A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors

Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, Manel Zarrouk

A supervised learning model for the automatic assessment of language levels*

Nicolas Ballier[1][0000−0003−2179−1043], Thomas Gaillat[2], Andrew Simpkin[3], Bernardo Stearns[3], Manon Bouyé[2], and Manel Zarrouk[3]

[1] Université de Paris-Diderot, CLILLAC-ARP, F-75013 Paris, France
[2] University of Rennes LIDILE, France
[3] Insight Centre for Data analytics, NUI Galway, Ireland
thomas.gaillat@univ-rennes1.fr, nicolas.ballier@univ-paris-diderot.fr, andrew.simpkin@insight-centre.org, bernardo.stearns@insight-centre.org, mbouye@eila.univ-paris-diderot.fr, manel.zarrouk@insight-centre.org

Abstract. This paper focuses on the use of technology in language learning. Language training requires the need to group learners homogeneously and to provide them with instant feedback on their productions such as errors [8, 15, 17] or proficiency levels. A possible approach is to assess writings from students and assign them with a level. This paper analyses the possibility of automatically predicting Common European Framework of Reference (CEFR) language levels on the basis of manually annotated errors in a written learner corpus [9, 11]. The research question is to evaluate the predictive power of errors in terms of levels and to identify which error types appear to be criterial features in determining interlanguage stages. Results show that specific errors such as punctuation, spelling and verb tense are significant at specific CEFR levels.

Keywords: CEFR level prediction · error tagset · regression · unsupervised clustering · proficiency levels.

1 Introduction

This paper focuses on the use of technology in language learning. For individuals, learning a language requires regular assessments for both learners and teachers to focus on specific areas to train upon. For institutions, there is a growing demand to group learners homogeneously in order to set adequate teaching objectives and methods. These two requirements rely on language assessment tests whose design and organization are labour-intensive and thus costly. Currently, language learning centres rely on instructors to design and manually correct tests. Alternatively, they use specifically designed short-context and rule-based online exercises in which a set of specific language errors are used as a paradigm

* This paper benefited from the support of the Partenariat Hubert Currien Ulysse 2019 funding for the project "Investigating criterial features of learner English and AI-driven automatic language level assessment" (ref 43121RJ).

for scoring. Both approaches retain certain error types over others, which may introduce a bias regarding the importance given to these errors. Even though it may be argued that the linguistic complexity of a student's essay and its quality rely on more than some errors, errors as a whole play a role in language assessment by experts. This raises the question of their importance in the overall process.

The literature on Automatic Scoring Systems applied to learner language shows that a comprehensive set of criterial features is necessary to obtain accuracy [7]. Many studies have focused on the use of various types of linguistic features such as syntactic and lexical complexity as well as word frequencies and lexicons [12]. In parallel, much effort has been invested in error-detection systems which also rely on linguistic features [15]. However, little work has been done to understand the role of errors in the assessment of levels by expert readers. Yet, such understanding could inform their potential use as features. Combining criterial features to CEFR levels could also inform on specific errors related to specific levels, hence unraveling aspects of Interlanguage [20].

Our research question is to investigate the predictive power of errors in terms of levels and to identify which error types appear to be criterial features in determining proficiency levels. To do so, a possible approach is to use error annotated corpora [9, 11] in which student writings are annotated in terms of proficiency level. By applying mathematical methods, it is possible to isolate significant error types in selecting proficiency levels. This paper analyses the possibility of automatically predicting Common European Framework of Reference (CEFR) language levels [5] on the basis of manually annotated errors in the EFCAMDAT [10] written learner corpus.

The paper is organised as follows: In Section 2, the literature related to automatic level assessment and language scoring is briefly discussed. In Section 3, we describe the data and the error tagset adopted for the EFCAMDAT corpus[4]. Section 4 reports on the prediction of the CEFR levels using regression analysis and clustering based on errors found for each level. In Section 5, we conclude on the possibility of automatically detecting errors that could be used as criterial features for a given CEFR level.

2 Automatic Essay Scoring Systems and second language learning

Automatic Scoring Systems (ASS), and more specifically Automatic Essay Scoring (AES) systems for open-ended questions, have been developed to automate student essay assessments. Early on, ASS focused on native English and applied probabilistic methods in which specific textual features were used in regression models. Page's PEG-IA system [18] included 30 features in a multiple-regression

[4] The EFCAMDATA is hosted by the University of Cambridge and data is accessible for academic and non-commercial purposes. Our scripts will be available on our github.

approach. With the recent advent of supervised learning methods, probabilistic models have become more complex in terms of features and thus more powerful. They also provide the benefit of consistency compared with human scorers.

More recently, AES systems have focused on learner language data [2, 21, 26], which has raised the need to use learner corpora to train models [3, 13]. Two shared-tasks organised in conferences have made use of learner corpora for the purpose of scoring. The two editions of the Spoken CALL shared Task [4] focused on the distinction between linguistically correct and incorrect short open-ended constructs in Swiss German learners' speech. Language level assessment, which can be seen as a sub-part of research on scoring, was the focus of the CAP18 conference. The conference included a shared task [1] on predicting CEFR levels. The distributed dataset was sourced from texts written by French L1 English learners and classified according to CEFR levels. Features were provided in the form of lexical and syntactic complexity and readability metrics. Specific studies have been conducted on automatic level assessment in learner English [2, 28] but also in other languages such as Estonian [24] and Swedish [25]. All papers report on different methods that use n-grams, errors, syntactic and lexical features to rank learner texts. They may focus on scoring specific language aspects such as text coherence or global proficiency levels of learners. Some of these approaches are deployed in commercial products[5].

Errors have been used as features in some learner-language AES. Nevertheless, their impact on proficiency levels has not received much attention. [28] reports on the classification of English as a Second or Other Language (ESOL) texts. Error rates are used as one type of features. Rates are computed automatically on the basis of syntactic patterns. The metric was found to improve correlation measures between predicted and annotated scores. [16] used spelling errors in a simple regression model but the feature significance was not evaluated. [23] implemented error features in a classification model. The set of error features included spelling and grammar errors which were automatically detected using the spelling and grammar check LanguageTool[6]. Results showed that the error features did not perform well (51% classification) when taken independently of the other features. [6] reports on an regression analysis linking various linguistic features to TOEFL-essay scores. They approached the issue of errors by comparing essays which were scored high by an AES and low by human raters, and vice versa. They observed that the AES misinterpreted spelling and syntactic complexity errors as positive features for predictions. Conversely, syntactic accuracy was not taken into account by the system, revealing the need to operationalise such features. Their study highlights the need to investigate the use of error features on a larger dataset including more error types.

Our contribution is to extend on [6] by using a larger dataset made up of 24 different error types extracted from Cambridge's EFCAMDAT corpus [10]. It also uses categorical levels of the CEFR as the outcome variable in learner

[5] For instance, see the IntelligentEssayAssessor developed at Pearson Knowledge Technologies; the IntelliMetricEssayScoringSystem developed by Vantage Learning

[6] http://languagetool.org

English. The classification task allows to quantify the effect and the significance of each error-type in the model. It also gives an insight in the error tagset used to annotate the essays.

3 Data and error sets

In this section, we present the EFCAMDAT corpus and the error codes used to annotate it.

3.1 Corpus description

The data used in this study are the French and Spanish L1 subsets of the EFCAMDAT corpus, an 83 million word learner corpus collected by Cambridge University [10]. The two subsets include writing essays of different Englishtown[7] levels ranging from 1 to 16, which were then mapped onto the six CEFR levels using the equivalence grid provided in [10]. A total of 49,813 annotated texts from 8,851 French and Spanish learners were downloaded from the EFCAMDAT database. Close analysis revealed that only 34,308 texts actually included errors, and there were 15,505 texts without error annotation. Those without errors were removed prior to modelling.

The EFCAMDAT corpus was processed and is freely available as an XML-format dataset containing text IDs, speakers' L1s and levels. It was also manually annotated for errors by [27], using an ad-hoc tagset which we describe in the following subsection.

3.2 The Cambridge tagset of errors

The Cambridge tagset consists of 24 types of errors, detailed in 1. As to September 2017, 66% of the whole EFCAMDAT corpus had been tagged by teachers using these codes [27].

Five tags in the tagset are linked to mechanic errors: they include punctuation, inappropriate or missing spaces, capitalization issues and spelling. Characteristic examples of spelling and typographic errors are illustrated in the examples below (respectively extracted from A1, B2 and C1 productions).

Example 1. I'm cleaning the living room and the kitcheen.

Example 2. Moreover that, they suscribe for you a full accident insurance and every year, you benefites of one month holiday every year.

Of particular interest are the tags used to label morphosyntactic errors, in particular Verb Tense (VT, see Example 3 below) or Plural (PL) and Singular (SI).

Example 3. She was recently catch by paparazzis drinking and smoking.

[7] See https://englishlive.ef.com

Table 1: EFCAMDAT error tagset

Code	Meaning	Code	Meaning
XC	change from x to y	NSW	no such word
AG	agreement	PH	phraseology
AR	article	PL	plural
AS	add space	PO	possessive
CO	combine sentences	PR	preposition
C	capitalization	PS	part of speech
D	delete	RS	remove space
EX	expression of idiom	SI	singular
HL	highlight	SP	spelling
IS	insert	VT	verb tense
MW	missing word	WC	word choice
NS	new sentence	WO	word order

Other tags include error categories which pertain to syntax (Missing word, Word order), information packaging (Combine sentences) and lexical or collocation errors (e.g. Expression of idiom and Phraseology). As stated by the authors, "the purpose of these corrections was to provide feedback to learners and as such it cannot be viewed as error annotation based on a specific annotation scheme developed specifically for annotating learner corpora" [27]. This raises a number of issues concerning the error codes used on the EFCAMDAT corpus. First, as inter-rater agreement was not a concern, errors were only hand-coded once by different annotators, which may explain why similar error types are sometimes coded differently, as illustrated in the following examples:

Example 4. This movement prepare the ways to the Abstract Art.

Example 5. The other have to hide. (...) When the person stopped counting, he try to find the others.

If *prepare* is coded as a subject-verb Agreement error in Example 4, *have* is coded as a Word Choice error in Example 5, while *try* has no annotation at all. Similarly, some errors which are coded as morphosyntactic violations in some essays are tagged as spelling mistakes or collocation errors in others. This is related to the second main problem arising from the tagset: the ambiguity and possible overlap between categories. While some tags are precise in their scope, like Preposition, Article, Plural, Singular and Spelling, which bear on specific part of speech or individual words, other broader categories seem to overlap with others. As no theoretical discussion backs up the different tag labels, the difference between some of them seems tenuous, as illustrated by the example below.

Example 6. I hope to see you again soon, maybe can we lunch together the next week?

The annotation file shows that the verb *lunch* is tagged as a *Word choice* error. Several codes from the tagset could have been equally appropriate here: Expression of idiom, or Phraseology (two categories which themselves appear to be very similar), since the error seems to stem from a lack of awareness of the collocation *have lunch*, which is expressed by a verb-noun collocation in English but by a single verb in both French and Spanish. The category *Insert* could thus also have been used. This example reveals that several types of categories can fit one type of error, and vice versa. The tag *Word Choice* (WC), in particular, is such a versatile, overarching category that it can either be substituted by more precise categories, as we have just seen, when in relation to collocational errors, or by morphosyntactic categories, as shown below.

Example 7. The other have to hide.

Here the subject-verb agreement error, which is a morphosyntactic violation, is tagged as Word choice and not Agreement, which demonstrates a difference in scope across the same tag (WC). This is also the case for the Spelling category, as we will now see.

Example 8. Timotie, the next door neighbor to Serena and Dave, he told us that Dave was an inestable man.

Example 9. If there are moving, he losed.

It could be argued here that *inestable* and losed could both be tagged as No such word (NSW), the first being so distorted that it hardly resembles its correct version *unstable* and the second constituting an unacceptable and ungrammatical preterit form of lose. They are, however, both tagged as spelling mistakes (SP), although they do not encompass exactly the same type of error. This is again due to the overarching scope of some error categories.

The ambiguities and inconsistencies of the error feedback, which was not, strictly speaking, designed as an annotation tagset, have to be kept in mind when processing the results further. These are, however, isolated examples which are by no means the result of a systematic assessment of the error tags. Our next section investigates the possibility to use these annotated errors as predictors for the CEFR levels.

4 Using the EFCAMDAT annotated errors as predictors for CEFR levels

4.1 Experimental design and model building

The aim of this study was to construct a classification model of learner levels (A1, A2, B1, B2, C1, C2), based on a corpus submitted by the learners. In order to test the efficacy of the error variables, we built a classification model using 24 error types. We report on the precision, recall, accuracy and F1-score of each model. To find the optimal classifier, we compared multinomial logistic

regression, random forests, linear discriminant analysis, k-nearest neighbours, Gaussian naive Bayes, support vector machine and decision tree classifier.

A second analysis used logistic regression to investigate the relative importance of the 24 error types across learner level. We split the data based on learner levels (A, B and C) and ran separate logistic regressions on these data using only the error variables. We report on the strongest positive and negative associated errors in terms of their Wald test statistic or z-score for each level, i.e. A2 v A1, B2 v B1 and C2 v C1. A positive association suggests that the error is more common in advanced learners, whereas a negative association suggests that the error is less common in advanced learners. A z-score comprised in the [-2;2] interval indicates non significant variables (p-value > 0.05). We report on the odds ratios of the errors to explore how much the occurrence of an error increases the odds of being an advanced learner.

We split the data into 75% training and 25% test data, resulting in 17,154 learners in the testing data. Among the seven model types tested here, the optimal classification performance in the testing dataset was found using a random forest model.

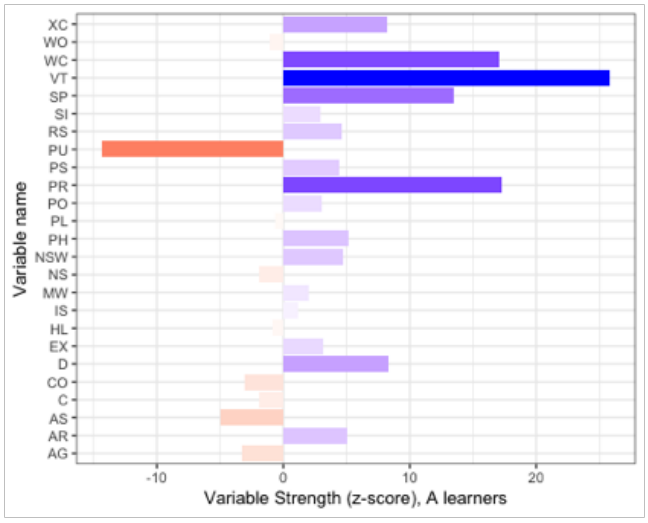
4.2 Results and discussion

Using the error variables, the classifier achieved 70% accuracy with results in full given in Tables 2 and 3. Classification performance using error variables shows that errors are a good predictor of CEFR levels given by human raters as they seem to account for 70% of the variance in their judgments. Results show that accuracy drops with higher levels of proficiency (C1 & C2). Nevertheless, precision shows that predictions are consistent as few essays classified as C2 are actually of another level.

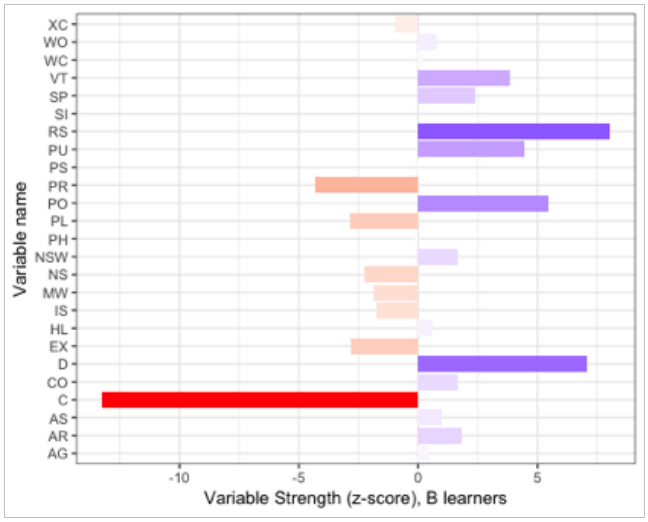
Table 2: Confusion matrix from the testing dataset using error variables

	Predicted					
Real	A1	A2	B1	B2	C1	C2
A1	5486	1227	878	467	111	11
A2	572	2918	383	211	33	4
B1	317	324	2398	177	46	3
B2	106	102	110	988	12	3
C1	10	13	15	8	196	0
C2	3	0	0	0	2	20

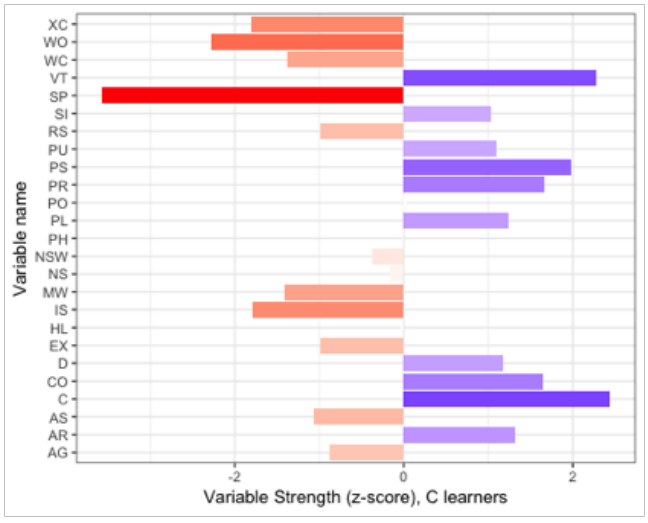
For level-A learners, the strongest variables are shown in Figure 1a. Verb Tense (VT) was the strongest positively associated variable. For every unit increase in VT there was a 80% increased odds of being an A2 learner (odds ratio 1.8, 95% CI 1.72 to 1.88). On the other hand, Punctuation (PU) was the strongest negative variable, with lower values more likely in A2 than A1 leaner



(a) Variable Importance for Level-A Learners



(b) Variable Importance for Level-B Learners



(c) Variable Importance for Level-C Learners

Fig. 1: Variable importance per CEFR level

Table 3: Classification performance on the testing dataset using error variables

Level	Precision	Recall	F1	Support
A1	0.67	0.84	0.75	6494
A2	0.71	0.64	0.67	4584
B1	0.73	0.63	0.68	3784
B2	0.75	0.53	0.62	1851
C1	0.81	0.49	0.61	400
C2	0.80	0.49	0.61	41
Mean	0.71	0.70	0.70	17154

on average. For every unit increase in PU there was a 11% decreased odds of being an A2 learner (odds ratio 0.89, 95% CI 0.88 to 0.91). In other terms, verb tense errors tend to predict A2 essays whilst punctuation errors tend to predict A1 essays.

For level-B learners, the strongest variables are shown in Figure 1b. Remove Space (RS) was the strongest positively associated variable. For every unit increase in RS there was a 6% increased odds of being a B2 learner (odds ratio 1.06, 95% CI 1.05 to 1.08). On the other hand, Capitalization (C) was the strongest negative variable, with lower values more likely in B2 than B1 essays on average. For every unit increase in C there was a 13% decreased odds of being a B2 learner (odds ratio 0.87, 95% CI 0.85 to 0.89). In short, errors on spaces between words seem to point towards B2 whilst errors on capitalization tend to suggest B1 writings.

For level-C learners, the strongest variables are shown in Figure 1c. Capitalization (C) was the strongest positively associated variable. For every unit increase in C there was a 13% increased odds of being a C2 learner (odds ratio 1.13, 95% CI 1.02 to 1.24). On the other hand, Spelling errors (SP) was the strongest negative error variable, with lower values more likely in C2 than C1 learners on average. For every unit increase in SP there was a 14% decreased odds of being an C2 learner (odds ratio 0.86, 95% CI 0.79 to 0.93). In a nutshell errors on capitalization lead to C1 whilst errors on spelling point to C2.

To summarize our regression analysis, the 24 error variables achieved 70% accuracy for classification of A1 - C2 learners. The approach also focused on the relative importance of error types across levels. The experimental setup operionalises Interlanguage stages in terms of CEFR levels. It allows the exploration of correlations between error types and specific levels. The analysis reveals that mechanic errors (see Section 3.2) are significant across all levels. Only sub-types correlate with specific levels. The results also show that some syntax-error types only correlate with the A level (Word Choice and Word Order). Conversely, the syntax error linked to Verb Tense is significant in the three models. This indicates that learners of all levels experience difficulties on this issue but the category does not distinguish tenses. It may be that learners face problems with different tense choices or constructions. In short, fine-grained tags appear to tie closely with levels while coarser grained categories do not.

Classifying C2 learners was difficult since very few C2 learners were available in the dataset. If data from more advanced learners were available, model accuracy would be improved, especially where features are calculated. We then tried another method to assess the possibility of predicting a CEFR level on the basis of clusters of error tags, in other words to predict CEFR levels on the basis of error clusters.

4.3 Using unsupervised Clustering of errors

To analyse the similarities in errors across texts, we used multivariate clustering to find an optimal number of groups of texts. We used model-based clustering through the mclust package in R v3.4 [19]. This clustering is unsupervised, i.e. learner level is unknown to the model. To investigate how well the errors cluster by level, we present the confusion matrix of learner level against group membership according to the model.

Table 4: Confusion matrix of cluster membership against learner level

	A1	A2	B1	B2	C1	C2
1	744	512	662	344	70	10
2	842	1020	1038	610	120	12
3	2998	2690	1868	882	138	12
4	17660	11262	8196	3798	788	66
5	1772	1280	1334	630	134	10
6	1332	1306	1302	644	146	10
7	492	526	790	364	190	12

The model is computed with the error-annotated texts (see Section 3.1). The optimal model found seven clusters in these data. Table 4 shows that that these clusters do not match the identified learner level, with no clear cross classification apparent. Table 5 shows a breakdown of the proportion of learners in the whole data compared to those who had any errors. Surprisingly, the main discrepancy is in A1 learners who make up 41% of the overall cohort, but are less well represented in those who made an error. This suggests they are less likely to make errors in their text. This may be explained by the fact that learners of level A were given prompts and examples prior to writing, hence facilitating their endeavours so much so that few errors, if any, were identified.

The 24 error variables achieve 70% accuracy for classification of A1 - C2 learners. Classifying C2 learners was difficult since very few C2 learners were available in the dataset. If data from more advanced learners were available, model accuracy would be improved. Unsupervised clustering of the multivariate error data does not map well to the learner levels, which bodes badly on the relevance of using error annotation for level prediction. Caution should be exerted, though, as some specific error-types have been found to be associated with

Table 5: Proportion of learner levels in the entire data compared with those in which errors were found

Level	All data	With errors
A1	0.41	0.38
A2	0.27	0.27
B2	0.20	0.22
B1	0.09	0.11
C1	0.02	0.02
C2	0.00	0.00

specific levels. This may be explained by the fact that the error tagset was not employed for level assignment by human raters but rather to provide feedback to the learners.

5 Conclusion and future research

In this paper, we have presented a predictive model for the prediction of CEFR levels in learner-English essays. The purpose was to test the predictive power of error types as features in a supervised learning approach. Even though errors appear to predict levels with significant accuracy, the clustering approach showed that not all errors help in the predictions. In other terms, only some error types defined in the tagset contribute to level assignment.

The experiment also shows that the tagset employed in error annotation must be carefully defined in terms of categories to avoid overlaps and to include error types which belong to the same dimension. For instance the capitalisation variable is significant but it is not comparable in nature with *Missing Word* errors. Some errors are indicative of Interlanguage stages whereas other reveal typos or spelling issues. This method could be applied on other error annotated corpora such as the NUCLE used in [17]. Other such tagsets may yield more consistency in terms of tags, which would support better classification. Another strategy might rely on making tagsets interoperable in order to apply a new tagset to an already annotated corpus prior to classification of the same texts.

Our next step is to build a fully automated prediction system for new texts. Hence the challenge is to have a workflow based on automatic detection of features, including errors. The present study highlights some error types which could be detected automatically. For instance Spelling errors appear to be an error type to consider for the implementation of an automatic detection heuristic. Lexicons could be used to exclude non-English words. Similarly, morphosyntactic error types may be identified by using POS patterns. [14] reports the robustness of parsers when analysing learner data, and that dependency parsing is more sensitive to errors than PoS-tagging. Conversely, error types such as verb tense remain challenging in terms of implementation due to the semantic value of contexts.

Transforming learning with meaningful technologies addresses how emerging and future learning technologies can be used in a meaningful way to enhance human-machine interrelations, to contribute to efficient and effective education, and to assess the added value of such technologies. AES applied to learner data can be a part of ICALL (Intelligent Computer-Assisted Language Learning) systems characterized by rich formative feedback [22]. Indicating level along with specific and goal-oriented feedback to learners would provide a strong incentive to motivation and learning performance.

Bibliography

[1] Arnold, T., Ballier, N., Gaillat, T., Lissn, P.: Predicting CEFRL levels in learner English on the basis of metrics and full texts. arXiv:1806.11099 [cs] (2018)

[2] Attali, Y., Burstein, J.: Automated Essay Scoring With e-rater V.2. The Journal of Technology, Learning and Assessment **4**(3) (2006)

[3] Barker, F., Salamoura, A., Saville, N.: Learner corpora and language testing. In: Granger, S., Gilquin, G., Meunier, F. (eds.) The Cambridge Handbook of Learner Corpus Research, pp. 511–534. Cambridge Handbooks in Language and Linguistics, Cambridge University Press (2015)

[4] Baur, C., Caines, A., Chua, C., Gerlach, J., Qian, M., Rayner, M., Russell, M., Strik, H., Wei, X.: Overview of the 2018 Spoken CALL Shared Task. In: Interspeech 2018. pp. 2354–2358. ISCA (2018)

[5] Council of Europe, Council for Cultural Co-operation. Education Committee. Modern Languages Division: Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press, Cambridge (2001)

[6] Crossley, S.A., Kyle, K., Allen, L.K., Guo, L., McNamara, D.S.: Linguistic Microfeatures to Predict L2 Writing Proficiency: A Case Study in Automated Writing Evaluation. (2014)

[7] Crossley, S.A., Salsbury, T., McNamara, D.S., Jarvis, S.: Predicting lexical proficiency in language learner texts using computational indices. Language Testing **28**(4), 561–580 (2011)

[8] Dale, R., Anisimoff, I., Narroway, G.: HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. pp. 54–62. NAACL HLT '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), event-place: Montreal, Canada

[9] Díaz-Negrillo, A., Fernandez-Domingez, J.: Error Tagging Systems for Learner Corpora. Spanish Journal of Applied Linguistics (RESLA) (19), 83–102 (2006)

[10] Geertzen, J., Alexopoulou, T., Korhonen, A.: Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In: Miller, R.T., Martin, K.I., Eddington, C.M., Henery, A., Miguel, N., Tseng, A., Tuninetti, A., Walter, D. (eds.) Proceeedings of the 31st Second Language Research Forum. Cascadilla Press, Carnegie Mellon (2013)

[11] Granger, S., Gilquin, G., Meunier, F. (eds.): The Cambridge Handbook of Learner Corpus Research. Cambridge University Press, Cambridge (2015)

[12] Hawkins, J.A., Buttery, P.: Criterial Features in Learner Corpora: Theory and Illustrations. English Profile Journal **1**(01) (2010)

[13] Higgins, D., Xi, X., Zechner, K., Williamson, D.: A Three-stage Approach to the Automated Scoring of Spontaneous Spoken Responses. Comput. Speech Lang. **25**(2), 282–306 (2011)
[14] Huang, Y., Murakami, A., Alexopoulou, T., Korhonen, A.L.: Dependency parsing of learner English (2018)
[15] Leacock, C.: Automated Grammatical Error Detection for Language Learners. Morgan & Claypool Publishers, California (2010)
[16] Nedungadi, P., Raj, H.: Unsupervised Word Sense Disambiguation for Automatic Essay Scoring. In: Kumar Kundu, M., Mohapatra, D.P., Konar, A., Chakraborty, A. (eds.) Advanced Computing, Networking and Informatics-Volume 1. pp. 437–443. Smart Innovation, Systems and Technologies, Springer International Publishing (2014)
[17] Ng, H.T., Wu, S.M., Briscoe, T., Hadiwinoto, C., Susanto, R.H., Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. pp. 1–14. Association for Computational Linguistics (2014), event-place: Baltimore, Maryland
[18] Page, E.B.: The use of the computer in analyzing student essays. International review of education **14**(2), 210–225 (1968)
[19] Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 205–233 (2016)
[20] Selinker, L.: Interlanguage. International Review of Applied Linguistics in Language Teaching **10**(3), 209 (1972)
[21] Shermis, M.D., Burstein, J., Higgins, D., Zechner, K.: Automated essay scoring: Writing assessment and instruction. International encyclopedia of education **4**(1), 20–26 (2010)
[22] Shute, V.J.: Focus on formative feedback. Review of Educational Research **78**(1), 153–189 (2008)
[23] Vajjala, S.: Automated assessment of non-native learner essays: Investigating the role of linguistic features. International Journal of Artificial Intelligence in Education (2017), arXiv: 1612.00729
[24] Vajjala, S., Loo, K.: Automatic CEFR Level Prediction for Estonian Learner Text. In: NEALT Proceedings Series. vol. 22, pp. 113–128 (2014)
[25] Volodina, E., Piln, I., Alfter, D.: CALL communities and culture short papers from EUROCALL 2016. In: Salomi Papadima-Sophocleous, Linda Bradley, Sylvie Thousny (eds.) Classification of Swedish learner essays by CEFR levels, pp. 456–461. Research-publishing.net (2016)
[26] Weigle, S.C.: English language learners and automated scoring of essays: Critical considerations. Assessing Writing **18**(1), 85–99 (2013)
[27] Yan, H., Jeroen, G., Rachel, B., Anna, K., Theodora, A.: The EF Cambridge Open Language Database (EFCAMDAT) information for users (2017)
[28] Yannakoudakis, H., Briscoe, T., Medlock, B.: 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 - Volume 1. pp. 180–189. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)