

Creation of an electronic learner corpus of French as a foreign language

Olympia Tsaknaki

Department of French, Aristotle University of Thessaloniki
tsaknaki@frl.auth.gr

Abstract. According to research findings, learner corpora can have many functions in the field of Applied Linguistics. This was the thinking behind the creation of the *KPG Learner Corpus of French*¹. This interlanguage corpus is comprised of written productions in French obtained from Greek-speaking learners/users of French who participated in the differentiated and graded certification examinations for the National Foreign Language Exam System (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ) in Greece. It is an ongoing project, which, in its present form, is comprised of 630 written productions of Levels A, B and C (165 000 words) covering a period of two years (2018-2019). As regards its potential applicability and value, useful observations could bring to the fore choices, problems and errors relating to the learners' interlanguage.

Keywords: *learner corpus, French as a foreign language, National Foreign Language Exam System (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ), language certification, written production*

Elektroninio prancūzų kaip užsienio kalbos mokinių tekstyno kūrimas

Santrauka. Ankstesnių tyrimų rezultatai rodo, jog mokinių tekstynai gali atlikti daugybę funkcijų taikomosios lingvistikos srityje. Būtent tokia nuostata buvo remtasi kuriant KPG prancūzų kalbos mokinių tekstyną (angl. *KPG Learner Corpus of French*). Ši tekstyną sudaro rašytiniai tekstai prancūzų kalba, surinkti iš graikiškai kalbančių prancūzų kalbos besimokančiųjų ir (arba) vartotojų, dalyvavusių diferencijuotuose ir įvertintuose Nacionalinės užsienio kalbų egzaminų sistemos (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ) sertifikavimo egzaminuose Graikijoje. Tai tęstinis projektas, kurio metu surinktą duomenų bazę šiuo metu sudaro 630 A, B ir C lygių rašto darbų (iš viso 165 000 žodžių). Tekstynas apima dvejų metų laikotarpį (2018–2019 m.). Šiuo tekstynu paremti tyrimai gali atskleisti mokinių kalbai būdingas vartosenos tendencijas, sunkumus ir tipines klaidas, kas gali būti vertinga prancūzų kalbos kaip užsienio kalbos mokymo praktikoje.

Raktažodžiai: *mokinių tekstynas, prancūzų kalba kaip užsienio kalba, Nacionalinė užsienio kalbų egzaminų sistema (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ), kalbos sertifikavimas, rašytinė kalba*

¹ A large part of the research was funded by the *Research Committee of the Aristotle University of Thessaloniki*. The procedure of data digitization and error annotation was assisted by the PhD researcher Natalia Sakellari.

1. Introduction

The aim of this work is to present and describe the newly created electronic *KPG² Learner Corpus of French*. This corpus is compiled from written productions in French of candidates who participated in the differentiated and graded examinations for the National Foreign Language Exam System (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ). First, we deal with some information issues related to this exam system. Then, we present the characteristics of the corpus. Finally, we refer to the benefits of the learner corpora and, especially, the utility of the corpus created.

2. The National Foreign Language Exam System³

The National Foreign Language examinations “which represent a ‘glocal’ multilingual examination battery, are the first of their kind in Europe, and take into account local needs, global conditions of knowledge and production, and international concerns regarding testing and assessment” (Karavas & Mitsikopoulou 2019: back cover).

Institutionalized by law in 1999, they are organized and administered semiannually⁴ by the Hellenic Ministry of Education. With a view to promoting multilingualism and plurilingualism in Greece, the languages for which levels of language proficiency are certified are English, French, German, Italian, Spanish and Turkish⁵. The examinations are based on the principles of the *Common European Framework of Reference for Languages* (Council of Europe 2001, 2020).

The overall aim of this exam system is to certify that candidates, wishing to be certified, have the competences and skills expected at different levels of proficiency, as defined by the Council of Europe’s 6-level scale, - to make socially purposeful use of the target language in different social contexts. The more specific objectives are as follows:

to measure the candidates’ a) competences to comprehend and produce oral and written discourse, b) skills to act as cross-linguistic mediators (in speaking and writing), and c) abilities to use their awareness of how language operates in different social contexts and discursive environments as to make socially purposeful meanings⁶.

The levels of proficiency certified include A1+A2 (Basic User), B1+B2 (Independent user) and C1+C2 (Proficient user). Competences are measured with the following modules: Reading comprehension and language awareness, Writing and written mediation⁷, Listening comprehension, and Speaking and oral mediation.

The overall responsibility of ensuring the quality and validity of the tests and the general supervision during the evaluation process of the written and oral productions of the candidates rests with the foreign language departments of the National and Capodistrian University of Athens and the Aristotle University of Thessaloniki. The Department of French Language and Literature of the latter institution is responsible for the certification in French.

² *KPG* is the transliteration of the initialism *KΠΓ*.

³ More information can be found at: <https://rcel2.enl.uoa.gr/kpg/>

⁴ There are two exam sessions, in November and in May.

⁵ The languages are presented by alphabetical order and not by the number of candidates or the chronological order of their integration into the exams.

⁶ *Overview of the State Certificate of Language Proficiency*: https://rcel2.enl.uoa.gr/kpg/files/KPG_Overview_2016.pdf

⁷ Candidates’ mediation performance is evaluated in Levels B and C.

The teams of experts who are responsible for the production of the tests take into account parameters such as sociocultural factors and topics of discussion and the age as well as the linguistic and cultural background of all candidates. The text types comply with the subject of the activity, the writing task, and the exam level. The texts included in the tests and the topics discussed are expected to be cognitively appropriate for the candidates so they are selected in accordance with this criterion. It is important to mention that Greek is the common language of the candidates, used either as a mother tongue or a language of schooling and the working environment. After each session, the collected data are statistically processed by experts of the field. The results of this processing are useful feedback provided to the test developers for features such as easiness/difficulty, reliability, and validity of language tests.

3. Corpus description

The corpus is comprised of digitized authentic textual data stored in a database. The digitized texts were reproduced manually word for word, including errors. All productions, i.e. written responses in the target language on the basis of a verbal or visual stimulus, were texts written in the framework of the module “Writing and written mediation”. Candidates are not allowed to bring to the location of the exams any kind of resources such as notes, reference materials, books, mobile phones, tablets or digital watches. The texts are hand-written in a notebook. When the candidates complete the exam, they hand over the notebook to the invigilators.

Files may be accessible in plain text or in annotated form. They may be processed by any text retrieval system, and search is possible through a concordancer. The target language is French and the first language is Greek, having predominantly the status of a mother tongue. According to the personal data protection rules, all productions were anonymized. In this respect, it should be underlined that the members of the research scientific team of French are the only ones that may have access to the written productions of French. Access to the written data is exclusively available to the research scientific team of each language.

The proficiency levels we are interested in are A (A1/A2), B (B1/B2), and C (C1/C2). The current size of the corpus is 165 000 words, which as follows within each broad proficiency level: 30 000 (A), 50 000 (B), and 85 000 (C). The component texts of the corpus were extracted from a sample collection of texts that originates in different exam sessions with the purpose of ensuring the corpus balance in respect of textual type (e.g. descriptive, narrative, and argumentative texts), textual genre (formal or informal mail, article, and questionnaire), medium (newspaper, magazine, and website) and thematic area (e.g. holidays, sports, hobbies, education, environment, tourism, and friendship). The corpus can be explored as a whole, but it is designed in such a way as to be also processable by proficiency level. Subcorpora consisting of texts corresponding to one of the three language proficiency levels are available. The learner’s proficiency level is a variable that may affect the nature of interlanguage (Gilquin 2015: 6). Candidates self-evaluate their competences and decide the level at which they wish to be certified. The material produced is classified according to this decision.

The corpus was processed through the open-source corpus processing suite Unitex⁸ by means of morphological dictionaries of French and transducers, i.e. grammars producing outputs. The results obtained were not always the expected ones. This was primarily due to the presence of erroneous forms in many of the written productions, which frequently resulted in the identified words being treated as unknown.

⁸ <https://unitexgramlab.org/>

In this corpus, text-related metadata includes the exam session and the year, the text's length (average of words: 100 (A) 180 (B), 350 (C)), the time given to produce the task (40 min. (A), 85 min. (B), or 120 min (C)), and, finally, the textual genre. The situation is always the same, namely language proficiency testing in the framework of the National Foreign Language Exam System.

For the time being, due to restrictions on personal data, the corpus is not publicly available to the broader research community. It must also be noted that, currently, information such as learning experience, languages known by the candidates or residence in a French-speaking country are elements that cannot be included in the candidate-related metadata of the corpus. In view of the future enrichment of the data, a website where authorized users could have access to the corpus can be envisaged. In addition to the above, it is important to mention that the development of the corpus is an ongoing project due to the continuously increasing data resulting from biannual examinations.

4. Error taxonomy

An error is defined as any deviation that contradicts the rules applying to the French language. The error taxonomy proposed follows the principles set out in the framework of linguistic error analysis, the theory of interlanguage and the classification of linguistic errors into intra- and interlinguistic errors (Corder 1967, 1971, 1980, Douglas Brown 2000, Dulay, Burt & Krashen 1982, Selinker 1972). Given the complex nature of an error annotation system, Granger's (2003: 467) proposal is adopted: the listing was set up with the purpose of being exhaustive in order to include all categories and flexible to allow rapid and efficient processing. The error annotation system is consistent, and it can also be reusable.

Errors were annotated manually or semi-manually with error tags in order to be properly identified. It must be noted that text tagging is a lengthy and complicated procedure and a key precondition for the correct interpretation of the results. Errors were assessed by two annotators with French teaching experience, members of the research scientific team of French, trained and experienced evaluators and test developers. According to the procedure followed, the annotators, based on an error taxonomy proposed according to the detailed evaluation criteria of the written productions (task completion, text grammar, sentence grammar and lexical features)⁹, first tagged an excerpt. Other members of the scientific team, evaluators, and test developers as well, were also consulted on how to enhance the error taxonomy. This led to several changes and additions in the error taxonomy classification, which includes language performance deviations on the grammatical, orthographic, syntactic and lexical level as well errors on appropriateness of discourse, i.e. the use of the appropriate structural forms that characterize the type of text or the textual genre, e.g. errors in the opening and closing lines of an email. The erroneous omission or addition of a lexical unit was also annotated. Our aim was to create an informative yet manageable error taxonomy which would not be either very simple and coarse or very complex and expansive. In practical terms, this enables a possible modification of the taxonomy and the adoption of a more detailed approach where applicable.

5. Learner corpora

Many studies so far have focused on the usefulness of learners' corpora in the area of Applied Linguistics. Corpora have described and analyzed ways in which they can support the educational procedure in teaching and learning a foreign language.

⁹ For more information on evaluation criteria, see: https://rcel2.enl.uoa.gr/kpg/gr_kpgcorner_jul_aug2009.htm

In contrast to other types of data that have traditionally been used in second language acquisition (SLA) research, learner corpora provide systematic collections of authentic, continuous and contextualized language use by foreign/second language (L2) learners stored in electronic format. (Callies 2015: 35)

There is extensive research literature on the particular importance and usefulness of learner corpora in the study and use of the material resulting from the observation of students' language errors (cf. Alonso-Ramos 2016, Gilmore 2009, Gotsoulia & Dendrinos 2011, Granger 2002, 2012, Han et al. 2010, Ionin & Díez-Bedmar 2021, Katinskaia et al. 2022, Lozano 2009, Nagata, Whittaker & Sheinman 2011, O'Sullivan & Chambers 2006, Yoon & Jo 2014).

During the period before the second half of the 20th century, language errors were considered as real obstacles to the realization of communication skills and the achievement of learning goals (Corder 1967). This belief has been questioned by the global scientific community, and language errors are now considered useful sources of data in the field of foreign language teaching-learning where the study of errors may become a very helpful instrument (Ellis 1994 : 20). Through a collection of learner texts, teachers and researchers can draw on authentic examples to answer research questions they have raised about the interlanguage of foreign language learners, retrieve information on erroneous and correct use of the language, and draw conclusions based on the quantitative and qualitative study of linguistic errors. The analysis of linguistic errors may concern the written output of one learner or of a group of learners.

Nevertheless, as Granger notes:

Although learner corpus research (LCR) and second language acquisition (SLA) studies both partake of the general field of L2 studies, it must be acknowledged that they are still essentially two different worlds. (2021: 243)

In other words, resources of this kind should be created in such a way that emphasizes their advantages, and teachers should make the best use of their possibilities in language teaching and interlanguage comparison (Granger 1996, 2003, Nesselhauf 2004).

6. Usefulness of the *KPG Learner Corpus of French*

Focusing particularly on the *KPG Learner Corpus of French* and potential benefits that could be obtained, we could argue that derived data can reveal the most common errors committed and assist learners of French as a foreign language to achieve a better performance. Learners are invited to demonstrate their ability to consciously use the foreign language in a certain context determined by the instructions and meet the requirements of a functional approach of language by acting as members of a society.

The study of the communicative skills that learners developed, as well as the detection, identification, and classification of language errors in relation to their language skills, such as syntax, vocabulary and spelling, may lead to important research results for the benefit of the students of every level of language proficiency. The feedback that can be derived from the comparative study on the differences in terms of the type and frequency of error production resulting from basic, independent, and proficient users of the language (Council of Europe 2001, 2020) is another possibility of similar character (Tsaknaki 2020, 2022).

The corpus offers the possibility to locate the errors committed, consider their importance, and discuss them. Data without errors may also be used to explore test-taking strategies selected by candidates and compare the ways the latter have coped with the task completion. A focus could also be on the evaluation criteria to observe if, how and to which extent these are satisfied.

Moreover, research results may be used in lexicography and the creation of textbook grammars that support both the teaching and learning of French as foreign language. The investigation of the Greek-speaking learners' language profile may also assist stakeholders in foreign language education to write innovative pedagogical materials and books that will meet the needs of the specific public and help develop detailed curricula or renew the existing ones. Specifically, regarding Greek-speaking learners/users of French, this corpus constitutes the necessary scientific and research basis for the continuous updating of the National Foreign Languages Curriculum¹⁰.

7. Innovative features of the *KPG Learner Corpus of French*

After searching information provided on the website “Learner corpora around the world” of the Institute for Language and Communication de l’Université Catholique de Louvain¹¹ and the report concerning the corpora that are part of the CLARIN infrastructure (Lenardič et al. 2018), we concluded that, in recent years, remarkable learner corpora have been created; however, these mainly concern English as a foreign language. Given that the protection and promotion of multilingualism go along with the policy pursued by the European Union, more corpora should include or provide texts where French or another language is the target language. There are multilingual corpora with different language combinations where French is one of the target languages. They are either spoken or written and composed of different text types. The written data originate from essays, narrative texts, academic texts, letters, journalistic texts, tests, forum posts, and other text types. There are also corpora with no specific information about their composition. As regards the levels to which written texts correspond, none of them explicitly mentions the existence of productions ranging between A1-C2 according to the *Common European Framework of Reference for Languages*. As Hong and Cao point out:

previous research has commonly targeted at advanced learners of English, whereas relatively very little research attention has been directed to young learners of English at the beginning and the lower-intermediate levels (see Tono et al. 2012 for an exception). (2014: 203)

They also add that:

Second, while argumentative writing has been well studied in previous learner corpus research, there is a lack of research on other text types, for instance, descriptive writing. (Hong and Cao 2014: 203)

These findings show the need for more corpora that cover a wider range of proficiency levels and text types. The Greek language is present only in one corpus: The *Chy-FLE* (Cypriot Learner Corpus of French), compiled by Freiderikos Valetopoulos at the Université de Poitiers in collaboration with the University of Cyprus, where the target language is French, and the first language is Modern Greek (and Cypriot Greek). The proficiency levels are intermediate and advanced. The corpus, currently consisting of 250 000 words, is under development.

The *KPG Learner Corpus of French* can be a linguistic resource serving (a) speakers of a less widely used and less taught language, in our case, Greek, and (b) teachers of French as a foreign language in

¹⁰ There is an interface between the National Foreign Languages Curriculum and the National Foreign Language Exam System, both of which are adapted to the proposals of the Common European Framework of Reference for Languages.

¹¹ Centre for English Corpus Linguistics (date of access 18-1-23): Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>

the country either in public schools or in the private sector. Teachers of French as a foreign language who are not necessarily Greek-speaking may also be defined as a target group.

Candidates of the National Foreign Language Exam System are distributed across different geographical areas of the country; they belong to different age groups, have different education levels, and differ in the way they have learnt the foreign language. With reference to the skill of producing written texts, the candidates must take into account, alongside the correct use of the language structural elements, the communicative role they must fulfil, considering that the linguistic production is carried out in concrete communicative contexts. Candidates should comply with the necessary contextual factors imposed every time as people do in real everyday life situations when they are invited to respond to the requirements of different social rules. Furthermore, they should consider the subject treated by the document that launches the production, the type of text and the textual genre. The achievement of the communication goals is assumed as a fundamental parameter during the assessment of the test papers.

The current article reports on the creation and description of a learner corpus composed of written productions in French obtained from Greek-speaking learners/users of French who participated in the differentiated and graded certification examinations for the Greek National Foreign Language Exam System (Κρατικό Πιστοποιητικό Γλωσσομάθειας, ΚΠΓ). The aim of the *KPG Learner Corpus of French* is to fill a void as regards the language pair Greek-French, to offer interesting crosslinguistic insights, and to provide valuable information on the interlanguage of learners of French as a foreign language.

References

- Alonso-Ramos, M. (ed.). 2016. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, Amsterdam: Benjamins.
- Callies, M. 2015. Learner corpus methodology. *The Cambridge Handbook of Learner Corpus Research*, S. Granger, G. Gilquin, F. Meunier (eds.). Cambridge: Cambridge University Press. 35-56. Prieiga internetu <https://doi.org/10.1017/CBO9781139649414.003> (žiūrėta 2023-01-18)
- Corder, S.P. 1980. Que signifient les erreurs des apprenants ? *Langages* 14^e année, 57, 9-15. Prieiga internetu <https://doi.org/10.3406/lgge.1980.1833> (žiūrėta 2023-01-18)
- Corder, S.P. 1971. Le rôle de l'analyse systématique des erreurs en linguistique appliquée. *Bulletin CILA (Commission Interuniversitaire Suisse de Linguistique Appliquée)* (« Bulletin VALS-ASLA » depuis 1994) 14, 6-15.
- Corder, S.P. 1967. The significance of learners' errors. *IRAL* 5, 161-170.
- Council of Europe, 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume with new descriptors*. Strasbourg: Council of Europe Publishing. Prieiga internetu www.coe.int/lang-cefr. (žiūrėta 2023-01-18)
- Council of Europe, 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge University Press. Prieiga internetu <https://rm.coe.int/1680459f97>. (žiūrėta 2023-01-18)
- Douglas Brown, H. 2000. *Principles of language learning and teaching*. England: Longman.
- Dulay, H., M. Burt, S. Krashen. 1982. *Language Two*. Oxford: Oxford University Press.
- Ellis, R. 1994. *Second Language Acquisition*, Oxford: Oxford University Press.
- Gilmore, A. 2009. Using online corpora to develop students' writing skills. *ELT Journal*. 64(3), 363–372.
- Gilquin, G. 2015. From design to collection of learner corpora. *The Cambridge Handbook of Learner Corpus Research*. S. Granger, G. Gilquin, F. Meunier (eds.), Cambridge: Cambridge University Press. 9-34. Prieiga internetu <http://hdl.handle.net/2078.1/145509> -- DOI : 10.1017/CBO9781139649414.002 (žiūrėta 2023-01-18)
- Gotsouli, V., B. Dendrinos. 2011. Towards a Corpus-based Approach to Modelling Language Production of Foreign Language Learners in Communicative Contexts. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics. 557–561.

- Granger, S. 2021. Have learner corpus research and second language acquisition finally met? *Learner corpus research meets second language acquisition*. B. Le Bruyn, M. Paquot (eds.). Cambridge: Cambridge University Press. 243-257.
- Granger, S. 2012. How to Use Foreign and Second Language Learner Corpora. *Research Methods in Second Language Acquisition. A Practical Guide*. A. Mackey, S.M. Gass (eds.). Chichester: Wiley-Blackwell, 7–29.
- Granger, S. 2003. Error-tagged learner corpora and CALL: a promising synergy. *CALICO* 20:3, 465–480.
- Granger, S. 2002. A Bird's-eye View of Learner Corpus Research. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. S. Granger, J. Hung, S. Petch-Tyson (eds.). Amsterdam: John Benjamins. 3–33.
- Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. *Lund Studies in English* 88, 37–51.
- Han, N., J. Tetreault, S. Lee, J. Ha. 2010. Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 763-770.
- Hong, H., F. Cao. 2014. Interactional metadiscourse in young EFL learner writing: A corpus-based study. *International Journal of Corpus Linguistics* 19(2), 201-224. DOI: 10.1075/ijcl.19.2.03hon.
- Ionin, T., M.B. Díez-Bedmar. 2021. Article use in Russian and Spanish learner writing at CEFR B1 and B2 Levels: effects of proficiency, native language, and specificity. *Learner corpus research meets second language acquisition*, 10-38.
- Karavas, E., B. Mitsikopoulou (eds.). 2019. *Developments in Glocal Language Testing The Case of the Greek National Foreign Language Exam System*, Peter Lang.
- Katinskaia, A., M. Lebedeva, J. Hou, R. Yangarber. 2022. Semi-automatically Annotated Learner Corpus for Russian. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 832–839.
- Lenardič, J., T. Lindström Tiedemann, D. Fišer. 2018. *Overview of L2 corpora and resources*. (CLARIN report). CLARIN ERIC. Prieiga internetu <https://office.clarin.eu/v/CE-2018-1202-L2-corpora-report.pdf> (žiūrėta 2023-01-18)
- Lozano, C. 2009. CEDEL2: Corpus Escrito del Español como L2. *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente*. Bretones Carmen M. et al. (eds.). Almería: Universidad de Almería, 197–212.
- Nagata, R., E. Whittaker, V. Sheinman. (2011) Creating a manually error-tagged and shallow-parsed learner corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1210–1219, Portland, Oregon.
- O'Sullivan, Í., A. Chambers. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15(1), 49–68.
- Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10, 209-241. <http://dx.doi.org/10.1515/iral.1972.10.1-4.209> (žiūrėta 2023-01-18)
- Tsaknaki, O. (2022) Some reflections on reference in the teaching of French as a foreign language context. *Proceedings of the 29th International Social Science Conference "Recent Research and Ideas"*, University of Lisbon, 141-143.
- Tsaknaki, O. (2020) Investigating the French article system in FFL written production: a learner corpus study. *24th International Symposium on Theoretical and Applied Linguistics (ISTAL 24)*, Aristotle University of Thessaloniki, 787-803.
- Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology* 18(1), 96–117. <http://llt.msu.edu/issues/february2014/yoonyoonjo.pdf> (žiūrėta 2023-01-18)