CONFERENCE PAPER
International Conference on Pedagogical Science and Digital Learning 2025

Accepted: October 5, 2025 | **Published:** October 24, 2025 | **Language:** English

Investigating Greek Learners' Interlanguage in Italian through a Custom-Tagged Learner Corpus

Andrea Malorgio*

National and Capodistrian University of Athens, Greece

✉ ***Corresponding author:** *Andrea Malorgio*

ABSTRACT: This paper describes the creation and development of a learner corpus of Italian written by Greek university students, conducted within the framework of an ongoing PhD project. The corpus aims to fill a notable gap in learner corpus research concerning less-represented language pairings, specifically Italian as a second language learned by Greek speakers. The data were collected from authentic written examinations of the State Certificate of Language Proficiency (КПГ) at the B and C CEFR levels, comprising 293 texts and approximately 53,000 words. To analyse learners' interlanguage, a custom error-annotation tagset was designed and refined through several trial phases. The tagset includes three hierarchical dimensions – error level, part of speech, and error type–allowing for a detailed categorisation of linguistic deviations. Preliminary observations reveal recurrent difficulties with prepositions, verb morphology, and syntactic agreement, alongside noticeable traces of cross-linguistic influence from Greek, English, and other languages. The project provides both a methodological framework and a valuable linguistic resource for further studies on Italian L2 acquisition in Greek-speaking and multilingual contexts.

KEYWORDS: Learner corpus, Italian L2, Greek learners, interlanguage, error annotation, corpus design, cross-linguistic influence, multilingualism.

How to cite: Malorgio, A. (2025). Investigating Greek Learners' Interlanguage in Italian through a Custom-Tagged Learner Corpus. *International Conference on Pedagogical Science and Digital Learning 2025*, (pp. 98-103). Futurity Research Publishing. https://doi.org/10.5281/zenodo.17473277

Introduction

Learner Corpora, defined as "electronic collections of texts produced by foreign language learners", share the fundamental characteristics of general corpora (Granger, 2008). However, their design and compilation may vary considerably from project to project, depending on the aspects of learner language under investigation and the specific objectives and interests of the corpus creators.

Tono (2003) outlines several design considerations for corpus construction, distinguishing three major sets of criteria: language-related criteria (e.g. mode, medium, genre, topic), task-related criteria (e.g. longitudinal vs. cross-sectional; spontaneous vs. prepared), and lastly learner-related criteria (e.g. age, sex, mother tongue, level of studies, L2 knowledge). Similarly, Granger (2008) proposes a detailed typology with many subcategories of learner corpora: commercial vs. academic; dimensions; English vs. non-English; writing vs. speech; immediate vs. delayed pedagogical use.

The degree of corpus processing also varies according to the project's aims and analytical needs. A corpus can employ additional linguistic annotation to enhance its research potential and provide an extra layer of information that can later be analysed and used in different ways (Granger, 2008). This practice is called "error tagging".

Error annotation plays a crucial role in interlanguage studies, as it enables researchers to observe developmental patterns and language transfer phenomena. Many attempts have been made to provide the linguistic community with a standardised tagset; no single framework has yet achieved universal adoption. Researchers often adapt or expand existing schemes to suit the specific linguistic or pedagogical focus of their corpus. According to Aarts & Granger (1998), a tag "consists of a label for a major, word class, plus any number of features indicating syntacto-semantic subclasses and/or morphological characteristics".

Research in Second Language Acquisition has extensively used learner corpora to understand better how languages are learnt and how interlanguage works. Beyond research, they also have significant pedagogical applications. As Granger (2008) notes, they can be used for material design, syllabus design, language testing, and classroom methodology.

Despite their growing importance, there remains a notable gap in the availability of learner corpora for less-represented language combinations. In particular, no publicly available corpus currently focuses on the Italian interlanguage of Greek learners. Within the Italian Language and Literature Department of NKUA, an attempt to collect written texts produced by students has previously existed to create a resource in this area. However, this corpus has not been made public.

Since July 2023, within the framework of my PhD research, I have been developing a learner corpus of Italian produced by Greek university students, accompanied by a custom-designed error annotation tagset. This project seeks to address the current gap in learner corpus resources and to contribute to a better understanding of Italian L2 acquisition within Greek-speaking contexts.

Research methodology

Corpus compilation

The first step in the corpus design and compilation process was to determine the most appropriate research context for data collection. The initial environment considered was the Language School of the National and Kapodistrian University of Athens, where Italian is taught across all CEFR levels. After establishing contact with the instructors and collecting a preliminary sample of written texts, it became evident that these materials did not fully meet the predefined criteria for text length and content maturity.

A second possibility was to gather written productions from two university courses, with the collaboration of NKUA's professors. However, the nature of the two classes ("Error Analysis in the Italian Language" and "Creative Writing in a Foreign Language") differed significantly. The texts produced in the former course were primarily reflective and autobiographical, whereas those from the latter were imaginative and narrative-based. This divergence in genre and communicative intent made the two datasets incompatible for the present corpus.

The third and ultimately selected option was to draw on data from the State Certificate of Language Proficiency (Κρατικό Πιστοποιητικό Γλωσσομάθειας). The written examinations for levels B and C were identified as a suitable and homogeneous source of learner language. These texts displayed comparable levels of cohesion and maturity, were sufficiently rich in grammatical and lexical variety, and, most importantly, were produced under uniform testing conditions and within a controlled time frame.

Table 1

The data collected from this source

Parameter	**Data**
Source	State Certificate of Language Proficiency (ΚΠΓ) – Written exams
Total texts collected	293
B level texts	193
C level texts	100
Number of participants	99
Total number of words	52,974
Average text length (Level B)	80-120
Average text length (Level C)	300-400

The data collected from this source are shown in the previous table.

The participants varied in age, educational background, and likely L1. Due to privacy constraints, no additional biographical or linguistic data could be collected. For the same reason, gender identification was inferred indirectly from morphological cues in the texts, such as adjective agreement and past-tense verb inflexion.

a. Custom tagset design

After an initial reading and assessment of all collected texts, an Excel database was created to organise the corpus. The spreadsheet included multiple columns detailing the source of each text, the text identifier, the declared and verified proficiency level, the writer's gender, the number of errors, the text's suitability for analysis, and the word count. This preliminary stage ensured systematic data management prior to the annotation phase.

The next step involved developing a custom error tagset for identifying and annotating learner errors. As a primary reference and inspiration, I adopted the tagset previously developed for the same university course on error analysis (Florou, 2025). However, during the annotation trials, it became clear that the range and complexity of the errors observed exceeded the descriptive capacity of the existing labels. Consequently, the tagset was expanded and refined into three hierarchical categories: error level, part of speech, and error type. For each category, I had to create subcategories. Error level distinguishes between *morphological* (M) and *syntactic* (S) errors. Part of speech comprises ten subcategories: *noun* (SO), *adjective* (AG), *adverb* (AV), *verb* (VE), *preposition* (PP), *pronoun* (PN), *article* (AR), *conjunction* (CO), *punctuation* (PU), and *sentence* (FR). Error type includes eight values: *absence* (ASS), *substitution* (SST), *repetition* (RIP), *phonetic* (FON), *not in use* (NIU), *wrong position* (PER), *influence* (INF), and *incomprehensible* (INC).

Examples of annotated learner sentences are shown below:

La conoscenza e la forte [M SO SST] che ci danno [M VE ASS] i libri sono grandi [M AG SST].

All'inizio dobbiamo rispettare la natura e non fare del male, come polverizzare, infastidire, sconvolgere, danneggiare, ecc.

All annotations were carried out manually. Since the original texts were handwritten, they were first transcribed into plain text (.txt) files using Windows Notepad before tagging. No automated annotation tools or software were employed.

b. Preliminary results

All texts in the corpus have been annotated, although statistical analyses are still in progress.

A preliminary review of the written data, accompanied by initial qualitative notes, reveals that the most frequent errors concern the use of prepositions, the formation and use of the subjunctive, and, more broadly, verb agreement and hypothetical constructions.

A comparison between the declared CEFR levels and the actual linguistic performance indicates certain discrepancies. Nonetheless, the errors observed in the B-level texts essentially correspond to those typically associated with intermediate interlanguage development. In contrast, the C-level

errors reflect more advanced yet persistent challenges in grammatical accuracy and lexical choice.

As expected, Greek exerts the most substantial cross-linguistic influence on the learners' Italian. However, traces of English, Spanish, and French interference were also identified, suggesting multilingual influence among the participants. In a few instances, syntactic patterns appeared to reflect transfer phenomena from Albanian and Georgian. Interestingly, only one participant displayed indications of Latin interference in their writing.

Conclusions

This study presented the creation and development of a learner corpus of Italian interlanguage based on Greek adults' written productions, filling an important gap in the field of learner corpus research for minority language pairings.

The corpus was grouped from authentic examination texts, ensuring homogeneity in production conditions and reliability for interlanguage analysis. A tagset was explicitly created and tested after several attempts to better capture the complexity of learners' productions through a multilayered error annotation system that includes error level, part of speech, and error type.

Preliminary observations are already offering interesting and valuable insights into Greek learners' interlanguage in Italian, disclosing challenges and difficulties with prepositional use, verb morphology, syntactic agreement, and the influence of many various linguistic systems beyond Greek, along with English, Spanish, and other linguistic varieties spoken by prominent immigrant populations in Greece.

These first results emphasise the usage of learner corpora to clarify cross-linguistic transfer and developmental features of L2 acquisition.

Future stages of this research will focus on enhancing the corpus's analytical potential and promoting further research on Italian L2 acquisition in multilingual and cross-linguistic contexts.

List of references

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132–141). London: Longman.

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 793–805). Vancouver: Association for Computational Linguistics.

Ferrari, S., & Pallotti, G. (2005). *Interlingua e analisi degli errori: Apprendimenti linguistici e valutazione. Un percorso sull'italiano L1 e la correzione di testi scritti.*

Florou, K. (2013). Comparing error tagged learner corpora and learners' variables. *Journal of Modern Education Review, 3*(9), 695–704. Academic Star Publishing Company.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London: Longman.

Granger, S. (2003). *The corpus approach: A common way forward for contrastive linguistics and translation studies.* Louvain: University of Louvain.

Tono, Y. (2003). *The role of learner corpora in SLA research and foreign language learning: The multiple comparison approach* (Doctoral dissertation). Lancaster University, Lancaster.

Županović Filipin, N., & Mardešić, S. (2013). Analisi dell'interlingua nell'apprendimento dell'italiano a livello universitario. *SRAZ, LVII*, 201–219. Zagreb: Sveučilišta u Zagrebu.