Learner Corpus of Creative Writing: An interdisciplinary challenge

Ioanna Tyrou1, Katerina Florou2,
1 National and Kapodistrian University of Athens, Greece, iotyrou@ill.uoa.gr
2 National and Kapodistrian University of Athens, Greece, kathyflorou@ill.uoa.gr

ABSTRACT (ENGLISH)

Over the past two decades, corpora have become fundamental tools in computational linguistics and foreign language teaching, giving rise to the field of Corpus Linguistics. Learner corpora, collections of texts produced by foreign language learners, have been extensively studied to enhance language teaching methodologies. At the University of Athens, five years ago, an innovative project was initiated to integrate creative writing into learner corpora. This resulted in the creation of a corpus designed to address research questions related to teaching Italian as a foreign language, creative writing structures, and cross-linguistic analysis. This Learner Corpus consists of written texts produced by adult students of Italian, primarily native Greek speakers, in the context of creative writing activities. The corpus is systematically enriched through assignments aligned with specific themes, genres, and techniques, such as point-of-view shifts, material symbolism, and the "What if..." scenario exploration. It currently includes eight sub-corpora based on thematic units developed over six academic years, with a total size that reflects variations in student participation and task nature. The pedagogical applications of SCRICREA Corpus (derived from *"Scrittura Creativa"*) are extensive, encompassing language proficiency enhancement, creative expression development, and interlanguage analysis. Through activities such as narrative continuations, genre transformations, and music-inspired writing, learners develop linguistic accuracy, creativity, and critical thinking. Additionally, SCRICREA serves as a repository for research on linguistic phenomena, such as n-gram frequency analysis, further contributing to the field of corpus linguistics and foreign language pedagogy.

Keywords: Creative writing, Corpus Linguistics, Italian as a Foreign Language

ABSTRACT (ITALIANO)

Corpus degli Apprendenti di Scrittura Creativa: Una Sfida Interdisciplinare. Negli ultimi due decenni, i corpora sono diventati strumenti fondamentali nella linguistica computazionale e nell'insegnamento delle lingue straniere, dando origine al campo della Linguistica dei Corpora. I corpora degli apprendenti, collezioni di testi prodotti da studenti di lingue straniere, sono stati ampiamente studiati per migliorare le metodologie didattiche. Cinque anni fa, all'Università di Atene, è stato avviato un progetto innovativo per integrare la scrittura creativa nei corpora degli apprendenti. Questo ha portato alla creazione di un corpus progettato per affrontare domande di ricerca relative all'insegnamento dell'italiano come lingua straniera, alle strutture della scrittura creativa e all'analisi cross-linguistica.Questo Corpus degli Apprendenti consiste in testi scritti prodotti da adulti che apprendono l'italiano, principalmente madrelingua greci, nel contesto di attività di scrittura creativa. Il corpus viene arricchito sistematicamente attraverso compiti allineati a temi specifici, generi e tecniche, come i cambiamenti di punto di vista, il simbolismo materiale e l'esplorazione dello scenario "E se...". Attualmente include otto sub-corpora basati su unità tematiche sviluppate nel corso di sei anni accademici, con una dimensione totale che riflette le variazioni nella partecipazione degli studenti e nella natura dei compiti.Le applicazioni pedagogiche di SCRICREA (derivato da "Scrittura Creativa") sono estensive e comprendono il miglioramento della competenza linguistica, lo sviluppo dell'espressione creativa e l'analisi interlinguistica. Attraverso attività come continuazioni narrative, trasformazioni di genere e scrittura ispirata alla musica, gli studenti sviluppano accuratezza linguistica, creatività e pensiero critico. Inoltre, SCRICREA funge da repository per la ricerca sui fenomeni linguistici, come l'analisi della frequenza degli n-grammi, contribuendo ulteriormente al campo della linguistica dei corpora e della pedagogia delle lingue straniere.

Parole chiave: Scrittura creativa, Linguistica dei Corpora, Italiano come Lingua Straniera

1. INTRODUCTION

Since the new millennium, corpora have established themselves as a cornerstone in the fields of computational linguistics and foreign language teaching, forming an independent discipline known as Corpus Linguistics (McEnery &Hardie, 2012). Different types of corpora are employed depending on the specific needs of each researcher. Learner corpora, for instance, are extensive collections of oral or written texts produced by students of a foreign language (Granger, 2004). As a result, an increasing number of

studies now rely on corpora as a primary source of data. Although corpus-based methodologies have grown in sophistication, the use of corpus data remains associated with several unresolved challenges (Arppe et al, 2010). In addition, creative inspiration often originates from analyzing textual corpora, which provides the vocabulary, expressions, and thematic ideas necessary for creative writing. This approach enhances lexical appropriateness and encourages linguistic complexity, as creative language tasks require users to "reanalyze and combine known utterances and structures to create new ideas and forms" (Tin, 2012, 2011). Using corpora in activities like unfinished stories, inventive problem-solving, and transitioning from controlled to free writing fosters gradual improvement in creative expression (McDonald et al., 1997, cited by Arshavskaya, 2015; Mansoor, 2010; Melvita, 2023). The pedagogical use of corpora bridges creative exploration and structured learning by encouraging the synthesis of ideas, cultural expression, and divergent thinking (Herawati, 2021). Through systematic study and innovative strategies like the "six words" technique or narrative perspective shifts, students can significantly enhance their creative writing abilities (Tin, 2011; Le, 2018). Language learners can transform their expressive abilities into an authentic, stimulating, and autonomous process through activities that develop both language proficiency and creative thinking. In foreign language education, we aim to inspire students with meaningful, authentic material, encouraging them to express ideas and emotions effectively while using language with precision and originality. Emphasizing authenticity, communication, interaction, research, and creativity, we can foster innovative teaching methods and collaboration, helping students create unique, meaningful work in the foreign language (Tyrou, 2024).

In Greece, within the context of higher education, and specifically at the National and Kapodistrian University of Athens, in the Department of Italian Language and Literature, an initiative was launched to integrate these two fields. As a result, learner corpora were transformed into learner corpora of creative writing, aimed at addressing research questions related to the teaching of Italian as a foreign language, the structure of creative writing in a foreign language, and cross-linguistic analysis. This corpus, the collection and development of which began as a project in 2019 and continues to the present day, is continually enriched with data derived from written productions collected in the Creative Writing course. Hence the name of the corpus, SCRICREA, which is derived from the Italian term "Scrittura Creativa" (Creative Writing).

2. CONSTRUCTION OF THE CORPUS

The case of SCRICREA represents an instance of a Learner Corpus (LC) that aligns with the parameters of learner corpora. Specifically, it documents the interlanguage of learners while simultaneously serving as a resource for extracting samples of authentic language use in the context of creative writing. It consists exclusively of written output produced by students of Italian as a foreign language, created within the framework of creative writing activities conducted in an academic environment.

2.1 The learners

The dataset comprises 339 individuals. The students' ages range from 18 to 52, indicating an adult learner population. These individuals are students of Italian Language and Literature at the University of Athens. All participants are either native Greek speakers or have Greek as their first language. Regarding specific audience characteristics, all students are proficient in at least one foreign language, typically English, which they learned prior to studying Italian. A significant proportion also knows an additional foreign language, making Italian their second or third foreign language. Approximately one third of the students hold prior university degrees. They are enrolled in the 4th semester of their studies, and their language proficiency level in Italian language ranges between B2 and C1 according to the Common European Framework of Reference for Languages (CEFR).

2.2 The texts of the corpus

The written outputs in creative writing stem from a prompt. The topic is provided and discussed with the students, and the writing is produced subsequently. Given this structure, there is no fixed time frame for completing the tasks; however, students are allotted one week from the assignment of the topic to submit their written production. Similarly, there are no specific restrictions on text length, such as a prescribed word count. Nevertheless, there are shared elements, including the language, the genre of the weekly text, and the initial prompt.

The commonality within texts from the same unit lies in the textual genre (e.g., fairy tale, song, poem) and the writing technique employed (e.g., continuation of a story, adaptation with new characters). The

overarching shared characteristic across all units is the authentic language produced by students under the same educational conditions within the framework of creative writing.

2.2.1 The topics/prompts

From the above, it becomes evident that it is crucial to divide SCRICREA into sub-corpora based on the topic of each unit, as this determines both the type of discourse and the textual genre. The division is made according to the weekly topic, which also corresponds to the subject of each lesson, as outlined below (see List 1):

LESSON	Topic/prompt
1st lesson	Continuation of the story (The Young Crab- "il giovane gambero")
2nd lesson	Inventing" a little man made of any material and making him act, creating relationships and randomness depending on the material he is made of (e.g., glass, plastic, wood, ice cream, etc.) (James of Crystal- "Giacomo di Cristallo")
3rd lesson	Imaginary interview of the protagonist or another character from the story, or a brief calendar with text elements, or the recording of the autobiography of the story's hero (The Well of Cascina Piana- "Il pozzo di Cascina Piana")
4th lesson	Changing the point of view, or interpolation within the story (The Beautiful Stranger- "la bella sconosciuta")
5th lesson	A short musical-themed film for writing a story (based on the theme of the video) aimed at young or older children. Alternatively, the choice of a "friend-object" for the protagonist of the story and writing a text for children, or narrating the story with the "voice" of the object-protagonist (a balloon) (Way on Clouds)
6th lesson	Genre change, or using techniques of insertions/additions in the story, or the "What would happen if..." technique (The Man Who Stole the Colosseum- "l'uomo che rubava il Colosseo")
7th lesson	Text alteration technique - replacing words (changing the verb in the title) (The King Who Had to Die- "il re che doveva morire")
8th lesson	Selection of a song and writing a story that precedes or follows the theme of the song, or free-writing techniques

List 1:Topics and tasks and prompt

2.2.2 Outcomes of the activities

In the first activity, students created their own endings to a given story to enhance language proficiency and creative expression. The task encouraged innovation by integrating personal experiences, imagination, emotional insights and expectations into the narrative. By incorporating diverse perspectives, the activity aimed to reveal how students interpreted the story and emphasized different aspects. Ultimately, it sought to foster narrative control, boosting creative confidence and ease in using the foreign language.
The second activity aimed to enhance students' language, critical thinking, and creativity. Students were expected to use descriptive language to detail material properties (e.g., "hard as stone," "fragile as glass") and action-oriented vocabulary (e.g., "the wooden figure creaks" or "the straw figure is vulnerable to wind"). They were to create stories that highlighted the figures' unique properties (e.g., a glass figure avoiding conflict due to fragility). Symbolic meanings related to materials (e.g., marble for strength, wood for naturalness) were also expected to emerge. Finally, students were encouraged to explore deeper themes, such as using strengths and managing weaknesses.
In the third activity, students were expected to demonstrate an understanding of the protagonist's motives, thoughts, and emotions through their writing, highlighting key aspects of their life or actions. If they choose to write an interview, they should explore unknown aspects of the story or imagine post-story events, using vocabulary related to emotions and philosophy. In writing pages from the protagonist's

diary, students were expected to adopt a direct, emotional tone, reflecting on their thoughts and dilemmas. If writing an imaginative autobiography, students were to create a cohesive narrative linking events from the story with fictional ones, showcasing character evolution and adding details about the protagonist's childhood, goals, or life after the story.
In the fourth activity, students explored the narrative by altering the point of view or intervening in the story. Changing the point of view required understanding different characters' perceptions, adjusting language, tone, and details to fit the narrator's personality while offering alternative explanations. Interventions involved creating new scenes, dialogue, or internal thoughts to deepen character insights or influence the plot. Students could alter the story's flow, add new meanings, or develop secondary characters, while ensuring coherence with the original story and introducing original ideas that enhanced the narrative.
The fifth activity combined creativity, imagination, music, and storytelling to achieve diverse educational objectives. Students used music as inspiration to convey emotions, tone, and plot, reflecting the mood, style, or message of the piece (e.g., joy, nostalgia, adventure). They employed metaphors, descriptions, and dialogues or monologues to connect emotions with actions and scenes. For narratives involving a companion-object or the "voice" of an object, students personified the object by attributing human qualities (e.g., personality, emotions, and voice). These stories explored relationships and themes such as loneliness, friendship, loss, or dreams. Students were expected to create original, engaging stories that combined music, imagination, and emotions, with a logical narrative flow. Additionally, themes relevant to children (e.g., dreams, friendship, adventure, innocence) were anticipated to emerge.
The "What if..." technique encourages students to explore alternative scenarios, fostering both creative and critical thinking. Students are expected to develop varied plots (e.g., a thief stealing a different historical monument, or the protagonist stealing the Colosseum to save Rome), moral dilemmas (e.g., choosing between saving the Colosseum or protecting their family), and shifts in time (e.g., the Colosseum already being stolen in a past era or the story set in the future). Through this activity, students will practice hypothetical reasoning and imaginative language, demonstrating originality and creativity in their alternative narratives.
In the seventh activity, students use verb substitution, particularly in the title, to change the premise, tone, and direction of the story. This technique sparks creativity and leads to entirely new narratives. Students are expected to adjust the plot to align with the new title, introducing new themes that may shift the narrative's message or focus. Additionally, students will explore various genres (e.g., drama, fantasy, comedy, thriller), examining the consequences of the verb change on plot, characters, and tone. They will also create original scenarios and future possibilities for the story, ensuring coherence, originality, linguistic accuracy, and thematic development in their writing.
In the final activity, students combine music and creative writing to explore the emotions, narratives, and ideas evoked by a song. They are expected to create an original story linked to the song, imagining its beginning, emotions, or what happens after its end. Students should draw themes and images from the song and transform abstract feelings into specific scenes, characters, or dialogues. Alternatively, through free writing, students can write without constraints, letting their thoughts flow based on musical inspiration. This encourages the expression of personal feelings and ideas in various styles (e.g., narrative, poetic, or essayistic), fostering creativity without concern for right or wrong.

2.3 The size of the Corpus

The size of SCRICREA is approximately 1 million words. Its articulation and the sizes of its sub-corpora are presented in the table below. As shown, the corpus is composed of sub-corpora derived from eight different topics, as they have been developed over the past six academic years (see Table 1):

ACAD. YEAR	1° unità	2a unità	3° unità	4° unità	5a unità	6a unità	7a unità	8a unità (canzoni)	TOTAL
2019	13.951	34.995	36.664	34.400	33.261	25.101	19.600	26.451	224.423
2020	17.584	33.853	38.404	37.250	36.893	28.681	37.878	29.451	259.994
2021	19.018	37.399	40.765	41.853	42.008	34.416	23.121	39.965	278.545
2022	8.485	20.524	19.966	18.302	20.244	17.396	17.390	16.307	138.614

2023	6.974	10.833	10.415	11.507	11.489	5.552	10.308	8.411	75.489
2024	7.255	9.716	10.878	8.693	7.562	8.934	9.470	9.248	71.756
TOTAL	73.267	147.320	157.092	152.005	151.457	120.080	117.767	129.833	1.048.821

Table 1: SCRICREA, Corpus and subcorpora

Although the number of students remains constant for each task within the same academic year, the size of the sub-corpora varies due to differences in the nature of the tasks. Regarding the overall size of each corpus per year, it differs because of the varying number of students who enroll and submit assignments. In recent years, there has been a declining rate of student admissions to the department, which is also reflected in the size of the corpus.

3. APPLICATIONS

The SCRICREA corpus is a paradigmatic example of interdisciplinarity applied to language research and education. The project integrates the perspectives of corpus linguistics, language education and creativity studies, enabling an articulated analysis of language learning processes. The incorporation of narrative techniques, imagination exercises and a variety of textual genres serves to activate students' cognitive, linguistic and expressive skills, thus laying the foundation for a holistic and innovative teaching approach. The pedagogical and instructional applications of corpora have been extensively explored in prior research. Such studies have investigated, for instance, the frequency of verb usage (Ringdom, 1998), the use of connectives (Altenberg & Tapper, 1998), vocabulary (Lenko-Szymanska, 2005), prepositions (Diez Bedmar & Casas Pedrosa, 2006), and modal verbs (Aijmer, 2002). A corpus of this kind, however, offers a range of additional advantages. It acts as a repository of data and ideas for students of Italian and creative writing, while simultaneously functioning as a learner corpus that facilitates the analysis of students' interlanguage and serves as a basis for error analysis. Notably, parts of SCRICREA have already been employed for such purposes. For the analysis of n-grams and their frequency, we employed sub-corpora corresponding to topics 2, 5, 6, and 7, covering the period 2019–2022 (Florou & Tyrou, 2024). Some initial conclusions have already been drawn regarding the usefulness of n-gram measurements and analyses in the study of learner interlanguage, both from the aforementioned study and from other research focused on learner corpora (Gries, 2015; Pezic, 2025). Nevertheless, this specific corpus is not only composed of texts produced by learners of Italian language but is also intended for their use. It could serve as a space for peer experimentation through contrastive text analysis and error analysis. Furthermore, if shared with other educational settings (e.g., secondary education schools where Italian is taught), the corpus may function as a valuable resource: both as a model for creative writing and as a repository of interlanguage data from learners of Italian whose first language is Greek. Finaly, SCRICREA, due to its extensive nature, gives us the opportunity to use NLP tools to statistically investigate the corpus as well as the language of students in Italian. Indicators such as readability and type/token ratio can serve as supplementary measures not only in recognizing the student's level of proficiency in the Italian language but also in detecting qualitative characteristics such as imagination and creativity.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude, above all, to the students of our department for granting us permission to use their written works and for the opportunity they provide us to research the subject that will be taught to future generations of students. We would also like to thank the anonymous reviewers for their comments.

REFERENCES

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 55-76.* John Benjamins Publishing Company, Amsterdam/ Philadelphia. https://doi.org/10.1075/lllt.6.07aij

Altenberg, B., & Tapper, M. (1998). *The use of adverbial connectors in advanced Swedish learners'*. In S. Granger (Ed.), *Learner English on Computer*, pp. 80-93. Longman, London and New York.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. In Corpora 5. pp. 1-27. https://doi.org/10.3366/cor.2010.0001

Arshavskaya, Ekaterina. (2015). Creative Writing Assignments in a Second Language Course: A Way to Engage Less Motivated Students. *InSight: A Journal of Scholarly Teaching*. 10. pp. 68-78. 10.46504/10201506ar.

Diez Bedmar, M., & Casas Pedrosa, V. (2006). The use of prepositions by Spanish learners of English at University level: Main problems. In *7th Conference on Teaching and Language Corpora (TaLC) Proccedings*, pp. 42-44, Paris.

Florou, K., & Tyrou, I. (2024). La particella "ne" nella scrittura creativa in italiano LS. *Indagine su un corpus di apprendenti greci. Umanistica Digitale*, (18), pp. 87-101. http://dx.doi.org/10.6092/issn.2532-8816/19939

Granger, S. (2004). Computer Learner Corpus Research: Curent Status and Futur Prospects. In M. Conn, & T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, pp. 123-145. Amsterdam and Atlanda: Rodopi. https://doi.org/10.1163/9789004333772_008

Gries, S. T. (2015). Statistics for learner corpus research. *The Cambridge handbook of learner corpus research*, pp. 159-182.

Herawati, H. (2021). Learners as Story Writers: Creative Writing Practices in English as a Foreign Language Learning in Indonesia. In: Bao, D., Pham, T. (eds*) Transforming Pedagogies Through Engagement with Learners, Teachers and Communities*. Education in the Asia-Pacific Region: Issues, Concerns and Prospects, 57. Springer, Singapore. https://doi.org/10.1007/978-981-16-0057-9_5

Le, P. T. (2018). Using six-word stories to trigger EFL learners' creative writing skills. Indonesian Journal of English Language Teaching, 13, pp. 175-188. https://doi.org/10.25170/ijelt.v13i2.1456

Lenko-Szymanska, A. (2005). The role of L1 influence and L2 instruction in the choice of rhetorical strategies by EFL learners. In B. Lewandowska-Tomaszczyk (Ed.), *PALC'2005*. Practical Applications in Language Corpora. Peter Lang.

Mansoor, A. (2014). Ekphrastic practices in catalysing creative writing in undergraduate ESL classrooms. *New Writing: The International Journal for the Practice and Theory of Creative Writing*. https://doi.org/10.1080/14790726.2014.904887

McDonald, B. A., Rosselli, J. A., & Clifford, J. E. (1997). Journal writing: Learning, reflections, and adjustments to American life. *ERIC Document Reproduction Service* No. ED 423 709.

Melvita, S. (2023). Implementing unfinished story to build students' creative writing. *Journal of Educational Study*, 3, pp. 25-31. https://doi.org/10.36663/joes.v3i1.436

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Pęzik, P. (2015). Using n-gram independence to identify discourse-functional lexical units in spoken learner corpus data. *International Journal of Learner Corpus Research (IJLCR)*, *1*(2).

Ringdom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer*, pp. 41-52, Longman, London and New York.

Tin, T. B. (2011). Language creativity and co-emergence of form and meaning in creative writing tasks. *Applied Linguistics*, 32 (2), pp. 215- 235.

Tin, T. B. (2012). Freedom, constraints and creativity in language learning tasks: new task features. *Innovations in Language Learning and Teaching*, 6 (2), pp. 177-186.

Tyrou, I. (2024). Suggestions and prompts of foreign language activities in expressive writing. *International Journal of Education*, 12 (2).