Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 67-74 recibido 31-03-2019 revisado 14-05-2019 aceptado 17-05-2019

Automatic proficiency classification in L2 Portuguese

Clasificación automática del nivel de proficiencia en Portugués Segunda Lengua

Iria del Río
University of Lisbon, Center of Linguistics-CLUL
igayo@letras.ulisboa.pt

Abstract: We present the first experiments on automatic proficiency classification for L2 Portuguese. For the experiments, we take advantage of a new version of the NLI-PT dataset, a compilation of L2 Portuguese texts written by learners. We use supervised learning and we approach the task as a classification problem, using the CEFR scale. Different linguistic features are tested, combined with different algorithms. With the best model, we get an accuracy of 72%, a result in line with previous experiments with other languages.
Keywords: Proficiency level, CEFR, L2 Portuguese, Supervised Learning

Resumen: Este trabajo presenta los primeros experimentos en clasificación automática del nivel de proficiencia en Portugués Segunda Lengua (L2). En los experimentos se usa una nueva versión del dataset NLI-PT, una compilación de textos escritos por estudiantes de Portugués L2. La tarea se aborda con aprendizaje supervisado, y se concibe como un problema de clasificación, usando la escala del MCER. Diferentes características lingüísticas son analizadas, así como diferentes algoritmos. Con el mejor modelo hemos obtenido una exactitud del 72%, un resultado en línea con previos experimentos realizados con otras lenguas.
Palabras clave: Nivel de proeficiencia, MCER, Portugués Segunda Lengua, Aprendizaje Supervisado

1 Introduction

This work has two main contributions. First, we present a larger and better version of the NLI-PT dataset[1], a compilation of L2 Portuguese texts with different types of linguistic annotations. Secondly, we describe the first experiments in automatic proficiency classification for L2 Portuguese, where we got similar results to previous works on the field.

The availability of data with linguistic annotations benefits different types of research, from theoretical analysis to statistical approaches like Machine Learning. Learner data is particularly difficult to gather, because of the specific context where this data is produced. For the English language there are big collections of learner data available, like the Cambridge Learner Corpus (16 millions of words) (Nicholls, 1999), but such type of collections are not common for other languages. The NLI-PT dataset aims to solve this gap for European Portuguese. We present a bigger and improved version, with more texts, better annotations and a different and more intuitive organization of the data.

As an example of the usefulness of the dataset, we present the first experiment for automatic proficiency classification of L2 Portuguese. Proficiency classification is a common task in second language learning. The development of the learner is usually defined in relation to a specific scale with different levels of linguistic complexity. One of the most common scales is the one described in the Common European Framework of Reference for Languages (CEFR) (Europe et al., 2009). The CEFR defines 3 broad divisions: A, basic user; B, independent user; C, proficient user, which are subdivided into 6 devel-

[1]http://www.clul.ulisboa.pt/en/resources-en/11-resources/894-nli-pt-a-portuguese-native-language-identification-dataset

ISSN 1135-5948. DOI 10.26342/2019-63-7

opment levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Each level is related to specific linguistic features and skills, establishing a progression from a very rudimentary language to a performance close to a native production. In this context, it is common that learners of a second language perform placement tests that define their proficiency level. The interest of an automatic system that can perform this task is, therefore, evident.

Automatic proficiency classification is commonly considered as a type of Automatic Essay Scoring (AES) task. AES systems are primarily developed for English (Burstein, 2003; Burstein and Chodorow, 2012; Yannakoudakis and Loo 2013), but in recent years systems for other languages have began to emerge (Vajjala and Loo, 2013). AES has been modeled in different ways, as a regres-sion (Yannakoudakis, Briscoe, and Medlock, 2011), ranking (Taghipour and Ng, 2016) or a classification problem (Pil´an, Vajjala, and Volodina, 2016). The features used are di-verse, from Bag-of-words (BOW) to more ab-stract representations that use higher levels of linguistic information (morphological, syn-tactic or even discursive). It is also very com-mon the use of descriptive metrics of the text related to word or sentence length, like av-erage syllable length, which have been con-nected to proficiency development in the area of Second Language Acquisition (SLA)(Lu, 2012). In general, AES is seen as a monolin-gual task, but recent works like (Vajjala and Rama, 2018) have explored multi and cross-lingual approaches.

Usually, the term AES is used as a general term for referring to different tasks: from proficiency classification of learner texts to readability assessment of teaching materials (Pilán, Vajjala, and Volodina, 2016). In fact, for the Portuguese language, and to the best of our knowledge, AES works have focused only on readability assessment (Branco et al., 2014) and (Curto, Mamede, and Baptista., 2015). We would like to differentiate the nature of these tasks, that is, readability assessment of input materials for the student, and proficiency classification of learners' texts because the linguistic parameters they involve are different. Readability assessment tasks usually focus on evaluating the linguistic complexity of potential input materials for the students. The goal is to automatically select materials that are appropriate for the learner. Therefore, systems working on readability assessment are trained with materials designed for learners and not written by them. Those materials use linguistic variants that are close to the target-native language, in some cases (for students with an elementary knowledge of the L2), they are simplified versions of this target-native language. However, these texts do not present the linguistic distinctive features that we can find in learner productions like, for example, orthographic or morphological errors or anomalous lexical or syntactic constructions influenced by the L1. These distinctive linguistic features constitute a challenge for automatically processing L2 texts since NLP tools are commonly built using native language models.

In our experiments, we apply the main three levels of the CEFR scale, A, B, and C, to automatically classify L2 Portuguese texts. Moreover, we try to answer the following question: What does it define the proficiency level of an L2 Portuguese text? To find an answer, we applied supervised machine learning techniques to build a classification model; we tested different algorithms and representations of the texts, from BOW to models that use descriptive features commonly used in the area and in SLA research (like average word length). With the best model, we got an accuracy of 72%, a result similar to those obtained in previous works. But, what is more relevant, we gained meaningful insights about the relation of certain linguistic features and automatic proficiency classification in L2 Portuguese.

The paper has the following structure: in section two we present related work; the NLI-PT dataset and the features used in the experiments are presented in section three; section four focuses on the description of our methodology and the results of the experiments; finally, in section five, we present our main conclusions and some future work.

2 Related Work

In this section, we present two types of previous work: SLA studies that have analyzed the relation between certain features and proficiency levels, and approaches that have used machine learning techniques to predict learner proficiency using the CEFR scale.

(Lu, 2012) analyses in detail the relationship between proficiency in L2 English and several lexical dimensions, concluding that the features linked to lexical variation (like Type-Token ratio) are the most correlated to the quality of an L2 essay. Several features identified as relevant in this work have been used by automatic approaches afterward. (Kyle and A. Crossley, 2014) explored lexical features too and showed that 47.5% of the variance in holistic scores of lexical proficiency in English as second language can be explained using a range of lexical sophistication indices. Other characteristics like syntactic complexity or error patterns have been studied too, mainly for English (Tono, 2000),(Lu, 2012), (Vyatkina, 2012), but also for other languages (Gyllstad et al., 2014).

Concerning automatic proficiency classification, (Yannakoudakis et al., 2018) is one of the most recent works for the English language. The authors used a subset of the Cambridge Learner Corpus with human proficiency annotations (levels A1 to C2), containing a total of 2,312 texts. They model the task as a ranking function and evaluate the quality of the predicted score by calculating Pearson's product-moment and Spearman's rank correlation coefficient against the scores assigned by a human expert. The features used include character sequences, POS, hybrid word, and POS sequences, phrase structure rules and error rates. The best model gets a Pearson r of 0.765 and a Spearman ρ of 0.773, with a κ of 0.738 (the standard error is 0.026) that indicates high agreement between the predicted CEFR scores and those assigned by humans.

In another recent study, (Vajjala and Rama, 2018) present the first multi and cross-lingual approach for proficiency classification. The authors use 2,286 manually graded texts (five levels, A1 to C1) from the MERLIN learner corpus (Boyd et al., 2014). It is an unbalanced dataset, with the following distribution of learner texts: German, 1,029 texts; Italian, 803 texts, and Czech, 434 texts. The authors compare different algorithms: logistic regression, random forests, multi-layer perceptron, and support vector machines for experiments with non-embedding features, and Neural Network models trained on task-specific embedding representations for other experiments. For non-embedding features, the best algorithm is Random Forests in most of the scenarios. They use a wide range of features: word and POS n-grams; task-specific word and character embeddings trained through a softmax layer; dependency n-grams (not used before); domain features mainly linked to lexical aspects (Lu, 2012); and error features. In their experiments, monolingual and multilingual models achieve similar performance, and cross-lingual classification yields lower, but comparable results to monolingual classification. For monolingual experiments, the best result (F1-score) is achieved with word n-grams plus domain features (German=0.686; Italian 0.837; Czech= 0.734). For multilingual tests, the best result is 0.726 with POS n-grams and information of the L1 as a feature. In cross-lingual experiments, they use German texts for training, and they get a 0.653 F1 score for Czech using dependency n-grams, and a 0.758 for Italian using POS n-grams. We can see that the features that allow for the best results vary in the experiments: words n-grams in monolingual; POS n-grams in multilingual; dependency n-grams for Czech and POS for Italian in cross-lingual experiments.

(Vajjala and Lõo, 2014) perform proficiency classification for Estonian. They use a corpus with 879 texts, belonging to four proficiency levels (A2 to C1) and also a balanced version of this dataset with 92 texts per category. They compare classification and regression models and use a rich feature set (78 features) that considers the morphological complexity of Estonian, as well as lexical richness features inspired by (Lu, 2012). Interestingly, POS models achieved poor performance and were not considered in the feature set. The best model is classification, with an accuracy of 79% in the whole dataset and 76.9% in the balanced one. For both datasets, the category with the poorest performance is B2. The authors perform a feature analysis and show that with the 27 best features they achieve a performance of 78.3%. The 10 best features in this group of 27 are lexical (Corrected Type Token Ration, Squared Verb Variation) and morphological (2nd person inflected verbs, distinct cases used in the document).

3 Dataset

3.1 Corpus

For our experiments, we use an updated version of the NLI-PT dataset (del Río, Zampieri, and Malmasi, 2018). The goal of this resource is to make available annotated data produced by L2 Portuguese learners. NLI-PT was originally created for running Native Language Identification (NLI) experiments, and it contains written texts compiled from different learner corpora of L2 Portuguese. Those texts are presented in a clean TXT version, together with versions annotated at two linguistic levels: morphological (POS) and syntactic. The annotation of the dataset was performed with freely available tools. For POS there is a simple POS representation, that is, only type of word, and a fine-grained POS, which is the type of word plus its morphological features. The annotations were performed using the LX Parser (Silva et al., 2010) for the simple POS and the Portuguese morphological module of Freeling (Padró and Stanilovsky, 2012) for detailed POS. Concerning syntactic annotations, NLI-PT includes constituency (from LX Parser) and dependency (DepPattern toolkit (Otero and González, 2012)) annotations.

The new version of the dataset is bigger and contains several improvements. We have corrected some tokenization issues and improved the constituency annotations. Besides this, we have enlarged the dataset with texts from the CAL2 learner corpus [2] (930 new texts). Additionally, we have modified the structure of the dataset. In the new version, the name of the files contains three types of information: the source corpus, the L1 and the proficiency level of the text. For example, for the file “ara_A_008CVETF_cop.txt”, the prefix *ara* corresponds to the native language, Arabic; the *A* corresponds to the CFR proficiency level and the suffix *cop* refers to the source of the file, the COPLE2 corpus (Mendes et al., 2016). The CEFR proficiency levels considered in the original learner corpora are not the same: two corpora consider five levels, A1 to C1, while other two consider only three major levels, A, B and C. For this reason, in NLI-PT we have homogenized the levels to three: A, B, and C. The final dataset contains a total of 3,069 texts, corresponding to 15 different native languages. The distribution by proficiency level is presented in Table 1.

As we can see, the distribution of texts by proficiency level is not balanced. For this reason, in our experiments we have used two different datasets: one containing the whole dataset and one with a balanced distribution of 466 texts per class. For the experiments, we split both corpora in training (80%) and testing (20%) sets.

[2] http://clunl.fcsh.unl.pt/en/online-resources/corpus-de-aquisicao-de-l2/

Proficiency	**Number of Texts**
A - Beginner	1,388
B- Intermediate	1,215
C- Advanced	466
Total	**3,069**

Table 1: Distribution of texts by CEFR proficiency level in the NLI-PT dataset

3.2 Features

We were interested in investigating the impact of different linguistic features in the classification task. For this reason, we have tested different types of features extracted from NLI-PT:

1. **Bag of words**, with different variations: using the word form, tokens and lemmas. We performed some initial experiments with the training set to check which representation produced the best results. We got similar results for word forms and tokens, being the word form representation slightly better. For this reason, for further experiments we kept only the word form representation.

2. **POS n-grams**: we used the fine-grained POS representation of NLI-PT, which contains the main POS and also morphological information, like gender or number. We consider that this information could be especially interesting because Portuguese has a rich morphology, and this feature is problematic for certain learners, especially at the initial stages. Agreement errors like *aréia branco* (*white*-MasculineSingular *sand*-FeminineSingular) can be captured with a POS n-gram representation, and we wanted to measure the impact of this feature. We evaluated n-grams of different sizes in the experiments.

3. **Dependency triplets n-grams**: we extracted dependency triplets with the form *head, relation, dependent* generated with DepPattern. Dependency re-

lations are not common in proficiency classification, and we were interested in checking their impact. We also evaluated different types of sizes for the dependency n-grams.

4. **Descriptive and lexical features of the text**: set of 39 features that the studies of SLA have proved as linked with proficiency. Those features are not present in NLI-PT, and therefore we extracted them using the software Pylinguistics (Woloszyn et al., 2016). The features include different types of measures:
   - **Lexical features**: number of nouns, number of verbs, number of connectives, lexical diversity, content diversity...
   - **Descriptive measures**: average syllables per word, syllable count, word count, etc. We also used the Portuguese adaptation of the Flesch reading index (Martins et al., 1996).

4 Experiments

As we have seen, the task of proficiency evaluation can be considered as a classification or a regression problem, depending on the way we consider the proficiency levels, that is, as discrete or continuous scales. For this first attempt, we explore the task as classification, considering that this model obtained better results in previous works (Vajjala and Lõo, 2014).

We used the scikit-learn package (Pedregosa et al., 2011) for training and testing the models and for feature selection. We divided both datasets into training and test sets. We performed some initial tests for feature selection (see above) and for evaluating different algorithms. In these previous experiments, we performed 10-fold cross-validation with the training set and the different sets of features, and we trained a different classifier for each type of features to support a comparison of them. We evaluated Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Random Forests, and LogitBoost. In general, we had the best results with three algorithms: Logistic Regression (LR), Random Forests (RF) and LogitBoost (LB). For this reason, we only used the models generated with these three algorithms against the test set.

We employed accuracy as the main measure to evaluate the performance of our trained models. We also report weighted-F1 score because the whole dataset is unbalanced. Weighted-F1 score is computed as the weighted average of the F1 score for each label, taking label support (i.e., number of instances for each label in the data) into account. As a baseline, we used text length, extracted with Pylinguistics.

4.1 Results and Discussion

Due to space restrictions, we report only the best-performing systems for each combination of features.[3]

Features	**Accuracy**	**F1**
Baseline_LR	0.58	0.54
BOW_LB	**0.70**	**0.7**
POS_LB	**0.66**	**0.65**
Dep_RF	0.64	0.59
Desc_RF	0.63	0.59
ALL(noBOW)_RF	**0.67**	**0.66**
ALL_LR	**0.72**	**0.71**

Table 2: Results for the whole dataset

Features	**Accuracy**	**F1**
Baseline_RF	0.48	0.46
BOW_LB	**0.66**	**0.67**
POS_RF	**0.63**	**0.63**
Dep_RF	0.54	0.53
Descriptive_RF	0.57	0.57
ALL(noBOW)_RF	**0.59**	**0.58**
ALL_RF	**0.65**	**0.64**

Table 3: Results for the balanced dataset

Features	**A-F1**	**B-F1**	**C-F1**
Baseline_LR	0.67	0.58	0
BOW_LB	**0.8**	**0.70**	**0.43**
POS_LB	**0.77**	**0.66**	**0.28**
Dep_RF	0.74	0.64	0.02
Desc_RF	0.72	0.63	0.13
ALL(noBOW)_LR	**0.77**	**0.67**	**0.3**
ALL_LR	**0.8**	**0.72**	**0.42**

Table 4: Results per class for the whole dataset

[3] For each set of features, the abbreviation after the underscore indicates the name of the algorithm employed: LR for Logistic Regression; RF for Random Forests; LB for LogitBoost.

Features	**A-F1**	**B-F1**	**C-F1**
Baseline_RF	0.62	0.45	0.31
BOW_LB	**0.75**	**0.62**	**0.63**
POS_RF	**0.73**	**0.57**	**0.58**
Dep_RF	0.7	0.47	0.43
Desc_RF	0.64	0.52	0.53
ALL(noBOW)_RF	**0.7**	**0.52**	**0.53**
ALL_RF	**0.71**	**0.6**	**0.59**

Table 5: Results per class for the balanced dataset

For all the models, the results obtained are better in the whole dataset than in the balanced one. The best result we got is **0.72** accuracy using an ensemble combination of all the features (ALL) with LR, although this value is very close to the one using a BOW representation, 0.7. Interestingly, in the balanced dataset the ensemble combination with all the features had worse results than the best model, BOW_LB, which uses only one feature. For the ensemble combination that does not use lexical information (it does not include the word forms), POS+Dep+Desc., the results are slightly better than the POS n-gram representation for the whole dataset and worse for the balanced dataset. Both results seem to indicate that adding more linguistic features to the best single-feature models (BOW and POS n-grams) implies only a small gain for the whole dataset and a drop in accuracy for the balanced dataset. In this case, simpler models work generally better.

If we compare the models that use only one type of feature, the best results are for the BOW representation in both datasets, followed by the POS n-gram representation. One of the reasons why the BOW representation captures better the proficiency can be the fact that it keeps the information concerning orthographic problems. A comparison between the results for each type of feature shows similar behavior for both datasets, with the only difference that the Descriptive feature set performs better than the Dep one for the balanced dataset. The algorithms with the best results differ among datasets: LR for the whole dataset; RF for the balanced one.

Concerning the results by class, we can see clear differences between datasets. In the whole corpus, the C class (the one with fewer texts) performs clearly worse than the other two, being the best result 0.43 (BOW_LB). In the balanced dataset, the F1 score is more equalized between classes, being the results for the B and the C classes pretty similar. In fact, in general, the C class gets better results than the B class. For both datasets and for all models, the best results are always for the A class. This pattern suggests that A texts exhibit certain specific features that make them easy to identify, in comparison with the other two levels.

Due to the lack of space, we cannot include the confusion matrix for each model, therefore, we include table 6 as a reference. False negatives for A are more frequent in the adjacent class, B, and the same happens with the B class, where more texts are classified as C than as A. For the C class, false negatives are more frequent in the previous class, B. This picture seems to show the expected progression of the learners as the proficiency level increases. Interestingly, B texts are more often confused with the adjacent class, instead of with the previous one.

	A	**B**	**C**
A	66	18	10
B	12	60	21
C	12	29	52

Table 6: Confusion matrix for the best model (Words_LB) in the balanced dataset

5 *Conclusions and future work*

We present a new improved version of the NLI-PT dataset, and we use it to perform the first experiments on proficiency classification for L2 Portuguese, using the CEFR scale. We modeled the task as classification, and, with the best model, we obtained an accuracy of 72%.

We were interested in answering the question: What defines the proficiency level of an L2 Portuguese text? With this goal, we tested the contribution of different linguistic features combined with different algorithms to the classification task. Additionally, we wanted to test the influence of the distribution of texts by class, and therefore we used two datasets: the whole NLI-PT corpus and a balanced set. We found that an ensemble model, with all the features combined, had the best accuracy (for the whole dataset), but also that a BOW model achieves a very similar performance (70% accuracy) in both corpora. A POS n-gram model, that does not use lexical information, gets a close result, 66%. This finding is particularly interesting because a POS n-gram model is a more abstract representation that can be less biased by topic or task variables and that can be applied to other L2 Portuguese corpora or even to other similar languages, like Spanish. The two ensemble combinations of features, one with all the features and the other without BOW, get slightly better results than the single-feature models in the whole dataset, but not in the balanced one. This result seems to indicate that simpler models work better in our datasets. A hypothesis that can explain the dominance of the BOW model is the fact that it may capture the orthographic particularities of the learners' writing, but further analyses are needed to prove this. For both datasets and all models, the class with the best F1 score is the basic user level. This fact seems to indicate that this is the proficiency level with the most characteristic traits.

We would like to investigate several aspects in future work. Since variables like task or textual genre have been proved to influence the linguistic complexity and accuracy of L2 texts (Alexopoulou et al., 2017), we would like to test our best models against L2 Portuguese texts from different sources with different topics and tasks, to check the influence of these variables. We also would like to run a cross-lingual experiment with a close language, like Spanish, following the approach of (Vajjala and Rama, 2018). Concerning the machine learning techniques we employed, we are curious about changing the approach and conceive the task as regression, as in (Yannakoudakis et al., 2018). Finally, we would like to test word embeddings and a neural network model, as in (Vajjala and Rama, 2018).

Aknowledgements

This work was supported by national funds through FCT - Fundacão para a Ciência e a Tecnologia postdoctoral program.

References

Alexopoulou, D., M. Michel, A. Murakami, and D. Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques: Task effects in a large-scale learner corpus. *Language Learning*, 03.

Boyd, A., J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, and C. Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC'14*, Reykjavik, Iceland, may.

Branco, A., J. A. Rodrigues, F. Costa, J. R. Silva, and R. Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. In *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR 2014*, pages 256–261.

Burstein, J. 2003. The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. Shermis and J. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, chapter 9, pages 113–121.

Burstein, J. and M. Chodorow. 2012. Progress and New Directions in Technology for Automated Essay Evaluation. *The Oxford Handbook of Applied Linguistics*, pages 487–497, 01.

Curto, P., N. Mamede, and J. Baptista. 2015. Assisting European Portuguese Teaching: Linguistic features extraction and automatic readability classifier. In *Computer Supported Education. Selected Papers from CSEDU2015*. Springer-Verlag.

del Río, I., M. Zampieri, and S. Malmasi. 2018. A Portuguese Native Language Identification Dataset. In *Proceedings of the 13th BEA Workshop*, pages 291–296, New Orleans, Louisiana, June. Association for Computational Linguistics.

Europe, C., C. Cultural Co-operation, E. Committee, and M. Languages Division. 2009. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.*

Gyllstad, H., J. Granfeldt, P. Bernardini, and M. Källkvist. 2014. Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, 14(1):1–30.

Kyle, K. and S. A. Crossley. 2014. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49, 09.

Lu, X. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2):190–208.

Martins, T. B. F., C. M. Ghiraldelo, M. d. G. V. Nunes, and O. N. d. Oliveira Junior. 1996. *Readability formulas applied to textbooks in Brazilian Portuguese*. Icmsc-Usp.

Mendes, A., S. Antunes, M. Janssen, and A. Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings LREC*.

Nicholls, D. 1999. The Cambridge Learner Corpus-Error coding and analysis.

Otero, P. G. and I. González. 2012. DepPattern: a Multilingual Dependency Parser. In *Proceedings of PROPOR*.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings LREC*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pilán, I., S. Vajjala, and E. Volodina. 2016. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *CoRR*, abs/1603.08868.

Silva, J. R., A. Branco, S. Castro, and R. Reis. 2010. Out-of-the-Box Robust Parsing of Portuguese. In *Proceedings of PROPOR*, pages 75–85.

Taghipour, K. and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891. Association for Computational Linguistics.

Tono, Y. 2000. A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora*, Bern, Switzerland. Peter Lang.

Vajjala, S. and K. Loo. 2013. Role of Morpho-Syntactic Features in Estonian Proficiency Classification. In *Proceedings of the Eighth BEA Workshop*, pages 63–72, Atlanta, Georgia, June. ACL.

Vajjala, S. and K. Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden, November. LiU Electronic Press.

Vajjala, S. and T. Rama. 2018. Experiments with universal CEFR classification. *CoRR*, abs/1804.06636.

Vyatkina, N. 2012. The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. *The Modern Language Journal*.

Woloszyn, V., S. Castilhos, D. Barone, and L. Wives. 2016. Pylinguistics: an open source library for readability assessment of texts written in Portuguese. 04.

Yannakoudakis, H. 2013. Automated assessment of English-learner writing. Technical Report UCAM-CL-TR-842, University of Cambridge, Computer Laboratory, October.

Yannakoudakis, H., T. Briscoe, and B. Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *HLT '11*, pages 180–189, Portland, Oregon, USA, June. ACL.

Yannakoudakis, H., Øistein E Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.