Native Language Identification on L2 Portuguese

Iria del Río(✉)

Centro de Linguística da Universidade de Lisboa, Lisbon, Portugal
igayo@letras.ulisboa.pt

Abstract. This study advances on Native Language Identification (NLI) for L2 Portuguese. We use texts from the NLI-PT dataset corresponding to five native languages: Chinese, English, German, Italian, and Spanish. We include the same L1s as in previous works, and more texts per language. We investigate the impact of different lexical representations, the use of syntactic dependencies and the performance of diverse classification methods. Our best model achieves an accuracy of 0.66 including lexical features, and of 0.61 excluding them. Both results improve previous works on NLI for L2 Portuguese.

Keywords: Native Language Identification · L2 Portuguese · Second language acquisition

1 Introduction

Native Language Identification (NLI) is the task of determining the native language (L1) of an author based on his second language (L2) linguistic productions [1]. The assumption behind NLI is that speakers of the same native language share a series of linguistic patterns in their L2 productions, influenced by their mother tongue. NLI works by identifying those patterns. A major motivation for NLI is studying second language acquisition (SLA). NLI models can enable analysis of inter-L1 linguistic differences, allowing us to study the language learning process and develop L1-specific pedagogical methods and materials.

NLI research is conducted using learner corpora: collections of learner productions in an acquired language, annotated with metadata such as the author's L1 or proficiency. These datasets are the foundation of NLI experiments and their quality and availability has been a key issue since the earliest work in this area.

A notable research trend in recent years has been the extension of NLI to languages other than English [2]. Recent NLI studies on languages other than English include Chinese [3], Norwegian [4], Arabic [5] and European Portuguese [6,7]. The present work extends previous approximations to NLI on European Portuguese. The novel aspects of our work include experimenting with

© Springer Nature Switzerland AG 2020
P. Quaresma et al. (Eds.): PROPOR 2020, LNAI 12037, pp. 87–97, 2020.
https://doi.org/10.1007/978-3-030-41505-1_9

a representation of lexical features that avoids topic bias, measuring the effect of syntactic dependencies on the task or using ensemble classification methods, among others.

The paper is organized as follows: Sect. 2 discusses related work in NLI, Sect. 3 describes the methodology and dataset used in our experiments, and Sect. 4 presents the experimental results. Finally, Sect. 5 presents a brief discussion and concludes this paper with avenues for future research.

2 Related Work

NLI is a recent area of research that connects Natural Language Processing with SLA. The first works in the field appeared in the early 2000s and most significant work has appeared over the last decade [8–12]. The research community has focused on aspects like improving classification [11], studying language transfer effects [13], and applying the linguistic features to other NLP tasks.

NLI is typically modeled as a supervised multi-class classification task. In this experimental design the individual productions of learners[1] are used as training and testing data while the author's L1 information serves as class labels. It has been shown that NLI is challenging even for human experts, with machine learning methods significantly outperforming humans on the same data [15].

There have been two shared tasks focusing on NLI, one in 2013[2] and the other in 2017.[3] In 2013 the dataset used was the TOEFL11 corpus [16], the first dataset designed for NLI. The winning entry was [17], which achieved an accuracy of 0.84. They used an L2-regularized SVM classifier and n-grams of words, Part-of-Speech (POS), and lemmas as features. In addition to normalizing each text to unit length, the authors applied a log-entropy weighting schema to the normalized values, which clearly improved the accuracy of the model.

Growing interest led to another edition of the shared task in 2017, where the task included speech data. The systems that achieved the best performance across the different tracks used ensembles and meta-classifiers. Participants using deep learning-based models and features (e.g. word embeddings) did not outperform traditional classification systems. The use of more sophisticated systems led to substantially higher results than in the previous edition. A detailed report on the findings of the task can be found in [18].

Regarding classification features, NLI employs a wide range of linguistic features, lexical, morphological and syntactic. A more detailed review of NLI methods is omitted here for brevity, but a comprehensive exposition can be found in [19,20]. Some of the most successful features used in previous work include **lexical features** like character n-grams [21], Function word unigrams and bigrams [22], Word and Lemma n-grams; **morphological features** like

[1] NLI is usually applied on whole texts, although [14] performs the task also at the sentence level.

[2] https://sites.google.com/site/nlisharedtask2013/home.

[3] https://sites.google.com/site/nlisharedtask/home.

Penn Treebank (PTB) POS n-grams or RASP POS n-grams [22]; and **syntactic features** as Adaptor Grammars (AG) [23], CFG Production Rules [9], Stanford Dependencies with POS transformations [11], and Tree Substitution Grammar (TSG) fragments [10].

Besides classification, another branch of NLI uses models based on these features to generate SLA hypotheses. In [24] the authors make use of L1 and L2 data to identify features exhibiting non-uniform usage in both datasets, using them to create lists of candidate transfer features. [13] proposes a different methodology, using linear SVM weights to extract lists of overused and underused linguistic features per L1 group.

Most English NLI work has been done using two corpora, the *International Corpus of Learner English* [25] and TOEFL11. The first one is a learner corpus of L2 English, fact that implies certain shortcomings for its use in NLI being widely noted [26]. On the other hand, TOEFL11 was specifically designed for NLI, although it only contains argumentative essays, limiting analyses to this genre.

In recent years, NLI research has extended to languages other than English [5,27]. [3] introduced the *Jinan Chinese Learner Corpus* [28] for NLI and their results indicate that feature performance may be similar across corpora and even L1-L2 pairs. Similarly, [4] proposed using the ASK corpus [29] to conduct NLI research using L2 Norwegian data. Recently, the NLI-PT dataset [6] was released for L2 European Portuguese, and [7] constitutes the first attempt to apply NLI techniques to this language. In that work, the authors try to identify five L1s: Chinese, English, German, Italian and Spanish. Since NLI-PT is not topic balanced, they use only non lexical features: functional words, POS and CFG rules, and a linear Support Vector Machine (SVM) classifier. They achieve an accuracy of **0.54** with a mean probability ensemble model. The present paper develops the work presented there, including more data, new linguistic features and classification methods.

3 Data and Method

3.1 Data

Similarly to [7], we used a sub-set of the NLI-PT dataset with texts for five L1 groups: Chinese, English, German, Italian, and Spanish. We chose these five languages because they are the ones with the greatest number of texts in NLI-PT. The dataset has been recently enlarged [30] and thanks to that the number of texts per language we use is much bigger than in [7], where the authors used 215 productions per L1. Table 1 shows the composition of our data.

It is important to note that NLI-PT is not topic balanced in terms of L1 [6]. The reason is that the dataset is the result of merging different learner corpora. Even after regrouping the thematic areas, there are more than 90 different topics in the dataset, with an unbalanced distribution by number of texts or L1.

Texts in NLI-PT have annotations at two levels: POS and syntax. There are two types of POS: a simple POS with only the type of word, and a fine-grained

POS with type of word plus morphological features. Concerning syntactic information, texts are annotated with constituency and dependency representations.

Table 1. Distribution of the five L1s in the NLI-PT dataset in terms of texts, tokens, types, and type/token ratio (TTR).

L1	Texts	Tokens	Types	TTR
Chinese	440	90,424	9,931	0.11
English	409	86,017	10,323	0.12
German	430	92,756	10,713	0.12
Italian	555	129,630	14,779	0.11
Spanish	607	121,452	14,018	0.12
Total	2,441	520,279	59,764	0.12

3.2 Classification Models and Evaluation

We model the task as a standard multi-class classification problem. We test different algorithms and feature vectors created using relative frequency values, in line with previous NLI research [19]. We also experiment with ensemble methods using multiple classifiers.

We perform two types of experiments. First, for testing the impact of linguistic features and algorithms, a single model is trained on each feature type. In these experiments, we use algorithms that have been used previously for NLI and generally for text classification. Multinomial Logistic Regression [31] and Support Vector Machines [32] showed good results in previous NLI work. For SVM we test two versions, one with a linear kernel and another with a rbf kernel (both with the one-vs.-rest (OVR) approach). We experiment also with Ridge Regression [33], and a Multi-Layer Perceptron classifier. For all the algorithms we use the default parameters in the scikit-learn package excepting:

- We set the random state to 7.
- For the RBF kernel SVM model we set gamma to 'scale'.
- With Logistic Regression we use a L2 regularization with a liblinear solver.
- We set the number of epochs to 10 for the Multi-Layer Perceptron.

Once we have identified the best combination of feature plus algorithm, we run experiments using ensemble combinations of classifiers. We test two different strategies: an ensemble method that uses mean probability rule[4] and classifier stacking.

Similar to the majority of previous NLI studies, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. For generating our folds we use randomized stratified cross-validation which aims to ensure that

[4] More details about this approach can be found in [19].

the proportion of classes within each partition is equal [34].[5] We use accuracy as our main metric and we also report per-class precision, recall, and F1 scores. We compare these results against a random baseline.

3.3 Features

Previous research in NLI has shown the importance of using datasets which are balanced in terms of topic and L1. This aspect is particularly relevant for the use of content-based features, which can be topic-related and inflate accuracy [20].

NLI-PT dataset is very heterogeneous and unbalanced in terms of distribution of topics. This is the reason why [7] do not use lexical features, achieving an accuracy of 0.54. In [6], a BOW representation gets an accuracy of 0.7, suggesting the influence of topic bias. It is then clear that the use or not of lexical features has a considerable impact for the NLI-PT dataset. To investigate this impact, we experiment with lexical features and we analyse their behaviour. We consider two types of features: one includes all the words in the text (WLP) and the second one all the words except nouns and adjectives (WLPmod). The rationale behind this decision is that nouns and adjectives carry most of the lexical content in a text and, therefore, it is expected that they are the words more influenced by topic bias. Removing them can therefore help to remove topic bias. To check this assumption, we also perform a chi-squared test to extract the most correlated unigrams and bigrams per L1 for both lexical representations. Finally, instead of a simple bag of words, we chose a richer representation which includes word+lemma+POS for each of the words in a text. In WLPmod, we remove the word and the lemma of all adjectives and nouns and we only keep the POS. For both lexical features we use n-grams of size 1–3.

Besides lexical features, we use a set of morphological and syntactic features that have been proved as useful for NLI. We employ the following topic-independent feature types: fine-grained POS tags, context-free grammar (CFG) production rules and dependency triplets. We extract the features from the annotations in the NLI-PT dataset. POS and CFG were used in [7] with good results. We include also dependencies, not tested before for L2 Portuguese. Grammatical dependencies have been found to be useful for NLI, as they capture a "more abstract representation of syntactic structures" [11,35]. NLI-PT dependencies include POS and lemma. For our experiments, we removed the word form information and we kept only the POS tag. For each of these non lexical features, we experiment with n-grams of different sizes. The maximum size is 4, except for POS, since previous work (and our own results) demonstrates that sequences of order 4 or greater achieve lower accuracy. For feature representation, we normalize the raw counts using TF-IDF weighting.

[5] Unfortunately, NLI-PT does not have a specific test set as other NLI datasets like TOEFL11. For this reason, we use 10-fold cv over the whole corpus for all the experiments. This method allows also for a direct comparison with [7].

4 Results

4.1 Individual Feature Types

We first report the CV results obtained using systems trained on different feature types. Results are presented in terms of accuracy in Table 2. These results are compared against a uniform random baseline of 0.20.

Table 2. Classification results under 10 fold cross-validation (accuracy is reported).

Features	LR	MLP	SVMrbf	SVMlin	Ridge
WLP1-3	0.65	**0.66**	0.63	**0.66**	**0.66**
WLPmod1-3	0.56	**0.58**	0.55	**0.58**	**0.58**
POS1	0.42	0.39	0.43	0.42	0.42
POS2	0.55	0.55	0.54	0.55	0.55
POS3	0.56	0.56	0.55	0.56	0.53
POS1-2	0.56	0.56	0.55	0.56	0.56
POS1-3	0.56	**0.58**	0.56	0.57	0.56
DEP1	0.41	0.41	0.41	0.40	0.41
DEP2	0.43	0.43	0.41	0.42	0.41
DEP3	0.41	0.40	0.36	0.39	0.40
DEP4	0.33	0.34	0.31	0.32	0.32
DEP1-2	**0.44**	**0.44**	0.43	0.43	0.43
DEP1-3	0.43	**0.44**	0.42	0.43	0.43
DEP1-4	**0.44**	**0.44**	0.42	0.43	0.42
CFG1	0.42	0.40	0.41	0.41	0.41
CFG2	0.44	0.43	0.44	0.44	0.45
CFG3	0.48	0.47	0.45	0.47	0.47
CFG4	0.45	0.46	0.44	0.46	0.47
CFG1-2	0.45	0.43	0.44	0.44	0.46
CFG1-3	0.48	0.47	0.47	0.47	0.48
CFG1-4	0.49	**0.50**	0.49	**0.50**	0.48
Random baseline **0.20**					

As expected, the best result is obtained with the lexical representation that includes nouns and adjectives. The number is close to the 0.7 obtained by [7] with a BOW representation over a different subset of NLI-PT, with the same L1s. When nouns and adjectives are excluded from the representation, and only their POS is kept, the accuracy decreases considerably, and it is in fact the same as for the best POS representation.

These results seem to confirm the intuition that the use of nouns and adjectives inflates the results and probably indicates a topic-classification instead

of a L1-classification. In order to investigate this aspect, we performed a test to extract the most correlated unigrams and bigrams per L1 for the two lexical features. For WLP, these n-grams always include proper names connected with the L1 of the text, like *China* or *Macau* for Chinese, *Inglaterra* ('England') for English or *Espanha* ('Spain') and *Madrid* for Spanish. However, in the representation without nouns and adjectives, WLPmod, the most correlated n-grams correspond to verbs or prepositional/verbal phrases which are not topic related. Examples are: *devemos_dever_vmis1p0* ('we should') for Chinese; *è_è_vmip3s0* ('is') (which contains an orthographic mistake) for Italian; *ele_ele_pp3ms00 vai_ir_vmip3s0* ('he goes') for German. Both the low accuracy and the correlated n-grams seem to indicate that excluding nouns and adjectives from a lexical representation reduces topic bias.

For the non lexical features, we can see how accuracy increases as we increase the size of the n-grams. For POS and as previous work has shown, the best representation is 1–3. On the other hand, DEP seems to benefit more of a 1–2 representation. It is interesting to note that the size of the n-grams affects particularly the CFG production rules, which achieve the best results with n-grams of range 1–4, which is not good for POS or DEP.

Concerning the performance of the algorithms, MLP is the algorithm with the best results for the features with the highest accuracy. On average and for the features with the best performance, MLP and SVMlin get the best results, followed by LR and Ridge, and finally by SVMrbf. As a reference, we have ran a test of significance using the results of the models with the best accuracy by type of feature: WLP1-3 with algorithms MLP, SVMlin and Ridge; WLPmod1-3 with the same algorithms; POS1-3 with MLP and SVMlin; DEP1-2 with LR and MLP; CFG1-4 with MLP and SVMlin. Comparing the performance of different algorithms through a test of significance is not a simple task [36], especially if the method used is cross validation, where the samples are not independent. Since the distribution of our data is not normal, and we want to compare more than two samples in some cases, we have chosen the Kruskal-Wallis H-test.[6] The test does not show a significant difference in accuracy for any of the single feature models compared ($p > 0.05$).

4.2 Ensemble Models

Since MLP shows the best results for the most relevant features, we use it as the base classifier for the ensemble experiments. As features, we select the best performing types in the previous experiments: WLPmod, POS1-3, DEP1-2 and CFG1-4. To test the impact of lexical features, we create two types of ensemble models: one including WLPmod (+lex) and another not including this feature (−lex). For classification stacking we use SVMlin as metaclassifier because it has shown good results in a comparative analysis of ensemble methods applied to NLI [20] (Table 3).

[6] We are aware that our data violates one of the assumptions of this test, that is, the independence of the samples.

Table 3. Accuracy of ensemble methods.

Ensemble+lex	Stacking+lex	Ensemble−lex	Stacking−lex
0.64	**0.66**	0.61	**0.61**

As expected, ensemble methods help to improve general accuracy, and stacking gets better results than the simple ensemble (using lexical features). The use of lexical features also helps to increase accuracy. We have applied the Kruskal-Wallis H-test to compare the results including and excluding lexical features. For both ensemble and stacking methods, the difference in accuracy is significant ($p \leq 0.05$). We perform a final test to check if the result of our best system was influenced by overfitting. We split the dataset into train (80%) and test (20%) sets, training and testing the system in different portions of data. The accuracy of the Stacking+lex on the test set was **0.7,** +0.04 points over the result obtained using the whole dataset with 10-fold cv.

Table 4. Stacking systems per-class results: precision, recall and the F1-score are reported.

Class	Precision	Recall	F1-Score
CHIN	0.80/0.78	0.86/0.84	**0.83/0.81**
ENG	0.59/0.54	0.54/0.50	0.56/0.52
GER	0.62/0.59	0.59/0.55	0.61/0.57
ITA	0.65/0.59	0.63/0.57	0.64/0.58
SPA	0.64/0.56	0.68/0.59	**0.66/0.57**
Average	0.66/0.61	0.66/0.61	0.66/0.61

In Table 4 we present the results obtained for each L1 in terms of precision, recall, and F1 score as well as the average results on the five classes. Each column shows the results of the two stacking systems, corresponding the first result to stacking+lex and the second to stacking−lex. Looking at individual classes, the results obtained for Chinese are clearly higher than those of other L1s, even when the number of texts is smaller than for Spanish and Italian. The same tendency was observed in [7], and it seems to illustrate the intuition that linguistic distance is directly related to level of performance per class in NLI. This idea is also supported by the confusion matrix in Table 5, which shows, for example, that Spanish and Italian, the two Romance languages, tend to be confused more frequently. On the opposite side, English is the L1 with the lowest accuracy. Again, [7] showed the same pattern. English is the class with less texts in our dataset, but the difference with German (the other Germanic language in the dataset), with +21 texts, does not seem big enough to justify the difference in performance (−0.05 points). If we take a look at the confusion

matrix in Table 5 we can see that English texts are confused with all the other L1s in a similar proportion, even with Chinese. One linguistic hypothesis for this behaviour could be that English is a Germanic language (then closer to German) with a high percentage of Latin vocabulary (then close to Spanish and Italian) and with an isolating morphology (then close to Chinese). German, on the contrary, is not so close to the Romance languages in vocabulary, and does have a rich morphological system.

Table 5. Confusion matrix for the Stacking+lex model.

		Predicted class				
		CHI	ENG	GER	ITA	SPA
Actual class	CHI	379	25	13	10	13
	ENG	48	210	53	31	67
	GER	19	53	253	58	47
	ITA	13	25	56	358	103
	SPA	13	50	33	94	417

Table 4 shows also that lexical features have a positive impact for all the L1s, being especially relevant for L1s that are more similar to Portuguese, Spanish and Italian. For Chinese, however, lexical features only increased F1 score in 0.2 points. This fact seems to indicate that, for L1s that are close to the target at all the levels (lexical, morphological and syntactic), the inclusion of lexical information makes a difference to improve accuracy. On the other hand, for L1s that are lexically unrelated and very distant morphosyntactically, the use of morphosyntactic information is enough to get good results.

5 Conclusion and Future Work

This paper presented new experiments on NLI for L2 Portuguese. Our results improve the best result previously obtained in [7], considering both general accuracy and results by class. The presented results are comparable to those of other NLI studies [2], but not as high as those on the largest and most balanced corpora [18]. This is likely a limitation of our data, mainly caused by topic distribution.

We proposed a linguistically motivated method to make use of lexical features while reducing topic bias. This method helped to increase the classification performance for all L1s, especially for those which are more similar to Portuguese. We tested different algorithms and features, defining the most effective combination for our dataset. We also found that n-gram size particularly affects the performance of the CFG production rules. We experimented for the first time with L2 Portuguese with dependencies and an ensemble method that uses stacking classification, obtaining the best results in our experiments.

This study opens several avenues for future research. One of them is investigating the influence of L1 in Portuguese second language acquisition. Such approaches, similar to those applied to English learner data [13], can have direct pedagogical implications. Particularly, we would like to investigate in more detail the impact of the different types of linguistic features in the classification task taking into account the linguistic distance between L1 and L2 Portuguese. We also would like to analyse the possible influence of L3 languages in the task.

Another important step will be the refinement and extension of our dataset, especially in terms of topic distribution by L1.

Acknowledgements. This work was supported by Fundacão para a Ciência e a Tecnologia (postdoctoral research grant SFRH/BPD/109914/2015). We would like to thank the anonymous reviewers for the suggestions and constructive feedback provided.

References

1. Malmasi, S.: Native language identification: explorations and applications. Ph.D. thesis (2016)
2. Malmasi, S., Dras, M.: Multilingual native language identification. Nat. Lang. Eng. **23**(2), 163–215 (2015)
3. Malmasi, S., Dras, M.: Chinese native language identification. In: Proceedings of EACL, Gothenburg, Sweden. Association for Computational Linguistics (2014)
4. Malmasi, S., Dras, M., Temnikova, I.: Norwegian native language identification. In: Proceedings of RANLP, Hissar, Bulgaria, pp. 404–412, September 2015
5. Malmasi, S., Dras, M.: Arabic native language identification. In: Proceedings of the Arabic Natural Language Processing Workshop (2014)
6. del Río, I., Zampieri, M., Malmasi, S.: A Portuguese native language identification dataset. In: Proceedings of BEA (2018)
7. Malmasi, S., del Río, I., Zampieri, M.: Portuguese native language identification. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 115–124. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_12
8. Wong, S.M.J., Dras, M.: Contrastive analysis and native language identification. In: Proceedings of ALTA, Sydney, Australia, pp. 53–61, December 2009
9. Wong, S.M.J., Dras, M.: Exploiting parse structures for native language identification. In: Proceedings of EMNLP (2011)
10. Swanson, B., Charniak, E.: Native language detection with tree substitution grammars. In: Proceedings of ACL, Jeju Island, Korea, pp. 193–197, July 2012
11. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: resources and empirical evaluations in native language identification. In: Proceedings of COLING, Mumbai, India, pp. 2585–2602 (2012)
12. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskes, T.: Improving native language identification with TF-IDF weighting. In: Proceedings of BEA (2013)
13. Malmasi, S., Dras, M.: Language transfer hypotheses with linear SVM weights. In: Proceedings of EMNLP, pp. 1385–1390 (2014)
14. Cimino, A., Dell'Orletta, F., Brunato, D., Venturi, G.: Sentences and documents in native language identification. In: CLiC-it (2018)
15. Malmasi, S., Tetreault, J., Dras, M.: Oracle and human baselines for native language identification. In: Proceedings of BEA (2015)

16. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., Chodorow, M.: TOEFL11: a corpus of non-native English. Technical report, Educational Testing Service (2013)
17. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of BEA (2013)
18. Malmasi, S., et al.: A report on the 2017 native language identification shared task. In: Proceedings of BEA (2017)
19. Malmasi, S., Dras, M.: Native language identification using stacked generalization. arXiv preprint arXiv:1703.06541 (2017)
20. Malmasi, S., Dras, M.: Native language identification with classifier stacking and ensembles. Comput. Linguist. **44**(3), 403–446 (2018)
21. Tsur, O., Rappoport, A.: Using classifier features for studying the effect of native language on the choice of written second language words. In: Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (2007)
22. Malmasi, S., Wong, S.M.J., Dras, M.: NLI shared task 2013: MQ submission. In: Proceedings of BEA (2013)
23. Wong, S.M.J., Dras, M., Johnson, M.: Exploring adaptor grammars for native language identification. In: Proceedings of EMNLP (2012)
24. Swanson, B., Charniak, E.: Data driven language transfer hypotheses. In: EACL 2014, p. 169 (2014)
25. Granger, S., Dagneaux, E., Meunier, F., Paquot, M.: International Corpus of Learner English (Version 2). Presses Universitaires de Louvain, Louvian-la-Neuve (2009)
26. Brooke, J., Hirst, G.: Measuring interlanguage: native language identification with L1-influence metrics. In: Proceedings of LREC (2012)
27. Malmasi, S., Dras, M.: Finnish native language identification. In: Proceedings of ALTA, Melbourne, Australia, pp. 139–144 (2014)
28. Wang, M., Malmasi, S., Huang, M.: The Jinan Chinese learner corpus. In: Proceedings of BEA (2015)
29. Tenfjord, K., Meurer, P., Hofland, K.: The ASK corpus: a language learner corpus of Norwegian as a second language. In: Proceedings of LREC (2006)
30. del Río, I.: Automatic proficiency classification in L2 Portuguese. Procesamiento del Lenguaje Natural **63**, 67–74 (2019)
31. Genkin, A., Lewis, D.D., Madigan, D.: Large-scale Bayesian logistic regression for text categorization. Technometrics **49**, 291–304 (2007)
32. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026683
33. Zhang, T., Oles, F.J.: Text categorization based on regularized linear classification methods. Inf. Retrieval **4**(1), 5–31 (2001)
34. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, vol. 14, pp. 1137–1145 (1995)
35. Bykh, S., Meurers, D.: Native language identification using recurring n-grams – investigating abstraction and domain dependence. In: Proceedings of COLING 2012, Mumbai, India, pp. 425–440. The COLING 2012 Organizing Committee, December 2012
36. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. **10**(7), 1895–1923 (1998)