Towards Automated Assessment of Scientific Explanations in Turkish using Language Transfer

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

Hacı 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].

T. Nazaretsky, H. H. Yolcu, M. Ariely, and G. Alexandron. Towards automated assessment of scientific explanations in turkish using language transfer. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 453–457, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115770

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[1]) and the BERT model pre-trained on Turkish language (DistilBERTurk[2]) [22], and the Hebrew AlephBert model[3] [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, *kardeşi* oksijenle bağlantılı 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[4] the positive examples for each category into 11 languages[5] 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 "kırmızı 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.

[1]https://huggingface.co/distilbert-base-multilingual-cased
[2]https://huggingface.co/dbmdz/distilbert-base-turkish-cased
[3]https://github.com/OnlpLab/AlephBERT

[4]by GoogleTranslator from *deep_translator* Python package
[5]English (En), Finnish (Fi), German (De), Greek (Gr), Hebrew (Iw), Italian (It), Japanese (Ja), Persian (Fa), Tatar (Tt), Ukrainian (Uk), Uzbek (Uz)

Table 1: The Instrument in Turkish and English.

Turkish version	English version
Anemia Item	
Kan testinde kırmızı kan hücre miktarının az olduğu kişiler anemi hastası olarak tanımlanır. Bu insanlar halsizlikten ve egzersiz yapmaktaki zorluktan şikâyet ederler. Az miktarda kırmızı kan hücrelerine sahip anemi hastalarının egzersiz yaparken zorluk yaşamalarının nedenini açıklayınız.	A person was found to have low levels of red blood cells in his blood test (anemia). This person complained to his doctor about weakness and difficulty to exercise. Explain how low levels of red blood cells make it difficult for people with anemia to exercise.
Smoking Item	
Sigara dumanı karbon monoksit (CO) gibi birçok zararlı maddeyi içermektedir. Sigara içerken CO salınımı olur. CO hemoglobine bağlanmada oksijenden daha etkindir. Sigara içenlerde yüksek CO seviyesi egzersiz yapmayı zorlaştırmaktadır, bu durumu nasıl açıklarsınız?	The smoke from cigarettes contains several harmful substances, including the gas carbon monoxide (CO). CO is released from cigarettes while smoking and has a stronger tendency than oxygen to bind to Hemoglobin. Explain how high levels of CO make it difficult for smokers to exercise.

Table 2: The categories of the analytic rubric for the Anemia and Smoking Items. The + and − signs represent if the category is relevant to the item. The percent indicated the percentage of the correct answers per category.

	Category Name	Anemia Item		Smoking Item	
a	Changes in oxygen levels that bind to Hemoglobin/RBC	−	−	+	40%
b	The role of Hemoglobin/RBC in oxygen transportation	+	45%	+	23%
c	Changes in oxygen levels in the body (general)	+	29%	+	24%
d	Changes in oxygen levels in the cells (micro level)	+	10%	+	8%
e	Oxygen is a reactant in energy production	+	7%	+	6%
f	Changes in energy/ATP levels	+	9%	+	9%
g	Using the term energy/ATP	+	23%	+	9%

Exp. 3 **Zero-shot with Turkish DistilBERTTurk.** The Hebrew training (n = 2007) was preprocessed and auto-translated into Turkish (Subsection 2.4.2). The Turkish test set (n = 139) was used as is.

Exp. 4 **Few-shot with Turkish DistilBERTTurk.** The training set consisted of the Hebrew training set as in Exp. 3 combined with the few-shot Turkish set (n = 2007 + 35 = 2042). The Turkish test set (n = 139) was used as is.

Exp. 5 **Few-shot by text augmentation with Turkish DistilBERTTurk.** 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.

Table 3: Example of student answers to Anemia and Smoking Items in Turkish and English with the corresponding gradings.

Typical Student Answers		a	b	c	d	e	f	g
Anemia Item								
Egzersiz yaparken enerjiye ihtiyaç duyarız ve bu enerjiyi oksijenli solunum yapar. Oksijeni hücrelere alyuvarlar taşır. Eğer hemoglobin az olursa oksijenli solunum az olur ve açığa çıkan enerji azalır.	When we exercise, we need energy and oxygen respiration makes this energy. Red blood cells carry oxygen to the cells. If the hemoglobin is low, aerobic respiration will be less and the energy released will decrease.	–	1	0	0	1	1	1
Smoking Item								
Sigara içenler sigaranın yanması sonucu açığa çıkan CO gazına daha fazla maruz kalır. CO gazı O2'nin yerine hemoglobinlere bağlanır. Hücrelere ihtiyacı olan yeterli O2 gazı taşınamaz. Hücre metabolizmasında aksaklıklar gözlenir. Bu sebeple kaslar daha çabuk yorulur. Çabuk yorulduklardından egzersiz yapmayı zorlaştırır.	Smokers are more exposed to the CO gas released as a result of cigarette combustion. CO gas binds to hemoglobins instead of O2. The cells cannot carry enough O2 gas that they need. Disturbances in cell metabolism are observed. For this reason, the muscles get tired more quickly. They get tired quickly, making it difficult to exercise.	1	1	1	1	0	0	0

Table 4: Responses resulting from back translation of the original student response, excluding the duplicates.

Response to Anemia Item	lang
Alyuvarlar küçük olduğu için yeterli o2yi taşıyamaz hücreler yeterli besin ve o2 alamaz (original)	Tr
Alyuvarları küçük olduğu için yeterince oksijen taşıyamazlar. Hücreler yeterli besin ve O2 alamazlar	Fa
Alyuvarlar küçük oldukları için yeterli O2 taşıyamazlar.Hücreler yeterli besin ve O2 alamazlar.	Uz
Alyuvarları küçük olduğu için yeterince O2 taşıyamazlar. Hücreler yeterli besin ve O2 alamazlar	Iw, En
Alyuvarları küçük olduğu için yeterince O2 taşıyamazlar. Hücreler yeterli besin ve O2 alamıyor	Gr, Tt
Kırmızı kan hücreleri küçük oldukları için yeterli O2 taşıyamazlar ve hücreler yeterli besin ve O2 alamazlar.	De, Fi
Kırmızı kan hücreleri küçük oldukları için yeterli O2 taşıyamazlar.Hücreler yeterli besin ve O2 alamazlar.	It, Uk
Kırmızı kan hücreleri küçük oldukları için yeterli oksijeni taşıyamazlar.Hücreler yeterli besin ve oksijeni alamazlar	Ja

Table 5: The results. Category a is not relevant for Anemia Item, so it was evaluated based on Smoking Item only. Kappa correlation values were interpreted using [13]: poor (< 0.00), slight ($0.00 - 0.20$), fair ($0.21 - 0.40$), moderate ($0.41 - 0.60$), good ($0.61 - 0.80$), and very good ($0.81 - 1$).

	Exp. 1			Exp. 2			Exp. 3			Exp. 4			Exp. 5		
Cat.	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
a	0.76	0.76	0.48	0.87	0.87	0.71	0.89	0.89	0.76	0.86	0.86	0.71	0.89	0.89	0.78
b	0.71	0.74	0.26	0.76	0.79	0.40	0.74	0.78	0.33	0.76	0.80	0.38	0.78	0.80	0.48
c	0.75	0.75	0.38	0.81	0.82	0.45	0.79	0.79	0.43	0.78	0.78	0.43	0.77	0.77	0.41
d	0.91	0.91	0.45	0.98	0.98	0.86	0.98	0.98	0.86	0.98	0.98	0.85	0.97	0.97	0.83
e	0.93	0.93	0.50	0.99	0.99	0.89	0.99	0.99	0.89	0.98	0.98	0.79	0.98	0.98	0.84
f	0.82	0.79	0.35	0.95	0.95	0.78	0.95	0.95	0.74	0.95	0.95	0.71	0.96	0.96	0.74
g	0.85	0.87	0.30	0.98	0.98	0.92	0.97	0.97	0.87	0.96	0.96	0.86	0.96	0.96	0.84
mean	0.82	0.82	0.39	0.90	0.91	0.72	0.90	0.91	0.70	0.90	0.90	0.68	0.90	0.90	0.70

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

[1] M. Ariely, T. Nazaretsky, and G. 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)*, pages 565–568, 2020.

[2] M. Ariely, T. Nazaretsky, and G. Alexandron. Personalized Automated Formative Feedback Can Support Students in Generating Causal Explanations in Biology. *The Proceeding of the 16th International Conference of the Learning Sciences (ICLS 2022)*, pages 953–956, 2022.

[3] M. Ariely, T. Nazaretsky, and G. Alexandron. Machine Learning and Hebrew NLP for Automated Assessment of Open-Ended Questions in Biology. *International Journal of Artificial Intelligence in Education*, 33(1):1–34, Mar 2023.

[4] A. M. Azmi, M. F. Al-Jouie, and M. Hussain. AAEE–Automated evaluation of students' essays in Arabic language. *Information Processing & Management*, 56(5):1736–1752, 2019.

[5] B. Beigman Klebanov and N. Madnani. Automated evaluation of writing–50 years and counting. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7796–7810, 2020.

[6] A. Çınar, E. Ince, M. Gezer, and Ö. Yılmaz. Machine learning algorithm for grading open-ended physics questions in turkish. *Education and information technologies*, 25(5):3821–3844, 2020.

[7] C. Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*, 2018.

[8] W. H. Gomaa and A. A. Fahmy. Automatic scoring for answers to Arabic test questions. *Computer Speech & Language*, 28(4):833–857, 2014.

[9] T. A. Grotzer and B. B. Basca. How does grasping the underlying causal structures of ecosystems impact students' understanding? *Journal of Biological Education*, 38(1):16–29, 2003.

[10] M. Ha, R. H. Nehm, M. Urban-Lurain, and J. E. Merrill. Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE—Life Sciences Education*, 10(4):379–393, 2011.

[11] M. Huggins, S. Alghowinem, S. Jeong, P. Colon-Hernandez, C. Breazeal, and H. W. Park. Practical guidelines for intent recognition: Bert with minimal training data evaluated in real-world hri application. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 341–350, 2021.

[12] C. Krist, C. V. Schwarz, and B. J. Reiser. Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. *Journal of the Learning Sciences*, 28(2):160–205, 2019.

[13] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[14] B. Li, Y. Hou, and W. Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 2022.

[15] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn. Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2):19–28, 2014.

[16] K. Moharreri, M. Ha, and R. H. Nehm. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(1):1–14, 2014.

[17] 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.

[18] R. H. Nehm and H. Haertig. Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, 21(1):56–73, 2012.

[19] J. F. Osborne and A. Patterson. Scientific argument and explanation: A necessary distinction? *Science Education*, 95(4):627–638, 2011.

[20] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*, 2019.

[21] K. Ryoo and M. C. Linn. Designing guidance for interpreting dynamic visualizations: Generating versus reading explanations. *Journal of Research in Science Teaching*, 51(2):147–174, 2014.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[23] A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R. S. Greenfeld, and R. Tsarfaty. AlephBERT: a Pre-trained Language Model to Start Off your Hebrew NLP Application, 2021.

[24] 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, Dec 2017.

[25] R. Tsarfaty, D. Seddah, S. Kübler, and J. Nivre. Parsing morphologically rich languages: Introduction to the special issue. *Computational linguistics*, 39(1):15–22, 2013.

[26] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (csur)*, 53(3):1–34, 2020.

[27] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

[28] X. Zhai, Y. Yin, J. W. Pellegrino, K. C. Haudek, and L. Shi. Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, 56(1):111–151, 2020.

[29] Y. Zhang, T. Ge, and X. Sun. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*, 2020.