**Online Appendices for "A comparison of automated and manual analyses of syntactic complexity in L2 English writing"**

Quang Hồng Châu and Bram Bulté

*International Journal of Corpus Linguistics*

**Appendix 1. L2 English studies using L2SCA between 2017 and 2020 in five major journals (*Journal of English for Academic Purposes*, *Applied Linguistics*, *Journal of Second Language Writing*, *The Modern Language Journal*, and *International Journal of Applied Linguistics*)**

| Study | Corpus data | L2SCA measures | Use of *L2SCA* |
|---|---|---|---|
| Bi (2020) | A subset of a corpus of Chinese learners' compositions<br>Proficiency levels: beginner, intermediate, and advanced (classified based on grade levels)<br>Corpus size: 360 essays (120 for each proficiency level) – 63,678 word tokens | Length-based: MLC, MLT, MLS<br>Subordination: C/T<br>Coordination: T/S, CP/C<br>Phrasal complexity: CN/C | The choice of *L2SCA* for automated analysis was not discussed.<br>No preprocessing of the texts was discussed. |
| Bi and Jiang (2020) | Narrative essays by Chinese-speaking EFL learners<br>Proficiency levels: beginner and intermediate<br>Corpus size: 410 essays – 54,472 word tokens | Length-based: MLC, MLT, MLS<br>Subordination: C/T<br>Coordination: T/S, CP/C<br>Phrasal complexity: CN/C | The authors accounted for the use of *L2SCA* following Jiang et al. (2019), who reported the tool's acceptable reliability for analyzing lower-level learner data. |

| | | | |
|---|---|---|---|
| Casal and Lee (2019) | A subset of the Corpus of Ohio Learner and Teacher English (COLTE) comprised of source-based research papers<br>Proficiency levels: high (based on proficiency test and course results); categorized into 3 grade tiers<br>Corpus size: 280 papers – 387,994 words | Global: MLT<br>Clausal (coordination): T/S<br>Clausal (subordination): C/T<br>Phrasal: MLC, CN/C | The authors accounted for the use of *L2SCA* using the high reliability reported by Lu (2010) and Yoon and Polio (2017).<br>The data were formally cleaned (e.g. headers, footers, appendices) without any changes to student's language.<br>Several *Tregex* commands in *L2SCA* were used to retrieve specific complex nominal structures for the measurement of 5 fine-grained indices, besides the *L2SCA* measures. |
| Hwang et al. (2020) | Written and spoken data by Korean-speaking child EFL learners<br>Proficiency level: beginner<br>Corpus size: 769 written sentences (76 samples) and 401 spoken utterances (76 samples) | Length-based: MLS<br>Sentence complexity: C/S<br>Subordination: DC/T<br>Coordination: CP/T, T/S<br>Particular structures: VP/T | The *L2SCA* component in *TAASSC 1.3.8* (Kyle, 2016) was used. The choice of the tool for automated analysis was not discussed.<br>Production data were modified, with mistakes corrected (unspecified) and sentences/utterances containing Korean words excluded. |
| Jiang et al. (2019) | Narrative essays by Chinese-speaking EFL learners<br>Proficiency levels: beginner and intermediate, grouped into 4 levels of writing proficiency based on writing scores.<br>Corpus size: 410 essays – 54,472 word tokens | Length-based: MLC, MLT, MLS<br>Subordination: DC/C<br>Coordination: T/S, CP/C<br>Phrasal complexity: CN/C | For *L2SCA* analysis, run-on sentences were split into two or more sentences.<br>The authors carried out an evaluation of *L2SCA* to examine whether the tool is reliable for data with grammatical issues. Forty randomly-selected texts, 10 from each level, were both manually coded and automated by *L2SCA* for the 7 measures. High correlations between the results computed by the annotator and the tool were reported for all 7 measures, and the tool was deemed reliable for the analysis of their data. |
| Jin et al. (2020) | Text samples from English language teaching materials in China<br>Proficiency levels: 12 grade levels | Overall sentence complexity: MLS<br>Clausal coordination: T/S<br>Overall T-unit complexity: MLT<br>Clausal subordination: DC/T | The authors accounted for the use of *L2SCA* using the high reliability reported by Lu (2010) and Yoon and Polio (2017). The tool was also chosen because the *L2SCA* measures are representative of |

| | | | |
|---|---|---|---|
| | Corpus size: 3,368 texts across 12 grade levels – mean text length: from 22.16 words (grade 1) to 565.96 words (grade 12) | Elaboration at clause level: MLC<br>Phrasal coordination: CP/C<br>Noun phrase complexity: CN/C<br>Non-finite elements/subordination: NFE/C (calculated by subtracting 1 from the VP/C value computed by *L2SCA*) | various syntactic complexity subconstructs and suitable for teaching material evaluation. |
| Larsson and Kaatari (2020) | Subsets of the Advanced Learner English Corpus (ALEC), the Varieties of English for Specific Purposes dAtabase – Swedish component (VESPA-SE), and BNC-15 (subset of the British National Corpus)<br>Proficiency levels: advanced and native<br>Corpus size: 131 learner texts (1,079,178 words in total), and 2,400,899 words from the native English corpus. | All 14 *L2SCA* measures | The authors reported the high reliability suggested by Lu (2010). The authors excluded conversation data from the analysis since punctuation is interpreted differently, which could lead to parsing issues.<br>They also raised concerns about the validity of the tool's results since "the analyst is forced to trust the developer's (grammatical) judgment" (Larsson and Kaatari, 2020: 12). |
| Lu et al. (2020) | Corpus of Social Sciences Research Article Introductions (COSSRAI)<br>Proficiency level: Advanced (expert writers)<br>Corpus size: 600 texts - 513,688 words | MLS | *L2SCA* was not directly employed for automated analysis, but relevant *L2SCA* code was adapted to calculate 4 out of 5 measures. |
| Khushik and Huhta (2019) | Argumentative essays by Sindhi-speaking and Finnish-speaking EFL learners<br>Proficiency levels: A1, A2, and B1 CEFR (beginner to lower-intermediate)<br>Corpus size: about 1,150 texts | All 14 measures | The corpus was cleaned before *L2SCA* analysis after potential issues were examined. Texts of fewer than 10 words, written in L1, or copied were excluded.<br>Minor spelling errors (unspecified) were corrected, missing sentence-final punctuation was added, and learners' comments were removed.<br>Linguistic errors apart from spelling errors were |

| | | | |
|---|---|---|---|
| | | | not corrected. It is pointed out that missing sentence-final punctuation affected all sentence length-based measures. |
| Kyle and Crossley (2018) | TOEFL argumentative essays<br>Proficiency levels: Scores range from 1 to 5. This could represent beginner to advanced writing proficiency levels.<br>Corpus size: 480 essays – 151,490 words | All 14 measures. However, only a selection of measures meeting certain statistical criteria was included for further analysis. | *L2SCA* was chosen to compute traditional, coarse-grained syntactic complexity measures, as opposed to fine-grained clausal and phrasal indices, computed by *TAASSC*.<br>The texts were not preprocessed even though lower-score texts could have "an accumulation of errors in sentence structure and/or usage" as specified in the TOEFL scoring rubric (Kyle & Crossley, 2018: 5). The same unprocessed texts were also analyzed by *TAASSC*. |
| Polat et al. (2019) | Descriptive paragraphs by Turkish EFL learners<br>Proficiency levels: elementary, pre-intermediate, and intermediate<br>Corpus size: 852 paragraphs (284 for each level) – 115,869 words | All 14 measures | The authors accounted for the use of *L2SCA* using the high reliability reported by Lu (2010) and Yoon and Polio (2017), although acknowledging that "like other research tools, the use of *L2SCA* is not without controversy" (Polat et al., 2019: 9).<br>The texts were not preprocessed before *L2SCA* analysis. |
| Wu et al. (2020) | Research articles from the SciELF corpus and the Corpus of Contemporary American English (COCA)<br>Proficiency levels: advanced and native<br>Corpus size: 300 texts (150 from each corpus) – 1,442,006 word tokens | All 14 measures | The choice of *L2SCA* for automated analysis was not discussed. |

**Appendix 2. Guidelines for manual annotation**

- Formulaic non-sentences are not annotated: *Hi!, The end*.

**1. Word:** a string of alphanumeric characters delimited by spaces or a space and a punctuation mark

Examples:

- *I like my new school, class, teachers en friends*: 9 words.

Exceptions:

- A wrongly joined or split word form is counted as the number of words in its correct form: *eachother* as two words, *every body* as one word.
- A contraction is counted as the number of words in its expanded form: *isn't* as two words, *can't* as one word.
- A proper noun is counted as one word: *New Zealand* as one word.
- A possessive marker is not counted as a separate word: *my friend's* as two words.

**2. Sentence:** a group of words delimited by sentence-final punctuation marks (period, question mark, exclamation mark, quotation mark, or ellipsis)

Examples:

- [Sentence I was skiing and it was very good snow.] [Sentence Better than ever!]

Exceptions:

- A run-on sentence is reanalyzed as two or more sentences, separated at the place of the missing conjunction or missing/inappropriate punctuation: [Sentence That was everything what I wanted to tell about XX] [Sentence I really want to recommend this holiday to you.]

**3. T-unit:** an independent clause together with its dependent clauses. A sentence fragment punctuated by the writer with no overt verb is also considered a T-unit.

Examples:

- [T-unit My favorite subject is Music because I love making music] and [T-unit I also like Dutch].
- [T-unit Better than ever].

Exceptions:

- T-units in a run-on sentence are annotated in accordance with how the run-on sentence is reanalyzed as sentences: [$_{\text{T-unit}}$ That was everything what I wanted to tell about XX] [$_{\text{T-unit}}$ I really want to recommend this holiday to you].

**4. Clause:** a structure with a subject and a finite verb. A sentence fragment punctuated by the writer with no overt verb is also considered a clause.

Examples:

- [$_{\text{Clause}}$ It would be a kind of present to my parents and sister to thank them] for [$_{\text{Clause}}$ what they did for me].
- [$_{\text{Clause}}$ Better than ever!].

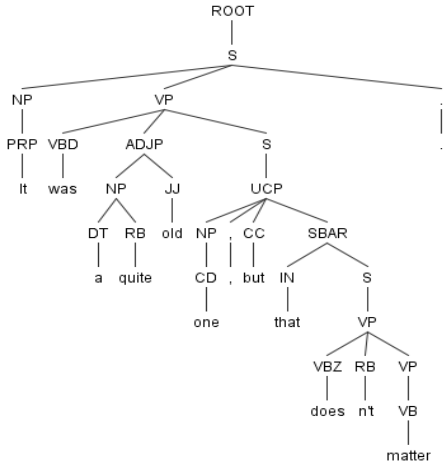
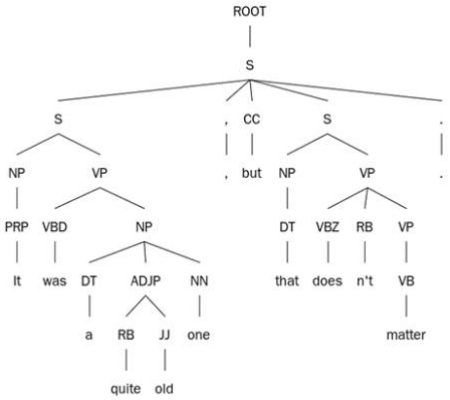
**5. Dependent clause:** a finite adjective, adverbial, or nominal clause

Examples:

- It would be a kind of present to my parents and sister to thank them for [$_{\text{Dependent clause}}$ what they did for me].

## Appendix 3. Parsing errors

| Problem | Example | Difference between *TAASSC* and *L2SCA* |
|---|---|---|
| **Tokenization and sentence segmentation** | | |
| Periods segmented as sentences | [$_{S1}$ the most children are going to the aula.] [$_{S2}$.]<br>The two periods at the end of the sentence were tokenized as two tokens by the system, leading to the identification of the second period as a separate sentence. | *L2SCA* identified two periods at the end of a sentence as a single token, leading to correct sentence segmentation. |
| Exclamation mark-separated sentences identified as one sentence | [$_{S}$ we walked for about 4 hours!! I didn't liked it but I have to came with my mom and dad.]<br>The token *!!* was not identified as sentence-final punctuation, leading to the two sentences together identified as one. Multiple exclamation marks were in general not identified as sentence-final punctuation by the system. | No |
| Ellipsis-separated sentences identified as one sentence | [$_{S1}$ [$_{S2}$ But half the year he had to go, because he get a new job] [$_{:}$ …] [$_{S3}$ we really didn't like that]]<br>The ellipsis was not identified as sentence-final punctuation, leading to the identification of S2 and S3 as one single sentence (S1). | No |
| Quoted material identified as multiple sentences | [$_{S}$ Then someone said "I know why the chair collapsed.] [$_{S}$ He is to fat!"]<br>The sentence was segmented into two sentences at the position of the period inside the quotation marks. | No |
| **POS tagging** | | |

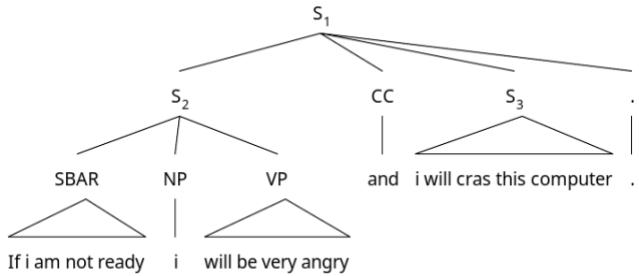
| | | | |
|---|---|---|---|
| Incorrect POS tags |  | *that* was incorrectly tagged as IN (subordinating conjunction), leading to the clause *that doesn't matter* identified as a subordinate clause (SBAR) instead of a coordinate clause. | Fewer POS tagging errors were observed for *L2SCA*. For instance, *that* in the same sentence was correctly tagged as DT (determiner) by *L2SCA*.<br>  |

**Syntactic parsing**

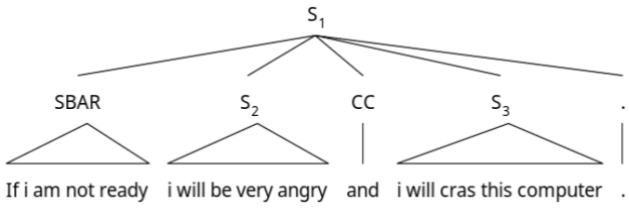
The following syntactic parsing errors were observed for both *TAASSC* and *L2SCA*, which employ the englishPCFG grammar for syntactic parsing. The parsing errors were observed in compound-complex sentences. It was noticed that most parsing errors were resolved with the use of englishFactored or englishRNN grammars.

Clause attachment errors

Parser output:



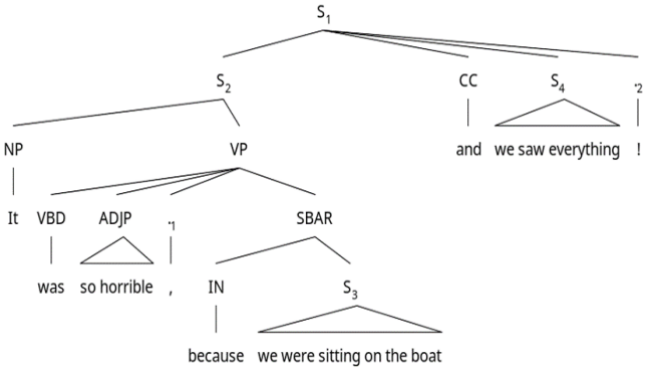

Gold tree:



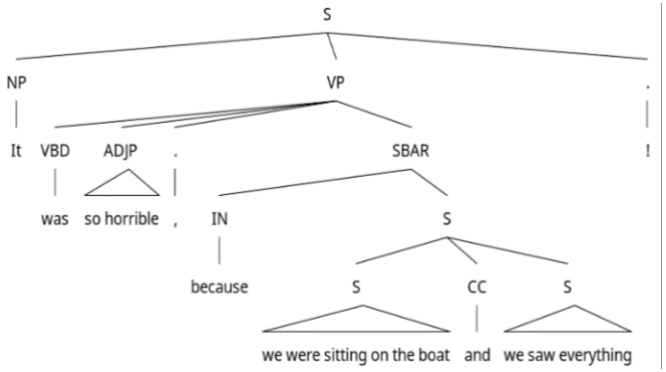

In the parser output, the SBAR node is located too low in the tree. To transform the parser output into the gold tree, this node needs to be moved out of the S2 node and attached to the immediate left of the new S2 node.

Coordination errors

Parser output:




Gold tree:




In the parser output, the S4 node is located too high in the tree. To transform the parser output into the gold tree, this node, together with its immediate left sister (CC), needs to be moved to under the SBAR node to the immediate right of the S3 node.

**Appendix 4. Learner errors causing system-annotator disagreement**

| Learner error | Level 1 | Level 2 | Level 3 | Level 4 | Total |
|---|---|---|---|---|---|
| Incorrect punctuation | 16 | 10 | 15 | 10 | 51 |
| Missing punctuation | 10 | 7 | 8 | 26 | 51 |
| Spelling error | 10 | 12 | 7 | 3 | 32 |
| Capitalization error | 9 | 8 | 1 | 3 | 21 |
| Missing verb | 5 | 3 | 2 | 4 | 14 |
| Missing pronoun | | | 7 | | 7 |
| Missing noun | 2 | 1 | 1 | 1 | 5 |
| Unnecessary punctuation | 1 | 1 | 1 | 1 | 4 |
| Unnecessary conjunction | 2 | | 1 | 1 | 4 |
| Wrong order | 3 | 1 | | | 4 |
| Missing conjunction | | | 2 | 2 | 4 |
| Wrong verb form | 3 | 1 | | | 4 |
| Wrong noun form | | 2 | | | 2 |
| Derivation of noun error | | 1 | | | 1 |
| Incorrect adverb | | 1 | | | 1 |
| Incorrect conjunction | | | 1 | | 1 |
| Replace | | | 1 | | 1 |
| Missing preposition | | | | 1 | 1 |
| Unnecessary verb | | 1 | | | 1 |
| Unnecessary pronoun | 1 | | | | 1 |
| Wrong determiner form | | | 1 | | 1 |
| Total | 62 | 49 | 48 | 52 | 211 |

## Appendix 5. Most frequent learner errors and their impact on syntactic parsing

| Error type | Example | Impact on syntactic parsing |
|---|---|---|
| **Punctuation errors** | | |
| Comma splices | I don't want to work at an office, I really hate that.<br>*Parser output:*<br>**[**S **[**S I don't want to work at an office] [, ,] [NP I] **[**ADVP really] [VP hate that]] | • For comma splices, *TAASSC*/*L2SCA* identified one T-unit and one sentence, as opposed to two T-units and two sentences by the annotators. |
| Fused sentences | I screamed to him he deserves it to be screwed.<br>*Parser output:*<br>**[**S [NP I] [VP screamed to him **[**SBAR he deserves it to be screwed]]] | • The *Stanford parser* was inclined to identify the second independent clause in a fused sentence as a subordinate clause (tagged as SBAR) nested inside the predicate of the first clause.<br>• *TAASSC*/*L2SCA* identified sentences such as the example as containing one T-unit and one sentence, as opposed to two T-units and two sentences by the annotators. |
| Missing space after sentence-final punctuation | It was at the start at school.I had got an one.<br>*Parser output:*<br>[S [NP It] [VP was at the start at [NP school.I] had got an one]] | • The *Stanford parser* identified sentence-final punctuation together with the words right before and after the punctuation as one token, leading to two sentences together being identified as one sentence by *TAASSC*/*L2SCA*. |
| Missing commas | (1) When I was running everything fell out.<br>*Parser output:*<br>[SBAR when I was running everything] [VP fell out]<br>(2) After a few days I met XX, XX, XX and XX and XX so I ended up with a lot of friends.<br>*Parser output:*<br>**[**S After a few days I [VP met XX, XX, XX and XX and XX [SBAR [IN so**]** I ended up with a lot of friends]]] | In several cases, missing commas led to structural misanalysis:<br>• Missing commas led to clause boundaries being misidentified, as in (1). The noun *everything* was identified as the object of the verb *running* rather than as the subject of the second clause.<br>• Missing commas before the coordinating conjunction *so*, as in (2), led to the identification of the *so*-clauses as dependent clauses rather than independent clauses. The *Stanford parser* labels *so* as |

(3) Then I would buy myself somethings I really need and maybe I buy myself a house or a car or something.

*Parser output:*

[$_{S1}$ Then I would buy myself some things [$_{SBAR}$ [$_{S2}$ [$_{S3}$ I really need] and [$_{S4}$ maybe I buy myself a house or a car or something]]]]

IN, which can function either as a coordinating or subordinating conjunction; the comma before *so* as a coordinating conjunction can facilitate correct parsing.

- Missing commas before the coordinating conjunction *and* when the conjunction followed a relative clause, as in (3), led to the identification of the *and*-introduced coordinating clause as a subordinate clause.

**Spelling errors**

Than we saw a boat with fishers.

*Parser output:*

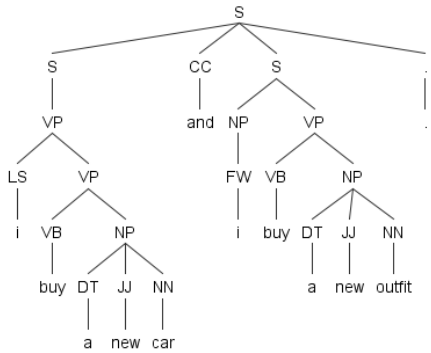
**[**$_{FRAG}$ **[**$_{SBAR}$ [$_{IN}$ than**]** [$_{S}$ we saw a boat with fishers**]].]**

- The most frequent spelling errors were *than* for *then* (7), *en/an/end/sand* for *and* (7), *al* for *all* (2) and *its* for *it's* (2).
- Most spelling errors that affected syntactic parsing were functional words located at clause boundaries. The misspelling of *then* (POS tag RB or adverb) as *than* (POS tag IN or subordinating conjunction) in the example led to the sentence being identified as a fragment by *TAASSC*/*L2SCA*.
- Most misspellings, especially minor errors, of lexical words did not lead to POS tagging issues, e.g. *diffucult* POS tagged as JJ (adjective), the same tag as its hypothesized correct form *difficult*.

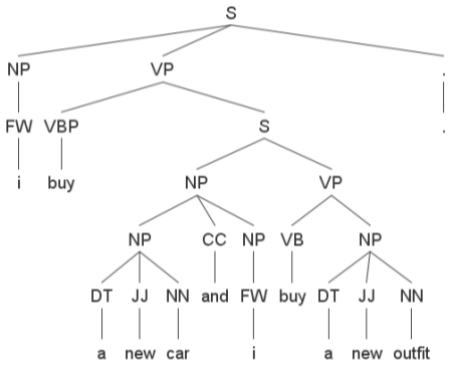
**Capitalization errors**

i buy a new car and i buy a new outfit.

*Parser output:*




- A large number of uncapitalized first-person singular personal pronouns, POS tagged either as LS (list item marker) or FW (foreign word), led to structural misanalysis. In the example, *buy* was POS tagged as VB (base-form verb), probably because the subjects of the clauses were incorrectly POS tagged; as a result, no clauses were recognized by *TAASSC*.
- The output by *L2SCA* in case of uncapitalized *i* was different from that by *TAASSC*. For the example

| | | |
|---|---|---|
| | | sentence, both uncapitalized pronouns were tagged FW, leading to a different parse. However, in general, fewer structural misanalyses were observed for *L2SCA*.<br>*Output by L2SCA:*<br>  <br><br>• It should be noted, however, that uncapitalized *i* is tagged as PRP (personal pronoun) in the most updated versions of the *Stanford parser (version 4.2.0)* and *CoreNLP (version 4.2.0)*. |
| **Missing verbs** | After I done things to entertain myself<br><br>*Parser output:*<br><br>[SBAR [IN after] [S [NP I] [VP done things to entertain myself]]] | • Most missing verbs in the corpus were auxiliary verbs, which are essential for clauses to be matched using the *Tregex* pattern defined by Lu. In the example, the missing auxiliary verb *have* led to the clause not being recognized by *TAASSC*/*L2SCA*. |

## References


Bi, P. (2020). Revisiting genre effects on linguistic features of L2 writing: A usage-based perspective. *International Journal of Applied Linguistics*, *30*(3), 429–444. https://doi.org/10.1111/ijal.12297

Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, *91*, 102248. https://doi.org/10.1016/j.system.2020.102248

Casal, J.E., & Lee, J.J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005

Hwang, H., Jung, H., & Kim, H. (2020). Effects of written versus spoken production modalities on syntactic complexity measures in beginning-level child EFL learners. *The Modern Language Journal*, *104*(1), 267–283. https://doi.org/10.1111/modl.12626

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, *46*, 100666. https://doi.org/10.1016/j.jslw.2019.100666

Jin, T., Lu, X., & Ni, J. (2020). Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *The Modern Language Journal*, *104*(1), 192–208. https://doi.org/10.1111/modl.12622

Khushik, G.A., & Huhta, A. (2019). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Kyle, K., & Crossley, S.A. (2018). Measuring syntactic complexity in L2 writing using finegrained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, *45*, 100850. https://doi.org/10.1016/j.jeap.2020.100850

Lu, X., Casal, J.E., & Liu, Y. (2020). The rhetorical functions of syntactically complex sentences in social science research article introductions. *Journal of English for Academic Purposes*, *44*, 100832. https://doi.org/10.1016/j.jeap.2019.100832

Polat, N., Mahalingappa, L., & Mancilla, R.L. (2019). Longitudinal growth trajectories of written syntactic complexity: The case of Turkish learners in an intensive English program. *Applied Linguistics*, *41*(5), 688–711. https://doi.org/10.1093/applin/amz034

Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a Lingua Franca academic writing. *Journal of English for Academic Purposes*, *43*, 100798. https://doi.org/10.1016/j.jeap.2019.100798