Towards a Data Analytics Pipeline for the Visualisation of Complexity Metrics in L2 writings

Thomas Gaillat
University Rennes 2 / France
thomas.gaillat@univ-rennes2.fr

Anas Knefati
ENSAI / France
@ensai.fr

Antoine Lafontaine
University Rennes 1 / France
@univ-rennes1.fr

Abstract

In this paper, we present the design of a tool for the visualisation of linguistic complexity in second language (L2) learner writings. We show how metrics can be exploited to visualise complexity in L2 writings in relation to CEFR levels.

1 Introduction

The analysis of educational data has been a growing field in the last decade. Learning Content Management Software (LCMS) platforms in education have provided the opportunity to collect and process large quantities of educational data supporting both data mining and analytics (Baker et al., 2016). As far as we know, the field of foreign language learning has not availed yet of projects with data analytics at their core. The proceedings from the Visualisation and Digital Humanities workshop series[1] and those of the Learning Analytics and Knowledge conference[2] fall short of studies focused on the automatic exploitation of linguistic data for learners of a language. This problem may be linked to the complexity of apprehending learner writings due to the multidimensional nature of this type of language (errors, usage, phraseology ...)

One way to approach the problem is to use data analytics methods in order to bridge the gap between the collection of learner productions and their automatic analysis resulting in meaningful feedback. To this end, it is necessary to identify quantifiable features of learner writings. Data could be sourced in one of the three dimensions of language proficiency, *i.e.,* Complexity, Accuracy and Fluency (CAF) (Housen et al., 2012). A data analytics framework could rely on measures that operationalise these three theoretical constructs. A selection of CAF measures could be the source of automatically generated linguistic profile reports of L2 writings.

As part of CAF, linguistic complexity is one of the constructs that lends itself well to computational methods. At theoretical level it 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 and Housen, 2012). The construct is already used in combination with corpora to achieve different tasks such as automatic proficiency level prediction. In these tasks, complexity metrics are exploited with supervised learning methods to predict levels (Ballier et al., 2020; Venant and D'Aquin, 2019; Pilán and Volodina, 2018; Yannakoudakis et al., 2011).Among all the metrics that have been tested (see (Bulté and Housen, 2012, p.31-33) for a review), several have been reported as predictive of proficiency (Kyle, 2016; Lu, 2012; Vajjala and Meurers, 2012). Some readability metrics have also been used in L2 studies (Lissón, 2017; Pilán et al., 2014).

A number of text analysis tools exist in education but are not focused on L2 learning. They provide environments for reading or writing assessment and training (Dascalu et al., 2013; McNamara et al., 2007; Roscoe et al., 2014; Attali and Burstein, 2006; Napolitano et al., 2015). They focus on providing quantified results in relation to internal scales for first language (L1) learning. In addition, and to the best of our knowledge, the tools do not provide visualisations for the textual measurements. In the field of L2 learning a tool called *FeedBook* (Rudzewitz et al., 2019) provides visualisations of linguistic features as part of the feedback given to students. One need that remains to be addressed is the ability for learners to position the linguistic properties of their productions with regard to proficiency levels.

[1] See http://vis4dh.org/

[2] See https://lak20.solaresearch.org/

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 123–129
April 20, 2021 ©2021 Association for Computational Linguistics

CEFR levels	Number of texts
A1	23
A2	72
B1	102
B2	43
C1	18
C2	16

Table 1: Cohorts in the CEFR-annotated corpus

Our proposal is to exploit state-of-the-art linguistic complexity metrics in the automatic analysis of L2 writings. NLP tools are used to annotate, compute metrics and display visualisations of learner productions compared with writings classified according to the CEFR[3] levels (European Council, 2001). Section 2 describes the method. Section 3 covers visualisation interpretation. Learner engagement is presented in section 4. We discuss issues and perspectives in Section 5.

2 Methodology

2.1 A CEFR-based reference data set

To compare new texts with existing texts, we exploit a learner corpus of written productions. Texts from English for Specific Purposes (ESP) university students are used. This corpus includes 274 third-level education writings. Two language certification experts assessed the writings in terms of CEFR proficiency levels. The first production task for learners consisted in describing an experiment/discovery/invention/technology of their choice and the second task was to give their opinion on the impact of the previously described item. Learners had 45 minutes to complete both tasks. Table 1 shows the breakdown of the texts according to the CEFR levels.

CEFR annotation was evaluated with a measurement of inter-rater agreement (Cohen's weighted Kappa = 0.71). Complexity metrics were computed and six subsets or cohorts were created according to the six CEFR levels. A comparative data set of metrics and CEFR levels was thus created[4].

2.2 Metrics

Three groups of metrics are computed at processing time. Syntactic complexity is operationalised with fourteen metrics. These metrics are grouped in five different types (Lu, 2014): Length of production unit (e.g. sentence), sentence complexity, subordination, coordination and particular structures (e.g.complex nominals). Each metric is a ratio of a frequency of a constituent over the frequency of all constituents of a higher-level scope.

Readability is operationalised with forty eight metrics. They are based on the morphological features of words used to compute different indicator values. The assumption is that indicators operationalise the level of maturity required for reading a specific text. This includes indicators such as the Coleman Liau, the Dale Chall readability score and the Flesch kincaid grade. They all rely on word length in terms of characters and syllables as well as predetermined lists of words judged as difficult[5].

Lexical richness is operationalised with thirteen metrics which provide information on lexical diversity, *i.e.,* the range of different words used in a text. Two types of lexical diversity are included. Diversity based on word type variation is accounted for with 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 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.

The metrics were selected for two reasons. Firstly, their significance is reported in the literature on L2 criterial features (Hawkins and Filipović, 2012; Lu, 2014; Kyle, 2016; Lissón, 2017) and analysing it in terms of CEFR levels is outside the scope of this paper. Secondly, it was decided to also include metrics linked to descriptive syntactic information. Complex Nominal and Coordinated Phrase indices were selected due to their meaningfulness. In total, 83 metrics are computed[7].

2.3 Data collection and cleaning

The learners' productions are collected via two types of MOODLE activities (Dougiamas and Taylor, 2003), *i.e., Assignment* and *Database*. The *Assignment* activity allows teachers to collect written assignments as they see fit within their course scenario. They can download all the assignments as

[3] Common European Framework of Reference in languages

[4] Available from IRIS database at https://www.iris-database.org/iris/

[5] For a detailed description of the formulae refer to https://quanteda.io/reference/textstat_readability.html?q=reada

[6] For the formulae see https://quanteda.io/reference/textstat_lexdiv.html

[7] A list of metrics is available as supplementary material

a batch file and transfer them as input into the data processing pipeline. The texts can also be collected via a learner-corpus building interface alongside student metadata. A file includes all the texts and metadata and can be imported into the pipeline.

Prior to processing the files, the texts are cleaned. All special characters are deleted. Punctuation symbols are spaced consistently. Accents (from expressions of other languages for instance) are removed. The pronoun ”I” is upper-cased for each text. The negative modal verb ”can’t” is replaced by ”cannot”. It permits to ensure a better parsing and a more accurate computation of the metrics.

2.4 The pipeline

The pipeline[8] is implemented in R (R Core Team, 2012) with a Creative Commons Share Alike licence. It includes our R implementation of L2SCA[9] (Lu, 2010) for syntactic complexity metrics. It also relies on Quanteda (Benoit et al., 2018) an R package used to compute readability and lexical diversity with `textstat\_lexdiv()` and `textstat\_readability()`.

The data workflow functions as described in Figure 1. Firstly, the input data is made up of new learner texts which are passed through the aforementioned processing tools to compute the metrics. Secondly, the reference corpus mentioned in Section 2.1 is also passed through the same processing tools. As a result, new texts can be compared with existing texts on the basis of the computed metrics. These can be visualised as of box-plots and radar charts.

2.5 Data transformation for visualisations

Prior to displaying metric values to users, these values are transformed to ensure comparability. First all the metric values are normalised to constrain them in a [0,1] interval for the radar chart. Yule’s K is transformed into its inverse for the radar chart to avoid confusions by learners. This is because, as opposed to all other indicators, K’s values drop as CEFR levels get higher. All the normalised indicators finally displayed, show increasing values as CEFR levels increase. In terms of statistics, the median and a shaded-grey strip for first and third quartiles are used to describe the control cohorts. Using an interval aims to show the variability of a metric within a CEFR level. Using the mean was not favoured to ensure robustness to outliers. Provision is also made for the rare cases in which metric values fall out of the interval. In this case, the value is not visualised on the graph and a warning is displayed: ”You are off radar for the following indicators:”.

3 Interpreting visualisations

In this section, we conduct an illustrative analysis of a sample text and compare some of its features with the visualised metrics. It was written by a French learner of English as part of the French National Language Certification Proficiency exam[10]. It was classified as B2 or higher. For reasons of space, we only provide the following exerpt.

> *With the development of new technologies such as smartphones, new questions are being caused about how to evaluate students. Indeed, using cellphones to cheat is common in high-schools. The first question we have to ask ourselves is wether we should authorize students to access their phone or not. Arguments against are well-know: ... But we also have to consider arguments in favor of it, in order to do what is best for our students. First, they will be working with these technologies in their professional lives, and we should be preparing them for that, by teaching them a proper use of smartphones and computers...*

In Figure 2 the metrics are compared with those of the cohort of B2 learners (see Section 2.1). The learner’s individual report is divided in two parts. On the left, a radar chart displays ratio-based metrics and, on the right, raw frequencies are reported. In addition to the metric acronym a categorisation label is provided in order to indicate the word, sentence or text scope (Anonymised reference 2019b) of the metric. For instance the Number of Different Words (NDW) is labeled *Text.size.type* and can be interpreted at text level in terms of types as a unit. The radar chart displays the cohort in a grey-shaded area representing the two central quartiles. The dark line represents the median of the cohort. The boxplots show the position of the learner’s metric in relation to the full B2 cohort.

The indicators in the radar chart show that the learner’s ratios globally correspond to those of the B2 cohort. For instance, the learner makes use of Complex Nominals in her text. This includes

[8] Available at https://github.com/LIDILE/VizLing

[9] Available at http://www.personal.psu.edu/xxl13/downloads/l2sca.html

[10] See CLES certification at https://www.certification-cles.fr/en/

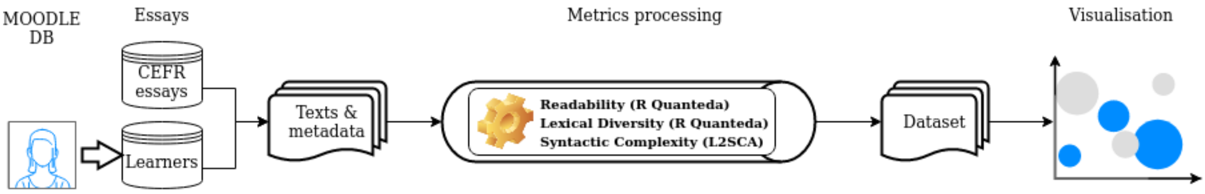

Figure 1: NLP pipeline - from data collection to visualisation

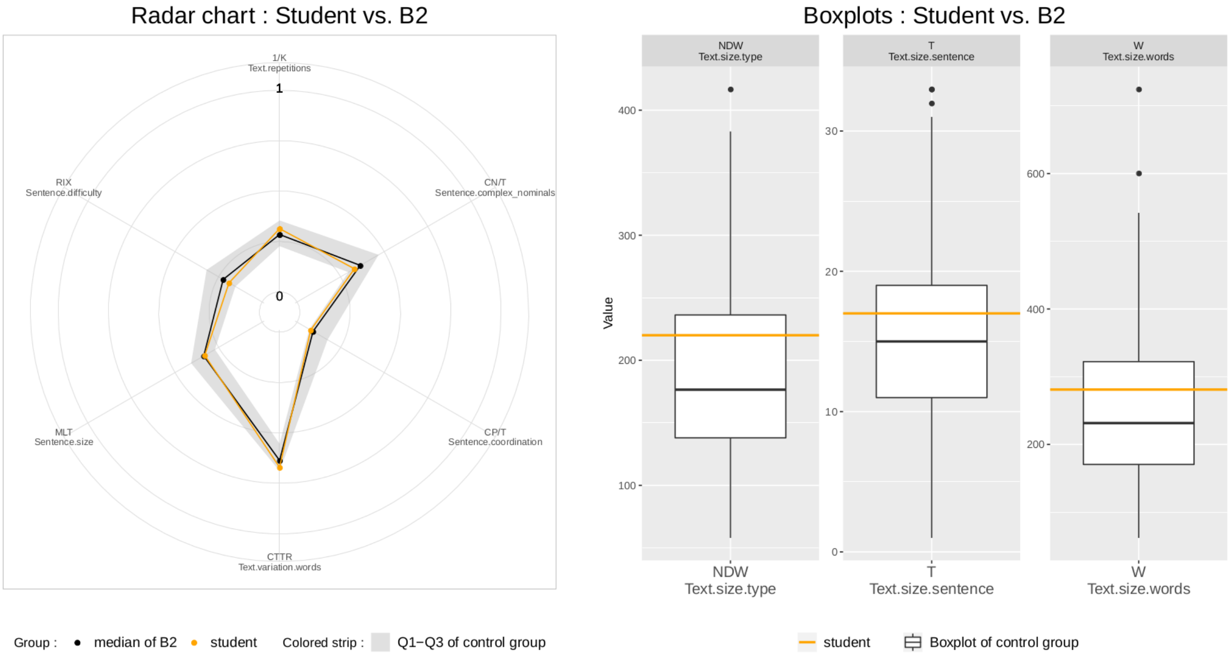

Figure 2: Individual report of a learner in comparison with a cohort of B2 learners

the adverbial clause ”With the development of new technologies such as smartphones” which is used in apposition to the main clause. It also includes the nominal clause ”The first question we have to ask ourselves” used in subject position of BE. The use of *adjective* + *noun* as in ”proper use” is another more simple example of nominal complexity. The three different cases are all accounted for by the system. It appears that the learner’s level of use is slightly under the B2 median (C1 and C2 radar charts show even higher values for the two central quartiles). The teacher and learner can analyse such structures more into details. The teacher could in turn note the lack of use of compounds and genitives in the text. In short, the metric helps learners and teachers identify an issue related to an objective criterion of linguistic complexity. Specific feedback and actions can then be undertaken.

4 Learner engagement

The efficacy of the tool needs to be evaluated. Learner engagement remains to be assessed thoroughly but a preliminary qualitative assessment was conducted in a class setting environment. We show results about the impact of the tool on learners’ engagement. Fifty-four first year higher-level students were given five individual writing tasks in five weekly waves. After each wave they were provided feedback within 24 hours. Notwithstanding the results, we measured the number of submitted writings (Figure 3) and the frequency of consultation of feedback reports (Figure 4). We use these measurements as a rough proxy to measuring learner engagement, i.e. how they respond to the feedback they receive (Ellis, 2010). The statistics are assumed to tap into the intensity of learners’ interest in the reports. Over time the number of submitted papers did not decrease in spite of the lockdown imposed on students in the midst of the COVID crisis. Following detailed explanations from their teacher to ensure comprehension, a majority of students consulted their reports three times or more, showing continuous interest.

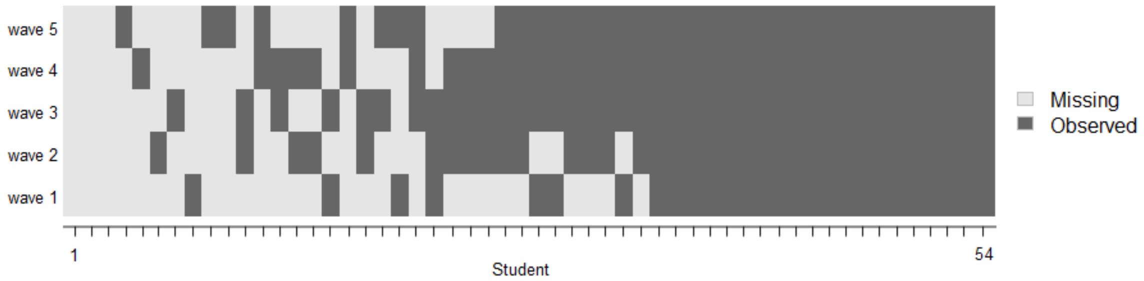

Figure 3: Writings available for each student

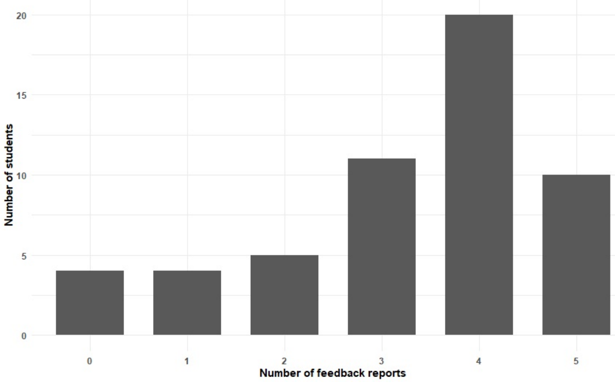

Figure 4: Consultation of feedback reports

5 Discussion and perspectives

In this paper we have presented a linguistic complexity visualisation tool. It displays learner writings according to several criteria and positions them in relation to cohorts of specific CEFR levels. More work remains to be done. Firstly, the visualizations used may be difficult to understand for learners who are not used to such types as radar charts. The tool aims primarily at helping trained language teachers analyse their students' writings in order to give them objective and specific feedback (Shute, 2008). By gaining access to these features, teachers can give specific answers regarding the mastery of certain concepts. They also become aware of features of language use that need to be addressed. Teachers can then provide evidence-based advice.

Secondly, the collected data shows limitations. The metrics on which the visualisations rely need to be evaluated on the data in terms of proficiency predictive power. Correlation analysis remains to be conducted in order to validate significant metrics to be displayed. The reference corpus is small and at the same time lacks diversity. All the texts belong to university students of specific fields, which may impact vocabulary and syntactic structures. More data needs to be collected in each field in order to support finer-grained analysis of third level education writings.

One last limitation is that some metrics remain difficult to interpret linguistically as argued by (Biber et al., 2020). For instance, readability formulas combine different features such as morphology and most common words. Specific advice on one of the features is therefore near impossible. Nevertheless, by interpreting the linguistic scopes (whether the measures apply at word, sentence or text level), it is possible to provide a certain degree of feedback.

The tool could be exploited in the learner module of an Intelligent Tutoring System dedicated to language learning. Because linguistic complexity measurements keep track of the evolution of systemic syntactic and lexical complexity, these data constitute part of the knowledge representation of the learner.

The tool gives access to learning analytics at linguistic level. In a context of distance learning, teachers are empowered with a rapid diagnostics tool that gives them an objective, although reduced, view of some of the features of their learners' language. Further developments will focus on identifying and evaluating more significant metrics in terms of proficiency and meta-linguistic influence on learners. Other types of charts could also be explored, and an aggregation functionality could provide group visualisations to reveal linguistic class patterns for teachers.

References

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

Ryan S. Baker, Taylor Martin, and Lisa M. Rossi. 2016. Educational Data Mining and Learning Analytics.

In *The Wiley Handbook of Cognition and Assessment*, pages 379–396. John Wiley & Sons, Ltd.

Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, and Thomas Gaillat. 2020. Machine learning for learner English. *International Journal of Learner Corpus Research*, 6(1):72–103.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Douglas Biber, Bethany Gray, Shelley Staples, and Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46:100869.

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Mihai Dascalu, Philippe Dessus, Stefan Trausan-Matu, Maryse Bianco, Aurélie Nardy, Mihai Dascălu, and Ștefan Trăușan-Matu. 2013. ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In *AIED 13 - 16th International Conference on Artificial Intelligence in Education*, volume 7926 of *Lecture Notes in Computer Science (LNCS)*, pages 379–388, Memphis, TN, United States. Springer.

Martin Dougiamas and Peter Taylor. 2003. Moodle: Using Learning Communities to Create an Open Source Course Management System. In *Proceedings of the EDMEDIA 2003 Conference, Honolulu, Hawaii*, pages 171–178, Hawaii. Association for the Advancement of Computing in Education (AACE).

Rod Ellis. 2010. EPILOGUE: A Framework for Investigating Oral and Written Corrective Feedback. *Studies in Second Language Acquisition*, 32(2):335–349. Publisher: Cambridge University Press.

European Council. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

John A. Hawkins and Luna Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, volume 1 of *English Profile Studies*. Cambridge University Press, United Kingdom.

Alex Housen, Folkert Kuiken, and Ineke Vedder, editors. 2012. *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching (LL&LT)*. John Benjamins Publishing Company, Amsterdam, The Netherlands, USA.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Paula Lissón. 2017. Investigating the use of readability metrics to detect differences in written productions of learners : a corpus-based study. *Bellaterra journal of teaching and learning language and literature*, 10(4):0068–86.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

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

Xiaofei Lu. 2014. *Computational Methods for Corpus Annotation and Analysis*. Springer, Dordrecht.

Danielle S. McNamara, Chutima Boonthum, Irwin Levinstein, and Keith Millis. 2007. Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In *Handbook of latent semantic analysis*, pages 227–241. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Diane Napolitano, Kathleen Sheehan, and Robert Mundkowsky. 2015. Online Readability and Text Complexity Analysis with TextEvaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, Denver, Colorado. Association for Computational Linguistics.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184, Baltimore, Maryland. Association for Computational Linguistics.

R Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rod D. Roscoe, Laura K. Allen, Jennifer L. Weston, Scott A. Crossley, and Danielle S. McNamara. 2014. The Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition*, 34:39–59.

Björn Rudzewitz, Ramon Ziai, Florian Nuxoll, Kordula De Kuthy, and Walt Detmar Meurers. 2019. Enhancing a Web-based Language Tutoring System with Learning Analytics. In *Proceedings of the Workshops of the 12th International Conference on Educational Data Mining (EDM 2019)*, volume 2592, pages 1–7, Montréal, Canada.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rémi Venant and Mathieu D'Aquin. 2019. Towards the Prediction of Semantic Complexity Based on Concept Graphs. In *12th International Conference on Educational Data Mining (EDM 2019)*, pages 188–197, Montreal, Canada.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.