A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners’ written productions
Nicolas Ballier, Paula Lissón

To cite this version:
Nicolas Ballier, Paula Lissón. A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners’ written productions. 4th LEARNER CORPUS RESEARCH CONFERENCE (LCR 2017), Oct 2017, Bolzano Italy. hal-01673699

HAL Id: hal-01673699
https://hal.science/hal-01673699
Submitted on 11 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners’ written productions

Nicolas Ballier¹, Paula Lissón²
Université Paris-Diderot (USPC), EA 3967 – CLILLAC-ARP, France
nicolas.ballier@univ-paris-diderot.fr, paula.lisson@etu.univ-paris-diderot.fr

This paper deals with the lexical assessment and classification of learners through the implementation of readability metrics as indices of syntactic 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 (Diaz-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)

<table>
<thead>
<tr>
<th></th>
<th>Bormuth</th>
<th>Dale.Chall</th>
<th>Spache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bormuth</td>
<td>1</td>
<td>0.824</td>
<td>-0.609</td>
</tr>
<tr>
<td>Dale.Chall</td>
<td>0.824</td>
<td>1</td>
<td>-0.834</td>
</tr>
<tr>
<td>Spache</td>
<td>-0.609</td>
<td>-0.834</td>
<td>1</td>
</tr>
</tbody>
</table>

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


