Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese

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

Abstract

This work explores the relationship between L2 proficiency levels and certain linguistic features through experiments in automatic proficiency classification. We use L2 Spanish and L2 Portuguese data and perform monolingual and cross-lingual experiments. We also compare native and leaner Spanish texts. To the best of our knowledge, this is the first work that performs automatic proficiency classification for L2 Spanish, as well as cross-lingual proficiency classification between L2 Portuguese and L2 Spanish. Our results for L2 Spanish are similar to the state-of-the-art, while our cross-lingual experiments got lower results than similar works. In general, all the experiments suggest new insights about the relationship between linguistic features and proficiency levels in L2 Portuguese and L2 Spanish.

1 Introduction

Proficiency classification is a common task in second language learning. The linguistic development of the learner is usually defined through a scale that accounts for 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. These classes are subdivided into 6 development levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Each level relates to specific linguistic features and skills, and the whole scale establishes a progression from a very rudimentary language to a performance close to a native production. CEFR has become the most common framework for second language learning in Europe, and in this context, it is common that learners perform placement tests that define their proficiency level according to the CEFR scale. The interest of an automatic system that can perform this task is, therefore, evident.

Automatic proficiency classification is considered as a type of Automatic Essay Scoring (AES) task. AES has been explored mainly for English (Burstein, 2003; Burstein and Chodorow, 2012; Yannakoudakis, 2013), but recent approaches have dealt with other languages (Vajjala and Loo, 2013). Researchers have modeled AES as a regression (Yannakoudakis et al., 2011), a ranking (Taghipour and Ng, 2016) or a classification problem (Pilán et al., 2016). Different types of features have been used in the task, from Bag-of-words (BOW) to more abstract representations that use higher levels of linguistic information (morphological, syntactic or even discursive). It is also very common the use of metrics that have been linked to proficiency development and/or linguistic complexity in the area of Second Language Acquisition (SLA), like lexical richness or syntactic complexity. Automatic proficiency classification has been approached mainly as a monolingual task, but recent approaches like (Vajjala and Rama, 2018) have explored multi and cross-lingual perspectives.

In our experiments, we use the main levels of the CEFR scale (A, B, C) and supervised machine learning techniques to classify L2 Portuguese and L2 Spanish texts. As features, we test different linguistic representations, from BOW to syntactic dependencies, and some complexity features. We perform monolingual and cross-lingual experiments, and we compare native to L2 productions

Iria del Río 2019. Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. Linköping Electronic Conference Proceedings 164: 31–40.

in Spanish. Furthermore, we try to answer the following questions: which linguistic features capture better the proficiency of a L2 text in Spanish and Portuguese? Are those features similar between these two close languages? When comparing L1 and L2 Spanish, which linguistic characteristics allow for predicting the level of linguistic development of a text? We present relevant related work in section 2, and our methodology in section 3. In section 4 we describe the experiments performed and discuss our results, while in section 5 we summarize our conclusions and future directions of work.

2 Related Work

In this section we focus on two types of research: SLA studies that have analyzed the relationship between certain linguistic features and proficiency levels, and approaches that have used machine learning to predict L2 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 a L2 essay. Several features identified as relevant in this work have been used by automatic approaches afterwards. Crossley and McNamara (2011) and more recently Eckstein and Ferris (2018) compare L1 and high proficiency L2 English texts through different metrics of lexical sophistication, syntactic complexity and cohesion. Both studies conclude that L2 texts can be clearly differentiated from L1 texts, and (Crossley and McNamara, 2011) shows also homogeneity between L2 learners with different native languages (L1). Other characteristics like error patterns have been studied too, mainly for English (Tono, 2000), (Lu, 2012), (Vyatkina, 2012), but also for other languages (Gyllstad et al., 2014).

Yannakoudakis et al. (2018) is one of the most recent works for automatic proficiency classification of L2 English. The authors use a subset of the Cambridge Learner Corpus with human proficiency annotations (levels A1 to C2), and model the task as a ranking function. They use features as character sequences, POS, phrase structure rules or errors 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) which 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. They 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: German, 1,029 texts; Italian, 803 texts, and Czech, 434 texts. 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). (Vajjala and Lõo, 2014) performs proficiency classification for Estonian. They use a corpus of 879 texts, with 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 set of 78 features that considers morphological aspects and lexical richness features inspired by (Lu, 2012). Interestingly, POS models achieved a 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 10 of the 27 best features they identified are lexical (like Corrected Type Token Ratio) and morphological.

3 Methodology

3.1 Corpus

For our experiments we used two different datasets: NLI-PT (del Río et al., 2018) for L2 Portuguese and CEDEL2 corpus (Lozano, 2009) for L2 Spanish. While CEDEL2 is a learner corpus with a planned design, NLI-PT is a compilation of learner texts that belong to four different L2 corpora. Because of this, CEDEL2 is more homogeneous in terms of L1, task and topic than NLI-PT. CEDEL2 has also native texts that

constitute a control corpus. We have used these texts for some experiments too, as we will see below. NLI-PT contains annotated versions of learners' texts with two types of linguistic information: morphological (POS) and syntactic (constituency and dependencies). CEDEL2 corpus is not annotated so, to extract the linguistic features that we needed for our experiments, we added similar annotations as the ones in NLI-PT. We incorporated fine-grained POS information using the Spanish tagger of Freeling (Padró and Stanilovsky, 2012), and syntactic dependencies using the DepPattern toolkit (Otero and González, 2012) for L2 and native texts. We also extracted several complexity metrics from both datasets using our own scripts (see section 3.2).

The way to conceptualize proficiency levels is also different in NLI-PT and CEDEL2. While NLI-PT texts are classified according to the CEFR scale, CEDEL2 uses a different classification system. In CEDEL2 the level of the text is determined through a placement test that uses a scale from 0 to 100. Since we are interested in using the CEFR scale as our reference, we converted the CEDEL2 scale to CEFR using the equivalences that the CEDEL2 team has established.[1] In our experiments, we consider the three major levels of the CEFR scale: A, basic user; B, independent user; C, proficient user.

In total, NLI-PT dataset contains 3,069 L2 Portuguese texts, and CEDEL2 1,778 L2 Spanish essays and 796 native Spanish texts. Tables 1 and 2 show the distribution of learner texts by proficiency level in each dataset.

Proficiency Level	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 NLI-PT.

3.2 Features

We were interested in investigating the impact of different linguistic features in the classification task. As a first approach to the task in L2 Spanish and cross-lingual L2 Spanish-Portuguese, we were interested in testing basic linguistic representations like BOW or POS n-grams. These features have been proved as useful in previous experiments like (Vajjala and Rama, 2018) or (Yannakoudakis et al., 2018), and they are already available in NLI-PT dataset (and they are easy to get for CEDEL2). Considering the evident importance of complexity features, we included some of them in our experiments too but, due to time and space limitations, we did not explore the wide spectrum of linguistic complexity.[2] We defined the following sets of features for our experiments:

Proficiency Level	Number of Texts
A - Beginner	456
B - Intermediate	675
C - Advanced	647
Total	**1,778**

Table 2: Distribution of learner texts by CEFR proficiency level in CEDEL2.

1. **General linguistic features:**
   (a) **Bag-of-words**: this is the simpler representation of a text. We used the original word form, keeping the case. Previous experiments in proficiency classification with NLI-PT for L2 Portuguese (del Río, 2019) showed that using tokens, word forms[3] or lemmas lead to similar results in the classification task. Considering this and the fact that the original word forms may indicate patterns of orthographic deviations in the L2 texts, we kept the original word forms for the BOW representation.
   (b) **POS n-grams**: we used the fine-grained POS representation from Freeling, which contains the main POS and also morphological information, like gender or number. We consider that this information could be especially interesting because Portuguese and Spanish have a rich morphology, and this is problematic for some learners, especially at

[1] We thank professor Cristóbal Lozano for providing us the equivalence table.

[2] There is almost not research in linguistic complexity for Spanish or Portuguese and, therefore, there are not available tools to extract lexical or syntactic complexity measures automatically.

[3] The difference between word form and token applies in special cases like contractions or verbal forms with clitics. For example, with the verb "ligar-te" ("to call you"), we have one word form "ligar-te", but two tokens: "ligar" and "te".

the initial stages. Agreement errors like *aréia branco* (*white*-MasculineSingular *sand*-FeminineSingular) can be captured with a POS 2-gram representation, and therefore we wanted to measure the impact of this feature. We evaluated n-grams of different sizes in the experiments.

(c) **Dependency triplets n-grams**: we extracted dependency triplets with the form *relationship, head, dependent* generated with DepPattern. Dependency relationships are not commonly used in proficiency classification, and we were interested in checking their impact. We also evaluated different sizes for the dependency n-grams.

2. **Complexity features**: as we have seen, complexity features have been proved to be useful to differentiate native and learner texts. Moreover, these type of features are commonly used in the task of proficiency classification. We have selected a set of 20 descriptive, morphological and lexical features linked to linguistic complexity (see related work). We have implemented different scripts to extract the features from NLI-PT and CEDEL2.[4] This group includes different types of metrics:

   (a) **Morphological metrics**: number of nouns, number of verbs, number of lexical words, etc.[5]

   (b) **Lexical metrics**: type-token ratio (ttr) (with different variations: rooted ttr, corrected ttr, mean segmental ttr...), hypergeometric distribution diversity (McCarthy and Jarvis, 2010), etc.

   (c) **Descriptive metrics**: average syllables per word, syllable count, word count, readability score (we used the Portuguese adaptation of the Flesch reading index (Martins et al., 1996)).

3.3 Classification and Evaluation

As we have seen, the task of proficiency testing 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 units. In this work we are interested in conceptualizing proficiency levels in the same way that the CEFR scale does, that is, as discrete entities. Therefore, we modeled the problem as a classification task. Another reason to choose classification over regression is presented in (Vajjala and Lõo, 2014), who compared both approaches and got better results with the classification model.

We used the scikit-learn package (Pedregosa et al., 2011) for training and testing the models and for feature selection. We split both datasets in training and test sets (20% of data) for all the experiments. In general, for each experiment we performed initial tests to check which algorithm worked better with each set of features. In these previous experiments, we performed 10-fold cross-validation with the training set, training a different classifier for each set of features to support a comparison of them. We evaluated a varied group of linear and nonlinear algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), KNeighborsClassifier (KNN), DecisionTreeClassifier(CART), GaussianNB (NB), Support Vector Clustering (SVC), LogitBoost (LB) and Random Forests (RF). For each set of features, we selected the best performing model and we evaluated it against the test set.[6]

We use accuracy as the main measure to evaluate the performance of our trained models. We also report weighted-F1 score because the datasets are 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. We also show F1 score per class, to analyze in detail the performance of the classifiers by level. We use text length as the general baseline.

4 Experiments and results

We have performed different classification experiments to investigate the relationship between the linguistic features selected and the main CEFR proficiency levels. We investigate this relationship in three different scenarios. First, we study the interaction of features and proficiency levels for L2 Spanish, using CEDEL2 texts. Our main research question is: which linguistic features in our two sets allow for an accurate classification of CEFR

[4] We will make the scripts available after the publication of the paper.

[5] Counts are normalized by text length with the following formula: number of nouns/total words in text*1000.

[6] We indicate the algorithm used for each model in the tables with the results.

proficiency levels in L2 Spanish? This is a monolingual approach similar to the ones presented in the related work section. Secondly, we investigate the same interaction in a cross-lingual scenario, from Spanish to Portuguese and vice versa. With this experiment, we try to reply to two research questions: (i) are the linguistic patterns linked to each proficiency level in our two L2 languages similar to the extent that a model trained in one language can be transferred to the other?; (ii) if so, which features work better in the cross-lingual model? This is an experiment similar to the one presented in (Vajjala and Rama, 2018), where a model trained with German texts is applied to a Czech and an Italian test set. From the typological point of view, German is not close to Czech or Italian, while Portuguese and Spanish are similar languages, with a close morphology and Latin vocabulary. Considering this fact, a priori we could expect good results in the cross-lingual experiments. Finally, we study the relationship between learner and native texts, using our sets of features. For those experiments, we were interested in replying to the following research question: which are the best linguistic features to differentiate a learner text from a native one? We present our results in the sections below.

4.1 L2 Spanish

For the monolingual scenario our best result is 74% accuracy and F1 score, a result similar to the ones we can find in the current literature for other languages.[7]

Comparing the two general types of features (linguistic and complexity), all the sets of features perform better than the baseline (text length), and the linguistic features perform better than the complexity ones. All the linguistic one-feature sets get a 70% of accuracy or more, while for the complexity group only the descriptive features achieve a 70%. Among the linguistic features, the best result is for POS, which in fact achieves an accuracy similar to the best result (73% for POS and 74% for the best result). The assembled set of linguistic features perform better than the assembled set of complexity features, but worse than the POS features individually. The best result is for the combination of POS and complexity features (POS+Co_RF) with a 74% of accuracy and weighted F1 score. This set performs even better than all the features together. Concerning the algorithms we can see that from the initial list of eight three got the best results: RF, LR and LB.

Features	Accuracy	F1-Score
Baseline_RF	0.60	0.58
BOW_LB	0.70	0.70
POS_RF	**0.73**	**0.72**
Dep_LB	0.70	0.70
LING_LR	0.72	0.71
CoLex_LR	0.63	0.61
CoMor_NB	0.49	0.47
CoDesc_LR	0.70	0.70
COMP_LDA	**0.70**	**0.70**
POS+Co_RF	**0.74**	**0.74**
POS+Dep+Co_LR	0.74	0.73
ALL_LR	0.72	0.72

Table 3: General results for L2 Spanish.

Features	A-F1	B-F1	C-F1
Baseline_RF	0.68	0.33	0.69
BOW_LB	0.71	0.59	0.79
POS_RF	**0.76**	**0.60**	**0.82**
Dep_LB	0.72	0.61	0.78
LING_LR	0.77	0.59	0.80
CoLex_LR	0.71	0.43	0.73
CoMor_NB	0.44	0.34	0.61
CoDesc_LR	0.74	0.60	0.77
COMP_LDA	**0.73**	**0.61**	**0.77**
POS+Co_RF	**0.77**	**0.62**	**0.83**
POS+Dep+Co_LR	0.76	0.60	0.80
ALL_LR	0.77	0.62	0.80

Table 4: Results per class for L2 Spanish.

Moving to the performance by class, we can see that the level that gets the best results for all the features is C, and the level with the worst results is B. Interestingly, CEDEL2 has more B texts than C texts (675 vs. 647). The A level, which has significantly less texts than B and C (456), gets in general a similar F1 score to that of level C. This seems to indicate that A and C texts are easy to classify, while B texts are difficult, no matter the number of training texts or the set of features we employ. This result makes sense from the linguistic point of view, because A and C levels are in the extreme of the proficiency development scale,

[7] It is important to note that, due to the unbalanced structure of our datasets, we were forced to use three classes that correspond to the main CEFR levels, while previous works generally use four or more classes.

while B is in the middle and, therefore, B texts can be close to A or to C levels.

Attending to the single features as predictors of each class, POS is the best for A and C, while dependencies are the best predictor for B (although very close to POS). POS is especially useful in the classification of the C class, with a 82% of F1 score.

Summing up, our results show that linguistic features are most effective than complexity ones for classifying proficiency levels in CEDEL2, being the POS features the most useful. Among the complexity features, the descriptive ones are the most efficient for all the levels, which a similar performance to the POS set. A and C levels are easy to classify, while B is difficult, no matter what type of feature we are using.

4.2 Cross-lingual experiments: Spanish and Portuguese L2

In this case, we used (L2) CEDEL2 and NLI-PT datasets and we performed cross-lingual experiments. We tested both directions: Spanish to Portuguese and vice versa. We performed the same type of experiments as for the monolingual dataset, with the only difference that, in this case, we use the whole monolingual corpus as the training dataset, and a section of the other dataset as the testing one.

Features	Accuracy	F1-Score
Baseline_LR	**0.57**	**0.54**
BOW_CART	0.47	0.40
POS_RF	**0.57**	**0.51**
Dep_LB	0.47	0.36
LING_RF	0.50	0.40
CoLex_NB	0.43	0.42
CoMor_SVM	0.39	0.22
CoDesc_NB	0.49	0.50
COMP_NB	0.44	0.44
POS+Co_RF	**0.57**	**0.52**
POS+Dep+Co_LR	0.55	0.48
ALL_RF	0.54	0.46

Table 5: General cross-lingual results for Spanish to Portuguese.

We can see that, in general, we get poor results in the cross-lingual models. For Spanish to Portuguese, none of the trained classifiers beats the baseline, and only the combination of POS and complexity features gets close. The only one-feature set that performs similar to the baseline is the POS one, but if we check table 6 we can see that the F1 score for level C is 0. All the combinations with complexity features get results below the baseline, being the descriptive metrics the ones with the best score. These numbers seem to be in line with the ones obtained for the monolingual dataset, with the POS and descriptive features as the most relevant in the classification task.

Features	A-F1	B-F1	C-F1
Baseline_LR	**0.67**	**0.55**	**0.54**
BOW_CART	0.61	0.33	0
POS_RF	**0.68**	**0.52**	**0**
Dep_LB	0.63	0.19	0
LING_RF	0.65	0.26	0
CoLex_NB	0.40	0.49	0.30
CoMor_SVM	0.48	0.48	0.25
CoDesc_NB	**0.60**	**0.49**	**0.25**
COMP_NB	0.48	0.48	0.25
POS+Co_RF	**0.66**	**0.55**	**0**
POS+Dep+Co_LR	0.65	0.30	0
ALL_RF	0.66	0.40	0

Table 6: Results per class for cross-lingual Spanish to Portuguese.

Concerning the results per class, POS features are the best to predict the A and B level, and lexical-complexity features the best for C. Interestingly, seven of the twelve models get a F1 score of 0 for this level, while they are able to classify the other two levels, and only the complexity features are useful to classify level C. On the other hand, all the models get the best results for the A level. Linguistic features are the more accurate for predicting this level, especially the POS features, which also get a high F1 score predicting the B level. However, they obtain a 0 F1 score predicting the C level. We can see in table 7 that for Portuguese to Spanish, POS features are also the best among linguistic features to predict level A and B, while they are the worst to predict level C. This fact seems to indicate that A and B level show certain recurrent morpho-syntactic patterns that allow for their identification cross-linguistically, while level C does not.

We could consider the different number of texts in both datasets as a possible factor for these poor results. CEDEL2 has almost half of the texts of NLI-PT, although CEDEL2 has more C texts than A texts, for example, and the performance of the

classification models for the A level is always better than for the C level. Another possible factor that can impact the results is the homogeneity of CEDEL2 versus the heterogeneity of NLI-PT in terms of L1, task or topic. More experiments are necessary to check the impact of these variables.

For Portuguese to Spanish the results are better, although still lower than expected. Only the complexity models beat the baseline, being the best result for the complexity lexical features, with a 60% of accuracy and a F1 score of 58%. The descriptive-complexity features beat also the baseline, while the morphological ones do not. All the linguistic features show a low performance, being the BOW model the best one. Interestingly, in this case POS features show a low performance. These results are similar to the ones obtained for monolingual L2 Portuguese in (del Río, 2019), where the best results in the classification task were for the BOW model. Unfortunately, that work did not use all the complexity features that we present, which makes difficult a complete comparison of the results.

Features	Accuracy	F1-Score
Baseline_NB	**0.56**	**0.54**
BOW_LB	0.50	0.49
POS_CART	0.47	0.46
Dep_LB	0.46	0.44
LING_RF	0.39	0.29
CoLex_NB	**0.60**	**0.58**
CoMor_NB	0.39	0.30
CoDesc_NB	0.57	0.55
COMP_NB	0.60	0.57
POS+Co_KNN	0.48	0.45
POS+Dep+Co_KNN	0.48	0.25
ALL_KNN	0.49	0.45

Table 7: General cross-lingual results for Portuguese to Spanish.

Analyzing the results per class, for most of the models A and C level perform better than B, being the A level the easiest to identify (as we saw for monolingual and cross-lingual experiments). Among the one-feature sets, lexical-complexity features are clearly the best to predict the A level; POS (as for cross-lingual Spanish to Portuguese) and morphological-complexity features are the best for predicting the B level; and lexical-complexity features are the best to predict the C level.

Features	A-F1	B-F1	C-F1
Baseline_NB	**0.69**	**0.37**	**0.62**
BOW_LB	0.57	0.40	0.52
POS_CART	**0.59**	**0.45**	**0.37**
Dep_LB	0.51	0.31	0.54
LING_RF	0.61	0.36	0
CoLex_NB	**0.74**	**0.38**	**0.67**
CoMor_NB	0.52	**0.45**	0
CoDesc_NB	0.71	0.38	0.63
COMP_NB	0.74	0.36	0.67
POS+Co_KNN	0.65	0.48	0.28
POS+Dep+Co_KNN	0.65	0.30	0
ALL_KNN	0.66	0.40	0

Table 8: Results per class for cross-lingual Portuguese to Spanish.

Considering that NLI-PT has more texts and is less homogeneous than CEDEL2 in terms of L1 languages, topics or even tasks, which theoretically implies more variation, it seems that complexity features are the most robust to support the adaptation from one L2 to another. POS features appear to be especially useful for predicting A and B level.

We would like to note that our results are quite different to the ones obtained by (Vajjala and Rama, 2018), where the cross-lingual model trained with German texts performs similarly when tested in Czech and Italian as the corresponding monolingual models do. There are two differences between our experiments and the ones presented in that work, though: first, all the texts used in their experiments belong to the same multilingual corpus, MERLIN, factor that allows for a higher homogeneity in terms of topic and task; second, they train the model on the language with more texts (German) and test with the languages with less texts (Czech and Italian). However, German Czech and Italian are more distant languages than Spanish and Portuguese, and even so their results are stable when the model performs cross-lingual, contrary to what we found. More tests will be necessary to investigate the possible causes of this difference.

4.3 Learner texts versus Native texts in Spanish

We were interested in measuring to what extent a machine learning algorithm is able to distinguish between a text written by a learner and a text writ-

ten by a native speaker, and also in knowing which of the features in our two sets are more useful in this task. CEDEL2 has a corpus control with 796 texts written by native speakers, covering the same topics as the L2 corpus. We created a dataset with the L2 and the native texts (L2+NAT), and we labeled the native texts with a new class, "N". Therefore, this time the classification model has to distinguish among four levels: A, B, C and N. For the selection of the algorithms, we used the same approach as before: we tested several algorithms with the training corpus, and we selected the best model to evaluate it against the testing set.

Features	Accuracy	F1-Score
Baseline_LR	0.50	0.43
BOW_RF	**0.73**	**0.73**
POS_NB	0.39	0.33
Dep_LR	0.37	0.30
LING_LR	**0.75**	**0.74**
CoLex_LR	0.62	0.61
CoMor_ NB	0.40	0.40
CoDesc_LR	0.60	0.59
COMP_LR	**0.65**	**0.64**
POS+Co_RF	0.74	0.74
POS+Dep+Co_LR	0.74	0.74
ALL_RF	**0.75**	**0.74**

Table 9: Classification including native texts.

The best result (LING and ALL) is slightly better than the best for the L2 Spanish dataset: 0.75 vs. 0.74 of accuracy. The one-feature set that performs better is BOW, with an accuracy and F1 score similar to the top result. If we compare the results for both experiments by sets of features, we can see that most of the sets get similar results, except for POS and Dep, that clearly got worse results in the L2+NAT dataset.

LING features allow for a increase in accuracy and F1 score using the native texts, while COMP features get worse results. If we analyze the results per class, it seems that the main cause of this is the behaviour of the C class. Using only learner texts, the C class got the best results, together with level A (see Table 4). However, including the native texts, the C level decreases in F1 score for all the sets of features. The combination that allows for a smaller decrease in F1 score for C is LING (80% vs. 70%). If we take a look to the confusion matrices included in Appendix A, we can see that when we include native texts many C instances

Features	A-F1	B-F1	C-F1	N-F1
Baseline_LR	0.64	0.53	0	0.58
BOW_RF	**0.78**	**0.62**	**0.67**	**0.83**
POS_NB	0	0.25	0.52	0.42
Dep_LR	0.45	0.29	0.48	0.06
LING_LR	**0.75**	**0.63**	**0.70**	**0.88**
CoLex_LR	0.75	0.52	0.49	0.70
CoMor_NB	0.47	0.27	0.40	0.46
CoDesc_LR	0.73	0.42	0.48	0.74
COMP_LR	**0.73**	**0.56**	**0.50**	**0.78**
POS+Co_RF	0.75	0.62	0.67	0.88
POS+Dep+Co_LR	0.73	0.63	0.66	**0.90**
ALL_RF	**0.76**	**0.65**	**0.67**	**0.88**

Table 10: Classification including native texts, per level.

are classified as native. However, when we have only learner classes, C texts "compete" only with B texts. In this scenario, linguistic features seem to be more effective to differentiate C texts from native ones than the complexity features, which indicates that C texts are probably more similar to native texts in terms of complexity metrics, but still different when we consider linguistic features.

POS features show a poor performance, especially if we compare them with the L2 model. POS features, which were the most informative feature there, are the less efficient here. Dependencies are not very useful either, although they work better than POS, especially for the A and the C level. None of the texts is classified as A using POS in the L2+NAT model. The system tends clearly to classify all the texts as C or N classes, although is able to classify at least 23 B texts. However, in the L2 model the system is able to correctly differentiate the three levels without favouring any of them.

5 Conclusions and future work

This work presents the first experiments on automatic proficiency classification for L2 Spanish and cross-lingual Spanish-Portuguese. We got similar results to the state-of-the art for L2 Spanish, and lower results for the cross-lingual approach. We investigated the relationship between different types of linguistic features and the three main levels of proficiency of the CEFR framework. We concluded that the linguistic features that work better for the L2 Spanish model are not the same for the cross-lingual models. POS

representation performs better for monolingual L2 Spanish and cross-lingual Spanish to Portuguese. Complexity features related to lexical and descriptive aspects perform better for cross-lingual Portuguese to Spanish. Morphological-complexity features show a low performance in all the scenarios. When comparing L2 and L1 Spanish texts, linguistic features work as better predictors than complexity features. The A level is generally the easiest to predict (together with C) and B the most difficult. When we mix native and learner texts, C level is usually confused with the native one, especially when we use complexity features.

In future experiments we would like to investigate in depth the causes for the low results in our cross-lingual experiments. Specifically, we would like to investigate the influence of factors like the homogeneity of CEDEL2 versus the diversity of NLI-PT. Secondly, we would like to explore new features like metrics of syntactic and discourse complexity, as well as the use of neural models in the classification task.

A Appendix A: Confusion matrices

Confusion matrices for the L2 Spanish monolingual experiment and L2+NAT Spanish experiment.

Comparison of results for the LING model.

	A	B	C	N
A	69	21	0	1
B	23	83	25	4
C	1	20	88	21
N	1	3	10	145

Table 11: LING model in L2+NAT Spanish.

	A	B	C
A	76	15	0
B	31	73	31
C	0	23	107

Table 12: LING model in L2 Spanish.

Comparison of results for the COMP model.

	A	B	C	N
A	65	18	7	1
B	21	72	33	9
C	1	27	63	39
N	2	7	17	133

Table 13: COMP model in L2+NAT Spanish.

	A	B	C
A	68	19	4
B	22	74	39
C	0	21	109

Table 14: COMP model in L2 Spanish.

Comparison of results for the POS model.

	A	B	C	N
A	0	18	6	67
B	0	23	64	48
C	0	1	94	35
N	0	7	70	82

Table 15: POS model in L2+NAT Spanish.

	A	B	C
A	70	20	1
B	23	73	39
C	0	15	115

Table 16: POS model in L2 Spanish.

References

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schone, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

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

Scott A. Crossley and Danielle McNamara. 2011. Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, 20(4):271–285.

Iria del Río, Marcos Zampieri, and Shervin Malmasi. 2018. A Portuguese Native Language Identification Dataset. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

Grant Eckstein and Dana Ferris. 2018. Comparing L1 and L2 Texts and Writers in First-Year Composition. *TESOL Quarterly*, 52(1):137–162.

Council Europe, Council Cultural Co-operation, Education Committee, and Modern Languages Division. 2009. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*.

Henrik Gyllstad, Jonas Granfeldt, Petra Bernardini, and Marie Kllkvist. 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.

Cristóbal Lozano. 2009. Cedel2: Corpus escrito del español como L2. In *Applied Linguistics Now: Understanding Language and Mind*, pages 197–212.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners oral narratives. *The Modern Language Journal*, 96(2):190–208.

T.B.F. Martins, C.M. Ghiraldelo, M. das Graças Volpe Nunes, and O.N. de Oliveira Júnior. 1996. *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Icmsc-Usp.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Pablo Gamallo Otero and Isaac González. 2012. DepPattern: a Multilingual Dependency Parser. In *Proceedings of PROPOR*.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2473–2479, Istanbul, Turkey. European Languages Resources Association (ELRA).

F. Pedregosa, 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.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *CoRR*, abs/1603.08868.

Iria del Río. 2019. Automatic proficiency classification in L2 Portuguese. Accepted.

Kaveh Taghipour and Hwee Tou 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.

Yukio Tono. 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.

Sowmya Vajjala and Kaidi Loo. 2013. Role of Morpho-Syntactic Features in Estonian Proficiency Classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia. Association for Computational Linguistics.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, Uppsala, Sweden. LiU Electronic Press.

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

Nina Vyatkina. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*.

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

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

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.