Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5356–5365
Marseille, 20-25 June 2022

Predicting the Proficiency Level of Nonnative Hebrew Authors

Isabelle Nguyen, Shuly Wintner
Humboldt University, Berlin, Germany; University of Haifa, Israel
isabelle.nguyen@pm.me, shuly@cs.haifa.ac.il

Abstract

We present classifiers that can accurately predict the proficiency level of nonnative Hebrew learners. This is important for practical (mainly educational) applications, but the endeavor also sheds light on the features that support the classification, thereby improving our understanding of learner language in general, and transfer effects from Arabic, French, and Russian on nonnative Hebrew in particular.

Keywords: Learner corpora, Hebrew, Proficiency

1. Introduction

The language of nonnative speakers (*L2*) differs from native language (*L1*), and learners take time to reach a level of proficiency that is comparable with native language. Assessing the proficiency level of nonnative speakers is essential for tasks such as determining university admittance, adaptation of language learning tools, tailoring foreign language training for students, etc.

We show that the proficiency level of nonnative Hebrew speakers (defined here technically, as scores in language aptitude tests) can be accurately predicted from the essays they author. On a recently-released corpus of Hebrew essays authored by prospective students with three different mother tongues (Arabic, Russian, and French), we present simple classifiers that can distinguish between native and nonnative speakers, identify the native language of the author, and predict the essay's proficiency score with accuracy that does not differ much from that of human graders. We then provide a detailed analysis of the features that enable the classification, thereby deriving insight into the characteristics of second language, transfer effects from the native language, and in general, a better understanding of bilingualism.

After discussing related work, we describe the experimental setup for this work in Section 3. The experiments' results are presented in Section 4, followed by an analysis in Section 5. We conclude with suggestions for future work.

2. Related work

Assessing the proficiency level of nonnative speakers is an important task that has great benefits for educational applications, and automating it is therefore an ongoing endeavor, typically aided by *learner corpora* (Lüdeling et al., 2008; Blanchard et al., 2013; Wisniewski et al., 2013).

Various measures have been proposed for assessing foreign language proficiency (Scarborough, 1990; Ortega, 2003; Attali and Burstein, 2005; Bulté and Housen, 2012; Zesch et al., 2015; Weiss, 2017, and many more). However, calculating these measures is complex, expensive and most importantly, does not always achieve good results.

Consequently, methods based on text classification have been widely adopted in recent years. Crossley et al. (2012) analyzed 100 English essays authored by speakers of 19 different L1s. They found that the strongest predictors of proficiency level were word imagability, word frequency, lexical diversity, and word familiarity. On this small dataset, their indices identified the proficiency level of authors (on a 2-class scale) with 70% accuracy.

Pilán et al. (2016) used L2 Swedish essays written by learners, but augmented them by texts written by experts, primarily intended as reading material for learners. Aiming to predict the 6-level categorization of the Common European Framework of Reference for Languages (CEFR, Hawkins and Filipović (2012)), they obtained an F1 score of 0.721, and an analysis revealed that the best-performing features are lexical. Amorim and Veloso (2017) worked on a dataset of 1840 essays in Brazilian Portuguese, classified to 5 levels. Lexical richness was the most important determinant of grade.

Vajjala and Rama (2018) focused on three different L2s (German, Italian, and Czech) for which CEFR-graded data are available through the Merlin Corpus (Wisniewski et al., 2013). With 2286 essays and six proficiency levels, they reached F1 scores of 0.686 to 0.837. The best-performing features were part-of-speech (POS) n-grams, but linguistic features specific to the task, including essay length, lexical richness, and spelling errors, boosted the accuracy somewhat.

More recent works use neural models rather than statistic classification. Hirao et al. (2020) worked on a dataset of essays written by Japanese learners, whose proficiency is scored between 1 and 6. They compared the performance of feature-based classification with neural models, and concluded that BERT (Devlin et al., 2018) yielded the best results. Standard neural models, however, use only the texts of the essays and cannot leverage linguistically motivated features. Dasgupta et al. (2018), working on the Automated Student Assessment Prize dataset, consisting of thousands

of English essays (Mathias and Bhattacharyya, 2018), combined a deep neural network with features such as lexical diversity, cohesion, well-formedness etc., and showed improved results. A similar approach was applied by Uto et al. (2020) to the same dataset; they showed that the best performing model is BERT, augmented (via simple concatenation) by essay-level features. These include length-based features, POS unigrams, number of spelling errors, and readability features.

We address two more tasks in this work. The binary classification task, namely distinguishing between native and nonnative authors, is known to be easy (Massung and Zhai, 2016; Rabinovich et al., 2016; Nisioi et al., 2016). The native language identification task (Koppel et al., 2005) gained much attention with the Native Language Identification Shared Tasks (Tetreault et al., 2013; Malmasi et al., 2017). The state-of-the-art in identifying the L1 of highly-fluent English nonnative authors, with over 30 different L1s, is currently achieved with statistical classification (Goldin et al., 2018).

3. Experimental setup

3.1. Data

Our data (Gafni et al., 2022) consists of 4000 short essays written in Hebrew. They are equally divided between the following four groups: Native speakers of Hebrew; and learners of Hebrew with L1s Arabic, French, and Russian. The natives' essays were produced as part of the Psychometric Entrance Test, a centralized exam for access to higher-education institutions in Israel. The learners' essays were authored as part of a specialized test for assessing Hebrew proficiency level, again in preparation for undergraduate academic studies. All essays were made available to us in sentence-scrambled format.[1]

The nonnative essays include anonymized meta-information about the author's L1 and a proficiency score. That score is the result of an averaged aggregate of four distinct factors: content, organization, lexical richness, and lexical precision (explained in Section 5). Each essay was graded by two human raters, who assign a score between 1 (very low) and 7 (very high) to each of the four factors. The sum of the four scores is the total score, which is then averaged over the two raters.[2] The scores thus range between 4 and 28; the essays in our dataset have scores between 17 and 28, distributed as equally as possible across the three L1s (see Figure 1; the Arabic and the Russian lines perfectly overlap).

The specific task definition and the choice of prompts are known to strongly affect facets of the written essays

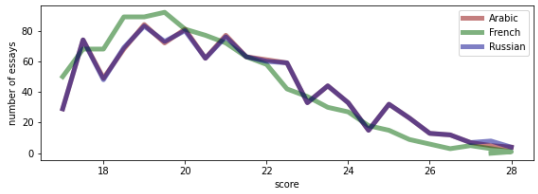

Figure 1: Scores are relatively uniformly distributed across L1s and follow the shape of a normal distribution (where the lower end was truncated by design when essays were selected).

(Alexopoulou et al., 2017). In our case, the allotted time for essay writing was 15 minutes in the nonnative test, but 30 minutes in the native test. In addition, there was a specific length requirement for each test: 10-15 lines in the former, and 25 lines in the latter. Expectedly, native speakers produced longer essays than their nonnative counterparts. At a mean length of 294 words, the texts by L1-speakers are more than twice as long as those written by the learners. Figure 2 shows the average essay length per language in words. Likewise, a higher score of a nonnative essay correlates with a longer essay, as can be seen in Figure 3 which contrasts the score with the number of words for each essay. It is also important to note that the prompts for the native and nonnative essays differ.

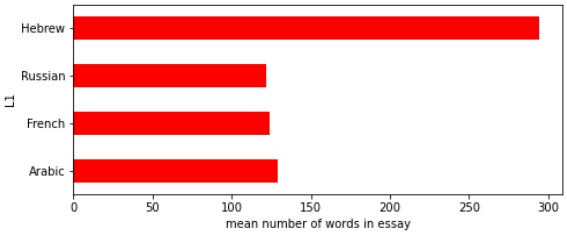

Figure 2: Essays by native Hebrew speakers are on average more than twice as long as those by learners.

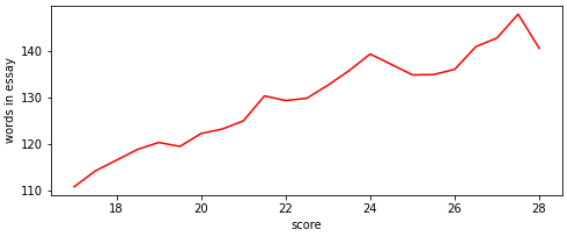

Figure 3: The higher the score, the longer the essay.

3.2. Methodology

We are motivated both by the practical goal of constructing a reliable predictor of Hebrew nonnative pro-

[1] The data, code, and parameters used in this work will all be made publicly-available for research purposes.

[2] Very few of the essays had a third human score, which we ignored.

ficiency level and by the research goal of better understanding the characteristics of nonnative Hebrew, in particular as a function of the writer's L1. Therefore, and also owing to the very small size of our dataset, we refrain from using contemporary deep neural models, which are typically opaque and do not lend themselves to easy interpretation, and resort instead to statistical text classification.

We define feature-based classifiers that predict the L1 and the proficiency level of authors, and analyze the features that perform best on these tasks. We hypothesize that such features can shed light on the characteristics of learner language in general, and transfer effects from L1 in particular. This basic methodology has been used extensively in computational linguistics to analyze the features that are indicative of translationese (Volansky et al., 2015; Rabinovich and Wintner, 2015), nonnative speakers (Rabinovich et al., 2017), speakers of various L1 backgrounds (Goldin et al., 2018), etc.

The native and the learners' essays were written in response to different sets of prompts. The use of content words as features may thus yield excellent results when distinguishing between native and nonnative authors, based on prompt-specific vocabulary rather than the native and nonnative linguistic traits that we seek to investigate. We therefore try, when we design features, to exclude any indication of content, and focus on the *style* of the authors.

3.3. Tasks

We define three classification tasks. **Identifying native Hebrew speakers:** we trained a binary classifier to distinguish between native and nonnative authors. **Native language identification:** using only the nonnative texts, we trained a classifier to distinguish between the three L1s Arabic, French, and Russian. **Assessing proficiency:** using the nonnative texts, we trained a regression model to predict the values that make up the Hebrew proficiency score. All the tasks are implemented using the XGBoost library for Python, with default parameters. In all tasks, the instance for classification is a single, complete essay.

3.4. Classification and evaluation

For each of the three experiments, we set aside 20% of the data as our evaluation set. The remaining 80% were used for training a model, using Gaussian processes to find the optimal hyperparameter settings. We ran the optimization function 40 times, after which we chose the hyperparameters that returned the best results to construct a final, deployable model.

We used the entire feature matrix, unless stated otherwise. The final models were evaluated with the help of the evaluation data set. As evaluation metrics, we used *accuracy* for the classification models and *absolute mean error* for the regression model. In addition, we made use of the fact that the score was annotated twice, by two different human raters, and report inter-annotator agreement between the annotators and between the human-assigned score and our model's predictions.

We work with the SHAP library to model feature importance. SHAP's TreeExplainer approximates Shapley values for every feature to explain that feature's impact on the model's predictions per datapoint. To showcase the overall effect of a feature, the library provides functions that aggregate and plot the SHAP values over the entire dataset.

3.5. Feature engineering

From the raw texts, we produced approximately 2300 features, ranging from trivial measurements (like text and average sentence length) to language-specific aspects such as lexical discourse markers. We now detail these features; understanding some of them requires familiarity with Hebrew morphology and orthography (Yona and Wintner, 2008; Fabri et al., 2014).

Length We computed the length of an article, the number of tokens and sentences, and mean word and sentence length.

Functional Morphemes As mentioned above, we wanted to exclude content words from the essays so as to alleviate prompt-dependent biases. Separating content from function words is non-trivial in our use case, since the data are not tagged for parts of speech.[3] We extracted function words through certain heuristics. Hebrew as a morphologically complex language uses various affixes and clitics to perform a host of morphosyntactic functions. For instance, objects, possession and prepositions may all be marked by attaching the appropriate morphemes to verbs and nouns.

We used a list of prefixes and suffixes that we stripped from a word whenever they occurred at the start or end of a word. We then kept the affix but not the rest of the word.[4] In addition, we used a list of 46724 stopwords to find all forms of standalone function words.[5] Finally, we converted the remaining feature matrix of function words and affixes to a Tf-Idf representation.

Lexical richness We used different type-token ratio metrics designed to measure lexical richness, which is known to be associated with fluency and proficiency. Specifically, we employed root type-token ratio (RTTR) (Guiraud, 1954; Guiraud, 1960), corrected type-token ratio (CTTR) (Carroll, 1964), and the measure of textual lexical diversity (MTLD) (McCarthy, 2005; McCarthy and Jarvis, 2010).

[3] Existing POS-taggers are trained on a register that is vastly different from our data. We therefore expect POS-tagging with currently available tools to yield poor accuracy.

[4] Some prefixes and suffixes used as features include:

ה־, ו־, וֹת־, י־, ים־, ית־, ך־, כם־, כן־, נו־, ת־, תי־, תם־,
תן־, א־, ב־, ה־, ו־, י־, כ־, ל־, מ־, נ־, ש־, ת־,

[5] The high number of stopwords again stems from the fact that Hebrew is morphologically complex and function words can inflect. All of these forms are included in the list.

Phonetics and orthography Hebrew has a number of letter pairs that are homophonic, i.e., may be used to represent the same sound. We hypothesized that learners might find the distinction difficult. We therefore counted occurrences of those letters (excluding those cases where they served as affixes), normalized by text length. Additionally, there are five letters whose form changes in word-final position. Since this is a concept foreign to writers of French and Russian (but not of Arabic), we included normalized counts of those letters as well.[6]

Discourse markers Various function words serve to highlight the logical structure of a text. Commanding a large repertoire of these connectors is likely an indicator of fluency. We created a list of 41 discourse markers and included their weighted counts in the feature set.[7]

Frequency Finally, we used several word-rank based metrics to account for lexical richness. Rank is computed on large corpora to account for the frequencies of words. A low rank corresponds to a high frequency of a word in a corpus. We used a rank list computed on the Hebrew Wikipedia (Itai and Wintner, 2008). For every essay, we computed its mean and median word ranks. We then divided the rank list into five quantiles of logarithmically increasing sizes and computed the normalized word count per quantile.

Note that our rank-based features take into account all the words in a text, including content words. If a word does not occur in a corpus, it might be because it is very rare, or because it has been misspelled. We decided to view any words that do not occur in the list of Wikipedia words as misspellings and ignore them. However, it should be noted that the abjadic nature of the Hebrew writing system, where vowels are often not written, contributes to a phenomenon whereby many misspellings result in actual, often rare, words, which will thus be counted in our framework.

4. Results

4.1. Native vs. nonnative authors

The nativeness classifier makes near-perfect predictions with an accuracy of 99%. Out of the 800 data points in our test set, the classifier predicts the wrong class for six essays, or 0.75%.

The most predictive feature for this model is the length of the essay. This is unsurprising, as essays written by native speakers of Hebrew are more than twice as long as those by nonnative speakers. Native essays wrongly classified as nonnative all have a lower-than-average essay length.

Interestingly, when we excluded essay length from the feature matrix, the classifier's accuracy only sank by one percent point and other features (RTTR, mean sentence length and the use of the homophonic letter ח and the discourse marker לסיכום *in sum*) come out more strongly as predictors. This points to the fact that the overall signal in the dataset is strong and does not rely solely on the length of an essay. This is reassuring since essay length is strongly dependent on the task and may be very different for another set of data points. In terms of homophonic letters, our model correlated lower use of the letter ח (/χ/) with nativeness. Nonnative speakers are more likely to overuse this letter in contexts where a same-sounding כ is called for. Other indicators of nativeness are a higher use of the suffix ־ית, used to form feminine nouns and adjectives; and a higher count of words that occur in the last logarithmic quantile, i.e., the rarest words.

4.2. Native language identification

In the second task, we sought to identify the L1 of an essay's author, using only the subset of the data produced by nonnative speakers. Our three-way classifier was able to predict an author's native language with an accuracy of 77%. The confusion matrix is depicted in Figure 4, where the true label is on the y-axis and the predicted label is on the x-axis. The plot shows the model mainly confused between authors whose L1 is French or Russian (approximately 16% errors). This may be explained by the typological closeness of these languages, in contrast to Arabic.

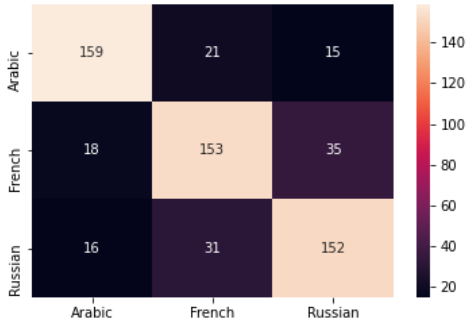

Figure 4: L1 identification confusion matrix.

Figure 5 graphically depicts some of the most indicative features for this task. The plot shows the most important features, ranked by overall relevance for

[6]The following Hebrew letters are used as features to capture phonological and orthographic effects:

י, ח, כ, ט, ת, א, ע, ס, ש, ק, פ, ב, ד, ף, ץ, ן, ם

[7]The following discourse markers are used as features; we used the number of markers per sentence and the number of unique discourse markers in an essay.

אכן, אמנם, בנוסף, בסיכומו של דבר, בשורה התחתונה, יחד עם זאת, יתרה מזאת, כאמור, כמו כן, לאור העובדה, לדוגמא, לדוגמה, לדעתי, להפך, לסיום, לסיכום, לעומת זאת, לעניות דעתי, לפי דעתי, מצד אחד, מצד שני, על כן, עם זאת, עקב כך, ראשית, שנית

the model, and coloured by the L1. Thus, the first row indicates that the model's top predictive feature—particularly for the classification of Arabic—is mean sentence length. However, the plot does not reveal whether this is due to overuse, underuse, or a different pattern entirely; we went back to the raw data and investigated a given feature's behavior by L1.

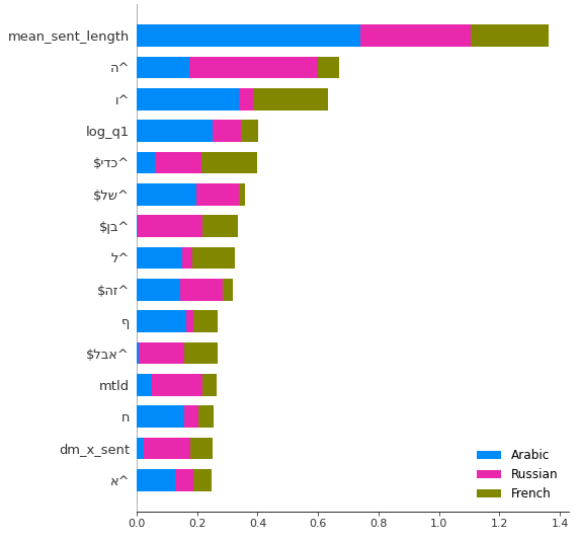

Figure 5: Top 15 predictive features by L1.

Unsurprisingly, we identified high mean sentence length as the most significant feature to predict L1 Arabic, in addition to overuse of the conjunctive proclitic ו־ *and*. This is likely because native speakers of Arabic tend to use punctuation sparingly, and instead chain sentences together with conjunctions. For Russian, on the other hand, underuse of the determiner ה־ *the* is the most predictive feature, which can be explained by the fact that Russian does not have definitive articles.

4.3. Predicting proficiency

For our final and primary task, we trained a regression model for each of the four score components outlined above, as well as a separate model for the overall score. We then evaluated the models on each of five targets separately. This problem turned out to be easier than we expected: the variance among scores is apparently not great, and since the lower portion of the score distribution is not included in our dataset, a baseline model that always predicts a conservative value for each component is going to be close to the truth much of the time. In Table 1 we report the mean absolute error (in terms of score points) of a baseline that always predicts the median of the values in the training set.

After hyperparameter tuning, our model for predicting the overall score yielded a mean absolute error of 1.67 points on the evaluation set (Table 1), reflecting a reduction of 12% of the error wrt. the baseline. The discrepancy between the two human annotators, for comparison, is 2.52 points on average. Table 1 shows the mean absolute error of the models on each component individually, as well as on the total score; in parentheses we report the error rate reduction with respect to the baseline, in percents. The table also depicts ablation results obtained by using subsets of the feature groups (Section 3.5) for prediction. Each feature set individually is close to the baseline, but their combination fares much better.

The models for the individual score components all selected very similar subsets of features as the highest-ranking for their predictions. This explains why the model for predicting score performs almost identically to the aggregated score, gleaned from the sum of the sub-components (1.67 and 1.69, respectively). We briefly overview these features here, and will refer to them in more detail in the discussion of the component models (Section 5).

To identify the most important predictors, we once again plot feature importance using the SHAP library. Figure 6 showcases the results for the general score regression model. The model's features are ranked by their impact, with the most important feature on top. For every feature (a row in the plot), the plot contrasts a data point's value—represented by a shade on the spectrum between red (high) and blue (low)—with that value's impact on the model (the SHAP value) which is high on the left and on the right, but low towards the middle line. For example, the first feature in Figure 6 shows that a high value for the RTTR feature is correlated directly with a high positive impact on the score prediction. Conversely, high usage of the letter ם correlates with a high negative impact on the score, as demonstrated by the penultimate feature.

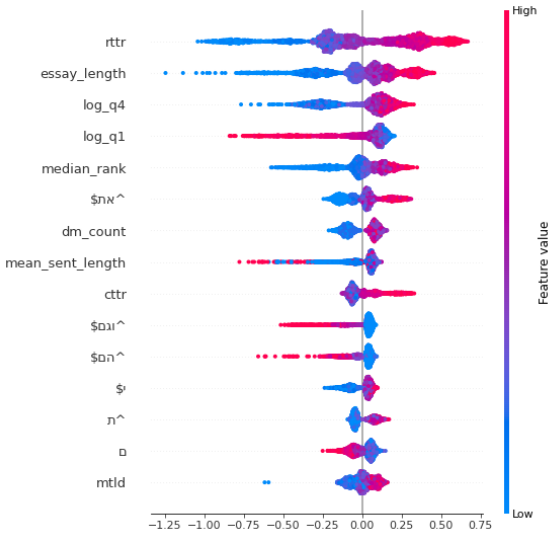

Figure 6: The score predictor's 15 top features.

To summarize our results, the most important features include at least one measurement of lexical richness (usually RTTR), text length, median word rank, the use of the accusative marker את, the use of discourse mark-

Features	Total	Content	Organization	Lex. richness	Lex. precision
Baseline	1.90	.51	.57	.52	.59
All	**1.67** (12%)	**.49** (4%)	**.53** (8%)	**.48** (8%)	**.56** (5%)
Length	1.89	.52	.59	.52	.57
Functional morphemes	1.92	.53	.57	.50	.59
Lexical richness	1.80	.50	.59	.51	.58
Frequency	1.87	.53	.59	.51	.57
Phonetics/Orthography	1.90	.53	.58	.51	.59
Discourse markers	1.81	.52	.55	.51	.58

Table 1: Results: Mean absolute error (the lower the better) and error rate reduction.

ers, and word-final use of י, ת, ה and ן (all of which can be morphological suffixes encoding functions such as gender and number). For all these features, a higher feature value correlates with a higher score. On the other hand, higher values for the following features correspond to lower scores: Words in the first quantile (i.e., the most common words), use of the pronouns אני *I* and הם *they*, the discourse markers וגם *and also* and אבל *but*, and the end-letter ם. Average sentence length is a mixed feature, in that longer sentences can correlate both with high and with low scores.

To further validate the results, we computed Krippendorff's Alpha (Krippendorff, 2011), a well-known inter-annotator agreement metric to be used with ordinal ratings (Artstein and Poesio, 2008), for each of the models. Agreement was computed both between the two raters, as well as between the raters' mean and the model prediction. We then compared the results to check whether agreement among the raters was higher or lower than agreement between the model and the human raters. In light of the above discussion, it is not surprising that the human raters receive a lower alpha score, 0.363, on the overall score task, than the model–raters alpha of 0.411 (Table 2).

5. Analysis

We now discuss the results, focusing on the features that best predict each of the proficiency score components. For each component, we first describe the rating guidelines and then present our model's results, followed by a feature analysis.

5.1. Content

To evaluate the *content* component, raters were instructed to check whether the text followed an overall idea and presented an adequate response to the prompt. Since no information about prompts is encoded in the features, we expected the model to fare rather poorly in comparison to human raters.

However, the model overtook the annotators. It produced a mean absolute error of 0.53 points for the prediction of the content score which, as a reminder, can lie anywhere between one and seven points. Krippendorf's alpha between the model's predictions and the actual values is 0.3, much higher than the inter-annotator agreement between the two annotators which is 0.185. While this is a relative measure, it shows that, in a situation where two humans find it hard to agree on a common assessment, the model is able to pick up on features that let it approximate the target value with higher accuracy. The most telling features in this model are essay length, RTTR, median word rank and the number of discourse markers used. Word-final ם and the use of יש *there is* and אתה $you_{sg.m}$ are markers for a lower score in this category. The use of the latter can simplify talking about a topic in an impersonal way. The importance of lexical features, coupled with the comparatively good performance of the model, raises the question whether the annotators actually followed the guidelines for this component.

5.2. Organization

This component evaluates the *structure* of the text. Raters were asked to assess whether the essay's sentences followed each other in a logical manner and built up a coherent argument. Here, we expected our model to pick up on some lexical cues, most notably the use of discourse markers, whose whole purpose is the logical structuring of a text.

Our model's performance on this component is again slightly above the raters'. The alpha coefficient for the model's agreement with the raters is 0.22, whereas that between raters it is 0.17. Again, RTTR, essay length and discourse markers are highly predictive features. According to the model, a higher count of discourse markers correlates with better scores (as we would expect). In addition, overuse of the conjunction וגם *and also*, the quantifier הרבה *many/much* and the existential יש *there is* indicates lower results in this component. Presumably, these are less-refined function words that help organize a text. For instance, a high use of וגם *and also* indicates a certain monotony in the writer's sentence-coordination arsenal. Finally, in our model for this component, greater mean sentence length correlates more clearly with lower scores than in the other components; this may indicate that longer sentences contribute to a less rigorous inner-textual organization.

Features	Total	Content	Organization	Lexical richness	Lexical precision
All	**.411**	**.297**	.217	.280	.224
Length	.222	.223	.157	.132	.073
Functional morphemes	.245	.233	**.225**	.269	.223
Lexical richness	.299	.239	.163	.177	.139
Frequency	.248	.164	.135	.108	.142
Phonetics/Orthography	.187	.113	.096	.147	.091
Discourse markers	.253	.228	.189	.066	.029
IAA human raters	.363	.185	.169	**.353**	**.355**

Table 2: Results: agreement (in terms of Krippendorff's alpha) between model predictions and human annotators.

5.3. Lexical richness

On this component, raters were asked to check whether the authors used a variety of words that were fitting to the topic and the style of the text. An overuse of repetition and the imprecise use of words was to be penalized.

Compared to the previous two components, our model this time displayed lower performance. While the raters agree at an alpha value of 0.35, the average agreement between our model and the raters is 0.28. For this component, our logarithmic rank values made for good prediction features. A high count of words in the first quantile—which represents the most frequent words in our rank list—was indicative of low lexical richness. Conversely, a high count of words in the fourth quantile—that is, rarer words—correlated with higher values for this score component. In addition, the model again identified a high RTTR and use of the accusative marker את as predictors for a higher result. Interestingly, the model picked up on *mean token length*, selecting it as the seventh-most important feature. According to the model, longer words correlate with a higher lexical richness score. At the same time, an overuse of the common conjunctions וגם *and also* and אבל *but* again point to a poorer vocabulary on the part of the author.

5.4. Lexical precision

This component evaluates the correct use of content words and function words in context. For example, raters were asked to check that authors make the right choice in terms of prepositions and look out for what is known as "false friends". Since we have attempted to exclude content words as much as possible, our model is handicapped by only being able to use function words for its predictions.

As expected, the agreement value between the model and the human raters is lower (0.22) than the agreement between the raters (0.36). The top features for predicting this category look very similar to the ones listed above for lexical richness. In terms of functional morphemes, the model actually ranked an overuse of the plural affix ים relatively high as an indicator for a lower score. On the other hand, higher use of the accusative marker את again correlates with higher scores. Interestingly, a higher count of words in the lowest rank quantile—i.e., very rare words—accounted for a lower result on this component. This is in line with the observation made above, that misspellings can easily give rise to actual words in a mostly vowelless writing system like Hebrew. This model does not like pronouns: higher counts of אני *I*, הם *they* and הוא *he* all correlate with lower scores.

6. Conclusions

We have demonstrated that simple, feature-based classifiers can accurately identify native Hebrew authors; predict the native language of nonnative authors; and predict their Hebrew proficiency scores. The models' predictions are often indistinguishable from those of human raters.

Evidently, our classifiers predict the total scores (as well as each component) with very high accuracy; in fact, using a small subset of the features suffices for highly accurate predictions. More surprising is the fact that extremely shallow features (e.g., length) suffice for accurate prediction of components such as *content*. We propose several explanations for these results.

First, recall that our dataset is biased: whereas each score component can be scored between 1 and 7, the actual scores in our corpus are higher, and the total scores, which can vary between 7 and 28, in fact begin at 17 (see the score distribution in Figure 1). Consequently, the actual score of each component are "squashed" between 3, sometimes even 4, and 7. Indeed, our model predicts values in a narrower distribution than the actual dataset. Figure 7 depicts the score distribution in the dataset vs. our model predictions. Evidently, our model tends to predict scores that are closer to the mean.

This is exacerbated by the fact that our test set is small, and its score distribution always has a flatter bell curve than the entire dataset, often with tails (especially the lower tail) cut off. The success of our models can be partly attributed to the fact that they tend to predict values that are closer to the mean. This is helpful since our test set has a flatter score distribution than the entire dataset, as can be seen in Figure 8.

Second, it is clear that the different feature groups are highly correlated (as all of them are correlated with

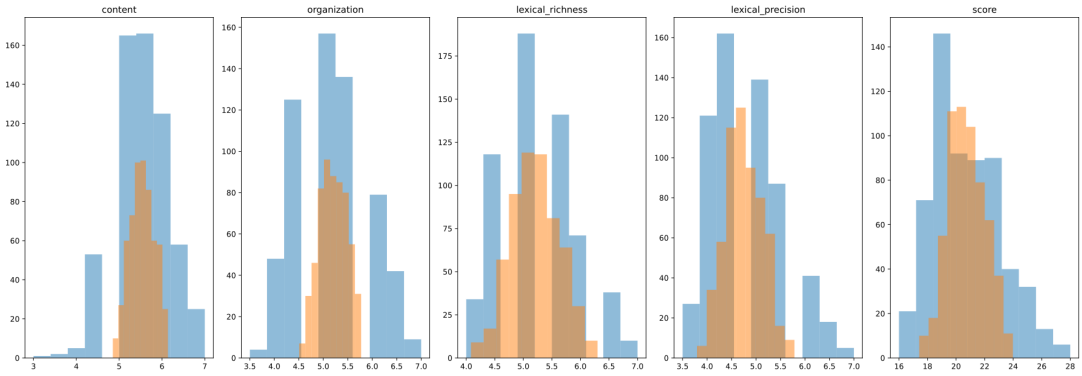

Figure 7: The distribution of scores in the dataset (blue) vs. our models' predictions (orange).

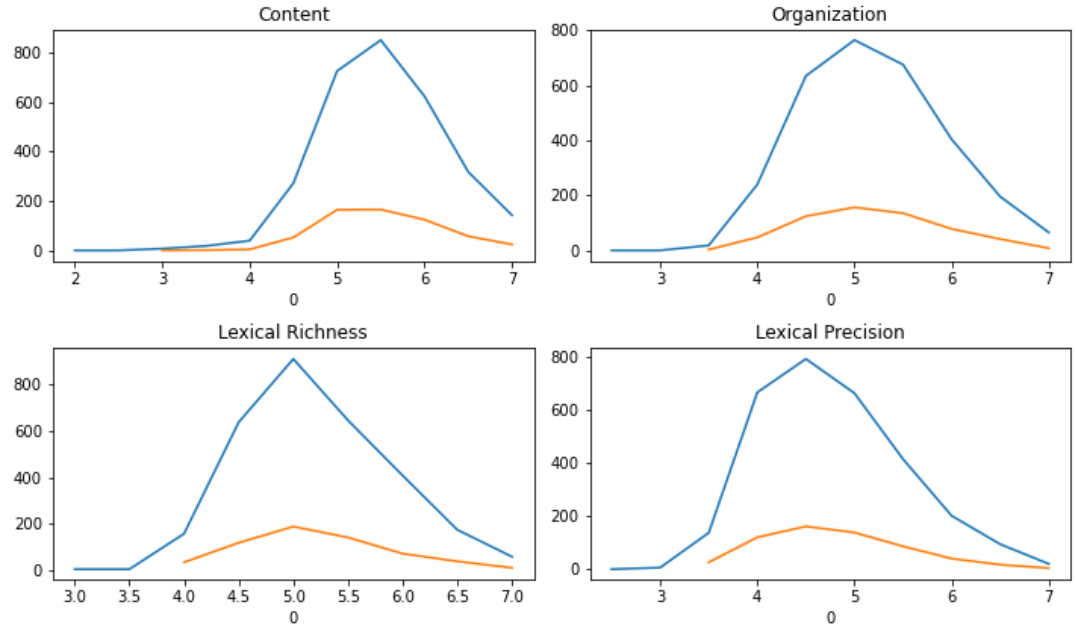

Figure 8: The distribution of scores in the full dataset (blue) vs. the test set (orange).

proficiency level). This explains the observed fact the each feature group individually is a good predictor of all score components.

Finally, observe that our predictions are in general more accurate than the scores of the human raters (in terms of agreement). We believe that human raters tend to form a holistic assessment of the essay they score, and may find it difficult to score each component individually. This is related to a more general question, namely the correspondence between standard aptitude tests and "real" language proficiency, which has often been disputed (Wisniewski, 2017). We leave this conjecture for more dedicated future investigation.

In the future, we hope to gain access to a larger dataset of essays, including ones on the lower end of the distribution, in order to fully validate the usefulness of our classifiers in mimicking human raters. With a more diverse dataset, we would like to investigate whether the same features enable predicting proficiency for the different L1s, or whether some are more L1-specific.

7. Acknowledgements

We are grateful to the Israeli National Institute for Testing and Evaluation for making the essays available. We are very grateful to Noam Ordan for generously sharing his ideas with us in the initial stages of the project and for his advice and support throughout. Many thanks to Chen Gafni for his help with the Hebrew corpus and for providing useful comments. Thanks are also due to Anke Lüdeling, Anat Prior, Sarah Schneider and Dominique Bobeck for advice and fruitful discussions. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 398186468 and by the Data Science Research Center at the University of Haifa.

8. Bibliographical References

Alexopoulou, T., Michel, M., Murakami, A., and Meurers, D. (2017). Task effects on linguistic complexity and accuracy: a large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1):167–209.

Amorim, E. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Valencia, Spain, April. Association for Computational Linguistics.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, December.

Attali, Y. and Burstein, J. (2005). Automated essay scoring with E-rater v.2.0. Technical report, Educational Testing Service (ETS), November.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. In Alex Housen, et al., editors, *Dimensions of L2 performance and proficiency: Complexity,accuracy and fluency in SLA*, pages 21–46. John Benjamins, Amsterdam.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, 2(1):80.

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2):243–263.

Dasgupta, T., Naskar, A., Dey, L., and Saha, R. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fabri, R., Gasser, M., Habash, N., Kiraz, G., and Wintner, S. (2014). Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In Imed Zitouni, editor, *Natural Language Processing of Semitic Languages*, Theory and Applications of Natural Language Processing, pages 3–41. Springer, Berlin Heidelberg.

Gafni, C., Prior, A., and Wintner, S. (2022). The Hebrew Essay Corpus. In *Proceedings of LREC-2022*, June.

Goldin, G., Rabinovich, E., and Wintner, S. (2018). Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601. Association for Computational Linguistics.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses Universitaires de France, Paris.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses Universitaires de France, Paris.

Hawkins, J. A. and Filipović, L. (2012). *Criterial Features in L2 English. Specifying the Reference Levels of the Common European Framework*. Cambridge University Press, Cambridge.

Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., and Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1250–1257, Marseille, France, May. European Language Resources Association.

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

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*, pages 41–76.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-reliability.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., and Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67–73.

Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., and Qian, Y. (2017). A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75. Association for Computational Linguistics.

Massung, S. and Zhai, C. (2016). Non-native text analysis: A survey. *Natural Language Engineering*, 22(2):163–186.

Mathias, S. and Bhattacharyya, P. (2018). ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Nisioi, S., Rabinovich, E., Dinu, L. P., and Wintner,

S. (2016). A corpus of native, non-native and translated texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), May.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4):492–518. Oxford University Press.

Pilán, I., Volodina, E., and Zesch, T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Rabinovich, E. and Wintner, S. (2015). Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Rabinovich, E., Nisioi, S., Ordan, N., and Wintner, S. (2016). On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1870–1881, August.

Rabinovich, E., Ordan, N., and Wintner, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics, July.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, 11:1–22.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, June.

Uto, M., Xie, Y., and Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification.

Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April.

Weiss, Z. (2017). Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects. Master's thesis, Eberhard Karls Universität Tübingen.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A., and Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT4LL 2013, 6th edition of the ICT for Language Learning Conference*, Florence, 10.

Wisniewski, K. (2017). The Empirical Validity of the Common European Framework of Reference Scales. An Exemplary Study for the Vocabulary and Fluency Scales in a Language Testing Context. *Applied Linguistics*, 39(6):933–959, 03.

Yona, S. and Wintner, S. (2008). A finite-state morphological grammar of Hebrew. *Natural Language Engineering*, 14(2):173–190, April.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado, June. Association for Computational Linguistics.