THE INFLUENCE OF RATING EXPERTISE AND LANGUAGE BACKGROUND ON JUDGMENTS OF SYNTACTIC COMPLEXITY IN L2 WRITING IN ITALIAN

Ineke Vedder[1]

1. INTRODUCTION

Syntactic complexity has often been considered a potential index of second language (L2) development and proficiency (Norris, Ortega, 2009; Kyle *et al.*, 2021; O'Leary, Steinkrauss, 2022). Defined in terms of the degree of variation, sophistication and structural elaboration (Bulté, Housen, 2012), syntactic complexity has been assessed both by quantitative measures as well as by judgments expressed on holistic rating scales. L2 writing studies have traditionally employed measures for overall complexity targeting the mean length of syntactic units (sentence, T-unit, AS-unit, clause) or the ratio of subordinate structures (clauses per T-unit, dependent clauses per clause; cf. Lu, 2010, 2011). More recent studies have, however, increasingly questioned the use of these global measures and they have advocated for a multidimensional approach to measuring syntactic complexity recommending the use of more fine-grained and usage-based measures (Biber, Gray, 2011; Yoon, 2017).

In other research, particularly in language testing and assessment, syntactic complexity has been measured through holistic rating scales used by human judges (Knoch, 2009; Pill, Smart, 2021). As has been observed in some of these studies, a possible source of variability that may influence raters' judgments on L2 performance are individual characteristics of raters, such as instructional and professional training, age and gender. Other factors that have been found to have an impact on raters' decisions are rating expertise and language background, on which the current study focuses (Bonck, Ockey, 2003; Kim, 2015). Research on the effects of these individual rater traits mostly concerns L2 speaking, whereas few studies have focused on L2 writing, specifically syntactic complexity.

This study examines the influence of rating expertise and language background on judgments of syntactic complexity in L2 academic writing in Italian. The main reason for conducting the research is that most studies that have investigated the development of syntactic complexity in L2 have been grounded in hypothesis-testing. Few researchers, however, have employed a bottom-up approach by exploring, from a pedagogic perspective, reflections of stakeholders in the language classroom.

Based on a quantitative and a qualitative analysis, the paper compares the complexity ratings provided by three groups of raters: language teachers whose first language (L1) is Italian, L2 learners of Italian (L1 speakers of Hungarian and Finnish), and L1 Italian university students. The comparison also includes the motivations and reasoning behind their scores and their feedback. An additional aim of the study is to explore the extent to

[1] Amsterdam Center for Language and Communication (ACLC), University of Amsterdam.

which raters' reflections are related to current views in the field of SLA (Second Language Acquisition) on the development of syntactic complexity in L2.

The paper first summarizes the main findings from the literature concerning the assessment of syntactic complexity in L2 by means of complexity measures and through teacher judgments. Subsequently, some studies on the impact of rating expertise and language background on raters' judgments of L2 performance are reviewed. The second part describes the experimental design, the methodology and the main outcomes of the study. Finally, the theoretical and pedagogical implications for L2 writing research and classroom practice are discussed.

2. ASSESSMENT OF SYNTACTIC COMPLEXITY

Syntactic complexity, as a potential predictor of L2 proficiency, has been approached from many theoretical angles addressing a wide range of learners with different target and source languages. An important concern of this strand of research has been how syntactic complexity develops over time and interacts with lexical complexity, accuracy and fluency (Housen *et al.*, 2012; Vyatkina, 2012; O'Leary, Steinkrauss, 2022). Recently, an increasing number of studies have also investigated morphological complexity (Brezina, Pallotti, 2019), phraseological complexity (Paquot, 2019), and propositional complexity (Vasylets, Manchón, 2019).

Syntactic complexity has been hypothesized to develop in three partly overlapping stages, reflecting the increase of L2 proficiency (Norris, Ortega, 2009). Complexity by coordination comes first followed by an increase of subordinate structures. At more advanced levels of L2 proficiency, phrasal complexity referring to the expansion and elaboration of the noun group has been observed to develop, whereas clause-level complexity (e.g. subordination) decreases (Wolfe-Quintero *et al.*, 1998; Pallotti, 2009). Higher proficiency writing in L2, particularly academic writing, is characterized by the use of longer and more diverse noun phrases and less clausal subordination when compared to speech and is more structurally "compressed" with phrasal (non-clausal) modifiers embedded in noun phrases (Biber *et al.*, 2011, 2016).

Whereas in language testing and assessment research syntactic complexity is usually measured through holistic rating scales (Koch, 2009; Pill, Smart, 2021), in L2 writing syntactic complexity has often been assessed by quantitative indices, focusing on the average length of syntactic units or the subordination ratio. Examples of measures of overall syntactic complexity are mean length of utterance, T-unit (C-unit, AS-unit)[2]. Measures for assessing subordination are number of subordinate clauses; number of clauses per T-unit; number of subordinate clauses per clause, dependent clause or T-unit (C-unit, AS-unit). Complexity by coordination has been measured, although to a lesser extent, by Bardovi Harlig's Coordination Index (Bardovi Harlig, 1992).

Criticisms were, however, soon raised against this reductionist approach of complexity indicating the risk of construct underrepresentation by using "coarse" overall complexity measures, and gaps and imbalances to complexity measurement in L2 research were identified (Norris, Ortega, 2009; Pallotti, 2009, 2015; Bulté, Housen, 2012). Other studies, conversely, have employed different complexity measures, tapping sometimes into the same sub-dimensions of the construct, as pointed out by Norris and Ortega (2009).

[2] The term T-unit refers to a main clause plus all subordinate and non-clausal structures that are attached to it. Typically, T-unit analysis is used for written language, whereas the C-unit and the AS-unit, closely related to the T-unit, are applied to spoken language.

In response to these criticisms, more fine-grained measures have been suggested addressing other syntactic levels, for instance the ratio of verb and noun phrases per T-unit or the number of pre-modifying and post-modifying noun phrases at the phrasal level (Biber, Gray, 2011; Biber *et al.*, 2011), or the use of measures that distinguish nominal subordination from subordination via subject/object relative clauses (Pallotti, 2009). Additionally, various automated tools have been proposed targeting the frequency of syntactic features, such as number/type of prepositional clauses, predicative or adverbial adjectives and conjunctions (Lu, 2010, 2011; Kyle, Crossly, 2018).

What remains to be seen is, however, the degree to which these complexity measures can function as an index of L2 proficiency (and more specifically, L2 writing), how valid and reliable they are, and to what extent they are redundant. Pallotti (2015, 2021) recommended, for reasons of simplification, to select only three of the many measures in this area, namely, the average number of words per phrase, number of phrases per clause, and number of clauses per higher-order syntactic unit (e.g. sentence, T-unit). In a similar vein, Bulté, Housen and Pallotti (2024) proposed, in a recently published theoretical and methodological overview of complexity measures in SLA, to establish a small set of measures to be used, in the interest of replicability and comparability of studies.

3. TEACHER JUDGMENTS OF SYNTACTIC COMPLEXITY

Differently from the numerous experimental studies grounded in hypothesis testing, fewer researchers have employed a bottom-up approach examining judgments and reflections of teachers (and/or learners) on various aspects of L2 learning. Although some studies have adopted a pedagogical perspective, directly or indirectly addressing teachers' reflections (Révész, Gurzynski-Weiss, 2016; Ait Eljoudi, 2018; Cucinotta, 2018), only a few have focused specifically on the syntactic complexification of L2 writing.

A study by Ågren, Granfeldt and Schlyter (2012) focused on the correlation between teachers' judgments of morphosyntactic development in L2 French and the automated profile analysis of the computer tool *Direkt Profil* (Granfeldt *et al.*, 2006). The results indicated relatively high degrees of correlations and showed that the analysis of developmental stage by *Direkt Profil* could explain 73% of variance. Another outcome was that teachers agreed with each other and with the computer tool when assessing morphosyntax.

In two studies by Kuiken and Vedder (2014, 2019b), teacher judgments of linguistic complexity by language teachers of, respectively, Italian and Dutch were examined based on a sample of argumentative texts written by L2 learners of Italian and Dutch (levels A2-B2). In Kuiken and Vedder's 2014 study, which involved four language teachers of Dutch and three of Italian, teachers' reflections on linguistic complexity (and functional adequacy) in Dutch and Italian L2 writing were explored in two retrospective panel discussions. In the 2019 study, including 16 teachers of Italian and 11 of Dutch, perceptions of syntactic complexity were specifically addressed. All participants were first language (L1) speakers of one of the two target languages. Teachers were first asked to evaluate individually the complexity of a sample of written texts on a six-point Likert scale. During the subsequent panel discussion, each of them articulated the motivations behind the assigned scores and the feedback they would offer to the writer.

Both studies showed that teachers tended to focus primarily on accuracy and to a lesser degree on comprehensibility, rather than on linguistic complexity. With respect to syntactic complexity, positive features to which they often referred were well-formed and complex but smoothly constructed sentences, use of relative clauses, verbal agreement

and appropriate use of grammatical connective devices. Occurrence of long and complex sentences affecting the readability of the text was viewed negatively.

4. RATING EXPERTISE AND LANGUAGE BACKGROUND

4.1. *Rating expertise*

Research on the role of rating expertise on judgments of L2 performance – one of the two variables of rater variability on which our study focuses – mostly regards the assessment of L2 speaking. No studies, so far, have specifically investigated the effects on judgments of syntactic complexity in L2 writing.

Studies on the influence of rating expertise on judgments of oral L2 performance have focused on the possible impact on internal rating consistency and on raters' severity or leniency. In a large-scale study on oral L2 proficiency conducted among 1000 Japanese English Foreign Language (EFL) learners, Bonk and Ockey (2003) observed large differences in rater severity and consistency between expert and non-expert raters. Experienced raters demonstrated greater severity and consistency, while novel raters were more lenient and showed much more inconsistency. Other studies have concentrated on interaction effects with other factors involved in the rating process which may affect raters' decisions, such as instructional context, task effects, type of learners (Lumley, 2005).

Kim's 2015 study involved nine Master's students and graduates in Applied Linguistics with varying educational backgrounds and rating experiences (*novice*, *developing* and *experienced* raters). These participants were asked to evaluate 18 EFL learners' oral responses on an analytic rating scale. Novice raters were defined as those who were entirely new to the field. Developing raters were those who had recently started a professional career and had some experience in teaching or rating. Experienced raters were those with an extensive background in teaching and rating of at least five years. Recorded verbal report data were analyzed to compare rating behavior and to examine the development of rating performance over time. The analysis revealed that the three groups presented varying levels of rating ability and different paces of improvement. The rating performance of the novice raters improved very slowly, and rater training seemed to have little effect. The developing raters were found to benefit most from repeated training and practice. The experienced raters showed high internal consistency throughout the sessions and needed little training.

Also, in Duym *et al.*'s (2018) study differences in rating behavior were observed between expert and non-expert raters. The purpose was to investigate whether experts and non-experts demonstrated similar responsiveness to fluency and linguistic accuracy in Dutch L2 speech. 55 linguistically trained raters and 41 non-trained raters (i.e., potential stakeholders, in an occupational context) holistically assessed 68 stimuli. The results showed that the expert raters rewarded linguistic accuracy, particularly morpho-syntactic accuracy, relatively higher than the non-experts. This effect was explained by the finding that compared to the novel raters, the linguistically trained raters seemed to be more preoccupied with errors, according to their responses to a questionnaire.

Although the results of these studies, in which various types of raters were involved and different individual traits were investigated, indicate that assessment of L2 proficiency appears to be impacted by rating expertise, it is difficult to compare the findings. As most studies, furthermore, have focused on oral performance, it is unclear to what extent the outcomes apply to L2 writing and, specifically, to the assessment of syntactic complexity.

What this type of research, nonetheless, demonstrates is the importance of rater training. As Pill and Smart (2021: 36) point out, «training is necessary to establish appropriate rater behavior and thereby develop rating expertise».

4.2. *Language background*

Another source of variability on which this paper concentrates is the influence of raters' language background (e.g. status of L1 or L2 speaker of the target language) on judgments of syntactic complexity in L2 writing. Studies in which the effects of linguistic background were examined have focused mostly on L2 speaking (particularly accent familiarity), similarly to the research on rating expertise discussed above. Only few researchers (Kuiken, Vedder, 2014, 2019b) have addressed the impact of raters' language background on judgments of syntactic complexity in L2 writing.

Bogorevich (2018), in a study on accent familiarity, examined the potential differences between raters with English as their L1 vs L2 English-speaking raters in how they assessed L2 students' oral performance in three different target languages. Two groups of raters, 23 North American and 23 Russian raters were asked to grade speech samples from learners with an Arabic (n = 25), Chinese (n = 25), and Russian (n = 25) L1 background. A subset of 16 raters (seven L1 vs nine L2 English raters) shared their scoring behavior through think-aloud protocols and interviews. The results revealed that although L1 and L2 raters overall behaved similarly, different rating patterns were observed. An interesting finding was that L2 English raters tended to be more lenient towards the students with whom their own L1 (i.e., Russian) matched.

Differences in rater behavior that could be ascribed to language background were found also by Winke *et al.* (2013). Based on evidence that listeners may favor certain foreign accents over others and that raters may better comprehend and/or rate the speech of test-takers whose first languages they are more familiar with, the study investigated whether accent familiarity may lead to a major leniency. The results indicated that when raters are familiar, to varying degrees, with the test-takers' L1, this may influence their ratings.

A study by Zhang and Elder (2011) addressed the question of whether holistic judgments of language proficiency by L1 speaking English raters correspond to those of L2 raters. Data for the study were derived from two sources: holistic ratings given by a group of 19 L1 English and 20 L2 English teachers to speech samples from 30 EFL Chinese test-takers, as well as written comments to justify the ratings assigned. Results revealed no significant difference in raters' judgments of the samples. The qualitative analysis of their comments, however, showed several differences in the way L1 and L2 teachers justified their scores.

Kuiken and Vedder (2014), in a cross-linguistic study on the relationship in L2 writing between linguistic complexity (syntactic complexity, next to lexical complexity) assessed on a 7-point Likert scale and functional adequacy, examined syntactic complexity based on a learner corpus (proficiency level A2-B2) for Italian, Dutch and Spanish collected by Kuiken *et al.* (2010). Data were assessed by language teachers with either Italian or Dutch as their L1. Variation in the gradual syntactic complexification of the L2 texts was found across proficiency levels and languages. Advanced Italian L2 learners (L1 Dutch) used more coordinate structures within T-units, more relative clauses, and longer post-modifying noun groups, whereas this was not the case for Dutch L2 learners (of various native languages) and neither for Spanish. An interesting finding that emerged from the retrospective interviews and panel discussions with two groups of L1 teachers of Italian

and Dutch were their different perspectives on syntactic complexity. Similar results were observed in Kuiken and Vedder's 2019 study, discussed in section 3 (Kuiken, Vedder, 2019b). Whereas the Italian teachers often referred to specific syntactic constructions as being overly basic and viewed a lack of variation in sentence structure negatively, the Dutch raters advised the writers to employ short and simple sentences. This shows that typological differences between a Romance language (Italian) and a Germanic language (Dutch) in word order, sentence structure and construction of the verb phrase may influence raters' perceptions and preferences of syntactic complexity.

What this strand of research suggests is that language background, status of L1 vs L2 speaker, familiarity with the target language, and differences between source and target language, may lead to L1-based score justifications of raters, preferences for specific L2 features and to major leniency or severity. It is unclear how far these outcomes, mostly derived from studies on L2 speaking, also apply to the evaluation of syntactic complexity in L2 writing, or which syntactic areas may specifically be influenced by raters' linguistic background.

5. Design and methodology

5.1. *Goal and research questions*

The current study aims to investigate by means of a quantitative and a qualitative analysis the influence of rating expertise and language background on their judgments and score motivations of syntactic complexity in L2 Italian writing. The study is an expansion of earlier research conducted by Kuiken and Vedder (2019b). The paper compares the L1 teacher data (*experts*) for Italian gathered by Kuiken and Vedder with new data collected from L1 and L2 student raters (*non-experts*): a group of L1 speakers of Italian and two groups of L2 learners of Italian, with Finnish or Hungarian as their native language. An additional purpose is to explore to what extent perspectives of expert and non-expert raters on syntactic complexity reflect the development of syntactic complexity hypothesized in the L2 literature (cf. section 2).

The following research questions were formulated:

RQ1: How do judgments and score motivations of experts (language teachers) on syntactic complexity align with ratings by non-experts (students)?

RQ2: To what extent are raters' judgments and score motivations on syntactic complexity influenced by language background (L1 vs L2)?

RQ3: To what extent are raters' judgments related to current views in the field of SLA regarding the hypothesized development of syntactic complexity in L2?

5.2. *Participants*

The data from previous research by Kuiken and Vedder (2019b) were collected among a group of 16 native Italian language teachers at an Italian university. Raters were asked to assess on a six-point Likert scale the syntactic complexity of six argumentative texts written by intermediate Dutch university students of Italian (proficiency level A2-B2), with Dutch as their native language. For each text the teachers had to indicate for which reasons they had decided to assign a particular score and which feedback they would give to the writers to improve the text. In the subsequent panel discussion, teachers discussed their motivations behind the assigned scores and the feedback they had proposed.

This study revisits the earlier teacher data collected by Kuiken and Vedder (2019b) and compares the judgments, score justifications, and suggestions for feedback with ratings from 38 non-expert raters, university students with Italian as their second or first language. The L2 raters (n = 18) were (high)intermediate Bachelor's students of Italian, with a proficiency level between B2 and C1. Data were collected in two distinct linguistic contexts: Finland (n = 10) and Hungary (n = 8). We contrasted the L2 findings with the judgments of a group of 20 L1 Italian Bachelor's or Master's students from various universities in Italy, studying in diverse fields such as Social Sciences, Linguistics, History, Architecture, Law, and Biology[3].

Originally, the experiment was designed to elicit data on location. However, the COVID-19 pandemic and the impossibility of traveling abroad to gather data onsite made it necessary to collect data online. Next to the teacher data from Kuiken and Vedder (2019b) and the new student data already collected online in Finland and Hungary, we had planned to gather also data from Dutch L2 teachers and L2 students of Italian in the Netherlands. Unfortunately, COVID-19 made this impossible.

Despite some minor adaptations, the organization of the online rating sessions and the procedures were identical to the ones followed by Kuiken and Vedder (2019b; see also section 3 and 5.4). The distribution of the participants can be found in Table 1 below.

Table 1. *Participants (n = 54)*

L1 Italian		**L2 Italian**	
Teachers	Students	Students	
16	20	10	8
		L1 Finnish	L1 Hungarian

5.3. *Materials*

The six argumentative text samples for Italian L2 from the earlier study by Kuiken and Vedder (2019b) were used again in the current study. The texts were taken from the written learner corpus of L2 Italian, Dutch and Spanish collected by Kuiken *et al.* (2010; see 4.2). The texts were written by Dutch Bachelor's students of Italian, with a proficiency level between A2 and B2. The writers had to indicate which of three non-governmental organizations (NGO) they would choose for receiving a grant. All texts were written in the classroom. To increase the comparability of the selected texts with respect to topic, content and lexis, the six texts employed in this study and in Kuiken and Vedder (2019b) refer to one of the three writing prompts, the organization *Ritorno alla Natura* (Back to Nature), an NGO aiming to protect the natural environment. See Appendix A for an example of one of the L2 Italian samples and its translation into English.

5.4. *Rating procedure and rating scale*

Online rating sessions of two hours were organized, one for the Finnish and one for the Hungarian student raters. Mirroring the method employed in Kuiken and Vedder

[3] All student data for Italian L1 and Italian L2 (L1 Finnish, L2 Italian) were collected by Sofia D'Alessandro, as part of her Master's thesis (University of Amsterdam, 2021).

(2019b) during the training session for the L1 teacher raters, we began by explaining the session's purpose. Following this, we introduced the rating scale for the assessment of syntactic complexity: a six-point Likert scale, derived from the CEFR (Council of Europe, 2001). The descriptors of the rating scale refer to complexity and control of syntactic structures (or lack thereof). Level 1, for instance, mentions the presence of "simple isolated phrases and sentences", "limited control of a few simple grammatical structures and sentence patterns" and "errors that may lead to misunderstanding". At level 6, it is expected that the outcome should be "a clear, highly accurate and smoothly flowing complex text", containing "a wide range of even the most complex language forms" (Appendix B).

After an individual rating round of the six texts, all student raters articulated the motivations behind the assigned scores and the feedback they would propose in a subsequent online panel discussion or retrospective interview. To familiarize the raters with the scale descriptors and rating procedure, two examples of writing samples were discussed. We instructed the participants to explicitly focus on syntactic complexity, rather than on vocabulary, accuracy or spelling.

For L1 Italian, due to the practical difficulty to schedule a joint online group session with the student raters, individual rating sessions were organized which followed the same procedure as for L2 Italian. Although the panel discussion had to be left out, a retrospective discussion between each participant and the researcher(s) took place.

5.5. *Data analysis*

Data were analyzed both quantitatively and qualitatively. We conducted a statistical analysis of intra-class correlation coefficients (ICC) to establish the inter-rater reliability. A one-way random effect model (ICC1) was chosen for the analysis. This decision was made because the raters were considered a random factor, and the coefficients were computed within each of the four groups. Pearson's correlation coefficient was then calculated to measure the strength of association between the variables and the direction of their relationship.

Subsequently, we gathered the comments justifying the assigned scores and proposed feedback from the individual rating forms and the transcriptions of the three panel discussions and the interviews with the L1 Italian students. Similarly to the procedure followed in Kuiken and Vedder (2019b) for the L1 data for the Italian language teachers, all score justifications and feedback were classified into five macro-categories that emerged from the data according to the different linguistic domains that were addressed by the raters (i.e., syntactic complexity, accuracy, comprehensibility, lexicon, text organization and coherence).

6. RESULTS

6.1. *Rating expertise and language background*

The first two research questions investigate to what extent the judgments of expert raters on syntactic complexity expressed on a six-point Likert scale align with the ratings by non-experts (RQ1), and how the participants' scores are influenced by language background (RQ2).

Table 2 shows an overview of means and standard deviations of the ratings of the six texts (A-F) given by the teachers (L1 Italian) and the students (L1 Italian and L2 Italian). The L2 students are subdivided into two subgroups (L1 Finnish and L1 Hungarian).

Table 2. *Teacher ratings vs student ratings in L1 Italian vs L2 Italian (L1 Finnish and L1 Hungarian)*

	L1 Italian				**L2 Italian**			
	Teachers (n=16)		Students (n=20)		Students L1 Finnish (n=10)		L1 Hungarian (n=8)	
Text	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	2.47	0.69	1.9	0.7	2.2	0.6	2	0.7
B	5.18	0.49	4.9	0.89	4.6	0.9	5	0.5
C	2.72	0.61	2.6	0.72	2.6	0.7	2.75	0.6
D	3.78	0.66	4	1.05	3.9	1	3.87	0.6
E	1.41	0.47	1.6	1.16	1.75	0.7	1.25	0.6
F	3.41	0.62	3.4	0.91	3.1	0.8	1.87	0.8

Although the scores assigned to the six texts A-F vary depending on the proficiency level of the writers, both experts and non-experts (teachers and students) and L1 and L2 raters of Italian largely agreed in their judgments of syntactic complexity, as shown in Table 2. Text B was considered by all raters as being the best, followed by D, F, C, A; text E unanimously received the lowest score. The only exception is represented by F, which scored low in the Hungarian sample (mean score = 1.87, SD = 0.8) compared to the other three groups, who assigned an average score between 3.1. and 3.41. The smallest standard deviations were found for the L1 teachers (SD between 0.47 and 0.69). This contrasts with the L1 Italian students, who varied more in their judgments (SD between 0.7 and 1.16). Table 3 summarizes the outcomes of the statistical analyses (intra-class correlation coefficients, ICC) that we conducted to answer RQ1 and RQ2.

Table 3. *Rater expertise and language background: Intra-class correlation coefficients (ICC)*

Variable 1	**Variable 2**	**ICC**
Finnish	Hungarian	$r = 0.921$
	L1 ITA Students	$r = 0.993$
	Teachers	$r = 0.977$
	Overall L1	$r = 0.992$
Hungarian	L1 ITA Students	$r = 0.891$
	Teachers	$r = 0.902$
	Overall L1	$r = 0.902$
L1 ITA Students	Teachers	$r = 0.973$
Overall L2	L1 ITA Students	$r = 0.961$
	Teachers	$r = 0.955$
	Overall L1	$r = 0.954$
Overall Students (L1 and L2)	Teachers	$r = 0.968$

As can be inferred from Table 3, the intra-class correlation coefficients between the total number of students (L1 Italian, Finnish, Hungarian) and the group of teachers (r = 0.968) showed that rating expertise did not seem to influence the scores for syntactic complexity assigned by experts and non-experts (RQ1). Similarly, the correlations (r = 0.954) for the overall comparison between L2 Italian (students with Finnish or Hungarian as their L1) and L1 Italian (students and teachers) or between the Finnish and Hungarian student raters (r = 0.921) demonstrated that L2 and L1 raters behaved similarly. This implies that L2 or L1 speaker status and language background did not seem to have an impact on raters' judgments (RQ2), as all correlations were high, ranging from 0.891 between the Hungarian group and the L1 Italian students, and 0.993 between the group of Finnish students and the L1 Italian students.

6.2. *Score motivations and feedback*

Next to the quantitative analysis we conducted a qualitative analysis to discover more insight into the reasons why teachers and students had assigned a particular rating score. We also investigated the type of feedback they would suggest for improving the texts, and possible differences between the four groups (L1 ITA Teachers, Students; L2 Students FIN, HUN). We grouped the score motivations and feedback into the five macro-categories that had been distinguished (syntactic complexity, accuracy, comprehensibility, lexicon, text organization and coherence). Although during the training session prior to the assessment raters had been instructed to explicitly zoom in on syntactic complexity when evaluating the six texts, they often justified their scores by referring to other linguistic domains, rather than the presence or absence of specific syntactic features. When raters explicitly addressed syntactic complexity to justify their scores, they concentrated on sentence structure, particularly subordination, topicalization, agreement of number/gender within verb and noun phrase, and variety of tense and mode.

Table 4 gives an overview of the score justifications per linguistic domain for each of the four groups, in descending order. The frequency counts between brackets (roughly) indicate the total number of occurrences for each category. Given the difficulty of reliably computing and confronting the rater's comments collected in different settings and conditions (onsite, online, group discussion, interview by researcher, individual rating form) we decided to refrain from reporting percentages, means and standard deviations.

Table 4. *Score justifications of teachers and students. Linguistic domains and total occurrences*

	Teachers ITA (n=16)	**Students** ITA (n=20)	**Students** FIN (n=10)	**Students** HUN (n=8)
1	Accuracy (75)	Synt. Complexity (120)	Synt. Complexity (61)	Synt. Complexity (40)
2	Comprehensibility (65)	Accuracy (59)	Accuracy (22)	Accuracy (31)
3	Synt. Complexity (49)	Comprehensibility (45)	Text Org. & Coh. (21)	Comprehensibility (27)
4	Lexicon (45)	Text Org. & Coh. (29)	Comprehensibility (19)	Lexicon (13)
5	Text Org. & Coh. (17)	Lexicon (18)	Lexicon (2)	Text Org. & Coh. (8)

An interesting finding is that teachers compared to students were more concerned with accuracy (in first position) than with syntactic complexity (ranked third). As can be deduced from Table 4, judgments of both the Italian L1 and the Finnish and Hungarian L2 student raters, were primarily motivated by syntactic complexity (first), followed by accuracy (second).

Regarding the other linguistic domains, the picture is less clear. Comprehensibility was addressed more often by teachers than by students (second position for teachers; third or fourth for students). Text organization and coherence are positioned, respectively, third (students L1 Finnish), fourth (students L1 Italian), or fifth (teachers; students L1 Hungarian). Lexicon was considered less important and was ranked fourth (teachers; students L1 Hungarian) or fifth (students L1 Italian, L1 Finnish).

Some examples of teacher and student comments for each of the five categories are reported in Table 5. Examples were translated from Italian into English.

Table 5. *Examples of score justifications by teachers and students (L1 and L2) per linguistic domain*

Teachers ITA (n=16)	**Students** ITA (n=20**)**	**Students** FIN & HUN (n=18)
Syntactic complexity		
There is some variation in syntactic structures, generally used however as unanalyzed, fixed formulas.	The sentences are coordinated in a controlled and correct way. There are some phrasal and orthographic mistakes, but they do not undermine the overall correctness of the text.	The writer produces coordinate sentences and uses subordination. (HUN)
Accuracy		
The text contains many basic errors including the lack of agreement between genitive and noun.	The text is simple but overall, fairly correct.	The writer shows good control of language use. (FIN)
Comprehensibility		
Systematic errors, but thanks to the context it is possible to understand the meaning.	It is not always easy to understand the message that the author intends to transmit. It is possible to grasp it, although it is not properly formulated.	The writer makes some mistakes which, however, do not affect comprehensibility. (FIN)
Lexicon		
I do not look at syntax, if there is still so much work to do on lexicon.	The author uses simple and frequent words which is probably a good strategy.	The writer has a rich vocabulary. (HUN)
Text organization and coherence		
The text is well structured and coherent.	The text still lacks some coherence, and the sentences are sometimes disconnected between each other.	The text is fluent and well structured. (HUN)

Interestingly, as demonstrated by these examples, the score motivations of the L1 and L2 student raters are largely aligned with the teachers' score justifications. However, they also differed; teachers paid more attention to the occurrence of frequently recurring errors, whereas the students' comments overall were framed more positively.

The suggestions that teachers and students give to the writers for improving their texts, not surprisingly, also address other linguistic domains beyond syntactic complexity. With respect to syntactic complexity their feedback concentrates on the same syntactic areas and features underlying their score justifications: verb and noun phrase, subordinate clauses, topicalization, agreement, tense and mode. Contrary to the comments of the student raters, teachers stress the need to focus on accuracy (agreement errors, use of subjunctive). An interesting difference, possibly due to raters' linguistic background, is that teachers and students with Italian as their mother tongue emphasize the importance of structural variation. This contrasts with the suggestions of the L2 students. They recommend their peers to play it safe and stick to the use of basic and simple structures (Table 6).

Table 6. *Feedback on syntactic complexity proposed by teachers and students (L1 and L2)*

Teachers ITA (n=16)	**Students** ITA (n=20)	**Students** FIN & HUN (n=18)
Syntactic complexity		
Pay more attention to variation. Only present and past tense are used. Many incorrect participles and one – incorrect – subjunctive.	It would be good to try to vary. Try not to use only simple structures, such as main clause plus coordinate or dependent clause.	Structures are simple and clear, and that's what counts. (FIN)
There are still problems with respect to agreement of gender and number in the NP. Work on agreement before trying to structure your sentences	It would be good to pay attention to number and agreement. In Italian this is important, but difficult.	You could link your sentences by means of a few conjunctions, to create a typically Italian text. (HUN)

6.3. *Judgments of raters vs L2 complexity research*

The third research question (RQ3) tentatively explores the extent to which the hypothesized development of L2 syntactic complexity in SLA (in three stages moving from co-ordination via subordination to phrasal complexity) is mirrored in the teacher and student raters' judgments and reflections on syntactic complexity. When focusing on the occurrence of different types of syntactic structures in the texts (i.e., complexity by coordination, subordination phrasal complexity), raters primarily referred to subordination, particularly relative clauses, relative pronouns and conjunctions, as has also been observed by Lambert and Kormos (2014). Coordinate structures were not often mentioned (by students, interestingly, more often than by teachers). Phrasal complexity, by means of clausal subordination and complexification of the noun group (Norris, Ortega, 2009; Biber, Gray, 2011; Biber *et al.*, 2011), was rarely considered. In the data, we found only a few cases of references to phrasal complexity, especially by some of the teachers (four out of 16) who were more familiar with L2 acquisitional research. As one

of the teachers put it, «The text contains various types of subordinate clauses, such as adverbial, relative, temporal and concessive clauses. There is also an elaborate complex noun phrase, comprising three embedded clauses. These are rather complex and difficult syntactic structures, even for native speakers» (Teacher, Italian L1).

7. Discussion

The quantitative analysis of the data showed that raters' complexity scores were not influenced by rating expertise or language background. Expert and non-expert raters, and L1 and L2 raters behaved similarly, as demonstrated by the highly significant intraclass correlations. Experts and non-experts thus all agreed in their syntactic judgments (RQ1), contrary to what has been observed in other studies on the role of rating expertise (Bonk, Ockey, 2003; Kim, 2015; Duym *et al.*, 2018). Language background did not seem to influence the complexity ratings either, as equally high correlations were found between the scores assigned by L1 and L2 raters, or between the Finnish and Hungarian group (RQ2). These findings are in line with the outcomes of the study by Zhang and Elder (2011), in which no differences in holistic judgments were established between the L1 and L2 raters involved in the research. They contradict the outcomes of Bogorevich (2018), who observed different rating pattens for L1 and L2 raters, as was the case in the study by Winke *et al.* (2013), who found that L2 speakers' ratings were influenced by accent familiarity with the test-takers L1.

The qualitative analysis of the comments by experts vs non-experts and L1 vs L2-raters showed some interesting differences. Concerning the linguistic domains underlying the raters' score motivations and feedback, the focus of the teachers turned out to be on accuracy – that is to say, on control rather than on sentence complexity (cf. Table 4). Teachers often referred to the (lack of) accuracy in syntactic structures. Lexical and orthographic errors were also mentioned, despite the explicit recommendation during the training sessions to concentrate on syntactic complexity. Compared to the teachers, student raters (L1 and L2) referred foremost to syntactic complexity to justify their scores, and overall, the students' feedback appeared to be more positively framed (Table 5-6).

How can we explain this different focus in the comments of the expert and non-expert raters? Language teachers in their daily classroom practice may be more accustomed to viewing L2 written production in terms of error correction ("what is lacking"), rather than in terms of amount of elaboration and complexity ("what is already there"), so their complexity judgments may have been impacted by "teacher bias". Unlike student raters, teachers are more likely to be familiar with the CEFR. Their emphasis on accuracy may (partly) have been triggered by the descriptors of the CEFR scale for syntactic complexity employed in our study, specifically those referring to the occurrence of syntactic errors (cf. Appendix B).

The qualitative analysis also revealed that score motivations and syntactic preferences of L1 and L2 raters in some cases differed, which was also established by Zhang and Elder (2011). While the Italian L1 raters emphasized the importance of structural variation in their comments, focusing on aspects such as variation in tense and mode, and the use of gerunds and subjunctives, the L2 group insisted on the importance of clarity and the necessity of starting with basic structures linked by a few simple conjunctions. Interestingly, unlike the feedback of the Finnish L2 group and comparable with the comments of their L1 Italian peers, the feedback of the Hungarian participants was characterized by a more sophisticated use of grammatical terminology. We may speculate that this could be due to a more grammar-oriented approach in the Hungarian language

class, but more information regarding the instructional Hungarian context would clearly be necessary. These findings regarding a possible impact of language background and instructional environment are aligned with the results from Kuiken and Vedder (2014, 2019a). In our earlier studies we found that compared to their Dutch colleagues, Italian language teachers who often come from a classical-literary background, placed greater emphasis on syntactic variation and the avoidance of both repetition and "prefabricated" chunks and routines (see also Palermo, 2013: 99).

8. Conclusion

What do the findings of our study imply for research on the assessment of syntactic complexity in L2 writing and for teaching practice? The different outcomes of the quantitative and qualitative analysis, firstly, show the benefits of a mixed-methods approach. Although in the quantitative analysis no impact of rating expertise and language background could be established, the investigation of the raters' comments and feedback demonstrated that experts and non-experts sometimes differed in rating behavior. The data furthermore revealed instances of differences in syntactic preferences between L1 and L2 raters for particular constructions in the target language.

The limitations of the study are the small sample of six short texts, the relatively small number of participants, and the different conditions in which data had to be collected because of COVID-19. The essays were written by learners at the A2-B2 levels, meaning that their structural sophistication and complexity may not be widely different. It would be important to employ a larger corpus comprising texts from L2 learners of Italian with a wider range of proficiency levels (A1-C2). Such a corpus would certainly be more representative of written language data that vary in syntactic complexity.

The pandemic made it also impossible to recruit more L2-learners of Italian, for instance L2-learners of Italian with Dutch as their first language. We had also planned to involve, in addition to the L1 Italian teacher group, a group of Dutch L2 teachers of Italian. Comparing the L1 teacher data with L2 Italian teacher data would have been important to get more insight into the influence of raters' language background and L1-speaker's status on syntactic complexity ratings in L2 writing. Finally, as the comparison of teacher raters (experts) and student raters (non-experts) might involve some confounding factors beyond rating expertise (e.g. age, hierarchy, power), it could be useful to compare also novice teachers with experienced teachers, in order to see how their experience plays a role in judgments of learners' syntactic development.

Another critical issue that should be addressed in follow-up research lies in the appropriacy of the CEFR rating scale employed in the study, with scale descriptors explicitly involving syntactic complexity and accuracy and connector use. The conflation of syntactic complexity and accuracy has probably affected, to some extent, the results of the study. It is, furthermore, possible that the CEFR descriptors, which are related to the A1-C2 proficiency levels, didn't fully align with the A2-B2 proficiency level of the writers.

8.1. *Pedagogical implications*

The study has some interesting implications for teaching practice. The finding that teachers did not often address syntactic coordination and even less so phrasal complexity, possibly due to the lower occurrences of elaborated noun groups at (low-)intermediate proficiency levels, suggests that teachers may not be sufficiently aware of how, and in

which order, syntactic complexity has been found to develop (Norris, Ortega, 2009). Research by Biber and Gray (2011) and Biber *et al.* (2011) has shown that academic language, particularly written academic prose (e.g. textbooks, scientific articles, reports), is characterized by the occurrence of both coordinate and subordinate clauses, and the employment of complex noun phrases. For these reasons, it is important, especially in the academic context, to pay attention in teacher training courses and writing classes to this development of syntactic complexity in three stages, from coordination via subordination to phrasal complexity. More attention should also be devoted not only to syntactic and lexical characteristics of written academic language but, more generally, to the acquisition of academic language skills of students and to the development of their *academic literacy*.

What's more, the finding that teachers tended to concentrate on accuracy, i.e., on what learners do wrong, instead of what already goes well, underlines the importance of teacher training. As has been pointed out, among others, by Pallotti (2017), mistakes often demonstrate a next step in the acquisition process, indicating that learners are progressing to a subsequent stage. During training sessions, it is therefore crucial to stress the need to consider L2 output from a different angle, focusing more on development and complexity and less on accuracy, and to have a "positive" attitude towards errors. That is to say, to remind teachers how important it is to compliment learners on what they have already achieved, instead of emphasizing their shortcomings (Pallotti, 2017).

The research finally shows the advantage of putting together a group of teachers (or learners) to discuss criteria of syntactic complexity and text quality in L2, giving way to an interesting exchange of perspectives and making them reflect on their teaching (or writing) practice. Last but not least, the study has made clear that complexity judgments of both L1 and L2 students may complete teachers' ratings and are often appropriate and relevant. Evaluating syntactic complexity and, more generally, text quality in L2 or L1 writing, may give students an insight into the development of their writing process and academic literacy. The precision and encouraging nature of peer feedback is illustrated by this recommendation of a Hungarian L2 learner of Italian:

> Try to link the sentences between each other. It is very important to use coordinate conjunctions and other textual connectors, such as adversative and temporal connectors, between sentences. This is a good way of writing a text that looks more Italian.

REFERENCES

Ågren M., Granfeldt J., Schlyten S. (2012), "The growth of complexity and accuracy in L2 French. Past observations and recent applications of developmental stages", in Housen A., Kuiken F., Vedder I. (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, John Benjamins, Amsterdam, pp. 95-119: https://doi.org/10.1075/lllt.32.

Ait Eljoudi Q. (2018), "Algerian teachers' and learners' beliefs about learner autonomy", in Mackay J., Birello M., Xerri D. (eds.), *ELT Research in action. Bridging the gap between research and classroom practice*, IATEFL, Kent, pp. 65-70.

Bardovi-Harlig K. (1992), "A second look at T-unit analysis: Reconsidering the sentence", in *TESOL Quarterly*, 26, 2, pp. 390-395: https://doi.org/10.2307/3587016.

Biber D., Gray B. (2011), "Grammatical change in the noun phrase: The influence of written language use", in *English Language and Linguistics*, 15, 2, pp. 223-250: https://doi.org/10.1017/S1360674311000025.

Biber D., Gray B., Poonpon K. (2011), "Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?", in *TESOL Quarterly*, 45, 1, pp. 5-35: https://doi.org/10.5054/tq.2011.244483.

Biber D., Gray B., Staples S. (2016), "Predicting patterns of grammatical complexity across language-exam task types and proficiency levels", in *Applied Linguistics*, 37, 5, pp. 639-668: https://doi/10.1093/applin/amu05.

Bogorevich V. (2018), *Native and non-native raters of L2-speaking performance: Accent familiarity and cognitive processes*, Doctoral dissertation, Northern Arizona University.

Bonk W. J., Ockey, G. J. (2003), "A many-facet Rasch analysis of the second language group oral discussion task", in *Language Testing*, 20, 1: https://doi.org/10.1191/0265532203lt245oa.

Brezina V., Pallotti G. (2019), "Morphological complexity in written L2 texts", in *Language Research*, 35, 1, pp. 99-120: https://doi.org/10.1177/0267658316643125.

Brown A., Iwashita N., McNamara T. (2005), *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks*, ETS Research Report Series, John Wiley & Sons Inc., Hoboken (N. J.): https://doi.org/10.1002/j.2333-8504.2005.tb01982.x.

Bulté B., Housen A. (2012), "Defining and operationalising L2 complexity", in Housen A., Kuiken F., Vedder I. (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, John Benjamins, Amsterdam, pp. 23-46: https://doi.org/10.1075/lllt.32.

Bulté B., Housen A., Pallotti G. (2024), "Complexity and difficulty in second language acquisition: A theoretical and methodological overview", in *Language Learning*, 20, pp. 1-42: https://doi.org/10.1111/lang.12669.

Council of Europe (2001), *Common European Framework of Reference for Languages. Learning, Teaching, Assessment*, Council of Europe, Strasbourg.

Cucinotta G. (2018), "Teachers' perceptions of motivational strategies in the language classroom. An empirical study of Italian Fl and Italian L2 teachers", in *ELLE*, 7, 3, pp. 446-472: https://doi.org/10.30687/ELLE/2280-6792/2018/03/006.

Duym K., Schoonen R., Hulstijn J. H. (2018), "Professional and non-professional raters' responsiveness to accuracy in L2 speech: An experimental approach", in *Language Testing*, 35, 4, pp. 501-527: https://doi.org/10.1177/0265532217712553.

Granfeldt J., Nugues P., Ågren M., Thulin J., Persson E., Schlyter S. (2006), "CEFLE and Direkt Profil: A new computer learner corpus in French L2 and a system for grammatical profiles", in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), pp. 565-570: http://www.lrec-conf.org/proceedings/lrec2006/pdf/246_pdf.pdf.

Housen A., Kuiken F., Vedder I. (eds.) (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. John Benjamins, Amsterdam: https://doi.org/10.1075/lllt.32.

Kim H. J. (2015), A qualitative analysis of rater behavior on an L2 speaking assessment, in *Language Assessment Quarterly*, 12, 3, pp. 239-261: https://doi.org/01.1080/1543303.2015.1049353.

Knoch U. (2009), "Diagnostic assessment of writing: A comparison of two rating scales", in *Language Testing*, 26, 2, pp. 275-304: https://doi.org/10.1177/0265532208101008.

Kuiken F., Vedder I. (2014), "Raters' decisions, rating procedures and rating scales, in *Language Testing*, 31, 3, pp. 279-284: https://doi.org/10.1177/0265532214526179.

Kuiken F., Vedder I. (2019a), "Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish", in *International Journal of Applied Linguistics*, 29, 2, pp. 192-210: https://doi.org/10.1111/ijal.12256.

Kuiken F., Vedder I. (2019b), "Investigating teachers' perceptions of syntactic complexity in L2 academic writing", in *Instructed Second Language Acquisition*, 3, 2, pp. 228-248: https://doi.org/10.1558/isla.37977.

Kuiken F., Vedder I., Gilabert R. (2010), "Communicative adequacy and linguistic complexity in L2 writing", in Bartning I., Martin M., Vedder I. (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, Eurosla Monograph Studies, 1, pp. 1-56.

Kuiken F., Vedder I., Michel M. (2019), "Linguistic complexity in second language acquisition: Introduction", in *Instructed Second Language Acquisition*, 3, 2, pp. 119-123: https://doi.org/10.1558/isla.39602.

Kyle K., Crossly S. A. (2018), "Measuring syntactic complexity in L2 writing. Using fine-grained clausal and phrasal indices", in *The Modern Language Journal*, 102, 2, pp. 333-349.

Kyle K., Crossly S. A., Verspoor M. (2021), "Measuring longitudinal writing development using indices of syntactic complexity and sophistication", in *Studies in Second Language Acquisition*, 43, 4, pp. 781-812: https://doi.org/10.1111/modl.12468.

Lambert G., Kormos J. (2014), "Complexity, accuracy and fluency in Task-based L2 research: Toward more developmentally based measures of second language acquisition", in *Applied Linguistics*, 35, 5, pp. 607-614: https://doi.org/10.1093/applin/amu047.

Lu X. (2010), "Automatic analysis of syntactic complexity in second language writing", in *International Journal of Corpus Linguistics*, 15, 4, pp. 474-496: https://doi.org/10.1075/ijcl.15.4.02lu.

Lu X. (2011), "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development", in *TESOL Quarterly*, 45, 1, pp. 36-62: https://doi.org/10.5054/tq.2011.240859.

Lumley T. (2005), *Assessing second language writing: The rater's perspective*, Peter Lang, Lausanne.

Norris J. M., Ortega L. (2009), "Towards an organic approach to investigating CAF in instructed SLA: The case of complexity", in *Applied Linguistics*, 30, 4, pp. 555-578: https://doi.org/10.1093/applin/amp044.

O'Leary J. A., Steinkraus R. (2022), "Syntactic and lexical complexity in L2 academic writing: Development and competition", in *Ampersand*, 9, 100096, pp. 1-10: https://doi.org/10.1016/j.amper.2022.100096.

Ortega L. (2003), "Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing", in *Applied Linguistics*, 24, 4, pp. 492-518: https://doi.org/10.1093/applin/24.4.492.

Palermo M. (2012), *Linguistica testuale dell'italiano*, il Mulino, Bologna.

Pallotti G. (2009), "CAF: Defining, refining and differentiating constructs", in *Applied Linguistics*, 30, 4, pp. 590-601: https://doi.org/10.1093/applin/amp045.

Pallotti G. (2017), "Applying the interlanguage approach to language teaching", in *International Review of Applied Linguistics in Language Teaching (IRAL)*, 55, 4, pp. 393-412: https://doi.org/10.1515/iral-2017-0145.

Pallotti G. (2021), "Measuring complexity, accuracy, and fluency (CAF)", in Winke P., Brunfaut T. (eds.), *The Routledge handbook of second language acquisition and language testing*, Routledge, London-New York, pp. 201-210.

Paquot M. (2019), "The phraseological dimension in interlanguage complexity research", in *Second Language Research* 35, 1, pp. 121-145: https://doi.org/10.1177/0267658317694221.

Pill J., Smart C. (2021), "Raters: Behavior and training", in Winke P., Brunfaut T. (eds.), *The Routledge handbook of second language acquisition and language testing*, Routledge, London-New York, pp. 135-144.

Révész A., Gurzynski-Weiss L. (2016), "Teachers' perspectives on second language task difficulty: Insights from think-alouds and eye-tracking", in *Annual Review of Applied Linguistics*, 36, pp. 182-204: https://doi.org/10.1017/S0267190515000124.

Vasylets O. R. G., Manchón R. M. (2019), "Differential contribution of oral and written modes to lexical and propositional complexity in L2 performance in instructed contexts", in *Instructed Second Language Acquisition*, 3, 2, pp. 206-227: https://doi.org/10.1558/isla.38289.

Vyatkina N. (2012), "The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study", in *Modern Language Journal*, 96, 4, pp. 572-594: https://doi.org/10.1111/j.1540-4781.2012.01401.x.

Winke P., Gass S., Myford C. (2012), "L2 background as a potential source of bias in rating performance", in *Language Testing*, 30, 2, pp. 231-252: https://doi.org/10.1177/0265532212456968.

Wolfe-Quintero K., Inagaki S., Kim H-Y. (1998), *Second language development in writing: Measures of fluency, accuracy, and complexity*, University of Hawai'i Press, Honolulu.

Yoon H.-J. (2017), "Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality", in *System*, 66, pp. 130-141: https://doi.org/10.1016/j.system.2017.03.007.

Zhang Y., Elder C. (2011), "Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?", in *Language Testing*, 28, 1, pp. 31-50: https://doi.org/10.1177/0265532209360671.

APPENDIX A

Ritorno alla Natura

I soldi devono andare alla Ritorna alla Natura. La natura è molto importante per tutti personi. Questa organizzazione sostegne la natura e lo suo scopo è auitare l'ambiente. Per sostegnere la natura credo che è bene per dare i soldi alla questa organizzazione. La natura è anche i animali. Quande fa bene con l'ambiente, fa anche bene con i animali! Per dare i soldi è anche auitare i animali, flori. Non è sola una case della gente. È davvero molto importante sostegnere Ritorno alla Natura, perchè sostegna tutto il mondo. Perchè nostra universita è una università importante nel mondo sarà bene sostegnere Ritorna alla Natura. L'universita lascia vedere che la natura è importante per lei. Un altro argomento perche l'Università deve sostegnere Ritorno alla Natura è perchè è una opportunità per fare il contatto con gli studenti è la natura. Perchè la natura è una cosa che sarà essistere di sempre è anche bene dare i soldi alla questa organizzazione.

Back to Nature [Translation into English]

The money must go to Back to Nature. Nature is very important to all people. This organization supports nature, and its purpose is to help the environment. To support nature, I think it's good to give money to this organization. Nature is also animals. When it's good for the environment, it's also good for animals! To give money is also to help animals, flowers. It's not just a house owned by people. It is really very important to support Back to Nature, because it supports the whole world. Because our university is an important university in the world it will be good to support Back to Nature. The university shows that nature is important to her. Another argument why the university must support Back to Nature is because it is an opportunity to contact students and nature. Because nature is a thing that will exist forever it is also good to give money to this organization.

APPENDIX B

Rating scale for syntactic complexity (CEFR, 2001)

Level	Descriptor
1	Can write simple isolated phrases and sentences. Limited control of a few simple grammatical structures and sentence patterns. Errors may lead to misunderstandings.
2	Can write a series of simple phrases and sentences linked with simple connectors (and, but, because). Use of simple grammatical structures is correct, but systematic, basic errors make comprehension difficult.
3	Can write a straightforward connected text. Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.
4	Can write a clear text with a relatively high degree of control. Uses occasionally less appropriate expressions, but errors do not lead to misunderstandings.
5	Can write a clear, well-structured text including complex constructions with a high degree of grammatical accuracy. Good control of connectors.
6	Can write a clear, highly accurate and smoothly flowing complex text. Use of a wide range of connectors. Good control of even the most complex language forms.