https://doi.org/10.18485/icgl.2024.15.2.ch30

Tantos Alexandros[1]
Aristotle University of Thessaloniki
Amvrazis Nikos[2]
Aristotle University of Thessaloniki
Drakonaki Elena[3]
Aristotle University of Thessaloniki
Despina Papadopoulou
Chrysanti Develaska
Gerakini Douka
Pinelopi Kikilintza
Ilia Papafilippou[4]

EMPOWERING L2-GREEK RESEARCH AND TEACHING: THE ROLE OF GLCII'S DESIGN AND WEB PLATFORM

This paper presents the CLCII Gateway, a web-based platform that enables users to access the Greek Learner Corpus II (GLCII). The platform offers a diverse set of functionalities for efficient searching, data retrieval, visualization, concordancing, and database downloading. By utilizing the annotated data and advanced search options, researchers and educators can investigate various aspects of second language acquisition and enhance their language teaching practices. The user-friendly interface and comprehensive features of the CLC Gateway make it a valuable resource for applied linguistics research, promoting effective exploration and analysis of learner language data.

Keywords: GLCII, Gateway GLCII, Learner Corpora, Second Language

[1] alextantos@lit.auth.gr

[2] amvrazis@lit.auth.gr

[3] chrysiel@lit.auth.gr

[4] Aristotle University of Thessaloniki; dpapa@lit.auth.gr; develaska@lit.auth.gr; pkikilin@lit.auth.gr; iliaki_pap@windowslive.com

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) (https://www.elidek.gr/en/call/6489/) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 3161).

1. INTRODUCTION

The field of Second Language Acquisition (SLA) research has witnessed significant advancements with the emergence of learner corpora, which provide valuable insights into the linguistic development of second language learners. These corpora serve as a rich source of authentic learner data, enabling researchers and educators to investigate various aspects of language acquisition.

In the context of Greek as a second language (L2), there have been several attempts to compile learner corpora. Tzimokas (Τζιμώκας 2010) and Tantos and Papadopoulou (2014) have created cross-sectional learner corpora aimed at collecting a representative sample of written productions to serve as the basis for error annotation that would illustrate the developmental stages of interlanguage. Tzimokas's (Τζιμώκας) corpus consists of 291 productions (around 65.000 words) by adult learners, while the Greek Learner Corpus (GLCI) by Tantos and Papadopoulou includes 450 productions (around 33.000 words) by adolescent students. Iakovou et al. (2016) compiled the SEPAME2, a longitudinal corpus of written and oral texts from adult learners, which incorporates the innovative feature of linguistic and metalinguistic comments on the productions. However, with the exception of the GLCI, these corpora are not freely available, thereby limiting in-depth exploitation of these resources. Recently, the Greek Learner Corpus II (GLCII) (Tantos et al. 2023) the largest corpus for Greek as a second language (L2) has become freely available through the Greek Learner Corpus Gateway (GLC Gateway), a web platform that integrates essential tools to support both research and pedagogical endeavors.

The GLC Gateway serves as a comprehensive and intuitive interface with functionalities that make access and navigation effective and beneficial even for those users who may not have extensive skills or experience with corpus technology. Researchers as well as educators can easily specify their search queries, apply filters, and locate specific linguistic feature patterns, or examples of interest. Search results are displayed through the use of visual cues that group the retrieval by categories thus aiding the user in understanding and interpreting the linguistic information effectively. Moreover, the integration of the results (e.g. errors) in the linguistic environment (i.e. the production) provides the user with in-depth insights into the elements produced by learners in their respective contexts.

The following sections present the way the GLC Gateway has been designed. Section 2 provides a brief overview of GLCII, which is the most recent version of the GLC corpus hosted by the GLC Gateway. Section 3 introduces the GLC Gateway and its array of tools, providing a comprehensive exposition of the platform's design. To enhance clarity and understanding, this section is further subdivided to present a detailed exploration of each component within the platform. Section 4 discusses the key features of the platform.

2. THE GREEK LEARNER CORPUS II

One of the key contributions of corpora to SLA research is their ability to capture learners' language production and development over the stages of interlanguage (IL). Whether longitudinal or cross-sectional, learner corpora offer valuable insights into learners' linguistic progress and shed light on the complex processes of second language acquisition (Gilquin 2015: 9–34). By examining the data from multiple proficiency levels, researchers can identify patterns of development, analyze the impact of instructional interventions and gain a deeper understanding of how learners acquire and internalize the target language. GLCII was compiled with the intention to foster research on several aspects of SLA by using Greek, a less commonly taught language, as a target language of, currently, more than 1000 L2-Greek learners. The Greek Learner Corpus I (GLCI), its predecessor, is a second language (L2) Greek learner corpus. It consists of approximately 450 written productions created by adolescent students, amounting to roughly 30.000 words. GLCI's error annotation scheme was the basis for the GLCII's error annotation scheme. In contrast to GLCI, which focused on written data from adolescent students, GLCII includes both spoken and written examples from adult students.

Diving deeper into the GLCII is a cross-sectional corpus based on a principled selection of productions that have been elicited by L2 learners enrolled in language courses across all levels of linguistic proficiency. The elicited productions feed the two main subcorpora, written and spoken, thus, allowing for a comprehensive understanding of language acquisition and development, considering the unique cognitive demands each medium imposes on learners (Myles 2015: 309–332, cited in Bell & Payant 2021: 53–67). In addition to the learner subcorpora, users of GLCII can benefit from contrasting the learners' linguistic patterns against a subcorpus of native Greek, which follows the same criteria (e.g. Sinclair 2005: 79–83). While the written subcorpus is currently under development, the spoken component is planned for future inclusion. In a nutshell, a clear-cut design has been followed (Figure 1) by considering representativeness of the contents.

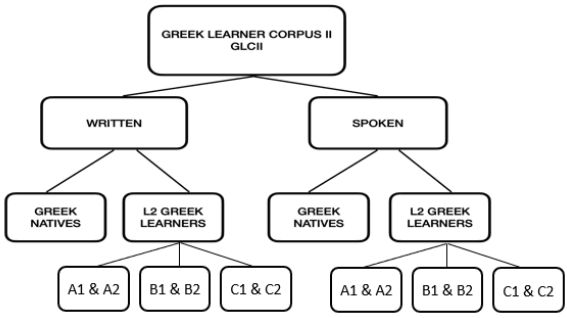

Figure 1. Design of GLCII

With the above design, the corpus comprises a collection of learner data including 1101 written texts and 318 spoken samples. The written subcorpus amounts to 422.360 words with the spoken parts currently being transcribed. The control subcorpus consists of 242 written texts amounting to 66.645-word tokens.

GLCII adopts an open-ended approach by using external criteria for representative and unbiased sampling. This approach ensures the inclusion of authentic tasks and topics, allowing for the observation of interlanguage features in natural communicative settings (Caines & Buttery 2017: 5–27; Lozano and Mendikoetxea 2013: 65–100). In the era of learner corpora natural communicative settings correspond to the second-language classroom conditions (Gilquin & Gries 2009: 1–26; Granger 2002: 3–33). The data collected in these settings is primarily generated through classroom tasks and instructional activities. GLCII aims to accurately represent the diverse communicative activities in which learners engage, encompassing a wide range of written and spoken tasks. As a learner corpus, GLCII reflects interaction within the classroom (Lozano 2021: 965–983; Tracy-Ventura & Myles 2015: 58–95) and facilitates a deeper understanding of the impact of the instructional practices on learners' language development.

Futhermore GLCII features a comprehensive repository of descriptive metadata and an error-annotation scheme that spans multiple layers, catering to both morphosyntactic and lexical level (Tantos et al. 2023). Their in-depth documentation and accessibility through the GLCII Gateway are critical for the end-user who can estimate whether the corpus is suitable for addressing specific research inquiries or teaching scenarios (e.g. McEnery, Xiao & Tono 2006). Secondly, the documentation of metadata and the annotation scheme enhances the interoperability and applicability of corpus-based outcomes beyond GLCII itself. Most importantly, the availability of comprehensive metadata alongside an error annotation scheme allows researchers to account for various learner related factors that may influence language production and development. The way these sources are accessible via the GLC Gateways is exemplified by the following section.

3. THE GLC GATEWAY

Designing a web interface for a corpus can be a complex and challenging task that requires careful consideration of a range of factors, from the searching system to the visualization tools and beyond. A user-friendly interface is critical for ensuring that the corpus is widely used and accessible to researchers with varying levels of technical expertise.

The GLC Gateway is a web-based platform that serves as a gateway to access the GLC. For seamless navigation within the GLC Gateway, it is hosted across three separate servers. These servers can be accessed through three distinct

URLs: https://glc.lit.auth.gr/app/GLC_Gateway,http://glcvm.lit.auth.gr/app/GLC_Gateway, and http://glc3.lit.auth.gr/app/GLC_Gateway. Both URLs landing page is the first red-labeled tab on the left of the homepage, called “About GLC”. The landing page provides users with introductory information about the two corpora that hosts, namely the GLCI and the GLCII. Users can click on the separate tabs on the left to choose the specific corpus they wish to access, and they are directed to the corresponding interface. In the case of the GLCII, the front page is divided into three sections indicated by respective ribbons, each serving a distinct purpose. They can easily navigate between different parts of the interface by expanding or collapsing the corresponding sections as needed. The first ribbon is linked to the annotation scheme developed within the GLCII. The remaining two ribbons represent the two distinct navigation levels: the corpus and text levels. Starting with the annotation scheme, these two levels are presented in the subsequent sections.

3.1. Annotation Scheme

When the user expands the ‘Annotation Scheme’ ribbon, they are shown the error categories used to annotate the corpus. The categories are presented in tables where each error type is attached to the relevant tag on the first left column with additional information and authentic examples from the corpus on the other two columns, illustrating the use of the specific tag (see Figure 2).

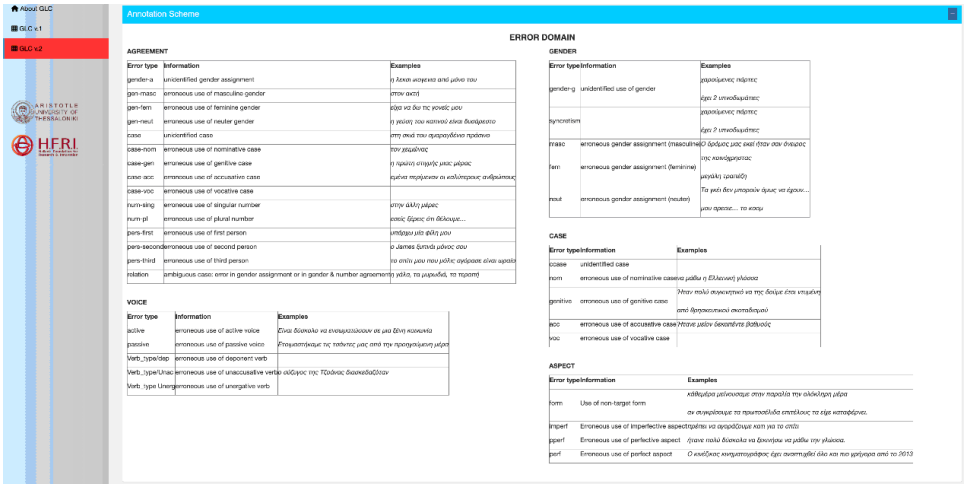

Figure 2. Expanding the “Annotation Scheme” ribbon

Documenting the annotation scheme is crucial for several reasons (Gries & Berez 2017: 379–409). Firstly, understanding the annotation protocol allows users

to assess the suitability of the error categories or linguistic features that have been annotated in the corpus. This knowledge enables researchers and educators to make informed decisions regarding the relevance of the GLCII to their specific research and pedagogical investigations. Secondly, detailed documentation of the annotation scheme enables the users of GLCII to critically evaluate the methodology and potential biases associated with the annotation process. This way GLCII promotes transparency in research practices since it fosters a more comprehensive understanding of the corpus data and encourages a critical examination of important aspects of the resource (i.e. annotation). Furthermore, documenting the annotation scheme enhances the replicability and reliability of the GLCII-based studies. Researchers can refer to the documented protocol to ensure consistency and accuracy in their own analyses, thereby contributing to the robustness and validity of their findings.

3.2. Corpus level

The aim of the interface at the "Corpus Level" is to assist users in intuitively prompting inquiries within it. To this end, the main area of the Corpus Level reflects the structure of the databases (see Figure 3) in the sense that the key options (i.e. filters) represent the way the corpus is composed and accessed by the user.

Figure 3. GLCII Gateway: Accessing the Corpus Level

At this level, the GLC Gateway incorporates a range of functionalities designed to facilitate essential interactions between the user and the interface. Serving as a platform for an annotated learner corpus, the GLC Gateway provides full access

to the set of error annotation tags and the corresponding text spans that they were applied on. Additionally, a second functionality is dedicated to conducting searches within the corpus for elements that extend beyond the annotated data. Notably, the integration of frequency analysis and search capabilities through part-of-speech (POS) tagging are essential components at this level. The following sections present a comprehensive overview of these functionalities and provide valuable insights into the rationale behind their incorporation within the Gateway.

3.2.1. Accessing annotated data

As depicted in Figure 3, through the "Error Profile Plot" tab the user can access a visual representation of the current state of the error annotation in GLCII. This representation includes five distinct barplot facets, each corresponding to the morphosyntactic domains of Agreement, Aspect, Case, Gender, and Voice. This representation allows users to grasp the architecture of the curated learner corpus and quickly assess the frequency distribution of error types within each error domain. Note that the error types have been encoded with a specific tag during the annotation process that is included in the "Annotation Scheme" ribbon. This first display of frequency distribution of error types enables a first assessment of the patterns and prevalence of learners' errors. From this, researchers can gain insights into learners' linguistic difficulties and target specific areas for further analysis.

Futhermore, one of the first beneficial features of the GLC Gateway is the quick access to three prominent variables, namely "Proficiency Level", "L1" and "Genre" of the production. These are presented on a separate panel on the left. Filtering can be conducted on multiple values of these variables simultaneously, thereby increasing the tool's versatility and the value of the search results. For example, the "Proficiency Level" filter allows users to choose one or more language proficiency levels (note that we follow the CEFR levels A1, A2, B1, B2, C1 & C2). By applying this filter, researchers and teachers can focus on specific learner groups and investigate developing L2 grammars as they are reflected on the emerging patterns and errors corresponding to different proficiency levels. By adding an option on the "L1" filter users can narrow down the data based on the learners' native language background. Users may select one or multiple L1s from the available options[5], allowing for comparative analysis between different learner groups or investigation of specific language transfer phenomena (e.g. Lozano & Mendikoetxea 2013: 65–100). The "Genre" filter enables users to select one or more of three genres: argumentative essays, narrations and descriptions. By choosing a specific genre, researchers and teachers can tailor their analysis or teaching materials to match the communicative demands and language features

[5] Currently, GLCII's productions belong to 36 typologically (un)related languages.

associated with different genres (e.g. Tavakoli and Foster 2011: 31–72).

The above filters are designed as drop-down menus, providing a user-friendly interface for making selections. Users can apply one or multiple filters simultaneously, depending on the variables they wish to explore or compare. Within the same block that integrates the above filters the user will find the "Download Corpus Texts" button so that the corpus is locally stored for offline use and analysis.

Moreover, an extended customization is offered under the "Error Profile Summary" tab. Researchers can search for specific error types by assigning relevant variables to them. Users can select from the five annotated error categories to focus their analysis on specific types of errors. Furthermore, they have access to a comprehensive range of metadata that was consistently collected throughout the process of compiling the GLCII. As displayed in Table 1, these variables encompass various crucial aspects for SLA (Brezina, Hawtin & McEnery 2021: 595–615).

Learner-related Variables		**Genre-and task-related variables**
Demographic Profile	*Linguistic Profile*	
Age	L1	Text type
Sex	L1 secondary	Use of external sources
Country of Origin	Proficiency level	
Educational Level	Knowledge of other languages	
	Total length of instruction	
	Total length of exposure to Greek	
	Greek language learning setting	
	Communication with Greek partner	
	Communication with Greek relatives	
	Communication with Greek friends	

Table 1. Variables accessible and customizable under the "Error Profile Summary"

All the variables listed in Table 1, which can be explored within the 'Error Profile Summary' tab, function as search filters, permitting the selection of one or more values from these variables. For example, the researcher can firstly set the Age range, then select one or more L1s from the drop-down menu, and finally set the Proficiency Level and the Length of Exposure to Greek to retrieve the errors across all the categories that are annotated in GLCII. By changing the search values for Length of Exposure, the researcher could potentially approach the differences in the learners' performance, at least based on the specific grammatical areas for

which error annotation is available. In addition, users can observe errors within both the sentence and the broader context. This capability aids in evaluating error patterns across different language levels, including morphosyntax, vocabulary, and discourse.

1.1.2. General search

Under the "Search" tab, users are allowed to engage with the data in a meaningful and powerful way. The user may enter a word, phrase, or even a regular expression, to trigger the extraction of all instances of the respective query within the corpus (Figure 4).

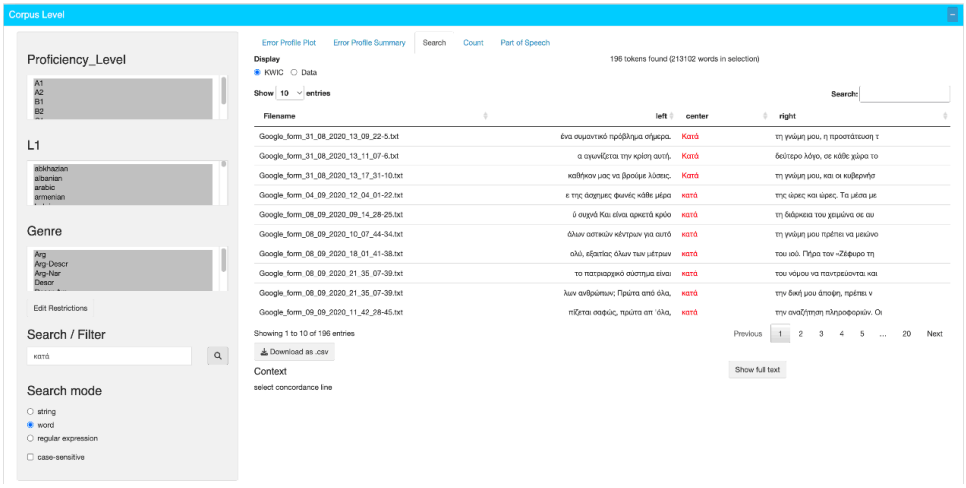

Figure 4. Searching and concordancing within GLCII Gateway

Data extraction in the GLC Gateway is facilitated in two distinct display formats: KWIC (Key Word in Context) and Data. The KWIC format presents the search results in a concise and user-friendly manner. It displays the searched word or phrase along with the five words preceding and following it, thereby creating a concordance line. Concordance tables, which are often featured in corpus linguistics, are notably useful in identifying lexical bundles, collocations and idioms and lay a groundwork for exploring domains beyond morphosyntax, such as formulaic language in learner language (Paquot & Granger 2012: 130–149). By examining concordance lines vertically for recurring patterns and horizontally for contextual interpretation, users can reveal lexical, grammatical and textual paradigms that would not have been noticeable through individual text readings. To this end, researchers can apply complex search queries to examine and classify morphosyntactic and lexical units, while educators can leverage this information to enhance language teaching practices and curriculum development.

The Data format includes the full range of metadata associated with each search result instance.[6] By presenting the metadata alongside the extracted instances, users can analyze the occurrences in a more detailed and contextually rich manner. This format facilitates a more thorough exploration of the relationship between linguistic phenomena and various contextual factors such as proficiency level, L1 or text type.

In sum, the "Search" functionality leverages patterns identification and evidence-based interpretation across a wide array of language phenomena in SLA. At the same time the availability of both KWIC and Data formats gives users flexibility in accessing and analyzing the extracted data.

3.2.3. Count

The "Count" functionality in the GLC Gateway further consists of two tabs, namely "Wordlists" and "Frequency table," which provide valuable insights into the vocabulary usage within GLCII. Users may utilize the search filter to query specific words or phrases, enabling them to gauge the frequency distribution of these linguistic elements. By entering a search query, the system promptly returns the total number of occurrences of the searched item in the corpus, shedding light on its prevalence within learner productions.

One of the key advantages of the "Count" functionality is its compatibility with the metadata filters integrated into the GLC Gateway. Users can refine their search by incorporating contextual factors (i.e., metadata). This customization allows researchers to conduct nuanced analyses and investigate vocabulary usage across specific learner profiles or language learning contexts. It provides a comprehensive view of how vocabulary is employed by learners, facilitating a deeper understanding of their lexical competence and growth over time. Importantly, the user can download the frequency tables in .csv format allowing researchers to access and analyze the frequency data offline, empowering them to perform further statistical analyses or integrate the data into their own research workflows. Furthermore, by downloading and utilizing these tables, teachers can enhance their teaching materials, create tailored vocabulary exercises, and provide their students with examples of common vocabulary usage.

Overall, the "Count" functionality serves as a powerful tool for researchers in applied linguistics, enabling them to explore vocabulary usage in the learner corpus through quantitative analysis. By leveraging this feature, researchers can gain valuable insights into vocabulary development, identify common lexical patterns, and inform language teaching practices with evidence-based findings.

[6] Unfortunately, displaying a screenshot of the Data format tab was not feasible due to the page's size.

3.2.4. Part of Speech

At the corpus level, the GLC Gateway provides the user with the capability to perform searches utilizing the "POS" (Part of Speech) tag associated with each word. Automatic POS tagging was accomplished with the "ILSP Neural NLP toolkit for Greek" that is hosted on the CLARIN infrastructure.[7] While automated POS tagging may lack -in the case of learner corpora- the same level of accuracy as in other L1 corpora, it still remains a practical and useful tool.[8] With the "POS" tag, GLCII users may retrieve information based on any part of speech (e.g., noun, pronoun, adjective, verb etc.) and indirectly draw useful conclusions as to the varying impact of parts of speech on learners' performance. Moreover, the teacher may identify areas of strength and weakness and tailor their interventions to enhance language development.

When using the "POS" tag, the interface presents the searched item within the context of six words on the left and six words on the right (see Figure 5). However, users have the option to individually select these arrows to gain direct access to a specific point within the text, which will open in a new window above the concordance table, as demonstrated in Figure 6.

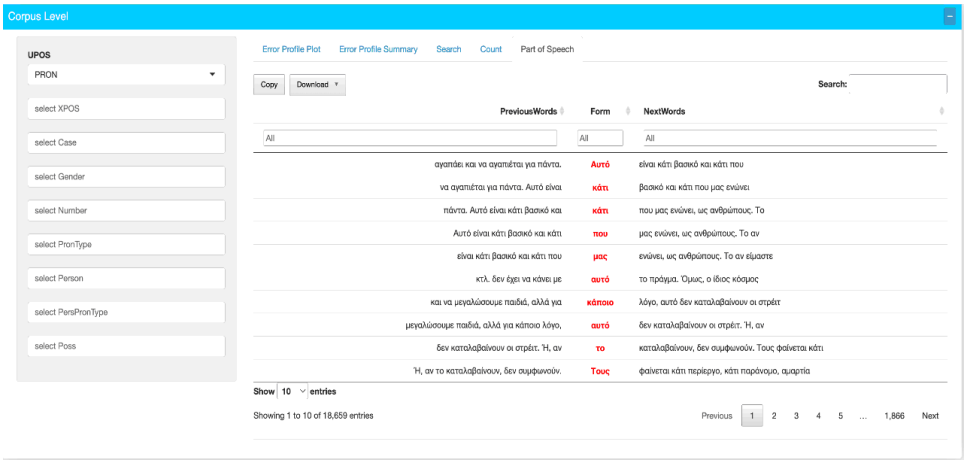

Figure 5. Part of Speech functionality in GLCII Gateway

The contextual information displayed in the figures 5 and 6 allows for a more comprehensive understanding of how the search item is used in context. However, it is essential for users to exercise caution and interpret the results with care, considering the limitations of automated POS tagging and the need for

[7] For more information on the resource, please refer to https://inventory.clarin.gr/tool-service/1129.
[8] However, due to the specific nature of L2 Greek data, the POS-tagged data currently requires further scrutiny to correct any inaccurate tag assignments.

additional linguistic analysis to ensure accurate interpretation of the POS-tagged data.

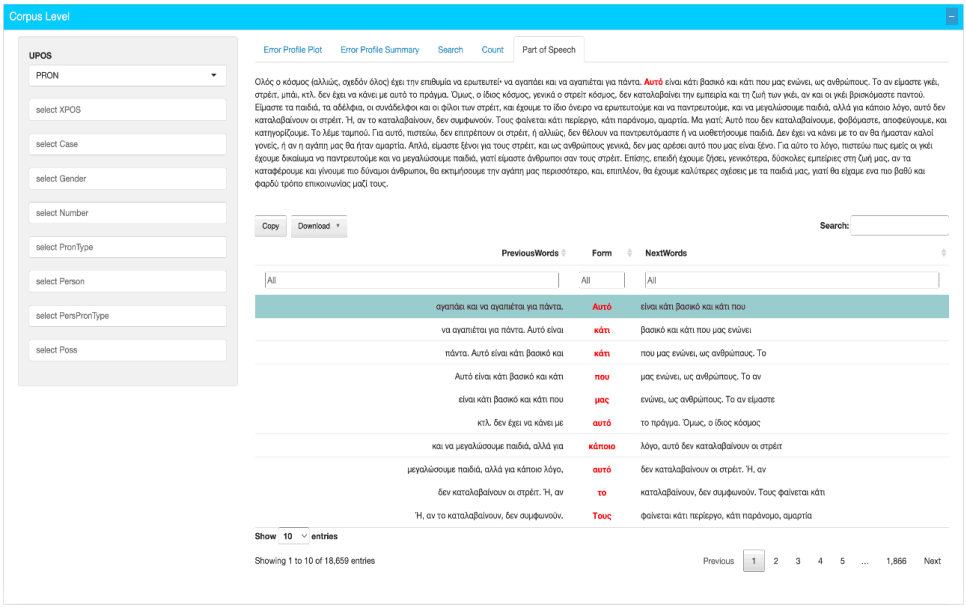

Figure 6. Selecting an individual row from the concordance table for PRON tags

3.3. Text Level

At the text level, the GLCII Gateway offers two essential functionalities for the user. The first functionality, known as the "Error Profile Plot," provides a graphical visualization of the error annotated categories within a selected production (i.e. text). This feature allows researchers to gain a comprehensive overview of the distribution and frequency of different error categories present in a selected text. Researchers can quickly identify areas that are prone to non-target performance and assess the relative prominence of different error categories.

The second functionality, under the "Error Profile Summary" tag, presents researchers with the raw numbers of errors for each annotated category on the same selected text, along with their ratio within the total occurrences of the same category in the text. This feature enables a more detailed analysis of the error categories, allowing researchers to compare and evaluate their occurrence frequencies. Both functionalities at the text level are indispensable for L2-learner language researchers, since they provide valuable tools for identifying and understanding areas of non-target performance within a specific text. Moreover, they allow them to focus on specific linguistic features or structures that may require further investigation.

4. DISCUSSION

As an open access resource, GLCII promotes reproducibility and transparency of research since it encourages researchers to test and verify or falsify hypotheses and to present new findings, thus strengthening the overall reliability and credibility of the field. The GLC Gateway is a web-based platform that provides free access to the GLCII, the largest corpus for L2 Greek. Moreover, it allows a powerful platform for exploring and analyzing learner language with an intuitive and user-friendly interface. Its wide range of searching tools allow users to explore GLCII and perform detailed analyses of developing interlanguage.

The inclusion of annotated error categories in the corpus enhances its utility for error analysis and investigation of developing L2 grammars. Researchers can delve into the frequencies and distributions of these error categories, gaining valuable insights into learners' linguistic challenges. Moreover, the GLC Gateway provides a comprehensive overview of the GLCII's annotation scheme, enabling users to understand the error annotation protocol followed during the annotation phase.

In summary, the GLC Gateway offers free access to GLCII, a comprehensive L2 Greek learner corpus and provides researchers and educators with indispensable tools for conducting further in-depth analysis and preparing targeted teaching materials. Its emphasis on ecological validity and enhanced contextualization of the learner's linguistic behavior make it a valuable resource for advancing our understanding of second language acquisition and informing language teaching practices.

References

Bell & Payant 2021: P. Bell, C. Payant, Designing learner corpora: Collection, Transcription and Annotation. In N. Tracy-Ventura & M. Paquot (eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 53–67). Abingdon: Routledge.

Brezina, Hawtin & McEnery 2021: V. Brezina, A. Hawtin, T. McEnery, The written British national corpus 2014–design and comparability. *Text & Talk*, *41*(5–6), 595–615.

Caines & Buttery 2017: A. Caines, P. Buttery, The effect of task type and topic on opportunity of use in learner corpora. In V. Brezina & L. Flowerdew (eds.), *Learner corpus research: New perspectives and applications* (pp. 5–27). London: Bloomsbury Publishing.

Gilquin & Gries 2009: G. Gilquin, S. Gries, Corpora and experimental methods: A state-of-the-art review. Corpus Linguistics and Lingustic Theory *5*(1), 1–26.

Granger 2002: S. Granger, A Bird's-eye View of Computer Learner Corpus Research. In S. Granger (ed.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Amsterdam & Philadelphia: Benjamins.

Gries & Berez 2017: S.T. Gries, A.L. Berez, Linguistic annotation in/for corpus linguistics. *Handbook of linguistic annotation*, 379–409.

Iakovou et al. 2016: M. Iakovou, O. Dima, I. Vasiliadi-Linardaki, S. Kitrou, F. Vlachou, B. Koutsoubou, T. Katsina, S. Perrea, F. Pappa, X. Kostakou, M. Kavvadia, SEPAME2: The first longitudinal corpus for Greek as an L2. In A. Moreno Ortiz & Ch. Pérez-Hernández (eds.). *CILC2016. 8th International Conference on Corpus Linguistics, 1*, 191–200.

Lozano 2021: C. Lozano, CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *38*(4), 965–983.

Lozano & Mendikoetxea 2013: C. Lozano, A. Mendikoetxea, Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds.), *Automatic treatment and analysis of learner corpus data* (pp. 65–100). Amsterdam: John Benjamins.

McEnery, Xiao & Tono 2006: T. McEnery, R. Xiao, Y. Tono, *Corpus-based Language Studies: An Advanced Resource Book*. London: Taylor & Francis.

Myles 2015: F. Myles, Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (eds.), *The Cambridge handbook of learner corpus research* (pp. 309–332). Cambridge: Cambridge University Press.

Paquot & Granger 2012: M. Paquot, S. Granger, Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Sinclair 2005: J. Sinclair, How to build a corpus. In M. Wynne (ed.), *Developing linguistic corpora: A guide to good practice* (pp. 79–83). Oxford: Oxbow Books.

Tantos & Papadopoulou 2014: A. Tantos, D. Papadopoulou, Stand-off annotation in learner corpora: Compiling the Greek Learner Corpus (GLC). In A. Díaz Negrillo & F. J. Díaz Pérez (eds.), *Specialisation and variation in language corpora* (pp. 15–40). Switzerland: Peter Lang International Academic Publishers.

Tantos, Amvrazis & Drakonaki 2023: A. Tantos, N. Amvrazis, C.E. Drakonaki, Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, *36*, 125–150.

Tavakoli & Foster 2011: P. Tavakoli, P. Foster, Task design and second language performance: The effect of narrative type on learner output. *Language Learning 61*, 37–72.

Tracy-Ventura & Myles 2015: N. Tracy-Ventura, F. Myles, The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, *1*(1), 58–95.

Τζιμώκας 2010: Δ. Τζιμώκας, Ηλεκτρονικό σώμα κειμένων εκμάθησης της νέας ελληνικής ως δεύτερης γλώσσας: προς ένα ερευνητικό και διδακτικό εργαλείο. *Μελέτες για την ελληνική γλώσσα 30*, 602–616.**Tantos**

Τάντος Αλέξανδρος - Αμβράζης Νίκος - Δρακωνάκη Έλενα - Δέσποινα Παπαδοπούλου - Χρυσάνθη Δεβελάσκα - Γερακίνη Δούκα - Πηνελόπη Κικιλίντζα - Ίλια Παπαφιλίππου

ΕΝΙΣΧΥΟΝΤΑΣ ΤΗΝ ΕΡΕΥΝΑ ΚΑΙ ΤΗ ΔΙΔΑΣΚΑΛΙΑ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΩΣ ΔΕΥΤΕΡΗΣ ΓΛΩΣΣΑΣ: Ο ΡΟΛΟΣ ΤΗΣ ΔΙΑΔΙΚΤΥΑΚΗΣ ΠΛΑΤΦΟΡΜΑΣ ΤΟΥ ΕΣΚΕΙΜΑΘΙΙ

Περίληψη

Το ΕΣΚΕΙΜΑΘΙΙ, το μεγαλύτερο σώμα κειμένων για την ελληνική ως δεύτερη γλώσσα (Γ2), αποτελεί τόσο για ερευνητές όσο και για εκπαιδευτικούς ένα πολύτιμο εργαλείο, το οποίο είναι ελεύθερα προσβάσιμο μέσω της Πύλης για το ΕΣΚΕΙΜΑΘ (Ι και ΙΙ). Η Πύλη είναι η διεπαφή του ΕΣΚΕΙΜΑΘ με τους χρήστες εξασφαλίζοντας τα απαραίτητα εργαλεία για μια σειρά χρηστικών λειτουργιών που αξιοποιούν τον πόρο. Το περιβάλλον της διεπαφής είναι διαθέσιμο μέσω της ιστοσελίδας https://glc.lit.auth.gr/app/GLC_Gateway και κατευθύνει μέσω της απλής δομής του και των ευδιάκριτων εργαλείων του στην αναζήτηση και ανάκτηση γλωσσικών δεδομένων από το ΕΣΚΕΙΜΑΘΙΙ. Μια βασική κατεύθυνση σε αυτή τη διαδικασία αφορά στην αξιοποίηση των επισημειωμένων δεδομένων. Πρόκειται για την επισημείωση λαθών που έχει εφαρμοστεί σε μια σειρά από μορφοσυντακτικά χαρακτηριστικά με κομβικό ρόλο στην ανάπτυξη της διαγλώσσας στη Γ2. Ερευνητές και εκπαιδευτικοί μπορούν να αναζητήσουν και να ανακτήσουν αυτά τα δεδομένα εφαρμόζοντας φίλτρα που ενσωματώνουν το ευρύ πλαίσιο μεταδεδομένων που έχει συλλέξει η ερευνητική ομάδα του έργου. Με αυτόν τον τρόπο η ερμηνεία των λαθών αναβαθμίζεται μέσω του πλαισίου αναφοράς της καθώς επιτρέπει τη συσχέτιση του λάθους με το γλωσσικό προφίλ του ομιλητή αλλά και το είδος της παραγωγής λόγου. Αυτή η προσέγγιση ενισχύει την οικολογική ισχύ της έρευνας διευκολύνοντας την εξαγωγή αξιόπιστων επιστημονικών συμπερασμάτων. Μία δεύτερη κατεύθυνση αφορά στην ευρύτερη αξιοποίηση της γλωσσικής παραγωγής των μη φυσικών ομιλητών πέρα από τα στενά όρια του γλωσσικού λάθους. Η Πύλη του ΕΣΚΕΙΜΑΘΙΙ ενσωματώνει μια σειρά από φίλτρα παραμετροποιημένης αναζήτησης που δίνουν τη δυνατότητα για ανάκτηση λεξιλογίου, δομών παγιωμένου λόγου και συμφραστικών πινάκων, ενώ παράλληλα είναι δυνατή και η εξαγωγή συχνοτήτων που αποτυπώνουν το ανάγλυφο της εμφάνισης των λέξεων ή των δομών στο σύνολο των παραγωγών. Είναι σημαντικό, επίσης, να σημειωθεί πως η εκάστοτε ανάκτηση δεδομένων συνοδεύεται και από γραφηματική απεικόνιση τόσο σε επίπεδο του συνόλου του σώματος όσο και σε επίπεδο παραγωγής.

Λέξεις-κλειδιά: ΕΣΚΕΙΜΑΘΙΙ, Πύλη ΕΣΚΕΙΜΑΘΙΙ, Σώμα Κειμένων, Δεύτερη Γλώσσα