Towards Automated Assessment of Scientific Explanations in Turkish using Language Transfer

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

Haci Hasan Yolcu  
Kafkas University  
Kars, Türkiye  
hasanyolcu@kafkas.edu.tr

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

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

ABSTRACT
The paper presents a preliminary study on employing Natural Language Processing (NLP) techniques for automated formative assessment of scientific explanations in Turkish, a morphologically rich language with limited educational resources. The proposed method employs zero and few-shot language transfer techniques for creating Turkish NLP models, obviating the need for extensive collection and annotation of Turkish datasets. The study utilizes multilingual BERT-based pre-trained transformer models. It evaluates the effectiveness of different fine-tuning approaches using an existing annotated dataset in Hebrew. The results indicate that, despite being trained using non-perfectly automated translations from Hebrew responses, the best-performing models demonstrated adequate performance when evaluated on authentic Turkish responses. Thus, this research may provide a useful method for building automated scientific explanations assessment models that are transferred between languages.

1. INTRODUCTION
Constructing scientific explanations is one of the core practices in science. Writing good causal explanations in biology requires students to provide a conceptual framework for the observed phenomenon, identify relevant information, infer the unobservable world, grasp underlying causes, and link the causes logically [19, 9, 12]. Biology teachers use open-ended constructive-response items to elicit students’ in-depth understanding of scientific concepts and mechanisms. However, answering such open-ended items is a challenging task. Students often struggle to write answers formulated in their own language [17]. Receiving formative feedback that is just and personalized is crucial in allowing students to relate to the missing or wrong parts of their answers and improve their responses accordingly [21, 24, 2].

Natural Language Processing (NLP) holds much promise for automation of this process[28, 5], especially in English [17, 10, 18, 15, 16]. However, for languages like Turkish, Hebrew, and Arabic, a combination of being morphologically rich (where each input token may consist of several functional units, e.g., multiple suffixes and prefixes added to the original word root), and relatively low resource in the educational domain, makes applications of NLP in such languages particularly challenging [25]. To our knowledge, little research exists in this area [1, 3, 6, 4, 8]. [3] proposed a method for automated formative assessment of scientific explanations in Hebrew based on analytic rubrics. [6] presented the first application of Turkish NLP for automated summative assessment of Physics open-ended questions. In the context of summative assessment of short essays in the Arabic language, which is morphologically rich too, [4] used latent semantic analysis and rhetorical structure theory, and [8] used human and automated translation to English to overcome the shortage in Arabic NLP educational resources.

We are unfamiliar with more recent research on NLP-based scoring of open-ended questions in Turkish or Arabic. This work is the first step towards NLP-based tools that can support K-12 science educators in providing formative feedback on scientific writing in Turkish. We propose and evaluate a method for creating Turkish NLP models with no need to collect and annotate large datasets in Turkish while using the corresponding annotated dataset in a different language (e.g., Hebrew). Based on this goal, our research questions are formulated as follows:

- Can our models accurately grade unseen responses in Turkish to an item after being trained on Hebrew responses to several items related to the same biological phenomenon?
- Can fine-tuning using a small number of Turkish responses improve the performance of our models?

2. METHODOLOGY
2.1 The instrument
The instrument consisted of two open-ended items about the effect of Smoking and Anemia on the human ability to exercise. Both items refer to the role of red blood cells (RBC) and Hemoglobin, blood circulation, and energy production in cells on humans’ physical activity ability. These topics are
part of the Israeli and Turkish high school science curricula. The instrument was constructed in English and Hebrew as part of our previous study [3]. One of the authors manually translated it into Turkish (Table 1).

2.2 Data collection
The research population for this study is high school students in Israel and Türkiye. The research sample included 669 Israeli students (25 schools), 10-12 graders, 70% females, and 84 Turkish students (2 schools), 11-graders, 61% females. The instrument was administered to the students by their teachers, who we contacted through teacher professional communities. The data was collected anonymously using an online Google form, the students were requested to fill in their gender, grade, and school name only. In both languages, several correct responses were written by the teachers. In total, 2007 responses in Hebrew and 174 in Turkish were collected.

2.3 Grading rubric and data annotation
This study used the analytic grading rubric created as part of our previous study [3] and aimed at assisting teachers in formative assessment tasks [2]. Each rubric category represents an essential element in the causal chain, constituting a complete scientific explanation. The original rubric consisted of 11 categories. In this study, we used 7 of them (Table 2), excluding the 4 categories challenging for Turkish students yielding highly unbalanced datasets with only 0 to 5 correct answers. We used the grading obtained in our previous study for the Hebrew responses. The Turkish responses were graded as follows. First, two raters (a biology high school teacher and one of the authors) graded all the answers separately according to the analytic rubric mentioned above. Next, the raters resolved all the conflicts and came to a complete agreement.

2.4 Turkish NLP pipeline
2.4.1 BERT language models
Using a transformer deep-learning architecture has led to the development of few shot learning - a method of fine-tuning ML models based on very small amounts of annotated data [26] using state-of-the-art language models pre-trained on enormous amounts of textual data in one or several languages. In this research, we employ the few-shot learning approach for sentence classification using several BERT models: the BERT multi-lingual language model pre-trained on the concatenation of Wikipedia in 104 different languages (DistilmBERT) and the BERT model pre-trained on Turkish language (DistilBERTurk) [22], and the Hebrew AlephBERT model [23].

2.4.2 Text preprocessing
All the original Hebrew responses were part of the training set. The pre-processing consisted of several steps. First, the Hebrew responses passed automated spelling corrections (e.g., the critical word "Hemoglobin" was misspelled in tens of different ways) and replacement of the Hebrew acronyms (e.g., "RBC" was replaced with "red blood cells") with the entire words. Second, the responses were Google-translated automatically into Turkish. Third, we examined the quality of the automated translation. Although the translation was not perfect and, in some cases, was even unsatisfactory (e.g., "Red blood cells contain Hemoglobin to which oxygen binds." was translated as "Kırmızı kan hücreleri, kardeş oksijene bağlanılı hemoglobin içerir." meaning "Red blood cells contain hemoglobin, which is associated with its sister oxygen."), we decided to proceed with the translated data as is.

2.4.3 Fine-tuning and text augmentation
Data augmentation is a typical solution to the problem of unbalanced and very small datasets (like our Turkish dataset) by generating new examples for the minority classes. The newly generated examples are supposed to be different from the original ones but carry the same semantic meaning and label as an original text. It is shown by previous research that text augmentation can significantly improve the resulting models’ performance [14]. This paper employed two standard paraphrasing augmentation techniques: back translation [29, 27] and using hand-crafted rules (fixed heuristics) [7]. The back translation was done by automatically translating the positive examples for each category into 11 languages and back (Table 4). In addition, the following rules were introduced for paraphrasing. First, we replaced the words with similar meanings (e.g., red blood cells "kırımzi kan hücresi" is a synonym to "alyuvar" and "eritrosit" and can also be replaced by "hemoglobin" in our context) and chemical acronyms and abbreviations with the words (e.g., CO, O2, and ATP were replaced by "karbonmonoksit", "oksijen" and "enerji" respectively). Second, we combined each positive example (per category) with several negative examples (e.g., the concatenation of the two responses in Table 3 can create an augmented answer with all positive categories.)

2.5 Experimental setup
To answer the research questions, we performed five experiments. To allow a fair comparison between zero-shot and few-shot models, we divided the Turkish responses dataset into 5 folds and ran each experiment 5 times per each fold and category. Each time 4 out of 5 folds were used as a test set (n = 139). The fifth fold (n = 35) was not used in the case of zero-shot experiments (Exp. 1-3) and was used as a source for fine-tuning (referred to as "few-shot set" below) using authentic Turkish responses (Exp. 4.5). Below we describe each experiment’s settings in more detail.

Exp. 1 Zero-shot with multilingual DistilmBERT. Both Hebrew training (n=2007) and Turkish test datasets (n = 139) were used as is, without preprocessing.

Exp. 2 Zero-shot with Hebrew AlephBERT. The Hebrew training (n = 2007) was used without preprocessing. The Turkish test set (n=139) was auto-translated into Hebrew.

3https://huggingface.co/distilbert-base-multilingual-cased
4https://huggingface.co/dbmdz/distilbert-base-turkish-cased
5https://github.com/OnlpLab/AlephBERT

4by GoogleTranslator from deep_translator Python package
5English (En), Finnish (Fi), German (De), Greek (Gr), Hebrew (Iw), Italian (It), Japanese (Ja), Persian (Fa), Tatar (Tt), Ukrainian (Uk), Uzbek (Uz)
Experiments by text augmentation with Turkish DistilBERT-Turk. The training set consisted of the Hebrew training set (n = 2007) as in Exp. 3 combined with the augmented by backtranslation and application of augmentation rules (Subsection 2.4.3) few-shot Turkish set. The size of the augmented few-shot set varied from 300 to 400 depending on the number of positive examples per category. The Turkish test set (n = 139) was used as is.

We fine-tuned the pre-trained models end-to-end (including all transformer layers, the pooling layer, and the final dense output layer) with the Adam optimizer (learning rate = 2e-6, learning warmup = 600) over 5 epochs to minimize the binary cross-entropy loss which is consistent with typical BERT fine-tuning for text classification [11].

### 3. RESULTS AND DISCUSSION

The performance of the Multilingual models (Exp. 1) was unsatisfactory (Table 5). We attribute the failure of multilingual models to generalize to the different subject (S), object (O), or verb (V) order in Turkish and Hebrew. Both Turkish and Hebrew have flexibility in word order. For example, the sentence “Red blood cells carry oxygen to the cells” can be written in Turkish in several ways depending on the connotation of emphasis on the importance of either the subject, object, or verb. However, the typical order in Turkish is SOV. For example, the authentic answer is Turkish “Alyuvarlar hücrelere oksijen taşır” when translated into English (preserving the word order) would be “Red blood cells to the cells oxygen carry” However, the typical order in Hebrew would be SVO, as in English. This typological dis-similarity and zero lexical overlaps between Hebrew and Turkish (which use entirely different scripts) possibly reduce the multilingual model’s power of zero-shot language transfer between Hebrew and Turkish[20]. Both zero-shot models based on the automated translation (Exp. 2 and Exp. 3) showed a significant improvement over the multilingual models (Table 5). They performed pretty similarly with a slight advantage towards the AlephBERT-based model (Exp. 2). However, in our context, the critical advantage of using DistilBERTTurk is automated translation in the training stage. After the training is completed, the real assessment systems based on the resulting models can work with authentic student responses in Turkish. The above guided our decision to try improving Exp. 3 models by fine-tuning using authentic Turkish responses.

The straightforward fine-tuning of the DistilBERTTurk models (Exp. 4) using a small number (n = 35) of authentic Turkish examples did not improve most models’ performance (Table 5). It even was a minor degradation compared to Exp. 3. The fine-tuning of the DistilBERTTurk models (Exp. 5) using augmentation performed similarly to vanilla DistilBERTTurk (Exp. 3). Yet, there was an improvement (from slight to moderate agreement) for the most problematic category b.

### 4. CONCLUSIONS AND NEXT STEPS

This paper presents the results of a study on the automatic scoring of scientific explanations in Biology conducted in Turkish using state-of-the-art language transfer methods.
Our models, trained based on a non-perfectly automated translated Hebrew training dataset, were analyzed on authentic responses written in Turkish. Using back translation for text augmentation, the best-performing models achieved good and very good agreement with human raters in 5 out of 7 and moderate agreement in 2 rubric categories. Notably, these two categories (b and c, see Table 2) were also the hardest to achieve satisfactory performance in the original Hebrew models [3].

The main limitation of this study is the size of the dataset used to evaluate the models. We plan to collect additional data in Turkish to check if the results are robust. Our previous study in Hebrew estimated the number of required responses to achieve the satisfactory performance of the models [3] as 500 – 900. Following the successful implementation of the back translation augmentation method in Turkish, we plan to investigate if the back translation can significantly reduce these numbers in original Hebrew models.

In Hebrew, our method is already implemented in PeTeL, a free learning management platform serving about a thousand science teachers in Hebrew and Arabic. We consider the presented results as a proof of concept of our ability to generalize our system to other (even very different, like Turkish) languages using language transfer, with no need to collect additional training data. Our next steps are to extend this study to the Arabic language.

5. ACKNOWLEDGMENTS

TN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.
6. REFERENCES


