A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners' written productions

Nicolas Ballier, Paula Lissón

A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners' written productions

Nicolas Ballier[1], Paula Lissón[2]
Université Paris-Diderot (USPC), EA 3967 – CLILLAC-ARP, France
nicolas.ballier@univ-paris-diderot.fr[1], paula.lisson@etu.univ-paris-diderot.fr[2]

This paper deals with the lexical assessment and classification of learners through the implementation of readability metrics as indices of syntactical complexity. The aim of the paper is twofold: first, delimiting which of the 30 readability metrics used in the study shows the most appropriate values for classifying learners into different proficiency groups; and second, validating the possibility of using readability metrics with frequency lists of difficult words generated from the learner corpus analysed.

With the expansion of learner corpora, many studies dealing with the automatic assessment of learner's language complexity have tackled lexical and syntactic complexity (Cobb & Horst, 2015). For example, Lu (2010) creates a computational system for the analysis of syntactic complexity in second language writing with 14 built-in metrics. These metrics present a high degree of reliability when used, for instance, as an index of ESL learner's writing development (Lu, 2011). Similarly, Vajjala (2016) shows how lexical and syntactic metrics help assessing learners' production; and Ballier & Gaillat (2016) use these type of metrics in order to classify French learners of English into different proficiency groups.

However, the domain of readability in relation with Learner Corpus Research (LCR) remains slightly less explored. Broadly speaking, the role of readability measures in SLA has been used to establish the difficulty of texts in reading tasks (Kasule, 2011; Vajjala & Meurers, 2012). Readability measures are typically used so as to determine if a text is appropriate or not for learners of a particular level (François, 2011; Gala *et al.,* 2014). Few studies combine the use of readability and lexical/syntactical metrics, the Vajjala & Meurers (2012) study is an example of the interconnection between traditional readability measurements and SLA complexity metrics.

In this paper, we aim at changing the traditional point of view of readability metrics; we are not using readability in order to see how difficult a text might be for a given level of proficiency; but rather applying readability formulae to learners' productions so as to see if the metrics can be used to classify learners into different levels. In order to do so, we assess the validity of 35 of the readability metrics implemented in the {koRpus}(Michalke, 2016) package of R (R Core Team, 2016) by applying them to randomly chosen samples taken from NOCE (Díaz-Negrillo, 2007), a written corpus of Spanish university students of English. Replicating Lu (2012), we assess the strength of the correlations among the metrics using Spearman's '*p*' (see Table 1).

Table 1: Correlations among 3 metrics with their original lists implemented ($p. < 0.001$ in all the cases)

	Bormuth	**Dale.Chall**	**Spache**
Bormuth	1	0.824	-0.609
Dale.Chall	0.824	1	-0.834
Spache	-0.609	-0.834	1

Some metrics (Spache, 1966; Bormuth, 1969; Chall & Dale, 1995) rely on the use/underuse of complicated words. These formulae rely on the implementation of lists of complex words which were originally compiled by and for native speakers of English, and its application to learner corpora might yield unsatisfactory results. Thus, the second aim of this paper is to create a list of complex words according to their frequency in the NOCE corpus, and to implement it in the readability formulae, instead of using the original lists.

By using a specific list generated from the corpus we are analysing, we can classify learners according to potentially more accurate criteria (see Figure 1). Our contribution to *widening the scope of learner corpus research* is to suggest that we should design learner-based frequency lists to adequately describe learner data. Taking learner output as the baseline for linguistic analysis raises issues in terms of L2 attainment that we also discuss.

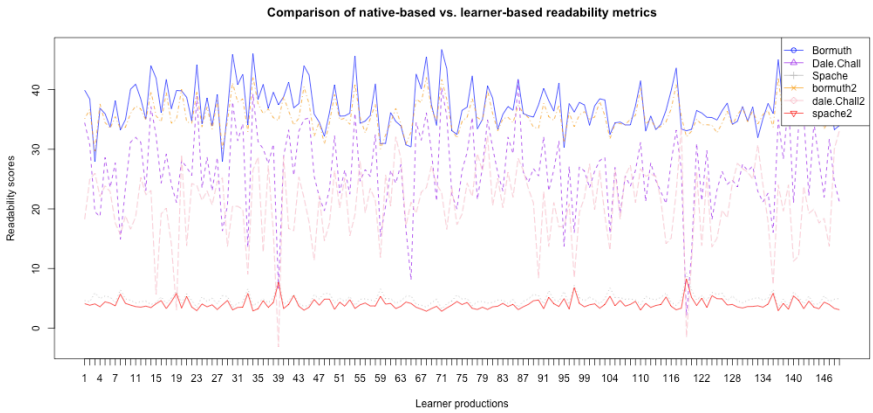

Figure 1: Learner output analysed with native-based vs. learner-based metrics

References

Ballier, N., & Gaillat, T. (2016). Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique (Vol. 9, pp. 1–14). Presented at the JEP-TALN-RECITAL 2016.

Cobb, T., & Horst, M. (2015). Learner Copora and Lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Díaz-Negrillo, A. (2007). *A Fine-grained Error Tagger for Learner Corpora* (PhD Thesis). University of Jaen, Jaen.

François, T. (2011). Les apports du traitement automatique des langues à la lisibilité du français langue étrangère.

Gala, N., François, T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots (pp. 91–102). Presented at the Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014).

Kasule, D. (2011). Textbook readability and ESL learners. *Reading & Writing*, *2*(1), 63–76.

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

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *Tesol Quarterly*, 36–62.

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

Michalke, M. (2016). koRpus: An R Package for Text Analysis (Version 0.06-5). Retrieved from http://reaktanz.de/?c=hacking&s=koRpus

R Core Team. (2016). R: A language and environment for statistical computing. (Version 3.3.1 (2016-06-21)). Vienna, Austria.: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Spache, G. D. (1966). *Good reading for poor readers* (Revised 9th edition). Champaign, Illinois: Garrard.

Vajjala, S. (2016). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *arXiv Preprint arXiv:1612.00729*. Retrieved from https://arxiv.org/abs/1612.00729

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition (pp. 163–173). Presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics.