Visualizing Linguistic Complexity and Proficiency in Learner English Writings

Thomas Gaillat, Antoine Lafontaine, Anas Knefati

Visualizing Linguistic Complexity and Proficiency in Learner English Writings
T Gaillat, A Lafontaine, A Knefati
CALICO Journal 40 (2), 178-197, 2023

Visualising linguistic complexity and proficiency in learner English writings

Abstract

In this paper, we focus on the design of a second language (L2) formative feedback system that provides linguistic complexity graph reports in the writings of English for Special Purposes students at the university level. The system is evaluated in light of formative instruction features presented in Shute (2008). Significance of complexity metrics is also evaluated. A CEFR-classified learner corpus of English was processed with a pipeline that computes 83 complexity metrics. By way of ANOVA, multinomial logistic regression and clustering methods, we identified and validated a set of 9 significant metrics in terms of proficiency levels. Validation with classification gave 67.51% (A level), 60.16% (B level) and 60.47% (C level) balanced accuracy. Clustering showed between 53.10% and 67.37% homogeneity depending on the level. As a result, these metrics were used to create graphical reports about the linguistic complexity of a learner writing. These reports are designed to help language teachers diagnose their students' writings in comparison with pre-recorded cohorts of different proficiency.

Keywords: Linguistic complexity; L2 English; automatic essay feedback; visualization

1. Introduction

Language writing activities are an essential part of learning tasks for second language learners of English at the university level. By repeated attempts to write short descriptive or opinion texts in class or at home, learners try various strategies to develop their discourse. In doing so their

productions may deviate from nativelike productions. It is at this point that feedback from teachers plays an important role. To be formative, the feedback should be specific and explicit.

Historically, quality formative feedback has relied on teachers who use their expertise in order to tailor their messages to their students. However, in resource-limited settings this reliance has led to increased correction time for teachers. To outbalance the amount of corrections due to the large number of students, teachers are likely to resort to giving fewer assignments to their students or assignments that are scaffolded differently from 'typical' composition practice. In addition, and due to the institutional pressure in providing grades, teachers tend to focus on summative assignments. As important as grades are, this pressure shifts the focus away from formative assessments or other assessments that deemphasize summative evaluation such as project/portfolio-based approaches. Yet, formative assessments are essential. They enable learners to produce the metalinguistic reasoning that is necessary for understanding forms and functions according to contexts. It is therefore important to find solutions to help teachers give more formative assessments to their students, which means giving them tools to process their learners' productions swiftly.

As argued by Meurers (2009), the automatic analysis of learner language is a possible response to this problem. Natural language processing tools can be used to consistently collect and evaluate learner language properties in order to design feedback messages. In L2, many tools focus on feedback about linguistic structures, such as grammatical and spelling errors (Leacock et al., 2015). When Corrective Feedback (CF) focuses on explicit metalinguistic information, it favours generalisation. In other terms, CF "contributes to system as well as item learning" (Ellis et al., 2006). However, this approach is partial as it only takes accuracy into consideration in terms of assessment. Yet, learner language proficiency is also composed of other dimensions

including complexity and fluency (Housen et al., 2012; Wolfe-Quintero et al., 1998). In addition, errors represent negative properties of L2 production. There are also positive properties which require examination when assessing L2 (Hawkins & Buttery, 2010). Specific patterns, such as verb co-occurrences, appear at different proficiency levels, suggesting that positive features are criterial.

To address these issues, a solution could rely on linguistic complexity, one of the three dimensions of language proficiency, alongside accuracy and fluency (Housen et al., 2012). We explore the feasibility of using complexity measures, such as the number of different words, to automatically generate reports on systemic aspects of learner writings, i.e. their global lexical and syntactic elaborateness. As learners progress in their acquisition, they introduce more complex syntactic structures and more sophisticated lexical forms, which correlates with improvement in proficiency. We select measures of complexity for their predictive power in terms of proficiency. We incorporate them in a system focused on formative feedback. This system targets teachers who can visualize linguistic complexity in their learners' productions across different proficiency levels.

2. From corrective feedback to learning analytics based on linguistic complexity

As Ellis et al (2006) point out, "corrective feedback takes the form of responses to learner utterances that contain an error". However, CF may also include the criterial dimension in which specific correct patterns correspond to specific proficiency levels. This dimension implies that deviations occur at systemic rather than structure level and may affect the L2 system in terms of systemic complexity (Housen et al., 2012). Complexity features may be seen as L2-positive properties (Hawkins & Buttery, 2010) that globally deviate from target-like structures. As a

whole, CF may include comments on the ‘criterialness’ and correctness of learners' production at global level.

However, in most studies, the focus on errors hinders the notion of criterial positive features in CF. In this case, linguistic systemic complexity metrics might be potential candidates to operationalise the notion. Linguistic complexity is one of the constructs that lends itself well to computational methods. At a theoretical level, it can be split into sub-constructs such as syntactic complexity based on grammatical constituents and dependencies, lexical complexity based on diversity and sophistication, and semantic complexity based on semantic-functional properties linking forms to meanings. Linguistic complexity informs on the elaborateness of the learner language. At operational level, there are a number of statistical measures in the form of frequencies, ratios and indices (Bulté & Housen, 2012) that operationalise the aforementioned sub-constructs. Among all the metrics that have been tested, some operationalise systemic complexity in L2 (see Bulté & Housen, 2012, pp. 31–33, for a review of grammatical and lexical metrics used in L2 complexity studies). By operationalizing linguistic constructs such as composition, coordination, subordination or cohesion, complexity metrics measure different syntactic, lexical, semantic and discourse dimensions of language. Systemic complexity measures have been exploited in proficiency predicting approaches (Gaillat et al., 2021; Ballier et al., 2020; Vajjala, 2018; Tack et al., 2017; Pilán et al., 2016; Yannakoudakis et al., 2011). The Type Token Ratio (TTR), the number of different words (NDW), also called *types*, and the Mean Length of Sentence (MLS) are some of the metrics that have been reported as significant (Kyle, 2016; Lu, 2012; Vajjala & Meurers, 2012).

The metrics can be extracted with many tools that implement lexical complexity (Lu, 2012; Kyle et al., 2018; Benoit et al., 2018), syntactic complexity (Kyle, 2016; Lu, 2010) and

discourse complexity (McNamara et al., 2010; Crossley et al., 2019; Dascalu et al., 2013). All these tools provide a wealth of objective measurements that can be exploited as positive properties of L2. As pointed out by Biber et al. (2020), these measures do not provide explicit grammaticality suggestions, but we argue that they provide some global information which can be turned into actionable recommendations by teachers. For instance, MLS, as simple as it may be, can be a point of comparison for learners who tend to write sentences well over the commonly admitted thresholds (e.g. the learner corpus used in this study includes sentences with more than 50 words).

Analytics tools are needed to provide feedback on L2-positive properties in relation to proficiency levels. In educational contexts, a number of text analytics tools exist, but they are focused on L1 learning (Attali & Burstein, 2006; Dascalu et al., 2013; McNamara et al., 2007; Roscoe et al., 2014). In the field of L2 learning, a tool called *FeedBook* (Rudzewitz et al., 2019) focuses on L2 learners of English at primary school level. Visualisations of linguistic features are part of the feedback given to students. One need that remains to be addressed in such a tool is the ability for learners to position the linguistic properties of their productions in relation to proficiency levels. *Write and Improve* (Yannakoudakis et al., 2018) is another L2-learning tool. It targets learners by giving them proficiency predictions and CF in texts. However, it does not provide systemic linguistic feedback. Overall, none of the aforementioned tools are designed to support teachers' work. Yet, teachers also need analytics tools that can help diagnose writing in a holistic manner. Such tools would favour feedback messages on the linguistic system underlying L2 productions, i.e. the L2 knowledge informing target-language composition. To do so, a solution could be to rely on measures of linguistic complexity.

Visualizing Linguistic Complexity and Proficiency in Learner English Writings
T Gaillat, A Lafontaine, A Knefati
CALICO Journal 40 (2), 178-197, 2023

We present a system that positions new learner writings according to i) specifically selected systemic complexity measures, and to ii) already classified writings with CEFR[1] levels. Instead of relying on statistical models to predict levels, our proposal is to contrast measurements of new writings with existing writings of specific proficiency levels. Reports, primarily aimed at teachers, are provided as graphics that display the measures along with their linguistic interpretation. The purpose is to highlight systemic criteria whose values need to be improved to make them comparable with cohorts of specific CEFR levels. Analysing the effects of this type of feedback in class contexts is outside the scope of this paper. Likewise, we do not cover the technicality of the measures regarding language-teacher competence. In this paper, we focus on the formative aspect of the system and the metrics it relies on.

Designing such a system raised two research questions:

i) What features make this system provide formative feedback?

ii) Which measures correlate with specific proficiency levels, and how do they help model language levels?

Section 3 addresses the first question and Sections 4 and 5 focus on the second question. Section 6 provides a discussion and conclusion in light of these questions.

3. A system for formative feedback messages

The purpose of the system is to use complexity metrics to design graphical reports to help visualize textual measurements. Firstly, we describe how the system creates feedback reports. Secondly, we conduct a qualitative assessment of why the reports are formative rather than what effects they have in class context.

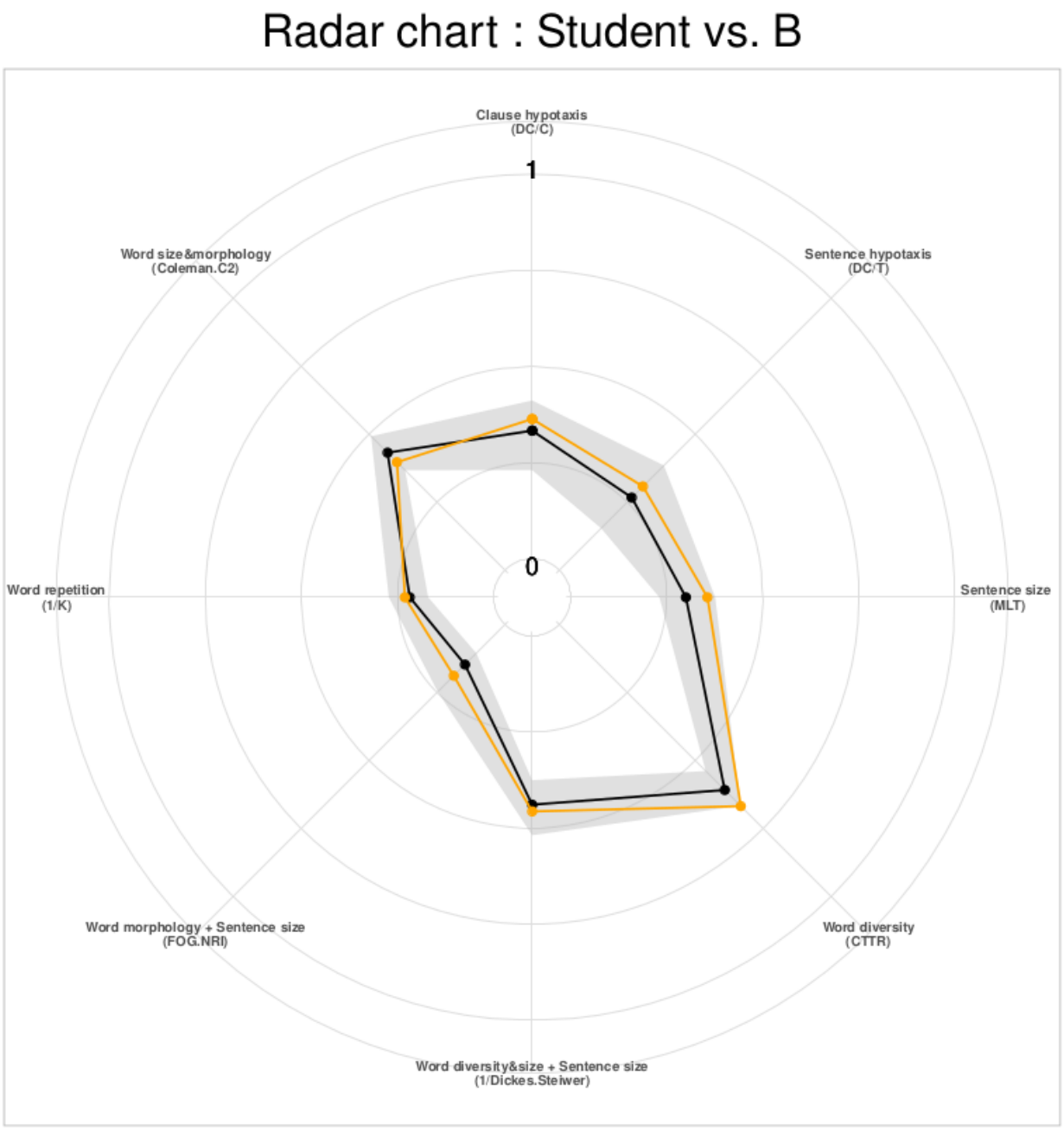
1 For a detailed description of the formulae refer to https://quanteda.io/reference/textstat_readability.html

To create the reports, the system imports learners' productions previously collected via two types of MOODLE[2] activities (Dougiamas & Taylor, 2003), i.e. *Assignment* and *Database*. Teachers can download all their students' writings and transfer them as input into the system's data processing pipeline. Once the students' texts have been processed, linguistic profiles can be visualised as graphical reports. A learner's individual report consists of six pages, each showing comparisons of the learner's writing with those of a specific CEFR cohort. Each page includes two types of graphs. A radar chart (see Figure 1) displays some ratio-based metrics, and boxplots show raw frequencies. In terms of statistics, the median and a shaded-grey strip for quartiles 1 and 3 show the values of a control cohort classified as B (see the description of the reference corpus in Section 4.1). As opposed to the mean, the median ensures robustness to outliers. The quartile strip shows the variability of a metric within a CEFR level. Provision is also made for the rare cases in which metric values fall out of the [0,1] interval. In this case, the value is not visualised on the graph and a warning is displayed: "You are off radar for the following indicator".

Figure 1

Radar chart presenting the various textual measures. A learner text is compared with measurements carried out on a control cohort of texts classified in the B level of the CEFR.

2 For details on lexical diversity formulae see https://quanteda.io/reference/textstat_lexdiv.html



To understand why the reports can be used as formative feedback, we conduct a qualitative assessment to identify the formative properties of the system. We analyse it in light of the features of formative instruction described by Shute (2008, pp. 177–178). Firstly, she specifies that a message needs to be offered in manageable units. The indicators in the radar

chart show different types of measurements. For example, the MLT indicator (Mean Length of T-units), i.e. essentially sentences as operationalised in (Lu, 2010), shows that the learner (orange line) creates sentences corresponding to the B cohort's sentence profile. The Corrected Type Token Ratio (CTTR) shows the diversity of words in the text. In this case, the learner is positioned on the rim of the B level. Each indicator gives a measurement of one facet of complexity, splitting the analysis into manageable units. We assume that the provision of such explanation rests within the instructor's prerogative.

Secondly, Shute shows that feedback must be multimodal as well as objective. The reports display feedback in several modes, including coloured graphs, textual labels and full text explanations. In addition, the measurements are computed automatically without any human intervention, and they are based on form counts.

Thirdly, the feedback messages focus on specific metrics, pointing out which dimensions of the textuality of the learners' texts are problematic. By way of comparisons with cohorts, the measurements suggest areas of improvement with the labels in the graphs. This tool aims to assist teachers and they choose how to effect such improvement. The labels describe the scopes of the metrics as presented in (Gaillat, In press). In short, the scopes correspond to the linguistic delineations implied by the metric formulae. For instance, the ratio of Dependent Clauses / T-Units (DC/T) (Lu 2010) is composed of two variables. In terms of scopes, the measure can be formally described as:

> Sentence.hypotaxis.rate (DC) where the measure compares the number of finite clauses, tagged as DC, to the number of T-Units, used as a denominator by way of a rate() method. The denominator delineates the measure and gives it a sentence scope. Looking at clauses in relation to sentences (T-units here) points to hypotaxis as an attribute to the sentence scope.

The DC/T value in Figure 1 shows that hypotaxis, operationalised as the proportion of dependent clauses per sentence, is comparable with that of the B-level cohort. This type of interpretation provides a specific descriptive explanation of the learner's text. This feedback is also linked to proficiency as operationalised by the CEFR strip shown in the graph. The combination of feedback and proficiency information can help teachers diagnose problematic dimensions of texts and provide suggestions on how to improve and get closer to the next level. A description of the metrics and their scopes is available as part of each report, including illustrative examples.

4. Methods for the selection of metrics displayed in the system

In this section, we present the selection methods of the features that were subsequently exploited in graphic reports. To do so, we conducted a data-driven approach using a reference corpus of learner English from which we identified and tested a number of numerical metrics.

4.1. Reference corpus

To identify features, we used a new specifically-designed learner corpus of written productions collected in 2018 and 2019 with L1 French learners. Texts from English for Specific Purposes (ESP) university graduate and post-graduate students were used. Their majors were in the fields of mathematics, biology, computer science, pharmacy and medicine. As part of their courses, the students followed classes focused on ESP English taught with a task-based language teaching approach (see Lai & Li, 2011, for a review). This corpus includes 274 writings that were classified in terms of CEFR-proficiency levels by two language certification experts following the CEFR-written assessment grid (European Council, 2018, appendix 4). The corpus includes educational metadata (Granger, 2015) about the characteristics of the subjects, such as domain of studies, age, number of years studying English and their learning behaviours, such as frequency

of exposure to English and travelling to L1 countries. The production task for learners consisted in describing an experiment/discovery/invention/technology of their choice and in giving their opinion on the impact of the previously described item. Learners had 45 minutes to complete the task. Table 1 shows the breakdown of the texts according to the CEFR levels.

Table 1

Breakdown of the number of learner texts in relation to their manually assigned CEFR level.

CEFR level	A1	A2	B1	B2	C1	C2
Number of writings	23	72	102	43	18	16
Average number of words (Standard deviation)	105 (65.6)	169 (78.6)	216 (110.2)	266 (147.9)	319 (122.9)	333 (91.7)
Min-Max	19-290	23-472	26-685	62-725	97-505	178-508
Median (IQR)	97.0 (81.0)	166.0 (86.8)	205.5 (120.8)	227.0 (149.5)	361.0 (142.5)	340.5 (101.5)

The CEFR classification was evaluated with a measurement of inter-rater agreement. Cohen's weighted Kappa was 0.71 (Gaillat et al., 2019).

4.2. Extracting metrics

The corpus texts were processed to compute different measures. Three tools, written in R, were used to compute three groups of metrics[3].

The first group corresponds to the construct of syntactic complexity metrics. It was operationalised with fourteen metrics by using a R version of L2SCA[4] (Lu, 2010) and relying on

3 The R version of L2 Syntactic Complexity Analyzer is available from http://www.personal.psu.edu/xxl13/downloads/l2sca.html

4 It is a common problem with unbalanced classes to predict the majority class too often as opposed to minor classes, which leads to lower recall for these classes.

Part-of-Speech tagging and syntactic parsing with CoreNLP (Manning et al., 2014). These metrics correspond to five different types of complexity: length of production unit (e.g. sentence), sentence complexity, subordination, coordination and particular structures (e.g. complex nominals). Each metric is a ratio of the frequency of a constituent over the frequency of all constituents of a higher-level scope. For instance the Mean Length of T-Unit is computed as:

$$MLT = \frac{number\ of\ words}{number\ of\ T-Units},$$

The second group of metrics corresponds to the construct of readability. It is operationalised with forty-eight metrics using the R Quanteda readability library (Benoit et al., 2018). The metrics are based on morphological features of words used to compute different indicators. The assumption is that the indicators--such as the Coleman Liau, the Dale Chall readability scores--operationalise the learners' linguistic proficiency as it is understood within a particular k-12 educational context. These indicators all rely on word length in terms of characters and syllables, as well as predetermined lists of words judged as difficult[5].

The third group of metrics describes lexical richness. The construct is operationalised with thirteen metrics computed with the R Quanteda lexical diversity library. Two types of lexical diversity are included. Diversity based on word-type variation is accounted for with Type-Token-Ratio (TTR) based formulae. Diversity based on type repetition is accounted for with Yule's K and similar formulae in which the frequency of word types, in a sample of size N,

5 This dataset is available from the IRIS database

is relative to the total number of words in a text[6]. We acknowledge that lexical sophistication and lexical density (content vs grammar words) are not taken into account in the lexical diversity metrics. However, the Dale Chall indicator relies on a list of difficult words, thus giving information about sophistication.

4.3. The dataset

A dataset[7] of 83 measures and CEFR levels per text was created, resulting in a 84x274 matrix. It includes six subsets according to the six CEFR levels.

To prepare the dataset, the metric values were transformed to ensure comparability. First, they were normalised as part of a [0,1] interval. The minimum and maximum values of the indicators are based on evidence from the reference corpus. All the normalised indicators, displayed in the radar chart, show increasing values as CEFR levels increase.

We also tuned the dataset. In our analyses, CEFR levels were initially scaled into 6 categories. Due to the small number of texts in the A1, C1 and C2 levels, the texts were collapsed in three groups corresponding to the three main CEFR levels, i.e. A (N=95), B (N=145), and C (N=34). Some metrics (K, Fucks, DRP, ...) were transformed into their inverse. This is because, as opposed to all other indicators, their values drop as CEFR levels get higher. The purpose was to have consistent scales in the radar charts of the reports (See Section 3). A description of the metric values in the dataset is available online[8].

6 The list of metrics includes their linguistic dimension and the tools used for computation. It is downloadable from the IRIS database.

7 Modular Object-Oriented Dynamic Learning Environment

8 The list of metric values from the reference dataset is downloadable from the IRIS database.

4.4. Identifying metric candidates with ANOVA

In order to select the most significant metrics, highly correlated pairs were sought in order to remove one of each pair and avoid redundant information in the dataset. We applied pairwise comparisons of all the metric values with the Spearman (ρ) correlation coefficient. Its values range from -1 to 1 (Eisinga et al., 2013). Highly correlated features ($\rho > 0.99$) were identified and only the most explanatory, according to our judgment, was kept.

Secondly, we conducted the Fisher test with Analysis of variance (ANOVA). The objective was to identify how sensitive the remaining features were in distinguishing classes (Baayen, 2008, p. 103; Levshina, 2015, p. 171). The following procedure was applied :

- Let :
    - Null hypothesis : all means are equal.
    - Alternative hypothesis : all means are not equal.
- For each feature, compute the F statistic, i.e. the ratio of the between-group variation with the within-group variation. This statistic gives the scores of the test (the groups reflect the proficiency levels: A, B or C).
- Calculate the p-values from the distribution function of scores.
- Set the significance level: α=5 %.
- Select the metrics whose p-values < α, i.e. the features for which the alternative hypothesis is true.

4.5. Measuring the predictive power of metric candidates

We performed a multinomial logistic regression to assess the goodness of fit of the selected metrics (see Section 6.1). This method was used to estimate the probability of belonging to a

certain class based on several metrics (Levshina, 2015, pp. 277–288). A 5-fold cross validation was performed to determine classification performance. In the case of small samples, cross-validation is favoured since results are less biased than a regular train/test approach. We chose K=5 as recommended in Breiman & Spector (1992). The steps of this method are the following :

- Randomise the dataset
- Split the dataset into k subsets of approximately equal size
- For each subset:
    a. Consider the subset as the test dataset
    b. Aggregate the other subsets to be used as the train dataset
    c. Fit a model on the train dataset, and evaluate it on the test data
    d. Calculate the performance

This method was repeated 100 times, with a new set of K folds for each repetition. For evaluation purposes, we computed the means of overall accuracy, balanced accuracy, recall, precision and F1-score (harmonic mean).

4.6. Measuring the splitting power of metric candidates

We also wanted to verify whether the metrics helped splitting the dataset, with good accuracy, into three clusters corresponding to the three proficiency levels: A, B and C. The purpose was to verify if metric values of the same cluster were more closely related to one another than values in other clusters. To do so, we used the k-means algorithm (Hartigan & Wong, 1979) (k=3).

5. Results

5.1. Metric candidates

The highly correlated features[9] were identified according to the Spearman correlation coefficient. The second variable in each pair is removed from our study in order to avoid the redundancy that could be caused by highly correlated metrics. For instance, results show a high correlation between the Coleman.Liau.ECP and the Coleman.Liau.grade metrics, leading to the removal of

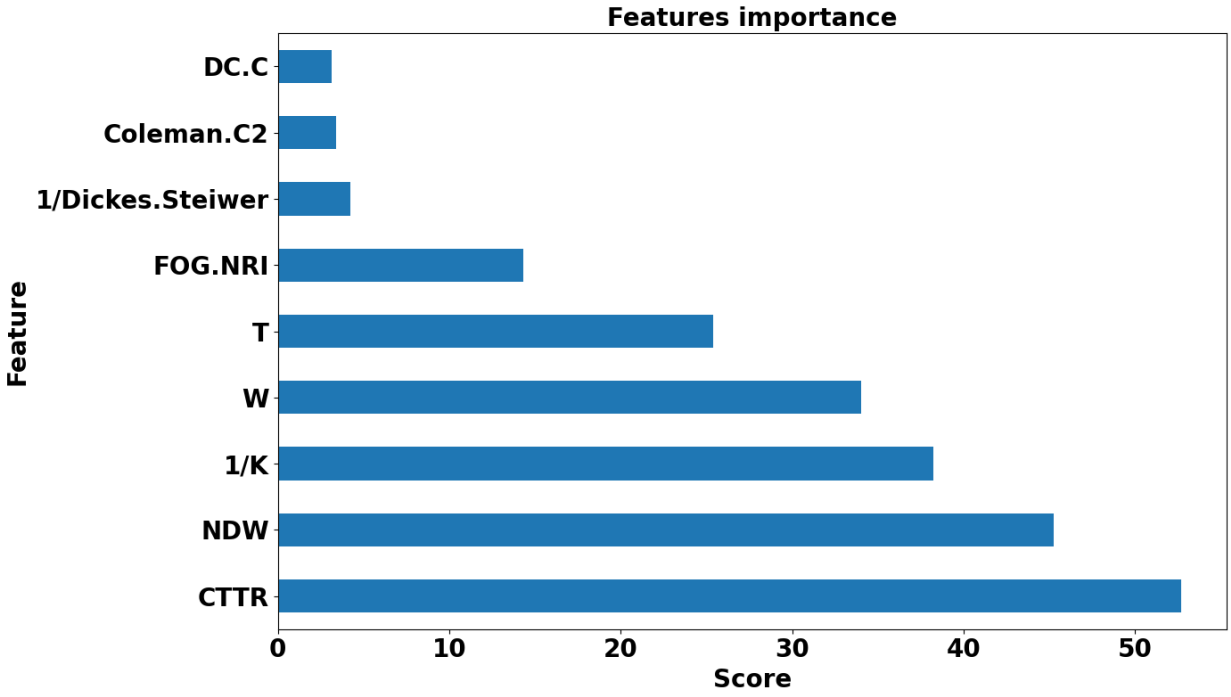
9 The values are available from the IRIS database.

the second metric from the model. As a principle, we kept the most interpretable metrics in terms of linguistic scopes as exemplified in Section 2.

The results of the Fisher test show which metrics are significant in the dataset (see bar chart in Figure 2). It provides two types of information including the F-statistics score on the x-axis showing the strength of the metric.

Figure 2

Score of significant metrics calculated from F-test in ANOVA ($p < 0.05$)



The selected metrics are described in Table 2.

Table 2

Description of the metrics that were selected for graphical reports, including the construct they operationalise, and the scopes of the variables that are combined in their formulae.

Metrics	Description	Constructs	Combined scopes in formula Scope.attribute.method(element)
Coleman.C2		Readability	Word.size.ratio(sentences) Word.morphology.rate(1syllables)
CTTR	Corrected TTR	Diversity	Word.diversity.ratio(types)
DC.C	Dependent Clauses/Clauses	Syntactic complexity	Clause.hypotaxis.ratio(DC)

Dickes.Steiwer		Readability	Sentence.size.mean(words) Word.diversity.ratio(types) Word.size.mean(characters)
FOG.NRI		Readability	Word.morphology.rate(3-syllables) Word.morphology.rate(3syllables) Sentence.size.mean(words)
MLT	Mean Length of T-Units	Readability	Sentence.size.mean(words)
T	T-Units	Syntactic complexity	Text.size.count(T-Units)
W	Words		Text.size.count(W)
K	Yule's K	Diversity	Word.repetitions.ratio(types)
NDW	Number of Different Words	Diversity	Text.size.count(types)

Table 2 shows metrics belonging to the three families (i.e. lexical diversity, readability and syntactic complexity) defined in Section 4.2. Nevertheless, only one syntactic complexity metric appears to be significant. The DC/C ratio indicates subordination at clause level.

5.2. Predictive power of the candidates

In order to measure the predictive power of the metrics in terms of CEFR levels, we applied a multinomial logistic regression method. Classification accuracy of our metrics is presented in Table 3 with a 5-fold cross validation repeated 100 times.

Almost half of A-rated productions were correctly classified (52.09%). Our model tends to offer a higher level precision but a lower recall resulting in a 55.79% F1-score. Three quarters of B-rated productions were correctly classified but 39.21 % of writings classified as B were not initially in this class. The B level had the highest classification performance with a 66.19% F1-score. Only one quarter of C-rated productions were correctly classified. Our model tends to behave in the same way as with class A regarding recall and precision. Nonetheless, the gap

between the two scores is larger for class C. Precision greater than recall means that the model is more careful when assigning a production to a C level. With a 34.92% F1-Score, the model had a lower performance for this class. This may be due to the reduced number of writings annotated in this category[10].

10 Common European Framework of Reference in languages

Table 3

Repeated 5-fold cross-validation classification performance metrics for the multinomial logistic regression method including three collapsed CEFR levels

Levels	Total	A (n=95)	B (n=145)	C (n=34)
Overall Accuracy	59.91%			
Balanced Accuracy		67.51%	60.16%	60.47%
Recall		52.09%	73.53%	25.24%
Precision		62.06%	60.79%	46.67%
F1-Score		55.79%	66.19%	34.92%

Only 0.92% of texts on average were classified in a non-adjacent level (level A classified as level C and vice-versa). This tends to show that the main default of the classification is to assign a text to an adjacent level instead of its actual level.

5.3. Splitting power of the metric candidates

To test the validity of the selected features, we applied a clustering method to examine how metrics help to group data into 3 levels. The k-means algorithm explained in section 4.6, assigns each data point to one of the three clusters obtained from the algorithm. Table 4 shows the contingency table of proficiency levels and data clusters, with raw frequency values followed by percentages between brackets.

The results show that the first cluster detects 67.37% of students having the A level. The second cluster detects 53.10% of students having the B level. The last cluster detects 55.88% of students having the C level. Also, when a cluster is not adjacent to the level (e.g. cluster 3 / level A and cluster 1 / level A) the error in affecting a level to a cluster is very small. For example, the error in affecting the level A to cluster 3 is 2.10%. Finally, when a cluster is adjacent to a level (e.g. cluster 1 / level B, cluster 2 / level A, cluster 3 / level B and cluster 2 / level C), the error is more important than the previous case.

Table 4

Three clusters broken down into three CEFR levels

Levels Clusters	A (n=95)	B (n=145)	C (n=34)
Cluster 1	64 (67.37%)	48 (35.10%)	3 (8.82%)
Cluster 2	29 (30.53%)	77 (53.10%)	12 (35.30%)
Cluster 3	2 (2.10%)	20 (13.80%)	19 (55.88%)

6. Discussion and conclusion

In this paper, we have explored the use of linguistic complexity metrics as a way to provide teacher-oriented reports on the systemic complexity of learner writings. In doing so, we have shown that criterial features, a part of CF, can be provided at global level rather than just at structure level, i.e. specific text segments. By providing elaborated, multimodal feedback messages the system is formative in nature. The impact of the tool in a class setting environment remains to be conducted.

Concerning the reference corpus, specific measures were selected for their correlation and predictive power in terms of proficiency. Significant features were selected by ANOVA and subsequently validated with multinomial logistic regression and a k-means clustering approach. The classification tasks showed poorer results than similar CEFR classification conducted in several other languages. For example, Vajjala & Lõo (2014) report Acc=79% (N=879) on Estonian essays with a 4-point scale, Pilán & Volodina (2018) report Acc=0.84 and F1=0.82 (N=867) on Swedish essays with a 5-point scale (more difficult than 3-point) and Gaillat et al. (2021) report Balanced Acc=0.81 (N=20,177) in English writings on a six point scale. The lower classification performance may be due to several factors. Most errors came from allocations to adjacent levels, showing a lack of discriminatory power. The size of the reference corpus (N=274) prevented a finer-grained approach as there were too few students in the A1, C1 and C2 categories. A larger sample would help build a model according to six levels instead of three. This may have tempered the tool's pedagogical impact in terms of formative assessment since most instructors operate within, and not between, these levels (i.e. from A1>A2 as opposed to A>B). Further work should focus on increasing the size of the dataset to 'calibrate' the tool for small proficiency gains (e.g. B1>B2, as opposed to B>C).

Other features such as spelling errors and semantic metrics could have also improved the results as exemplified in (Ballier et al., 2020; Yannakoudakis et al., 2011). The k-means clustering approach showed over 50% accuracy in the three A, B, C proficiency classes. It is important to stress that classification and clustering were implemented as part of an explanatory approach. The purpose was to select the best metrics in the reference dataset.

Concerning the writing tasks, it is essential to make sure that new writings, which are passed through the system, match most of the conditions in which the reference corpus was collected. The type of task, the length of production time, the genre (description and opinion) are factors that need to be controlled in order to offer a fair comparison between different writings. Alternatively, the architecture of the system allows for the introduction of more reference data with a greater variety of tasks, L2 English of different L1s and genres.

Using systemic metrics for linguistic feedback is also a challenge. As explained in (Biber et al., 2020), metrics are omnibus measures that "combine multiple aspects of structure and syntax", which hinders explicitness. Despite their predictive power in terms of proficiency, some of the readability metrics (i.e. FOG.NRI, 1/Dickes.Steiwer) make linguistic interpretation difficult. This is why we adopted a linguistic scope approach (Gaillat, In press). These scopes were defined on the basis of the variables found in the metrics' formulae. They indicate the linguistic delineations of the formulae. Combining the scopes of a readability metric helps clarify the systemic dimensions it operationalises. Other metrics offer simpler interpretation. Syntactic complexity and lexical diversity metrics operationalise systemic complexity by comparing ratios of constituents or lexical forms. In this respect, it is possible to assess a specific grammatical or lexical feature in the light of its global application in the writing. The limitation in this approach is that it does not identify specific error occurrences, which would help error-focused corrective

feedback. Instead, it highlights linguistic areas of interest at global level. The feedback is focused on how appropriate it is to repetitively use a type of textual strategy. Due to the technicality of the metrics, the reports are meant to assist English teachers. With proper training on the measures and their scopes, teachers can identify the problematic linguistic areas which are linked to scopes.

Overall, the system provides a form of CF in terms of criterial features. By way of global measures on the text, specific types of linguistic issues are detected rather than errors. Following Shute's guidelines (Shute, 2008), the system's feedback messages can therefore be classified as formative. By including criteria and CEFR levels, it links performance to goal, i.e. what learners do, and what they ought to do. Our approach contributes to the move towards the design of learning analytics tools for the language teaching community. Integrating such tools in Content Management Software platforms would empower language teachers with real-time analyses of their students' writings.

7. Bibliography

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3), 3–29.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., & Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, *6*(1), 72–103.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018).

quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, 100869.

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, *60*(3), 291–319. https://doi.org/10.2307/1403680

Bulté, B., & Housen, A. (2012). *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors*. Council of Europe. http://www.coe.int/t/dg4/linguistic/Source/Framework_FR.pdf

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., Nardy, A., Dascălu, M., & Trăușan-Matu, Ștefan. (2013). ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *AIED 13—16th International Conference on Artificial Intelligence in Education* (Vol. 7926, pp. 379–388). Springer. https://doi.org/10.1007/978-3-642-39112-5_39

Eisinga, R., Grotenhuis, M. te, & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642.

Visualizing Linguistic Complexity and Proficiency in Learner English Writings
T Gaillat, A Lafontaine, A Knefati
CALICO Journal 40 (2), 178-197, 2023

https://doi.org/10.1007/s00038-012-0416-3

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339–368. https://doi.org/10.1017/S0272263106060141

Gaillat, T. (In press). Investigating the Scope of Textual Metrics for Learner Level Discrimination and Learner Analytics. In A. Lenko-Szymanska & S. Götz (Eds.), *Complexity, Accuracy & Fluency in Learner Corpus Research*. Benjamins.

Gaillat, T., Janvier, P., Dumont, B., Lafontaine, A., & Kerfati, A. (2019, December). *CELVA.Sp: A corpus for the visualisation of linguistic profiles in language learners*. PERL 2019, Paris, France. https://hal.univ-rennes2.fr/hal-02496713

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, *34*(2). https://doi.org/10.1017/S095834402100029X

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 485–510). Cambridge University Press.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hawkins, J. A., & Buttery, P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, *1*(01). https://doi.org/10.1017/S2041536210000103

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and*

proficiency: Complexity, accuracy and fluency in SLA (Vol. 32). John Benjamins Publishing Company.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Dissertation, Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lai, C., & Li, G. (2011). Technology and Task-Based Language Teaching: A Critical Review. *CALICO Journal*, *28*(2), 498–521.

Leacock, C., Chodorow, M., & Tetreault, J. (2015). Automatic grammar- and spell-checking for language learners. In F. Meunier, G. Gilquin, & S. Granger (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 567–586). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.025

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company. http://www.jbe-platform.com/content/books/9789027268457

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. http://acl2014.org/acl2014/

McNamara, D. S., Boonthum, C., Levinstein, I., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In *Handbook of latent semantic analysis* (pp. 227–241). Lawrence Erlbaum Associates Publishers.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Meurers, D. (2009). On the Automatic Analysis of Learner Language: Introduction to the Special Issue. *CALICO Journal*, *26*(3). https://journals.equinoxpub.com/index.php/CALICO/article/view/23054

Pilán, I., & Volodina, E. (2018). Investigating the importance of linguistic complexity features across different datasets related to language learning. In L. Becerra-Bonache, M. D. Jiménez-López, C. Martín-Vide, & A. Torrens-Urrutia (Eds.), *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing* (pp. 49–58). Association for Computational Linguistics. http://aclweb.org/anthology/W18-4606

Pilán, I., Volodina, E., & Zesch, T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2101–2111. https://aclanthology.org/C16-1198

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The

Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition*, *34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Rudzewitz, B., Ziai, R., Nuxoll, F., Kuthy, K. D., & Meurers, W. D. (2019). Enhancing a Web-based Language Tutoring System with Learning Analytics. In Luc Paquette & C. Romero (Eds.), *Joint Proceedings of the Workshops of the 12th International Conference on Educational Data Mining co-located with the 12th International Conference on Educational Data Mining, EDM 2019 Workshops* (Vol. 2592, pp. 1–7). CEUR-WS.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Tack, A., François, T., Roekhaut, S., & Fairon, C. (2017). Human and Automated CEFR-based Grading of Short Answers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 169–179. https://doi.org/10.18653/v1/W17-5018

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*, 79–105. https://doi.org/10.1007/s40593-017-0142-3

Vajjala, S., & Loo, K. (2014). Automatic CEFR Level Prediction for Estonian Learner Text. *NEALT Proceedings Series*, *22*, 113–128.

Vajjala, S., & Meurers, D. (2012). On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 163–173. http://dl.acm.org/citation.cfm?id=2390384.2390404

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Second Language Teaching & Curriculum

Center, University of Hawaii at Manoa.

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education, 31*(3), 251–267. https://doi.org/10.1080/08957347.2018.1464447

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002472.2002496