Automatic Error Detection: Comparing AI vs. Human Performance on L2 Italian Texts

Irene Fioravanti[1], Luciana Forti[1] and Stefania Spina [1]

[1] *University for Foreigners of Perugia, Piazza Fortebraccio 4, 06123 Perugia, Italy.*

Abstract

This paper reports on a study aimed at comparing AI vs. human performance in detecting and categorising errors in L2 Italian texts. Four LLMs were considered: ChatGPT, Copilot, Gemini and Llama3. Two groups of human annotators were involved: L1 and L2 speakers of Italian. A gold standard set of annotations was developed. A fine-grained annotation scheme was adopted, to reflect the specific traits of Italian morphosyntax, with related potential learner errors. Overall, we found that human annotation outperforms AI, with some degree of variation with respect to specific error types. We interpret this as a possible effect of the over-reliance on English as main language used in NLP tasks. We, thus, support a more widespread consideration of different languages.

Keywords

Error detection, error correction, artificial intelligence, large language models, L2 Italian.

1. Introduction

Identifying errors in texts written by second language (L2) learners is a relevant task in several research areas, which can also have practical applications in a variety of fields. Error analysis is a traditional approach adopted in second language acquisition research for decades (Corder 1967), which learner corpus research has more recently revisited in light of the availability of learner corpora and corpus-based methods of analysis (Dagneaux et al. 1998). In addition, acquisitional research on learners' errors has relevant pedagogical implications involving error-related feedback: appropriate corrective feedback can lead to improved writing skills in both L1 and L2 writing (Biber et al. 2011). Furthermore, automatic error detection and categorisation is key in language testing and assessment research and practice, with reference to automated essay scoring (e.g., Song 2024), which has important implications for rubric descriptors.

The interest of Natural Language Processing (NLP) in grammatical error correction (GEC) and grammatical error detection (GED) relies on the creation of systems used in Intelligent Computer-Assisted Language Learning (ICALL), Automated Essay Scoring (AES) or Automatic Writing Evaluation (AWE) contexts. ICALL systems integrate NLP techniques into CALL platforms, providing learners with flexible and dynamic interactions in their learning process. AES systems automatically grade written texts with machine learning techniques, as well as AWE systems, which also provide learners with feedback.

Identifying and annotating errors in the performance of L2 learners, while beneficial for both pedagogical and research purposes, presents considerable challenges. This process is typically conducted manually in the case of learner corpora due to the inherent nature of errors as latent phenomena. The manual identification of learners' errors requires a substantial degree of subjective judgment by human annotators (Dobrić 2023), as well as a considerable investment in terms of time.

The present study aims to contribute to the evaluation of the performance of Large Language

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

* irene.fioravanti@unistrapg.it (I. Fioravanti).

† The present article is a joint effort by the co-authors. IF wrote the following sections: Method; Sample texts; Error type identification; Annotation; LLMs; Human annotator groups; Evaluation; Overall error detection; and Error type detection. LF wrote Discussion and conclusion section. SS wrote the Introduction and Related works sections.

irene.fioravanti@unistrapg.it (I. Fioravanti); luciana.forti@unistrapg.it (L. Forti); stefania.spina@unistrapg.it (S. Spina).

0000-0001-5182-9394 (I. Fioravanti); 0000-0001-5520-7795 (L. Forti); 0000-0002-9957-3903 (S. Spina).

Models (LLMs) in the task of automatic grammatical error detection (GED) in written texts produced by L2 learners. In particular:

1. it evaluates the behaviour of different LLMs in relation to an error detection task in written texts produced by L2 learners of Italian, a language other than English, in line with recent studies criticising the over-reliance on English in NLP research (Søgaard 2022) and seeking to contribute to the very few studies that do consider languages other English (e.g., MultiGED-2023; Volodina et a. 2023);
2. it targets specific error types and grammatical categories in order to mitigate the problems arising from the broadness of the notion of error, focusing on clear-cut and possibly unambiguous error categories;
3. it relies on a high degree of accuracy in error annotation, which was manually performed by three researchers on a small learner dataset serving as the test set on which the systems are evaluated;
4. it assesses the performance of LLMs in error detection and categorisation, through a comparison with the performance of native Italian students and advanced learners of L2 Italian on the same task.

2. Related works

Research on automatic error detection in L2 written texts, mainly adopting machine learning approaches, has significantly developed in recent years (Bryant et al. 2023), especially within the framework of shared tasks focused on GED and GEC. For instance, Di Nuovo et al. (2019; 2022) implemented a novel Italian treebank which includes texts written by learners of Italian. An annotation scheme suitable for L2 production was proposed encompassing UD and error annotation.

The CoNLL-2014 Shared Task on Grammatical Error Correction (Ng et al. 2014) was based on the identification of 28 error types involving major grammatical categories as well as spelling and punctuation errors. The test set consisted of 50 essays on two different topics, written by 25 learners of L2 English, that were error-annotated by two native speakers. The BEA Grammatical Error Correction shared task (Bryant et al. 2019) used a larger dataset (350 essays written by 334 learners and native speakers of English) and a similar taxonomy consisting of 25 error types. More recently, the NLP4CALL shared task on Multilingual Grammatical Error Detection (MultiGED-2023; Volodina et al. 2023) was the first multilingual shared task including four languages in addition to English: Czech, German, Italian and Swedish. The datasets used for the task varied across languages: the Italian dataset consisted of 813 written learner texts. Participants mainly used systems based on pre-trained LLMs.

A recent study by Kruijsbergen et al. (2024) focused on L1 and L2 Dutch and explored the capabilities of LLMs in written error detection, with both a fine-tuning and a zero-shot approach through prompting a generative language model (GPT-3.5). Results highlight that the fine-tuning approach largely outperforms zero-shotting, both for L1 and L2.

3. Method

To evaluate AI performance in automatic GED on L2 written texts, we designed our study based on the following stages: selection of the text sample; error type identification; definition of the gold standard (henceforth, GS); evaluation of LLMs' annotations; comparison between LLMs and human performance.

3.1. Sample texts

We used authentic L2 data derived from a learner corpus of Italian, the CELI corpus (Spina et al., 2022; Spina et al., 2024). It is a pseudo-longitudinal corpus of L2 Italian, representative of written Italian produced by intermediate and advanced learners. The CELI corpus is made of four subcorpora, one for each proficiency level (B1; B2; C1; C2) equally designed in terms of tokens. Eleven texts were randomly selected from the B1 subcorpus, of the total size of 1,335 tokens. We focused on morphosyntactic errors only. We chose to extract our texts from the B1 level, assuming they would be characterised by a higher number of morphosyntactic errors compared to higher proficiency levels. To make the annotation task easier, we divided each text into sentences. Details about the sentences' sample can be found in Table 1.

Total number of sentences	67
Average and range of sentences' length (in tokens)	79; 29-143
Range of number of sentences in each text	5-7

Table 1. Description of the sentences' sample.

3.2. Error type identification

Contrary to previous study (Ng et al. 2014; Bryant et al. 2019) that employed a broad notion of error, we focused only on specific morphosyntactic errors (selection (S), addiction (A), omission (O), ending (E)) within four Parts of Speech (PoS; articles (A), prepositions (P), nouns (N),

verbs (V)), for a total of eight error types (Table 1). This choice was due to the fact that Italian is a morphologically rich language, and that the four selected grammatical categories are a frequent source of errors for learners.

Type	PoS	Description	Example
AS	article	selection of the wrong article	*In montagna ci sono ****i*** *alberi.*
AA	article	unnecessary use of the article	*Ho fatto ****la*** *fatica a salire le scale.*
AO	article	absence of the article although necessary	*Maria ha fatto * compiti ieri.*
NE	noun	incorrect ending of the noun	*Ho comprato tre ****mela****.*
VE	verb	incorrect ending of the verb	*Ieri Luca ****andavo*** *in spiaggia.*
PS	preposition	selection of the wrong preposition	*Domani parto ****a*** *Roma.*
PA	preposition	unnecessary use of the preposition	*Ho comprato ****a*** *un libro.*
PO	preposition	absence of the preposition although necessary	*Anna è andata * casa.*

Table 1. Description of the eight error types.

3.3. Annotation

The outputs of the four LLMs were compared to a benchmark (GS) obtained from the annotation of three researchers. Three Italian trained linguists (i.e., the three authors of this paper) manually annotated the sample texts. The three researchers annotated only the error types described above. Initially there was a substantial agreement between the three linguists (k = 0.61). The three annotators disagreed mostly on the PA error (k = 0.39). Any inter-annotator disagreements were resolved through negotiation until a partial agreement (i.e., two annotators out of three) was reached. The agreement turned out to be improved (k = 0.81). Then, all the remaining disagreements (i.e., the cases that reach a partial agreement) were resolved reaching a perfect annotator agreement prioritising the two annotators' decision over the third one (k = 1). In the GS, 47 grammatical errors were identified with an average of 4 errors per text, while no errors were found in 32 sentences. On average, each sentence contained 2 errors.

3.3.1. LLMs

ChatGPT-4o (July 2024 version), Copilot, Gemini and Llama3 were evaluated. Several steps were followed to arrive at the final prompt, which can be found in Appendix A. We started giving the prompt in Italian and presenting all the texts together. However, the four LLMs, which were not pre-trained, were able to find a small number of errors. We, then, proposed the prompt in Italian again, repeating the instructions for each text. In this case, the LLMs identified types of errors that were not required. Following the recommendations from Kruijsbergen et al. (2024) on the prompt's language, the entire prompt was then given in English. The performance improved as a greater number of errors were identified, but still types of errors that were not required. Therefore, we gave a more detailed prompt in English following the recommendation of Coyne et al. (2023). Definitions of the four Italian PoS were provided. Further, we listed the eight error types with descriptions and examples. The texts were presented in numbered sentences. LLMs were instructed to classify each detected error and were informed that there could be more than one error in a sentence as well as no errors at all. The entire prompt was repeated for each text. This last version of the prompt was used for this study. Subsequently, we calculated the inter-annotator agreement between the four LLMs, which resulted to be weak (k = 0.21).

3.3.2. Human annotator groups

LLMs' performance was also compared to two human groups. Twenty-two L1 (age range: 19-50) and Twenty-seven L2 speakers (age range: 22-40) of Italian took part in the annotation task. They were undergraduate and postgraduate students in humanities and social studies. They were asked to annotate only the error types described above, with definition and examples provided for each type of error. They were also asked to report the incorrect form and to provide the correct one. Then, we calculated the inter-annotator agreement between the raters of the two groups. L1 speakers reached a good agreement (k = 0.52), while the agreement between L2 speakers was poor (k = 0.33).

4. Evaluation

Four measures were used to compare the performance of LLMs and human annotators in detecting errors: Accuracy, Precision (P), Recall (R) and F-score (F_β). Accuracy was calculated by dividing the number of correctly identified errors by the total number of annotated errors. To be consistent with previous works in GED (Volodina et al. 2023), F-score was set to 0.5 given that it weights P twice as much as R (i.e., it is more important that a system makes a correct prediction, than being able to detect all errors).

4.1. Overall error detection

Gemini outperformed the other three systems, demonstrating the highest accuracy (65,52%). In contrast, Llama 3 turned out to be less accurate in comparison to the others (51,72%). ChatGPT and Copilot behaved similarly in terms of accuracy (57,47%). LLMs were less accurate than human groups in detecting errors, as L1 and L2 speakers reached much higher values of accuracy (89,66% and 78,16% respectively).

When looking at AI performance, Copilot and Llama3 showed worse P than ChatGPT and Gemini, indicating that they had low ability in detecting true error instances. Conversely, Gemini and Copilot were able to detect a higher number of errors compared to ChatGPT and Llama3. ChatGPT made the best predictions, while Gemini had better R. Human groups outperformed AI systems for R, P, and F-score (Table 2). L1 speakers were able to detect almost all errors and to make correct predictions. On the contrary, L2 speakers had better P and worse R, suggesting they had lowest number of FP but a reduced ability to detect TP.

Figure 1 shows the performance of each group in terms of P and R.

	P (%)	R (%)	F_{β}
ChatGPT	65.22	58.82	63.83
Copilot	34.78	69.56	66.75
Gemini	58.69	71.05	60.81
Llama3	45.65	55.26	47.29
L1	93.02	89.96	92.39
L2	93.55	63.04	85.29

Table 2. Groups' performance in the detection of the overall errors.

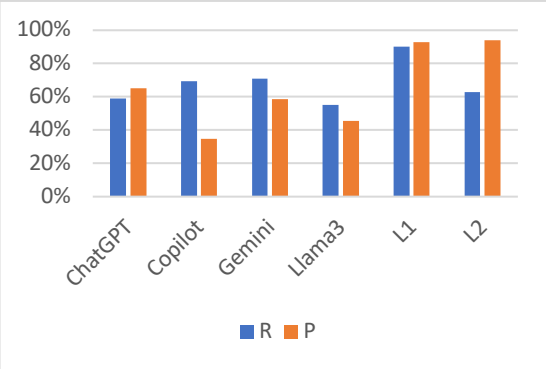

Figure 1. R and P for each group in the detection of overall errors.

4.2. Error type detection

To examine thoroughly the performance of the LLMs in GED, we calculated R, P and F-score metrics for each of the eight error types (Table 3).

Error type	P(%)	R(%)	F_{β} (%)
ChatGPT			
AO	50	100	55.56
AS	20	60	23.08
AA	/	/	/
NE	20	100	23.81
VE	46.15	54.55	47.62
PO	100	25	62.50
PA	50	50	50
PS	28.57	11.76	22.22
Copilot			
AO	/	/	/
AS	/	/	/
AA	/	/	/
NE	/	/	/
VE	50	30	44.12
PO	/	/	/
PA	/	/	/
PS	20	6.25	13.89
Gemini			
AO	100	100	100
AS	/	/	/
AA	/	/	/
NE	/	/	/
VE	40	44.44	40.82
PO	/	/	/
PA	/	/	/
PS	9.09	5.88	8.20
Llama3			
AO	/	/	/
AS	16.67	20	17.24
AA	/	/	/
NE	/	/	/
VE	35.71	50	37.88
PO	/	/	/
PA	/	/	/
PS	14.29	6.67	11.63

L1 speakers			
AO	100	100	100
AS	100	80	95.24
AA	100	80	95.24
NE	100	100	100
VE	90	90	90
PO	100	75	93.75
PA	100	100	100
PS	88.24	88.24	88.24
L2 speakers			
AO	100	100	100
AS	100	100	100
AA	100	40	76.92
NE	100	100	100
VE	88.89	80	86.96
PO	66.67	50	62.50
PA	100	50	83.33
PS	100	47.06	81.63

Table 3. Human vs. AI in detecting different error types.

Copilot, Gemini, and Llama3 failed to detect various error types exhibiting a high number of FP without detecting true instances. Copilot showed a fair prediction of VE and PS errors. Gemini had better R and P in detecting and correctly predicting AO and VE errors. However, it performed worse on PS errors in terms of both P and R. Llama3 was able to predict AS, VE, and PS errors but showing low values of R. ChatGPT turned out to be the best in predicting all error types, except for the AA error. ChatGPT showed high values of P in the prediction of AO, PA, and PO errors and showed low values of P and R for PS errors.

Human groups performed better than LLMs in detecting each error type. L1 speakers exhibited high values of R and P in detecting all error types but performed less well in making correct predictions on PS errors. L2 speakers demonstrated better R and P in detecting AO and AS errors. Conversely, they were unable to identify all AA errors. Furthermore, they showed a reduced ability in detecting all PO errors and in predicting them correctly.

5. Discussion and conclusion

The main aim of our paper was to investigate whether AI can be a valid support for second language acquisition research, in learner error detection, with specific reference to a language other than English, i.e., Italian. Our study compared the performance of four LLMs among them and also compared with L1 and L2 annotators. A GS, produced by the annotations of three trained linguists, was adopted as benchmark. Given the richness of Italian morphosyntax and the variety of possible morphosyntactic errors that L2 Italian learners may produce, contrary to the few other studies on Italian, this study considered three different error types for two of the parts of speech listed in Table 1, i.e. article and preposition. This methodological novelty can potentially lead to much more fine-grained results, while counterbalancing, like in our case, the low number of annotated texts.

The general finding about human annotators performing better than LLMs, both in terms of overall error detection and in terms of error type detection, is particularly significant if we consider the structural differences between English and other languages. Italian, like many other languages, is characterised by rich morphosyntatic traits, which inevitably have a considerable impact on the computational processing of L1 and L2 texts. Our findings may thus be a reflection of the well-known language bias in NLP, linked to the dominance of English, which then leads to a number of scientific but also social inequalities (Søgaard 2022; Volodina et al. 2023). Repeating the study with pre-trained LLMs might improve their performance. At present, pivotal tasks such as automatic error detection and classification, performed on a morphologically rich language such as Italian, does not seem to be viable with LLMs, as they do not add effectiveness to the same task performed manually.Future developments of this study may also include fine-tuned models, which are generally indicated as potentially better-performing than non-tuned ones (Kruijsbergen et al. 2024), as well as an increased number of annotated texts and an even more fine-grained and extended error annotation scheme. Automatic error detection and classification can be crucial for both the development of online language assessment systems and for second language acquisition research as a whole. This is especially true for languages other than English, which continue to be severely under-represented in all domains of language sciences, including NLP.

Acknowledgements

This study was conducted in the context of *CARLA – Corpus Approaches to Research on Language*, a research group affiliated with the Department of Italian language, literature and arts in the world (University for Foreigners of Perugia, Italy).

References

[1] D. Biber, T. Nekrasova, B. Horn, The Effectiveness of Feedback for L1-English and L2-Writing Development: a Meta-Analysis, ETS Research Report Series (2011), 1, i–99.

[2] C. Bryant, M. Felice, Ø. E. Andersen, T. Briscoe, The bea-2019 shared task on grammatical error correction, in: H. Yannakoudakis et al. (Eds.), Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2019, pp. 52–75, doi: 10.18653/v1/W19-4406.

[3] C. Bryant, M. Felice, T. Briscoe, T. (2017). Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 793–805, doi: 10.18653/v1/P17-1074.

[4] C. Bryant, Z. Yuan, M. Reza Qorib, H. Cao, H. Tou Ng, T. Briscoe, Grammatical Error Correction: A Survey of the State of the Art, Computational Linguistics (2023), 49 (3), 643–701, doi: 10.1162/coli a 00478.

[5] S. P. Corder, The Significance of Learners' Errors, International Review of Applied Linguistics in Language Teaching (1967), 5, 161-170, doi: 10.1515/iral.1967.5.1-4.161.

[6] S. Coyne, K. Sakaguchi, D. Galvan-Sosa, M. Zock, K. Inui, Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction (2023), arXiv:2303.14342.

[7] E. Dagneaux, S. Denness, S. Granger, Computer-Aided Error Analysis, System (1998), 26 (2), 163-174, doi: 10.1016/S0346-251X(98)00001-3.

[8] E. Di Nuovo, A. Mazzei, C. Bosco, M. Sanguinetti, Towards an Italian learner treebank in universal dependencies, in: R. Bernardi et al. (Eds.), CLiT: CEUR Workshop Proceedings (Volume: 2481), 2022.

[9] E. Di Nuovo, M. Sanguinetti, A. Mazzei, E. Corino, C. Bosco, Valico-UD: Treebanking an Italian Learner Corpus In Universal Depencies, Italian Journal of Computational Linguistics (2022), 8 (1), doi: 10.4000/ijcol.1007

[10] N. Dobrić, Identifying errors in a learner corpus – the two stages of error location vs. error description and consequences for measuring and reporting inter-annotator agreement, Applied Corpus Linguistics (2023), 3 (1), 1-11, doi: 10.1016/j.acorp.2022.100039.

[11] J. Kruijsbergen, S. Van Geertruyen, V. Hoste, O. De Clercq, Exploring LLMs' capabilities for error detection in Dutch L1 and L2 writing products, Computational Linguistics in the Netherlands Journal (2024), 13, 173-191.

[12] C. Leacock, M. Chodorow, M. Gamon, J. Tetreault, Automated Grammatical Error Detection for Language Learners, Morgan & Claypool, 2014.

[13] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, C. Bryant, The CoNLL-2014 shared task on grammatical error correction, in: H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, C. Bryant (Eds.), Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 2014, pp. 1–14, doi: 10.3115/v1/W14-1701.

[14] Y. Song, Q. Zhu, H. Wang, Q. Zheng, Automated Essay Scoring and Revising Based on Open-Source Large Language Models, IEEE Transations on Learning Technologies, 2024, 17, pp. 1920-1930, doi: 10.1109/TLT.2024.3396873.

[15] A. Søgaard. Should we ban English NLP for a year? In: Y. Goldberg, Z. Kozareva, Y. Zhang, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5254–5260, doi: 10.18653/v1/2022.emnlp-main.35.

[16] S. Spina, I. Fioravanti, L. Forti, V. Santucci, A. Scerra, F. Zanda, Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2, Italiano LinguaDue (2022), 14(1), pp. 116-138, doi: 10.54103/2037-3597/1. I.

[17] S. Spina, I. Fioravanti, L. Forti, F. Zanda, The CELI Corpus: design and linguistic annotation of a new online learner corpus. Second Language Research (2024) 40(2), 457-477, doi: 10.1177/02676583231176370.

[18] E. Volodina, C. Bryant, A. Caines, O. De Clercq, J.-C. Frey, E. Ershova, A. Rosen, O. Vinogradova, MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection, in: D. Alfter, E. Volodina, T. François, A. Jönsson, E. Rennes (Eds.), Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023), 2023, 1–16, https://aclanthology.org/2023.nlp4call-1.1.

Appendix A

The Prompt

In this task, we present a text in Italian, produced by a learner of L2 Italian at B1 proficiency level.

The text is numbered and divided into numbered sentences. For each sentence, you will have to identify specific errors, if any.
The errors considered in this task involve articles (in Italian "il, lo, la, i, gli, le, un, uno, una"), prepositions (in Italian "di, a, da, in, con, su, per, tra, fra", in their simple forms or associated with articles "del, dalla, negli, etc.", nouns, and verbs).
For each error, you will have to indicate the type, which you can choose from the following list:
1a: Article addition: the learner has added an article where it was not necessary (e.g. "Ho fatto la fatica a salire le scale": "la" should not have been used);
1b: Article omission: the learner did not use the article even though it was necessary (e.g. "Maria ha fatto compromesso con il suo capo": "un" should have been used before "compromesso");
1c: Article choice: the learner used the wrong article (e.g. "In montagna ci sono i alberi sempreverdi": "i" is wrong, the correct article is "gli");
2: Verb ending: the verb ending is incorrect (e.g. "Ieri Luca andavo al mare": "andavo" has the wrong ending "o", the correct one is "a" --> "andava");
3: Noun ending: the noun ending is incorrect (e.g. "Ho comprato tre mela gialle": "mela" has the wrong ending "a", the correct one is "e" --> "mele");
4a: Preposition addition: the learner added a preposition where it was not necessary (e.g. "Ho comprato a un libro": "a" should not have been used);
4b: Preposition omission: the learner did not use a preposition even though it was necessary (e.g. "Anna Ë andata casa": the preposition "a" is missing before "casa");
4c: Preposition choice: the learner used the wrong preposition (e.g. "Questo Ë il libro a mio professore": "a" is wrong, the right preposition was "del").
It is possible that there is more than one error in a sentence, but also that there are no errors at all.
If you find no errors, do not indicate anything and move on to the next sentence.
Here is the text with the numbered sentences.