the formative feedback needs to be personalized and specific [3, 14]. It should also provide applicable recommendations for improvement, and explanations as to why such improvements are needed [16].

Proper writing instruction demands a significant amount of time from teachers, for preparing materials, reading, editing, and providing feedback. The educational reality is that teachers are faced with large class sizes that limit their ability to find the necessary time to devote to this process, resulting in a considerable delay in the feedback that students receive, and in its quality [1]. Another challenge is designing guidance that motivates students to engage in substantial writing revisions. Consequently, revising written explanations based on personalized guidance rarely occurs in science classrooms [19].

Technology holds much promise for improving this process, by supporting teachers in providing formative assessment. Automated computer scoring systems are being developed in order to address the challenges of assessing students' writing (e.g., [22, 19, 21, 18, 12, 20]). Among these, automated essay scoring technologies can enhance both large scale assessment and classroom instruction [3], as they have many advantages in the fields of assessment and instruction including objectivity, standardization and efficiency [5]. However, these technologies were mostly employed for *summative*, rather than *formative*, purposes [21].

In addition, while automated supporting tools for revising texts on the micro-level (such as grammar and spelling) are well represented [18], tools that support the development of writing strategies including self-monitoring and improving

Moriah Ariely, Tanya Nazaretsky and Giora Alexandron "First Steps Towards NLP-based Formative Feedback to Improve Scientific Writing in Hebrew" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 565 - 568

macro-level of text quality (such as argumentative structure and rhetorical moves) are infrequent [18]. In the transition from essay scoring to writing instruction, feedback design is of great importance, since it is the primary means through which students can evaluate and improve their writing [17].

In Hebrew, the use of NLP in computer-supported writing instruction was until recently hindered by the lack of publicly available resources. Hebrew is a morphologically rich language (MRL). It is complex, characterized by a highly productive inflectional morphology, with prefixes, suffixes and circumfixes, and also nouns, adjectives and numerals inflections for number (amount) and gender [8]. Following recent developments in Hebrew NLP, the high-level goal of our research is developing NLP-based techniques to assist teachers in providing personalized feedback in scientific writing in Hebrew, which might be applicable to other languages as well.

The task that we study is writing scientific reports in inquiry projects in Biology. A fundamental component of the report is a well-formulated research question(s). Formulating research questions that can be answered empirically is one of the practices needed in order to become scientifically literate [11]. In fact, by 'composing questions', students attend to the main ideas and check if the content is understood [13]. Since the research questions are defined on the early stages of the project, failure to properly define them can have a long-term effect on the quality of the project and report.

With this rationale in mind, we focus on studying NLP-based means to provide personalized feedback on the quality of the questions that students define. A precondition for an automatic assessment of the quality of the research questions is detecting them automatically in the text. The identification of the research questions in students' essays serves as a proof-of-concept (POC) for learning a formative assessment grading scheme for major components of the report.

Our work is the first step towards NLP-based tools that will support K-12 science educators in teaching and assessing scientific writing in Hebrew. To the best of our knowledge, there is no published work on NLP-based formative assessment in Hebrew, and this research has the potential to pioneer this exciting domain.

2. METHOD AND RESULTS

This section describes the experimental setup, how the data was collected and annotated, the NLP pipeline and features, the machine learning algorithm, and the results.

2.1 Research context

Over 20,000 high-school students in Israel major in biology each year [4]. The Israeli Biology curriculum includes an inquiry project that constitutes 30% of the final grade [7]. It is conducted collaboratively in groups of 2-3 students. The students conduct an inquiry on a biological issue, ask research questions, design and carry out an experiment, collect data, and analyze it. Students are required to document their work in a scientific report. Within this process, the writing task was reported by teachers and students to be the most challenging part [6]. It is an iterative process, which often takes up to 10 iterations to complete.

2.2 Data Collection

The data include 705 scientific reports, collected from 520 student groups that belong to 33 classes.

The reports are submitted in Hebrew as Word documents. In the first phase of the project the Introduction part, which is where the research question should be defined, was separated from the rest of the text. The Introduction typically consists of 2-5 pages. Well-written introduction section should contain the following discourse categories:

- Biological process.
- Research question. One research question if the work is submitted by two students, two research questions if the work is submitted by a group of three students.
- Research hypothesis.
- Description of the organism.

Following is an example of a typical well-written research question: "Our research question is how does alcohol concentration influence cell respiration rate in yeasts".

2.3 Creating a supervised data set

In order to create a supervised data set that can be used as an input to the machine learning algorithm, we annotated students' texts. The goal of the annotation was to mark the segments of the texts that represent the aforementioned four discourse categories. The relevant parts of the texts were encoded with <tagname> and </tagname> tags that preceded/succeeded the relevant segments. For example, each research question was preceded with an <rq> tag, and succeeded by an </rq> tag, which were inserted into the text. The annotation was performed at a sentence or multiple sentences level. Each sentence was labeled with at most one discourse category. Our annotation scheme does not allow overlapping of the categories, but the same category may appear multiple times (e.g., two different research questions). We note that the majority of the sentences do not belong to any category and are not labeled at all.

The process was conducted by two domain experts (including one of the authors). The experts first created a grading rubric and then tagged the texts accordingly. In the first stage of the annotation process, both judges worked together to create a protocol for detecting the discourse elements in the text. Next, they worked independently to label 147 texts (from 44 student groups that belong to 6 classes), and the resulting labels were discussed until disagreements were resolved. Finally, additional 56 texts were labeled by one of the experts.

To create a training and test sets we chosen randomly one report from each student group, so the chosen reports represent different stages of report readiness. This means that some of the reports do not contain research questions at all and some research questions are ill defined. In total, the data set includes 100 texts containing 5513 sentences and 197 research questions.

2.4 Research Question identification

We consider the task of research question identification as a sentence-level classification task. Each sentence is classified as a research question or not. The data set was divided into training and test sets, as presented in Table 1.

Table 1: A summary of the annotated data.

	Number of texts	Number of sentences	Number of research questions
Training set	70	4013	139
Test set	30	1500	58
Total	100	5513	197

One of the challenges is that the data set is highly imbalanced. The ratio between examples in the minority class (research question sentences) and the majority class (non-research question sentences) is less than 1:25. Thus, a naive classification algorithm returning a negative answer for all the sentences will achieve 96.4% accuracy, but it is of no practical value. To evaluate the goodness-of-fit of our algorithm, we use the following measures:

- Precision = $\frac{TruePositive}{(TruePositive+FalsePositive)}$
- Recall = $\frac{TruePositive}{(TruePositive+FalseNegative)}$
- F-measure = $\frac{2\times Precision\times Recall}{Precision+Recall}$

2.5 Parsing and feature engineering

2.5.1 Parsing

We use the Hebrew morphological parser developed by the National Institute for Testing and Evaluation (NITE) [2]. It is used to resolve morphological and parts of speech (POS) disambiguity. The reported accuracy of the NITE parser is 90% for the full morphological analysis and 95% for POS analysis.

Running the parser on the annotated student texts generates a tab-separated value file. Each row in the file corresponds to one word in the text, and contains the following information:

- isResearchQuestion: $True/False$ - indicates whether the word is part of a research question sentence
- word original form: the word as appears in the text
- word basic form: the base form of the word
- POS: part of speech of the word

2.5.2 Bag of Words and feature set

First, we construct a Bag of Words (BOW) dictionary as follows:

1. Divide the data set randomly into training and test sets as presented in Table 1.
2. Build a BOW dictionary containing the basic form of each word that appears at least three times in a research question text segment (within the training set), and its corresponding POS.
3. Remove stop words: numbers, punctuation marks except for question mark, prepositions, pronouns, auxiliary verbs, all forms of the word "the" (could appear in a number of forms in Hebrew)

Then, for each sentence in the data set, we compute the following set of features:

- We introduce a feature for each BOW dictionary entry. The value of the features is defined as the number of appearances of the corresponding dictionary entry in the sentence.
- In addition, human experts composed a list of phrases that can be used as markers for a research question, such as "what is the connection", "what is the relation", etc. (in Hebrew, due to word agglutination, these phrases consist of two words only). We introduce an additional Boolean feature to represent the appearance of any of these phrases.

2.6 Results

We used the training set to train three types of classifiers: SVM, Logistic regression, and Naive Bayes. Their performance, computed over 500 5-fold cross-validation iterations, is presented in Table 2 (mean values). The best performance was achieved by the Logistic Regression classifier. To evaluate the performance on unseen data, we then trained a logistic regression classifier on the entire training set, and measured its performance on the test set. The results are presented in Table 3.

Table 2: The results of 500 5-fold cross validation runs on the training set

	Precision	Recall	F-measure
Logistic Regression	86.9%	74.3%	79.9%
SVM	75.9%	77.3%	75.9%
Naive Bayes	62.0%	88.6%	72.8%

Table 3: The results of the Logistic Regression model on the test set

	Precision	Recall	F-measure
Logistic Regression	84.2%	94.1%	88.9%

To understand the source of the errors we examined the sentences missed by the classifier. The main source of the errors is related to the failure of the parser to treat correctly a point sign '.' inside Latin names of organisms (e.g., 'E. Coli', 'St. Albus'). As a result, sentences containing such names were considered by mistake as two separate sentences and the classifier failed to identify them as a research question.

3. NEXT STEPS

Next, we plan to extend our model to identify the internal structure of the research question, as defined in the grading rubric. To support this step, the annotation scheme was extended to identify the required components of the research question: *opening* (e.g., "Our research question is:"), *independent variable* (e.g., ethanol concentration), *dependent variable* (e.g., cellular respiration rate), *connection between the variables*, and *organism* (e.g., bacteria, yeast). As this rubric is designed to be the basis for generating formative feedback, the experts gave a score (0-2) to each of these components, as well as an additional score for the *location* of the entire sentence in the text. This scheme was applied to 115 texts in a process similar to the one reported in Subsection 2.3. We also used the texts to create synthetic examples. In case the final version of a particular report was not well-written, the judges fixed the writing and inserted

the fixed version as an additional example. In total, 32 additional examples were created in this manner.

Based on this, we intend to create a computational model for identifying the internal structure, and use it to conduct an intervention study, in which students will receive formative feedback that is based on the computational analysis of the research question structure. In parallel, we will extend our method to the identify the remaining three discourse categories (biological process, research hypothesis, and description of the organism).

4. CONCLUSIONS

This paper presents preliminary results from a study that aims to develop NLP-based tools to assist teachers in providing formative feedback on scientific writing in Hebrew. Specifically, we demonstrate that our model can accurately identify the *research question* (or its absence), which is a key component of the specific writing task that we study (scientific report of inquiry project in Biology). Our results, although very preliminary, are a first step towards using NLP to provide formative assessment on scientific writing in Hebrew. To the best of our knowledge, there is no prior work that applies Hebrew NLP to provide formative feedback in K-12 science education.

5. ACKNOWLEDGMENTS

The authors thank Cipy Hofman and Yona Dolev for their contribution, and the National Institute for Testing and Evaluation (NITE) for providing access to the Hebrew morphological parser, and for partially funding this project. This research is supported by The Willner Family Leadership Institute for the Weizmann Institute of Science, Iancovici-Fallmann Memorial Fund, established by Ruth and Henry Yancovich, and by Ullmann Family Foundation.

6. REFERENCES

[1] L. K. Allen, M. E. Jacovina, and D. S. McNamara. Computer-based writing instruction. In *Handbook of writing research*, pages 316–329. The Guilford Press, New York, 2016.

[2] A. Ben-simon and Y. Cohen. The Hebrew Language Project : Automated Essay Scoring & Readability Analysis. In *IAEA Annual Conference*, January 2011.

[3] J. Burstein, D. Marcu, and K. Knight. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39, 2003.

[4] Central Bureau of Statistics. Trends in Math and Science in Upper Secondary Education, 2006-2016 [Press release], 2018.

[5] Y. Cohen, E. Levi, and A. Ben-Simon. Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, 31(3):241–250, 2018.

[6] B. Galia Zer-Kavod Advisor and A. Yarden. *Thesis for the degree Doctor of Philosophy*. PhD thesis, Weizmann Institute of Science, 2017.

[7] Israeli Ministry of Education. Syllabus of Biological Studies (10th-12th grade). Technical report, 2011.

[8] A. Itai and S. Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, 2008.

[9] G. J. Kelly and C. Bazerman. How students argue scientific claims: A rhetorical-semantic analysis. *Applied Linguistics*, 24(1):28–55, 2003.

[10] H. McGarrell and J. Verbeem. Motivating revision of drafts through formative feedback. *ELT Journal*, 61(3):228–236, 2007.

[11] National Research Council (NRC). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Technical report, 2012.

[12] R. H. Nehm, M. Ha, and E. Mayfield. Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21(1):183–196, 2012.

[13] J. Osborne. Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, 25(2):177–196, 2014.

[14] N. Pendar and E. Cotos. Automatic identification of discourse moves in scientific article introductions. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–70, 2008.

[15] R. Porter, K. Guarienti, B. Brydon, J. Robb, A. Royston, H. Painter, A. Sutherland, C. Passmore, and M. H. Smith. Writing better lab reports. *The Science Teacher*, 77(1):43–48, 2010.

[16] E. Riedel, S. L. Dexter, C. Scharber, and A. Doering. Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *Journal of Educational Computing Research*, 35(3):267–287, 2006.

[17] R. D. Roscoe, L. K. Varner, S. A. Crossley, and D. S. McNamara. Developing pedagogically-guided algorithms for intelligent writing feedback. *Grantee Submission*, 8(4):362–381, 2013.

[18] C. Strobl, E. Ailhaud, K. Benetos, A. Devitt, O. Kruse, A. Proske, and C. Rapp. Digital support for academic writing : A review of technologies and pedagogies. *Computers & Education*, 131:33–48, 2019.

[19] C. Tansomboon, L. F. Gerard, J. M. Vitale, and M. C. Linn. Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, 27(4):729–757, 2017.

[20] J. Wilson, R. Roscoe, and Y. Ahmed. Automated formative writing assessment using a levels of language framework. *Assessing Writing*, 34:16–36, 2017.

[21] B. Woods, D. Adamson, S. Miel, and E. Mayfield. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 2071–2080, 2017.

[22] M. Zhu, O. L. Liu, and H.-S. Lee. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers Education*, 143:103668, 2020.

[23] M. Zion, S. Cohen, and R. Amir. The spectrum of dynamic inquiry teaching practices. *Research in Science Education*, 37(4):423–447, 2007.