Comparing Syntactic Complexity Measures Counted by Tregrex-based Tagging and UD-based Tagging for Evaluating L1 Japanese EFL Paragraph Writing

SPRING Ryan[1] JOHNSON Matthew[2]
[1]Tohoku University [2]Unaffiliated
{spring.ryan.edward.c4}@tohoku.ac.jp {mjokimoto}@gmail.com

Abstract

Automatically calculated measures of syntactic complexity can provide useful insight in L2 learners' writing ability and general proficiency. Many tools have been developed for this purpose, but so far many of them depend on Tregrex tagging. However, recent NLP tools provide universal dependency tags which can also be used to calculate some similar measures of syntactic complexity. This paper compares the ability of a tool on parsing provided by SpaCy [1] for this purpose, the SSCC, to another popular tool that provides measures based on Tregrex tagging, the L2SCA. We found that the SSCC and L2SCA calculated many measures similarly, but that for congruent measures (i.e., MLS, MLC, DC_C), the SSCC measures were more associated with L2 writing ability and general proficiency. Furthermore, some measures only provided by the L2SCA correlated with TOEFL® ITP scores, but not writing scores (i.e., C_T, VP_T, CT_T, CN_T, CN_C), whereas some measures that only the SSCC could provide were correlated with both (i.e., DC_S, and CSTR_S). We conclude that both the SSCC and L2SCA have advantages and disadvantages and that more study is required to see under what conditions certain measures from the two tools are most associated with L2 writing and general proficiency.

1 Introduction

Many objective measures of syntactic complexity have been found to be correlated to writing ability, albeit in differing ways and amounts [2, 3, 4]. Due to the large and steadily increasing number of syntactic complexity measures that exist, automatic calculation of such measures through reliable tools can be helpful to researchers and educators. One such tool is the L2 Syntactic Complexity Analyzer (L2SCA) [5] which automatically provides a number of measures based on the results of Tregrex tagging provided by the Stanford Parser [6]. Recently one NLP tool of interest is SpaCy, which has been shown to be both quicker and, in some cases, more effective than many other similar NLP tools at tagging noun types [7], and providing measures of lexical complexity [8]. However, SpaCy uses universal dependency (UD) tagging [9] as opposed to Tregrex, and therefore the measures of syntactic complexity that can be provided by SpaCy may differ both theoretically and in usefulness as compared to previous systems such as the L2SCA. Therefore, this paper aims to determine what sorts of meaningful syntactic complexity measures can be calculated with SpaCy, how comparable they are to Tregex-based measures, and how well they correlate to human-based writing assessments and general L2 proficiency.

2 Previous Studies

2.1 Syntactic Complexity and L2 Writing

Syntactic complexity has been linked to L2 writing ability and proficiency in a wealth of studies [5, 10, 11]. However, the measures that correlate the most can vary depending on factors such as task [4] and learner L1 [12]. Furthermore, the syntactic features that assist L2 students in academic writing are not always congruent with those that will result in them receiving higher human rating [2]. Researchers have been continuing to invent, refine, and compare various measures of syntactic complexity in a variety of contexts in order to determine which are most

universal and practical in certain situations.

2.2 Automatically Calculated Measures based on Tregex

One popular tool that has been created to automatically calculate several measures of syntactic complexity is the L2SCA [5]. It takes a folder of text files and runs them through the Stanford Parser and then analyzes the resulting Tregex tree sentence by sentence to provide nine different counts of syntactic elements and 14 transformed measures based on average length of syntactic units, ratios or counts of subordinate clauses, t-units, verb phrases, noun phrases, complex nominals, and subordination. The measures that this tool calculates have been shown to be rather reliable as evidenced by several studies which found them to be associated with L2 proficiency and writing ability, although there is variance regarding which measures are associated and to what degree [4, 5, 10, 12, 13].

2.3 Syntactic Complexity Measures Based on UD

It is not yet known how well automatically calculated measures of syntactic complexity based on UD syntax will correlate with L2 proficiency. Though various UD tags have been used for a more "fine-grained" analysis of syntactic complexity and demonstrated correlation with L2 writing proficiency, there are still many UD tags that have not been explored for L2 writing evaluation purposes, and the tags in initial works were human augmented, rather than purely automated calculations [10]. Here, the NLP tool SpaCy [1] could be potentially useful, as it has been shown to have superior parsing ability to similar tools such as NLTK [7, 8] and provides UD tagging, which could then be counted and used for calculating transformed measures similar to those of the L2SCA. However, simply counting UD tags does not allow for certain syntactic relationships to be observed as with Tregex parsing, and thus it is impossible to accurately measure the number of verb phrases or t-units with SpaCy in the same way that the L2SCA does. Thus, it is unclear exactly which measures of syntactic complexity SpaCy can be used to reliably calculate and how much these measures would be associated with L2 writing ability and general proficiency.

2.4 Research Questions

Based on the body of work introduced above, SpaCy should be able to calculate some measures of syntactic complexity similarly to the L2SCA, but it is unclear exactly which measures can be calculated based on UD tagging or how associated the measures produced with these tags will be with L2 ability. This paper thus seeks to offer new insights by answering the following research questions:

1. What measures of syntactic complexity can be calculated with SpaCy UD tagging, and how do these measures compare to the measures calculated from a Tregrex-based tool?
2. To what degree are the measures of syntactic complexity calculated by SpaCy associated with the L2 writing ability and general proficiency of L1 Japanese EFL learners?

3 Methods

3.1 Creating a UD Tag-based Tool

We created a SpaCy based Syntactic Complexity Calculator (SSCC) tool for automatically calculating syntactic measures using Python 3.9 that can run on both UNIX-based and windows-based machines. The user provides the SSCC with a set of text files, and the tool then parses each file using SpaCy, calculates measures based on the UD tags[i], and then exports the results into a csv file. However, since there is no Tregex pattern to follow, the SSCC can calculate only some of the same syntactic elements that the L2SCA does, while some must be counted in a different fashion, and others (i.e., T-units and verb phrases) become problematic or impossible to count. Based on SpaCy's outlined UD tagging, we designed the SSCC to calculate the following measures of syntactic complexity, as given in Table 1 and explained below.

Some measures that are theoretically the same or highly similar between the SSCC and the L2SCA include: W, S, C, CO, DC and CP. The SSCC calculates the number of words and sentences in almost an identical fashion to the L2SCA: the words are tokenized and then counted minus punctuation, and the number of sentences is calculated by counting the number of ROOT tags. The

[i] SpaCy UD tag explanation at universaldependencies.org

number of clauses in the SSCC is obtained by counting the number of subject tags, since every clause requires a subject. The SSCC determines the number of dependent clauses by counting UD tags for dependent clauses: ACL, RELCL, and ADVCL. The number of coordinates, which is similar, but not identical to coordinate phrases counted by the L2SCA, is counted as the number of coordinate tags: CC and CONJ.

Major differences between the SSCC and L2SCA are the former tools' inability to count T-units and verb phrases. Navigating Tregrex parsing allows the L2SCA to define and reliably count T-units, verb phrases and complex nominals by comparing their relationship to other syntactic elements within each sentence. However, the SSCC creates a string of UD tags, and thus counting certain syntax units only when they appear in relation to other units is not possible. Nevertheless, the UD tags do allow the counting of other syntactic units, such as modifications and compliments. Therefore, though the SSCC cannot count "complex" units in the same way as the L2SCA, it can provide a count of complex structures by summing the tags of all coordinates, compliments, modifications, etc. We programmed the SSCC to also give a count of all theoretically complex structures in a count called CSTR.

The SSCC then calculates transformations based on the UD tag counts. Following the L2SCA [5], the mean number of the various counts were divided by the syntactic units that the SSCC can count (e.g. sentences and clauses).

Table 1 SSCC Transformed Measures

Measure	Formula	Measure	Formula
MLS	W / S	DC_S	DC / S
MLC	W / C	DC_C	DC / C
C_S	C / S	CP_S	CP / S
CST_S	CSTR / S	CP_C	CP / C
CST_C	CSTR / C		

3.2 Participants and Writing Task

The used the same data set from a previous study [8]: 135 paragraphs written by 2nd year L1 Japanese university students on the topic of whether or not they thought tobacco should be made illegal in Japan. The participants performed the task under a 15-minute time constraint and the 135 paragraphs were obtained after excluding any responses that were under 50 words or written off-topic from a wider participant set. According to the results of students' TOEFL® ITP scores (383-677, M=519.5, SD=43.3), 12 should be considered CEFR A2 level, 84 should be considered B1, 38 should be considered B2 level, and one should potentially be considered C1 level [8, 14]. Five humans rated the paragraphs from 1 to 3 based on the fact that the students mostly belonged to one of three CEFR levels, and highly substantial agreement amongst raters was achieved; $\kappa = .74$, $p<.001$ [8].

3.3 Data Analysis

To compare the measures produced by the tools, we made pairwise Pearson's correlation analyses between those calculated by the L2SCA and the SSCC. The summed rater score was used as an ordinal measure of writing ability, and thus Spearman's correlation analyses were used to check for association between writing ability and individual measures of syntactic complexity given by the two tools. Pearson's correlation analyses were also used to check for association between the measures and TOEFL® ITP scores.

4 Results

4.1 Tool Comparison

The correlations between the measures produced by the SSCC and the L2SCA are shown in Table 2.

Table 2 SSCC Transformed Measures

Measure	Corr. (*r*)	Measure	Corr. (*r*)
W	.928**	C_S	.825**
S	.914**	DC_T/S	.640**
C	.899**	DC_C	.408**
DC	.626**	CP_T/S	.650**
CP	.683**	CP_C	.752**
CN/ST	.605**	CN/ST_T/S	.412**
MLS	.873**	CN/ST_C	.452**
MLC	.686**		

*$p<.05$, **$p<.01$

4.2 Association with L2 Proficiency

The correlations between writing and TOEFL® ITP scores and various measures provided by the SSCC and L2SCA are given in Table 3.

Table 3 SSCC Transformed Measures

Measure	Corr. to SSCC		Corr. to L2SCA	
	Writing	TOEFL	Writing	TOEFL
W	.672**	.333**	.657**	.339**
S	.272**	-.028	.303**	.014
T	N/A	N/A	.427**	-.013
VP	N/A	N/A	.475**	.288**
C	.378**	.164	.392**	.164
CT	N/A	N/A	.310**	.203*
DC	.305**	.352**	.185*	.195*
CP	.171*	.079	.146	.129
CN/CSTR	.481**	.250**	.338**	.311**
MLS	.343**	.336**	.282**	.280**
MLT	N/A	N/A	.140	.353**
MLC	.200*	.201*	.206*	.167
C_S	.169	.167	.126	.120
C_T	N/A	N/A	-.031	.178*
VP_T	N/A	N/A	.053	.295**
CT_T	N/A	N/A	.063	.202*
DC_T/S	.170*	.311**	.006	.153
DC_C	.157	.326**	.012	.167
CP_T/S	.189*	.028	.038	.067
CP_C	.151	.010	.037	.026
CN/ST_T/S	.275**	.206*	.073	.308**
CN/ST_C	-.029	.095	.088	.219*

5 Discussion and Conclusion

The results of this study suggest that some measures of syntactic complexity provided by the SSCC and L2SCA are quite comparable. The pure counts of words, sentences, and clauses are similar between the two tools, as expected, as are the levels of correlation between their respective calculated measures and both writing and TOEFL® ITP scores. The transformations of these variables, i.e., MLS, MLC and C_S, are also quite similar, although they show slightly less correlation than the pure counts, and the SSCC calculated measures show slightly higher correlation to human rating and TOEFL® ITP scores than those of the L2SCA. While both tools count dependent clauses and coordination, these counts and their transformations (i.e., DC_T/S, DC_C and CP_T/S or CP_C) show more variation than other measurements. These differences resulted in the SSCC's measures being generally more associated with L2 writing ability and proficiency than those of the L2SCA.

One of the largest drawbacks to the SSCC, as compared to the L2SCA, is that it is unable to count VPs and T-units. However, despite the theoretical importance of these measures and transformations based on them [5, 13, 15], calculating them seems only somewhat important for the data set in this study. Specifically, the number of T-units, VPs and CTs showed significant correlation to human rating, and VPs and CTs showed correlation to TOEFL® ITP scores, but most of the transformations based on them did not. For example, MLT, C_T, VP_T, and CT_T were correlated with TOEFL® ITP scores, but not human rating. Furthermore, though CN_T was correlated with TOEFL® ITP scores, so was the similar transformation CSTR_S given by the SSCC, which correlated to writing scores as well. Finally, the correlation between T and S, given by the L2SCA was r=.859, p<.001, and there was significant correlation between measures given by the SSCC divided by S as similar measures given by L2SCA divided by T (i.e., CP_T and CO_S, DC_T and DC_S). Therefore, it is questionable how important calculating T-units is for L2 syntactic complexity analysis, when transformations based on sentences seem to be just as, if not more, associated with greater L2 writing ability and general L2 English proficiency.

The results of this study vary from those of Lu [5], who found all of the transformations provided by the L2SCA showed significant differences in the average scores of three different level groups of L1 Chinese learners' essays. The discrepancies could be partially due to the fact that Lu [5] used ANOVA analyses to find average differences, whereas this study used Spearman's correlation tests to determine how closely each variable correlated with summed rater scores. Another possibility for the differences could be due to the task type - this study utilized single paragraphs rather than full essays.

In summary, both the SSCC and L2SCA provide some very similar measurements of syntactic complexity, but those given by the UD-counting system of the SSCC seem to be more associated with the L2 paragraph writing ability of L1 Japanese EFL learners. However, the L2SCA provides some measures that the SSCC does not which seem to be correlated to general L2 proficiency, but not writing ability. Therefore, which tool and measurements to use will likely vary depending on user intent.

References

[1] M. Honnibal and I. Montani, "spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017. [Online]. Available: https://spacy.io/

[2] S. A. Crossley and D. S. McNamara, "Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners," *Journal of Second Language Writing*, vol. 26, pp. 66–79, 2014.

[3] L. Ortega, "Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing," *Applied Linguistics*, vol. 24, no. 4, pp. 492–518, 2003.

[4] W. Yang, X. Lu, and S. C. Weigle, "Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality," *Journal of Second Language Writing*, vol. 28, pp. 53–67, 2015.

[5] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[6] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pp. 3–10. MIT Press, 2003.

[7] X. Schmitt, S. Kubler, S., J. Robert, M. Papadakis, and Y. LeTraon, "A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate," *Sixth International Conference on Social Networks Analysis, Management and Security,* pp. 338-343, 2019.

[8] R. Spring and M. Johnson, "The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK, and SpaCy tools," *System*, forthcoming, 2022.

[9] C. M. de Marnefe, C. D., Manning, J. Nivre, and D. Zeman, "Universal Dependencies," Computational Linguistics, vol. 47, no. 2, pp. 255–308, 2021.

[10] J. Jiang, P. Bi, and H. Liu, "Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus," *Journal of Second Language Writing*, vol. 46, pp. 1–13, 219.

[11] K. Wolfe-Quintero, S. Inagaki, and H. Y. Kim, *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*, University of Hawaii Press, 1998.

[12] X. Lu and H. Ai, "Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds," *Journal of Second Language Writing*, vol. 29, pp. 16–27, 2015.

[13] X. Lu, "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development," *TESOL Quarterly*, vol. 45, no. 1, pp. 36–62, 2011.

[14] Educational Testing Service, "Interpreting TOEFL® ITP scores," 17 December 2021. [Online]. Available: https://www.ets.org/toefl_itp/scoring/interpret/

[15] K. W. Hunt, "Do sentences in the second language grow like those in the first?," *TESOL Quarterly*, vol. 4, no. 3, pp. 195–202, 1970.