LangLearn at EVALITA 2023: Overview of the Language Learning Development Task

Chiara Alzetta[1], Dominique Brunato[1], Felice Dell'Orletta[1], Alessio Miaschi[1], Kenji Sagae[2], Claudia H. Sánchez-Gutiérrez[2] and Giulia Venturi[1]

[1] *ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy*
[2] *University of California, Davis*

Abstract

Language Learning Development (LangLearn) is the EVALITA 2023 shared task on automatic language development assessment, which consists in predicting the evolution of the written language abilities of learners across time. LangLearn is conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios. A total of 9 systems were submitted by 5 teams. The results highlight the open challenges of automatic language development assessment.

Keywords
Language learning development, student essays, shared task, multilingual language learning assessment

1. Introduction

Over the last twenty years, there has been a growing interest in exploiting the potential of Natural Language Processing (NLP) tools to characterize the properties of learners' language and to study how it evolves over time, both in first (L1) and second language (L2) acquisition scenarios. A similar concern has been paid to turning theoretical considerations into educational applications, such as Intelligent Computer-Assisted Language Learning (ICALL) systems [1] and tools for automatically scoring learners' writing with respect to language proficiency and writing quality [2, 3], and more generally systems able to automatically assign a learner's language production to a given developmental level [4, 5] or to operationalize sophisticated metrics of language development thus alleviating the laborious manual computation of these metrics by experts [6, 7, 8].

Generally, a greater number of studies has been carried out in the field of L2 learning where the study of L2 writings is seen as a proxy for language ability development [9]. In this respect, much work is devoted to predicting the degree of L2 proficiency according to expert-based evaluation [10] or to modeling the evolution of grammatical structures' competence with respect to predefined grades, e.g. the Common European Framework of Reference for Languages (CEFRL) [11, 12, 13]. On the other side, fewer studies have focused on exploiting NLP techniques in the context of L1 development. Among these, we can mention studies devoted to assessing syntactic development in preschool children [8, 14] and to examine overall writing ability and its development during later language acquisition [15].

It is worth noting that most research has been focused on English. Few exceptions are represented by e.g. the works by [16] and [17], which investigated writing development in German-speaking students across the elementary and secondary school, and [18], and [19], who proposed a methodology for tracking the evolution of written language competence of L2 Spanish and L1 Italian learners, respectively.

The Language Learning Development Task (LangLearn) organized at EVALITA 2023 [20] continues this scenario and it represents the first shared task on automatic language development assessment. It was aimed at developing and evaluating systems to predict the evolution of the written language abilities of learners across several time intervals. Additionally, the task was conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios.

2. Definition of the task

The assessment of language development is cast in LangLearn as a binary classification problem: it consists in predicting the relative order of two essays written by the same student. We started from the assumption that, given a set of chronologically ordered essays written by the same student, a document d_j should have a higher

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

chiara.alzetta@ilc.cnr.it (C. Alzetta); dominique.brunato@ilc.cnr.it (D. Brunato); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); alessio.miaschi@ilc.cnr.it (A. Miaschi); sagae@ucdavis.edu (K. Sagae); chsanchez@ucdavis.edu (C. H. Sánchez-Gutiérrez); giulia.venturi@ilc.cnr.it (G. Venturi)

quality level with respect to the ones written previously (d_i). Specifically, we followed the approach devised by [19]: given a randomly ordered pair of essays (e_1, e_2) written by the same student, we ask to predict whether e_2 was written before e_1.

LangLearn was articulated in two sub-tasks based on the resources allowed for training the models.

- **Sub-task 1** consists in predicting the order of essays using only the official training data released for the task;
- **Sub-task 2** consists in predicting the order of essays using information acquired from the training data released for the task and also from additional external resources.

3. Datasets

In line with the aim of having a multilingual shared task, we distributed two datasets composed of essays written by learners of the Italian and Spanish languages. Notably, the two datasets reflect an additional dimension of variation, which is the different learning scenarios from which the written productions were obtained. Specifically, the collection of Italian essays was written by students learning Italian as their first language, while the Spanish essays were produced by L2 learners.

For each corpus, LangLearn participants were provided with two files:

- a .tsv file containing the following information pertaining to a pair of essays (e_1, e_2) written by the same student: IDs of e_1 and e_2 in the correct chronological order, and t_1 and t_2 corresponding to the time of writing of e_1 and e_2.
- an XML file containing the text of the essays with randomly generated document IDs, as in the example below:

```
<dataset>
<doc id="9843"> Essay </doc>
<doc id="7432"> Essay </doc>
</dataset>
```

3.1. Corpus Italiano di Apprendenti L1 (*CItA*)

CItA (*Corpus Italiano di Apprendenti L1*) [21] is a longitudinal corpus of essays written by the same L1 Italian students in the first (2012-2013) and second year (2013-2014) of lower secondary school. The original corpus contains a total of 1,352 essays written by 156 students. The essays belong to five textual typologies, which reflect the different prompts students were asked to respond to, i.e. reflexive, narrative, descriptive, expository and argumentative.

For the purposes of the LangLearn shared task, we selected a subset of 882 essays authored by 133 different students at different time intervals. A time interval is identified by the year and specific period during which each essay was produced (e.g., the label 1_4 denotes the fourth essay written during the first year). Specifically, we considered 11 intervals, six for the first year and five for the second one. As it can be seen in Figure 1, the essays feature diverse linguistic characteristics across the considered time intervals[1]. In fact, essays written in the first year tend to be shorter in terms of the total number of tokens than those produced in the second year. Interestingly, the length of the document is a raw text feature highly related to various linguistic aspects that shape the writing style of an essay. Furthermore, the essays are increasingly lexically richer across time, as emerged from the Type/token ratio (TTR) values calculated for the first 100 tokens of the texts. It is worth noting that the last essays of the second year (interval 11) deviate from this trend. This is possibly due to the fact that they are mostly related to similar prompts that involved completing a history. In this case, students tend to write shorter and less lexically varied essays.

In order to build the training and test sets of LangLearn, essays from each student were paired based on their chronological order of writing, ensuring that the first essay in each pair was written prior to the second. This process resulted in 2,673 essay pairs: 2,366 were assigned to the train set, and the remaining 307 were placed in the test set. The distribution of pairs across time intervals is reported in Table 1. Note that some time interval pairs (e.g. 1_2 – 1_5) appear only in the test set. This is done to challenge participants since they do not have any corresponding pairs within the train set. Similarly, we isolated 4 students whose essays appear only in the test set, while the essays of 49 students appear only in the train set and the essays of 80 students appear in both sets. Indeed, it is possible for the same essay to appear in both the training and test sets, but it would appear in different pairs, ensuring that a specific pair occurs exclusively in either the train or test set.

3.2. Corpus of Written Spanish of L2 and Heritage Speakers (*COWS-L2H*)

The COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*) corpus [22] consists of 3,498 short essays written by second language (L2) students enrolled in one of ten lower-division Spanish courses at a single American university. Student compositions in the corpus are

[1] Note that these two linguistic characteristics are those used to compute the baseline scores.

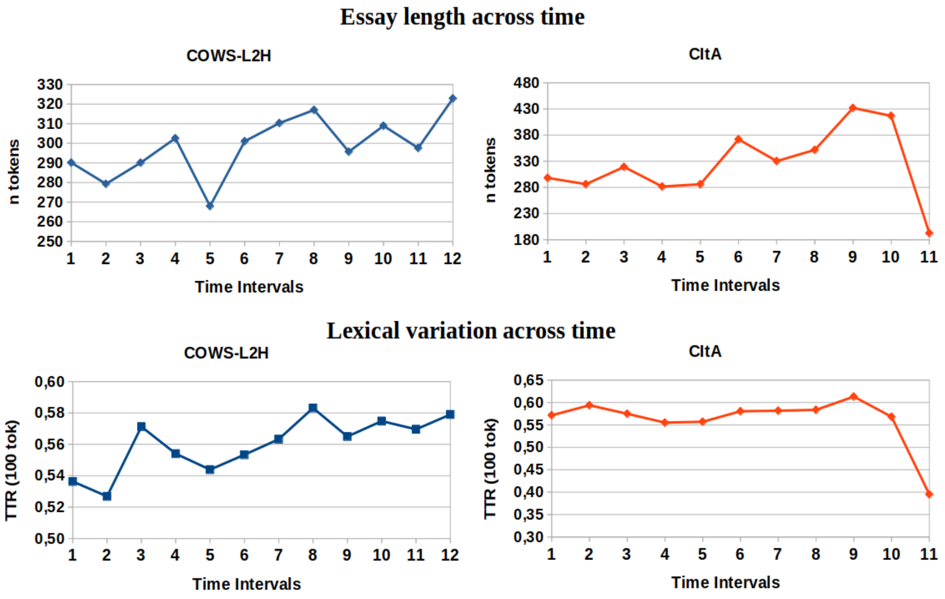

Figure 1: Distribution of two linguistic characteristics of the LangLearn datasets across the considered time intervals.

written in response to one of six writing prompts, which are changed periodically. According to these prompts, in each essay, the student is asked to write about: a famous person, your perfect vacation plan, a special person in your life, a terrible story, a description of yourself, or a beautiful story. During each period (an academic quarter, which consists of ten weeks of courses) of data collection, students are asked to submit two compositions, approximately one month apart, in response to the previously mentioned prompts. These composition themes are designed to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation by the writer.

To select essays from the original COWS-L2H dataset for the LangLearn task, we considered only essays written by students who wrote essays in two separate academic terms. This way, we can pair essays written at different points in time by the same student. To reduce the possibility that factors independent of language learning could systematically differentiate between essays in a pair, we considered only pairs of essays written in response to the same prompt. With these constraints, we were left with 1,329 pairs of essays written by 440 students. To split these essay pairs into training and test sets, we selected the essays written by 330 students to be in the training set, and the essays written by the remaining 110 students to be in the test set. This means that, in contrast with the CItA dataset used in LangLearn, there is no overlap in essays or authors between the training and test sets. The resulting training set contains 1,009 essay pairs, and the test set contains 320 essay pairs. The time interval between essays in a pair usually consists of one, two or three academic terms, with each term corresponding to 10 weeks of courses (Table 2). It is important to note that these intervals are not easily comparable across datasets, since COWS-L2H deals with highly structured L2 instruction, which progresses differently from L1 writing.

4. Evaluation

Baseline The baseline scores were calculated by training a LinearSVM using, for each pair (e_1, e_2), the number of tokens per document (in each pair) and the type/token ratio of the first 100 tokens in each document as input features.

Metrics The models' performance achieved on the CItA and COW-L2H test sets have been evaluated independently using Accuracy (A) and F1-score (F-score).

5. Submitted Systems and Participants

Following a call for interest, 5 teams registered for the task and submitted their predictions for both datasets, for a total of 18 runs (namely, 9 for each language tackled in the shared task). Eventually, one team (i.e. aroyehun_angel) did not submit a system report, thus we included

CItA - Train set											
	1_1	1_2	1_3	1_4	1_5	1_6	2_1	2_2	2_3	2_4	2_5
1_1		86	35	50	53	11	85	72	86	80	31
1_2			34	50			87		87	79	31
1_3				36	36	15	37	37	36	45	
1_4					56	16	57	53	55	50	
1_5						18	62	58	60	52	
1_6							22	15	17	19	
2_1								71	95	84	30
2_2									76	65	
2_3										80	28
2_4											28

CItA - Test set											
	1_1	1_2	1_3	1_4	1_5	1_6	2_1	2_2	2_3	2_4	2_5
1_1		2					2	2	2	1	6
1_2			2	1	54	15	4	74	4	4	6
1_3				1	2	2	2	4	2	3	
1_4					1	1	1	2	1	1	
1_5						2	2	4	2	3	
1_6							2	4	2	3	
2_1								6	4	4	6
2_2									6	7	35
2_3										4	6
2_4											3

Table 1
Distribution of essay pairs with respect to time intervals in the train and test sets of CItA. The rows correspond to the time points of the first essay in a pair, and the columns the time points of the second essay in a pair.

their scores in the overall dashboard, but we excluded them from the system description and error analyses. As shown in Table 3, all teams participated only in sub-task 1.

BERT_4EVER [23] proposed three different systems based on the base Italian BERT[2] model [24]. For fine-tuning the models, the team augmented the CItA and COWS-L2H datasets by reversing essay pairs to obtain negative examples and generating new positive examples by constructing transitive pairs. In the first system, BERT, BERT was fine-tuned performing simultaneous training on the augmented CItA and COWS-L2H datasets. The second model, Sequential, employs a novel sequential information attention mechanism to capture the interaction between the essays in a pair, which allows for incorporating the attention weights derived from the last-written essay in the representation of the pair relying on the [CLS] token and using average pooling. This pair representation is then fed into a linear classifier with a softmax function. The third model proposed is the Merge one, which fuses BERT and Sequential by averaging their output probabilities.

COWS-L2H - Train set										
	SU17	F17	W18	S18	W19	S19	F19	S20	F20	W21
S17	2	45	24	12						
SU17		2								
F17			51	25						
W18				74						
S18										
SU18										
F18					153	65	37			
W19						109	45			
S19							61			
SU19										
F19										
W20								109	24	36
S20									35	35
SU20										
F20										65

COWS-L2H - Test set										
	SU17	F17	W18	S18	W19	S19	F19	S20	F20	W21
S17		17	13	5						
SU17										
F17			17	11						
W18				24						
S18										
SU18										
F18					43	19	11			
W19						45	18			
S19							21			
SU19										
F19										
W20								29	6	6
S20									9	11
SU20										
F20										15

Table 2
Distribution of essay pairs with respect to time intervals in the train and test sets of COWS-L2H. The rows correspond to the time points of the first essay in a pair, and the columns the time points of the second essay in a pair.

bot.zen [25] tackled LangLearn as a regression problem, where the goal was to determine the stage of the learning process at which a student wrote a text. To achieve this, the team first pre-processed the official training sets in order to acquire the absolute order of each essay written by a student. Then, they performed predictions relying on an ensemble of decision tree algorithms. The model was trained using 125 normalised features capturing lexical and morpho-syntactic properties for each essay. By using MALT-IT2 [26], the team was able to include a set of features measuring text complexity in terms of document length, and lexical, syntactic, and morpho-syntactic properties. These features, however, are available only for the Italian language, thus they were used only for CItA predictions.

[2] https://huggingface.co/dbmdz/bert-base-italian-uncased

Team	Members	Affiliation	Sub-task	Runs per Language
BERT_4EVER	4	Guangdong University of Foreign Studies, Guangdong University of Technology, China	1	3
aroyehun_angel	2	Instituto Politecnico Nacional, Mexico	1	2
bot.zen	6	Eurac Research, Italy	1	1
IUSSnets	3	Iuss Pavia, Italy	1	1
ExtremITA	4	Universitá degli Studi di Roma Tor Vergata, Universitá di Torino, Italy	1	2

Table 3
Teams participating in EVALITA 2023 LangLearn shared task. For each team, we detail the number of team members, their affiliations, the sub-task(s) they participated in, and the number of submitted runs for each language of the shared task.

IUSS-Nets [27] approached LangLearn using linguistics features (e.g. density of various part-of-speech categories, frequency of different kinds of syntactic constituents, mean sentence length, etc.) extracted using the existing Common Text Analysis Platform, or CTAP [28], and surprisal-based metrics derived from token probabilities obtained using pretrained language-specific BERT models. These different pieces of information were encoded in features used in random forest classifiers. Interestingly, unlike most systems in LangLearn, which obtained better performance on the CItA dataset than on COWS-L2H, this approach produced higher accuracy and F-Score on COWS-L2H. In fact, it produced the strongest results on the COWS-L2H dataset among those submitted. Although its performance on CItA was not among the strongest submitted, it was still substantially above the baseline.

ExtremITA [29] team participated in the task with two Language Models trained in a multi-task learning framework. The first model is an encoder-decoder based on IT5-small [30], while the second model was a decoder based on Camoscio [31], the Italian version of LLaMA [32]. These models show substantial differences in terms of parameter count, with IT5-small comprising around 110 million parameters, whereas the utilized version of Camoscio encompasses 7 billion parameters. Both models underwent joint fine-tuning on all EVALITA 2023 tasks and sub-tasks, leveraging prompting techniques. Specifically, for the LangLearn task, the extremIT5 model received each instance of the dataset with the task name preceding it as input, and it produced the predicted label as output. Conversely, the extremITLLaMa model, which requires a structured prompt, was provided with a textual description of the task and the desired output format specification, as follows: *"Questi due testi separati da [SEP] sono presentati nell'ordine in cui sono scritti? Rispondi sì o no".* As regards the dataset treatment, some preprocessing steps were adopted: firstly, the dataset was segmented into sentences, allowing a maximum of 100 tokens per sentence. Additionally, in order to augment the dataset, inverted sentence pairs were incorporated, resulting in an expansion of the dataset from 3,377 to 6,438 examples.

CItA		
Team	**Accuracy**	**F-Score**
BERT_4EVER-BERT	0.932	0.934
BERT_4EVER-Merge	0.925	0.927
BERT_4EVER-Sequential	0.925	0.926
aroyehun_angel-system2	0.863	0.865
aroyehun_angel-system1	0.840	0.845
bot.zen	0.834	0.837
IUSS-Nets	0.645	0.673
ExtremITA-camoscio-lora	0.596	0.613
Baseline	0.550	0.549
ExtremITA-it5	0.606	0.410

COWS-L2H		
Team	**Accuracy**	**F-Score**
IUSS-Nets	0.753	0.752
aroyehun_angel-system1	0.703	0.708
BERT_4EVER-Sequential	0.641	0.663
Baseline	0.659	0.663
BERT_4EVER-Merge	0.616	0.631
BERT_4EVER-BERT	0.609	0.620
aroyehun_angel-system2	0.588	0.569
ExtremITA-camoscio-lora	0.575	0.553
bot.zen	0.497	0.517
ExtremITA-it5	0.506	0.160

Table 4
LangLearn shared task leaderboard.

6. Results

Table 4 reports the leaderboard of systems participating in the LangLearn shared task. Most systems outperformed the baseline when tested on CItA dataset while surpassing the baseline proved to be more challenging on COWS-L2H dataset. The team BERT_4EVER submitted the best-performing systems in the L1 scenario, while the highest score for the Spanish dataset was achieved by the IUSS-Nets team. ExtremITA obtained the lowest scores on both datasets.

Overall, we observe varying system rankings across the two learning scenarios. We discuss such variation in more depth in the next Section.

7. Discussion

Upon examination of system performance, we notice differences in model performance between the CItA and COWS-L2H datasets. Considering that each dataset reflects a different learning scenario, this might indicate that the challenges posed by these scenarios were distinct. One notable finding is that models leveraging stylistic properties of essays, such as the IUSS-Nets model, were more effective in the L2 setting. On the other hand, teams that employed Neural Language Models achieved higher results on the CItA dataset.

The observed differences in performance might be attributed to two main factors: model architectures and specific properties of the two learning scenarios. Concerning the former, we highlight, for instance, that the BERT model used by the BERT_4EVER team was pretrained only on Italian texts. This choice likely contributed to its lower performance on COWS-L2H, despite the simultaneous fine-tuning on both CItA and COWS-L2H. In fact, while BERT was the best-performing model of the BERT_4EVER team and overall on CItA, it was surpassed on Spanish essays by their Sequential model, which incorporates information about the interaction between the essays in a pair. Similar observations can be made for the bot.zen and IUSS-Nets teams. Both teams employed classification models that leverage a set of explicit features capturing linguistic properties of the texts. While both teams exploited features measuring raw text properties and the distribution of part-of-speech and syntactic dependencies for both languages, they differed in terms of features that captured deeper textual properties. Specifically, IUSS-Nets achieved the highest score on the COWS-L2H dataset thanks to a wide set of features measuring text complexity, sophistication, refinement, lexical variety, and cohesion. Conversely, the bot.zen team was unable to compute features capturing text complexity for Spanish, resulting in lower scores for that language.

These results reflect also specific properties of the two learning scenarios of LangLearn, which clearly affected all systems submitted to the shared task. As observed by [27], the evolution of writing abilities in a second language shows greater variation in terms of style within a shorter time period compared to a first language. We can assume that during the learning phase of an L2, new linguistic structures are acquired by the students in a highly structured schedule dictated by the L2 learning environment, gradually becoming more complex in a somewhat uniform way. Consequently, the essays produced by L2 learners may primarily serve as a measure of their progress in acquiring these new, more complex structures. On the other hand, L1 learners may face challenges from their teachers to enhance their proficiency in accurately using linguistic structures they have already acquired. As a consequence, L2 essays may exhibit more significant stylistic variations as learners are faced with the acquisition of new language structures. In contrast, L1 essays over time may show a more accurate use of already familiar linguistic structures, highlighting the learners' mastery of these elements.

Team	Accuracy	F-Score
BERT_4EVER-BERT	0.855	0.865
BERT_4EVER-Merge	0.841	0.850
BERT_4EVER-Sequential	0.841	0.846
bot.zen	0.717	0.717
IUSS-Nets	0.648	0.683
ExtremITA-camoscio-lora	0.600	0.608
Baseline	0.566	0.559
ExtremITA-it5	0.655	0.537

Table 5
Results on the CItA Test set considering only unseen students.

To deepen our analyses on the CItA dataset, we compared the system performance on a subset of essay pairs that correspond to the most challenging prediction scenario, i.e. considering pairs involving students whose essays appear only in the test set. The results on this subset are reported in Table 5. As can be noted, the system ranking remains unvaried, but the bot.zen and BERT_4EVER systems suffer a drop in their performance on this setting. The main cause of the decline in scores is due to the increased complexity of this particular setting. In fact, systems cannot rely on information extracted from essays present both in the training and test sets, although paired with different essays. As a result, the systems must rely solely on their generalization abilities to discern significant variations within each essay pair. However, it is important to acknowledge that even in this particular setting, the scores achieved by the BERT_4EVER team significantly surpass the baseline. This further highlights the potential of language models, particularly in the L1 classification scenario, as previously mentioned.

As a final remark, it is worth discussing the performance of ExtremITA systems. This team employed two Large Language Models to tackle all shared tasks proposed in the EVALITA 2023 campaign and explored the applicability of a single model in solving multiple different tasks. Although extremITLLaMA achieved the top position in 41% of all EVALITA sub-tasks (i.e., 13 out of 22 sub-tasks) and a top-three placement in 14 sub-tasks, the results on LangLearn were just slightly above the baseline on CItA and below the baseline on COWS-L2H. Such a result lays the foundation for an interesting and highly timely discussion on the effectiveness of these large and powerful models on real-world tasks. It appears, in fact, that tasks that are strongly affected by stylistic properties, such as language learning development assessment, still pose challenges to these models.

8. Conclusions

In this report, we introduced LangLearn, the first shared task dedicated to the development of systems able to automatically predict the development of language learning starting from learners' essays, in two learning scenarios and in a multilingual setting. Analysis of the results from the 9 submitted models indicates that the task of language learning development assessment continues to present numerous unresolved challenges. Notably, models that relied on explicit stylistic features demonstrated superior performance in Spanish as an L2 learning scenario. Conversely, Large Language Models showcased greater effectiveness in Italian as an L1 learning scenario.

These findings shed light on the complex nature of language learning assessment and suggest possible future directions for future evaluation campaigns. On the one hand, by leveraging insights from the LangLearn task, researchers can devise new approaches that incorporate both explicit stylistic features and the strengths of Large Language Models. On the other hand, the comparably lower scores achieved by the ExtremITA in our task seem to prompt a new typology of evaluation campaigns devoted to putting under pressure the potential of Large Language Models, pushing the boundaries of their language comprehension and generation capabilities.

Acknowledgments

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] S. Granger, Error-tagged learner corpora and call: A promising synergy, CALICO journal (2003) 465–480.

[2] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, J. Dai, A hierarchical classification approach to automated essay scoring, Assessing Writing 23 (2015) 35–59.

[3] P. Deane, T. Quinlan, What automated analyses of corpora can tell us about students' writing skills, Journal of Writing Research 2 (2010) 151–177.

[4] K. Sagae, A. Lavie, B. MacWhinney, Automatic measurement of syntactic development in child language, in: Proceedings of the ACL, ACL, 2005, pp. 197–204.

[5] X. Lu, Automatic measurement of syntactic complexity in child language acquisition, International Journal of Corpus Linguistics 14 (2009) 3–28.

[6] B. Bram, A. Housen, Conceptualizing and measuring short-term changes in l2 writing complexity, Journal of Second Language Writing 26 (2014) 42–65.

[7] S. A. Crossley, D. McNamara, Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners, Journal of Second Language Writing 26 (2014) 66–79.

[8] S. Lubetich, K. Sagae, Data-driven measurement of child language development with simple syntactic templates, in: Proceedings of COLING: Technical Papers, 2014, pp. 2151–2160.

[9] S. Crossley, Linguistic features in writing quality and development: An overview, Journal of Writing Research (2020).

[10] S. A. Crossley, D. S. McNamara, Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication, Journal of Research in Reading 35 (2012) 115–135.

[11] S. Vajjala, K. Loo, Automatic CEFR level prediction for Estonian learner text, in: Proceedings of the third workshop on NLP for computer-assisted language learning, 2014, pp. 113–127.

[12] E. Volodina, I. Pilán, D. Alfter, et al., Classification of Swedish learner essays by CEFR levels, in: CALL communities and culture–short papers from EUROCALL, 2016, pp. 456–461.

[13] L. Zilio, R. Wilkens, C. Fairon, An SLA corpus annotated with pedagogically relevant grammatical structures, in: Proceedings of LREC, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[14] K. Sagae, Tracking child language development with neural network language models, Frontiers in Psychology 12 (2021).

[15] A. S. Crossley, J. Weston, S. McLain Sullivan, D. S. McNamara, The development of writing proficiency as a function of grade level: A linguistic analysis., Written Communication, Written Communication, vol. 28, no. 3, pp. 282–311 (2011).

[16] Z. Weiss, D. Meurers, Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school, in: Proceedings of BEA, 2019, pp. 380–393.

[17] E. Kerz, Y. Qiao, D. Wiechmann, M. Ströbel, Becoming linguistically mature: Modeling English and German children's writing development across school grades, in: Proceedings of BEA, ACL, Online, 2020, pp. 65–74.

[18] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, Tracking the evolution of written language competence in L2 Spanish learners, in: Proceedings of BEA, ACL, Online, 2020, pp. 92–101.

[19] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105.

[20] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] A. Barbagli, Quanto e come si impara a scrivere nel corso del primo biennio della scuola secondaria di primo grado, Nuova Cultura, 2016.

[22] S. Davidson, A. Yamada, P. Fernandez Mira, A. Carando, C. H. Sanchez Gutierrez, K. Sagae, Developing NLP tools with a new corpus of learner spanish, in: Proceedings of the 12th LRE Conference, ELRA, Marseille, France, 2020, pp. 7240–7245.

[23] H. Wu, N. Lin, S. Jiang, L. Xiao, BERT_4EVER at LangLearn: Language development assessment model based on sequential information attention mechanism, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[24] T. Wolf, L. Debut, V. Sanh, alii, Transformers: State-of-the-art natural language processing, in: Proc. of EMNLP, ACL, Online, 2020, pp. 38–45.

[25] E. W. Stemle, M. Tebaldini, F. Bonanni, F. Pellegrino, P. Brasolin, G. H. Franzini, J.-C. Frey, O. Lopopolo, S. Spina, bot.zen at LangLearn: regressing towards interpretability, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[26] V. Santucci, F. Santarelli, L. Forti, S. Spina, Automatic classification of text complexity, Applied Sciences 10 (2020) 7285.

[27] M. Barbini, E. Zanoli, C. Chesi, IUSS-Nets at LangLearn: The role of morphosyntactic features in language development assessment, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[28] X. Chen, D. Meurers, CTAP: A web-based tool supporting automatic complexity analysis, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119.

[29] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[30] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[31] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.