



Științele limbajului

**Roxana Patraș
Loredana Cuzmici**
(editors)

**Tools, Methods, and Solutions
for the Exploration of Romanian Corpora**

INSTITUTUL EUROPEAN

Roxana Patraș, Loredana Cuzmici (editors), *Tools, Methods, and Solutions for the Exploration of Romanian Corpora*

© 2024 Institutul European Iași, pentru prezenta ediție

INSTITUTUL EUROPEAN

Iași, Str. Bălușescu nr. 2, et. I, O P 6, CP 1309

euroedit@hotmail.com.; www.euroinst.ro

Descrierea CIP a Bibliotecii Naționale a României

Tools, methods, and solutions for the exploration of Romanian corpora /

ed.: Roxana Patraș, Loredana Cuzmici. - Iași : Institutul European, 2024

Conține bibliografie

ISBN 978-606-24-0395-9

I. Patraș, Roxana (ed.)

II. Cuzmici, Loredana (ed.)

811.135.1

All rights reserved. No part of this work covered by the copyrights hereon may be reproduced or copied in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping – without written permission of publisher.

PRINTED IN ROMANIA

Roxana Patraş, Loredana Cuzmici
(editors)

TOOLS, METHODS, AND SOLUTIONS FOR THE EXPLORATION OF ROMANIAN CORPORA

INSTITUTUL EUROPEAN
2024

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS–UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

The revision and editing of this collection was performed by BolderWords translation & revision and was supported by the Altissia Chair in Digital Cultures and Ethics held by Chris Tanasescu at Université catholique de Louvain.

Table of Contents

INTRODUCTORY NOTES	7
Tactics: your DH is not my DH Roxana PATRAȘ	9
CREATING TOOLS FOR ROMANIAN CORPORA	23
RELATE: a modern processing platform for Romanian language Vasile PĂIȘ, Radu ION, Andrei-Marius AVRAM, Maria MITROFAN, Dan TUFÎȘ	25
Tools and Resources for Digitizing Historical Romanian Documents Victoria BOBICEV, Tudor BUMBU, Ludmila MALAHOV, Alexandru COLESNICOV, Svetlana COJOCARU, Liudmila BURTSEVA, Cătălina MĂRÂNDUC	53
Digitalization of Romanian Lexicography Elena Isabelle TAMBA	85
ADAPTING TOOLS FOR ROMANIAN CORPORA	97
Postdigital Creativity: modeling poetry translation with multiplexes, neural networks, and large language models Raluca TANASESCU, Chris TANASESCU	99
Challenges of Digitization and Digitalization for the Study of Writing: from archival survey to process analysis Georgeta CÎȘLARU	125
Towards a Digital Corpus-Based Method for Assessing Language Level in EFL Student Writing: a case study on Romanian undergraduate literary analyses Alexandru ORAVIȚAN, Mădălina CHITEZ	139
DIGITIZING ROMANIAN HERITAGE	151
A Computer-Assisted Analysis of Eminescu's <i>Doină</i> Ioana GALLERON	153
Mapping Eastern European Political Discourse and Inequalities through Public Monuments: a digital cartography crowdsourcing project Voica PUȘCAȘIU	175
The Title of the Feuilleton Novel: a model of semantic annotation Lucreția PASCARIU	193
Exercises in Literary Geography: where hajduks roam in the second half of the nineteenth century Alexandra OLTEANU	227

<i>DIGITUS DEI EST HIC!</i>	251
Proxy Structures in Computational Analyses of Text Corpora	253
Laura PRICOP	
The Many Faces of Virtual: charting the family resemblance network for conceptual understanding	267
Matei Alexandru STOENESCU	
CARTOGRAPHIES OF THE ROMANIAN NOVEL: THEN AND NOW	297
The Recontextualization of a Genre	299
Loredana CUZMICI	
Notes on the Contributors	321

Towards a Digital Corpus-Based Method for Assessing Language Level in EFL Student Writing: a case study on Romanian undergraduate literary analyses

Alexandru ORAVIȚAN
Mădălina CHITEZ

Introduction

In Romania, the issue of using digital tools and corpus-based methods to assess student writing remains largely unexplored. As interest is on the rise with regard to producing new assessment frameworks targeting writing among EFL (English as a foreign language) students in particular, the Common European Framework of Reference (CEFR) for languages constitutes a powerful authority, as it continues to be used as the main point of reference for evaluating student languages levels. Throughout Central and Eastern Europe, there is significant demand for assessment practice updates that integrate digital tools and that contribute to streamlining the evaluation process and to producing reliable CEFR-based scores. This chapter treads new ground by proposing a digital corpus-based method for assessing language levels in EFL student writing in the aggregate, while paying particular attention to the case of literary analyses written by students who study English in Romania.

Research on student writing in Eastern Europe, and Romania in particular, has intensified only over the last decade, and systematic inquiries into related regional and national data are scarce and still ongoing (Doroholschi *et al.* 2018, 6–7). One area where significant strides are being made is the humanities, since writing is an integral part of university curricula within such disciplines. In Romanian philology faculties, for example, student writing has been used as a tool with which students can practice their language skills, and instructors can assess student progress. At the undergraduate level, writing practice now primarily serves to create the conditions under which students can develop the independent research skills required to

complete the final BA paper and to pursue postgraduate studies. Perhaps nowhere are these skills more in-demand than in the case of literary analyses, which students produce during seminars or term examinations.

Craşovan and Rogobete (2020, 94) provide a valuable overview of how literary analyses can be used as tools for assessing Romanian students of philology. In this survey, they stress the flexibility of the genre in terms of probing and developing students' skills in understanding and interpreting literary texts. Based on a corpus of student texts written in Romanian (L1), the study reveals some of the main issues that arise when undergraduates produce literary analyses of Romanian texts in academic settings: the inaccurate contextualisation of the literary work under scrutiny, the misuse of terms, and a subservience to the marking scheme whereby the rubric is treated as the definitive structural model with which to comply rather than as a template that facilitates the development of more original critical thought (Craşovan and Rogobete 2020, 111).

Indeed, awareness of the specificities of the genre remains a crucial issue when attempting to implement pedagogical practices for academic writing in Romanian universities, especially in the case of L1 writing in Romanian. Arguably also playing a pivotal role in the context for achieving performance is the discipline, which may expose students to a rich range of writing practices while also helping them to develop familiarity with academic language. In the context of Romanian universities, much of this pedagogical effort has been left to the practical courses and modules offered by foreign language departments, which has led to a paradoxical situation: "it may be a greater challenge for the Romanian students to handle academic texts in their mother tongue than in a foreign language" (Chitez 2018, 135).

However, while Romanian students of foreign languages may benefit from more systematic and structured exposure to reading and producing academic writing, one issue that is decidedly connected to the quality of the texts submitted for assessment—but which has yet to be the object of substantial analytic scrutiny—is the level of linguistic competence demonstrated in students' L2 writing. It is arguably this dimension that achieves the connection between the receptive and productive halves of literary analysis, since analyzing a literary text in L2 implies having both a clear understanding of the set foreign-language text as well as the ability to express

an informed critical opinion in a foreign language in writing. This conjunction of the productive and the receptive may also be viewed as one of the main aims of foreign language study programs at Romanian universities: to foster among students the ability to build practical communication skills in the languages studied in the program, cultivate critical analysis skills for literary studies, and develop intercultural skills (Facultatea de Litere, Istorie și Teologie, UVT 2021).

Indeed, assessing language levels in the case of L2 productive and receptive skills has been a preoccupation in the Central and Eastern European educational context ever since the Common European Framework of Reference (CEFR) for languages—which was formally introduced in 2001—was piloted in the 1990^s (Little 2011, 2). This significant development led to fundamental changes in foreign language teaching across educational systems, with one of the ripple effects being the introduction of criteria for quality assessment and test design, and a decisive shift towards skills-based teaching (Bérešová 2011, 178–181). In the case of the Romanian university system, foreign language study programs have integrated the CEFR into their everyday practice, with notable efforts being made to impose a minimum required level of linguistic proficiency for BA graduates. For example, in the case of the Faculty of Letters, History and Theology at the West University of Timișoara (Romania), the reference level B2—which denotes an upper-intermediate independent user—is used as the minimum threshold for the BA graduation examination in both of the study programs that are conducted in foreign languages: Languages and Literatures, and Applied Modern Languages. One of the exam requirements for final-year students in the Languages and Literatures undergraduate program, for instance, is to produce an on-the-spot analysis of a literary text from the curricula (Facultatea de Litere, Istorie și Teologie, UVT 2021, 5).

Therefore, language level assessments of literary analyses produced by students writing in L2 should be a primary concern when it comes to developing pedagogical frameworks that are appropriate for foreign language programs and that seek to integrate advances in digital technology and corpus-based methodologies. Focused on literary analyses written in L2 English by undergraduate students, this paper draws on the Corpus of Romanian Academic Genres (entitled ROGER)—a bilingual corpus which consists of three years' worth of academic papers by L1 Romanian and

L2 English students (2018–2021) and which was compiled with the help of 27 collaborators from 9 Romanian universities. The paper draws on the ROGER corpus in attempt to provide a framework for assessing the CEFR language level of academic writing among L2 English students, by using digital tools to diagnose language level efficiently and to obtain an aggregate view of language level across the entire corpus. The working hypotheses of this case study are: (1) that the language level of L2 English students is more or less on par with that of L1 English students; (2) that digital tools and corpus-based methods can be used to gain valuable insights into more extensive studies on student writing pedagogy or on the washback effect on teaching and learning practices associated with student writing.

Methods

1. Assessing language levels

Born primarily out of a belief in a plurilingual Europe, the CEFR has—since its formal introduction in 2001 (Little 2011, 2)—become a legitimized permanent fixture in the assessment of linguistic proficiency both within and without the borders of the European Union, including when it comes to L2 language levels. Divided into three broad divisions (A: Basic User; B: Independent User; C: Proficient User) and six competence levels (A1: Breakthrough; A2: Waystage; B1: Threshold; B2: Vantage; C1: Effective Operational Proficiency; C2: Mastery), the CEFR enjoys broad international recognition and impacts language education on a global scale, particularly regarding evaluation and marking criteria. Its impact on assessment has occurred despite the fact that CEFR levels were not originally developed to rate learner performance (Zheng, Zhang and Yan 2016, 6). Moreover, with respect to how university language courses are designed and evaluated, the CEFR has sparked a series of debates, a pivotal one of which is connected to the minimum level of language proficiency a student is expected to attain upon completion of a given course (Little 2011, 11–13). Significant attention has been paid to this aspect, especially in the larger context of international language certification through standardized testing (such as Cambridge, TOEFL, IELTS etc.).

Over the past two decades, the goal of automating language proficiency assessment by aligning user-generated messages to the CEFR has been a constant

preoccupation for scholars both within and beyond the realm of digital humanities. Some of the research ventures in this area have had a considerable degree of success, both at the written and spoken levels (*see* Ballier *et al.* 2019, Wang *et al.* 2018, Schmalz and Brutti 2021). In the case of various digital tools that have been developed, the assessment is based on learner dictionaries that use the CEFR levels to target vocabulary relevant to specific learner categories. Some of these tools have been branded as self-assessment tools (*see* English Vocabulary Profile Online 2015, CEFR Checker 2021, Self-assess your language skills with our CEFR tool 2020) or employ the advantages of NLP (Bengoetxea, Gonzalez-Dios and Aguirregoitia, AzterTest: Open Source Linguistic and Stylistic Analysis Tool 2020). Using such tools to assess a corpus of student writing may offer the benefit of accessing swift generalizable information about language proficiency at the macro level.

2. Corpus

The corpus used for this study is a subset of the bilingual Corpus of Romanian Academic Genres (ROGER). The research project also dubbed ROGER, conducted at the West University of Timișoara (Romania), aimed to develop a corpus-based methodology to analyze academic writing genres at Romanian universities from a contrastive perspective, comparing genres written in Romanian and genres written in English (Bercuci and Chitez 2019, 737–738). The searchable ROGER corpus is part of the deliverables stemming from this project (Chitez, Bercuci *et al.* 2021). The subset used for this research venture is divided into two study corpora: the ROGER_LIT_EN corpus, which consists of literary analyses produced by BA-level students; and the ROGER_LIT_EN_BA corpus, which consists of BA papers on literary topics. ROGER_LIT_EN contains short literary analyses mostly about authors (such as William Shakespeare, Charles Dickens, and modernist writers like James Joyce and Virginia Woolf) who have become central to English-language literature syllabi; these texts have been produced within the setting of BA programs in English from three Romanian universities: the West University of Timișoara (65 texts), the “Alexandru Ioan Cuza” University of Iași (98 texts), and the University of Bucharest (2 texts). The ROGER_LIT_EN_BA corpus comprises BA papers on

literary topics written within the last five years at the West University of Timișoara (16 texts) and at the University of Bucharest (4 texts).

To gain a comparative understanding of L2-L1 dynamics, we have compared the two ROGER study corpora against an L1 reference subset, taken from the British Academic Written English Corpus (BAWE). This subset contains 93 L1 student texts that pertain to the discipline of English and that mainly consist of essays on and critiques of literary topics (Nesi *et al.* 2008).

Corpus	Number of texts	Word count	Average
ROGER LIT EN	165	263,942	1599
ROGER LIT EN BA	20	366,210	18310
BAWE English L1	93	212,964	2290

Table 1: Corpora profiles

3. Tool

An online and open-source NLP tool, MultiAzterTest has been chosen as the most suitable digital tool for this study, because it automates the process of rating the English texts in digital format according to the CEFR, and because it makes it possible to establish the incidence of words pertaining to a specific language level. MultiAzterTest is capable of providing over 125 linguistic measures for texts in English, Spanish and Basque, but these can also be easily adapted to other languages (Bengoetxea and Gonzalez-Dios 2021). In its first iteration as AzterTest, the tool was designed to assess the readability of texts and outline their main stylistic features (Bengoetxea, Gonzalez-Dios and Aguirregoitia 2020). The newer multilingual version, however, can calculate scores for a range of features, such as lexical diversity, classical readability formulae, word frequencies, and morphology and syntax. The feature relevant to this research project is vocabulary knowledge, as processing an English text with MultiAzterTest yields the number and incidence of words at each CEFR level (Bengoetxea and Gonzalez-Dios 2021, 11). The wordlist that we used was taken from the Oxford Learners Dictionary, which consists of 3,000 core words, at the A1 through B2 levels, and an additional 2,000 words, at the B2

through C1 levels (Bengoetxea and Gonzalez-Dios 2021, 7). Following experimentation and analysis that used a sequential minimal optimisation (SMO) classifier, the MultiAzterTest obtained a 69.29% accuracy score for its vocabulary knowledge feature (Bengoetxea and Gonzalez-Dios 2021, 22).

Results

All the texts making up the corpora under study have been processed using the MultiAzterTest's descriptive and vocabulary knowledge modules. *Figure 1* outlines the results obtained (in absolute values), in terms of the number of words identified in each CEFR level.

It should be noted that the word count associated with the “not in A1-C1” category merely implies that these words are neither part of the 3,000 core A1-B2 nor part of the 2,000 supplementary B2-C1 wordlists provided by the Oxford Learners Dictionary; arguably, these words reflect specialized vocabulary or proper names discussed in the texts themselves or in the in-text citations.

A broader understanding of the language level present in the corpora can be gained from the incidence of vocabulary per 1,000 words classified according to the CEFR levels, as shown in *Figure 2*.

An interuniversity comparison according to the incidence per 1,000 words, contrasted with the BAWE subset, yields the results for ROGER_LIT_EN and ROGER_LIT_EN_BA presented in *Figure 3*.

Discussion and conclusion

Overall, both when it comes to word count (*Figure 1*) and incidence per 1,000 words (*Figure 2*), the hypothesis of the closeness between undergraduates' L2 English language levels and that of L1 students is partially confirmed by the results. These similarities are particularly evident at the intermediate and advanced levels, where the values scored by the ROGER study corpora are increasingly similar to those recorded for BAWE_English_L1. More specifically, the combined incidence per 1000 words for the B1-B2 thresholds is 59.02 for ROGER_LIT_EN, 64.58 for ROGER_LIT_EN_BA, and 71.04 for BAWE_English_L1, which situates the L2 BA papers in close proximity to the texts by L1 students. For level C1, the results are all

the more striking in that ROGER_LIT_EN_BA scored a higher incidence (16.99) than BAWE_English_L1 (16.1).

The results thus confirm the assumption that implementing CEFR-based pedagogy in the context of L2 writing among students will lead to success in language proficiency, particularly in the context of an existing language level threshold for final-year examinations. Moreover, while limited in scope, this study demonstrates that digital tools and corpus-based research methods can be used for the purposes of assessing language level not only in the case of undergraduate student writing but at all stages of developing EFL productive skills according to the CEFR. Thus, new grounds are established for implementing a new corpus-based method in the assessment of EFL-learner writing, and such implementation can be used to gain insight into the teaching and assessment of academic writing among learners, at the interuniversity level.

Annex 1: Figures

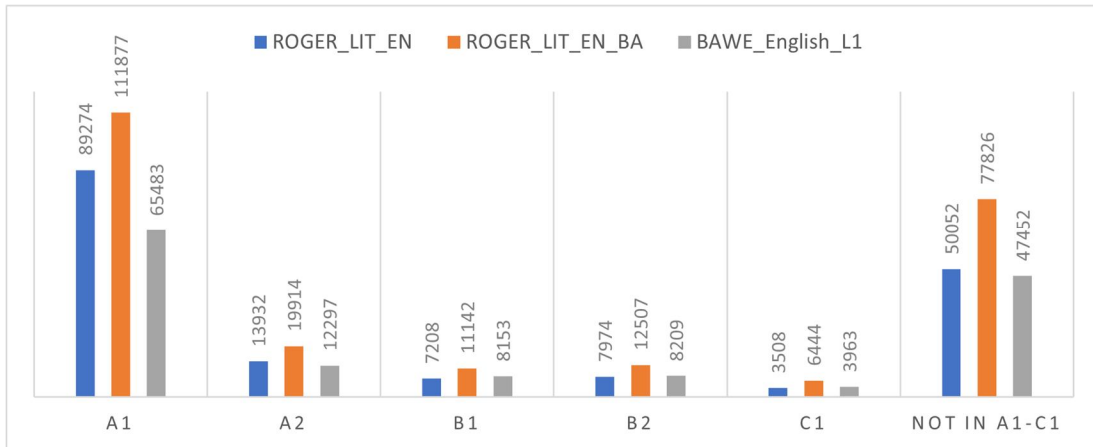


Figure 1: Word count - CEFR levels

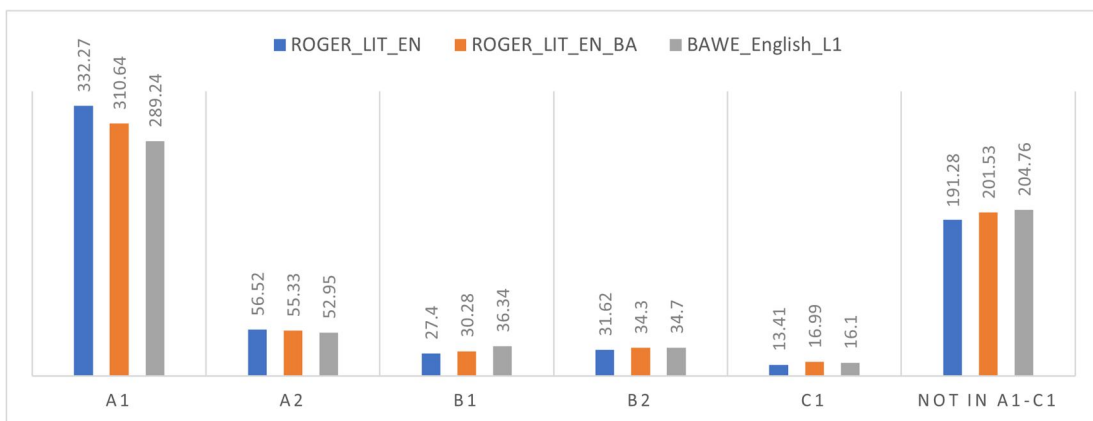


Figure 2: Incidence per 1k words

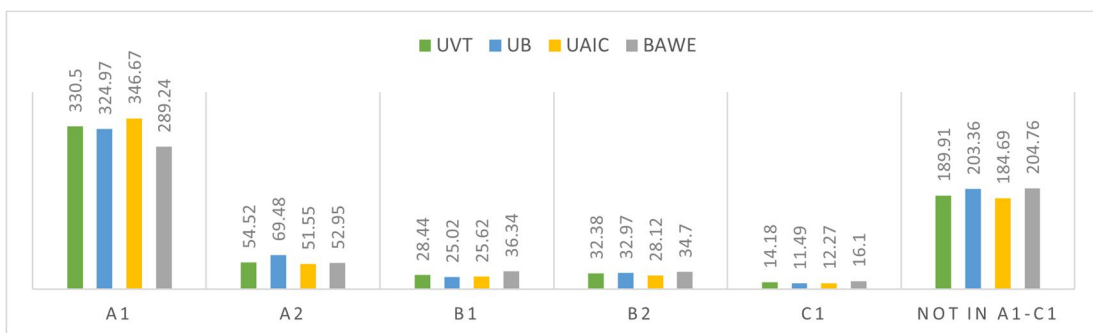


Figure 3: ROGER_LIT_EN Interuniversity incidence per 1k vs. BAWE_English_L1

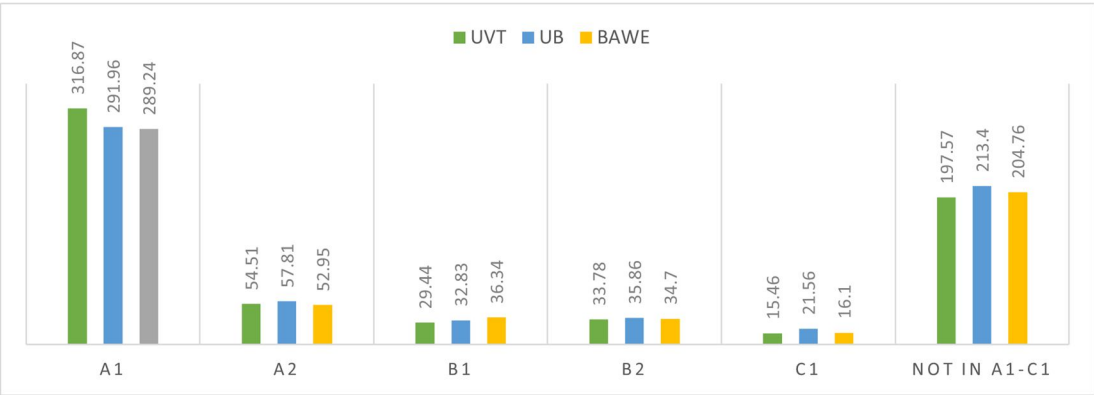


Figure 4: ROGER_LIT_EN_BA Interuniversity incidence per 1k vs. BAWE_English_L1

References

- Ballier, Nicolas, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. "A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors." In *Transforming Learning with Meaningful Technologies. EC-TEL 2019. Lecture Notes in Computer Science*, edited by Scheffel M, Broisin J, Pammer-Schindler V, Ioannou A and Schneider J. Cham: Springer. doi:10.1007/978-3-030-29736-7_23.
- Bengoetxea, Kepa, and Itziar Gonzalez-Dios. 2021. "MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment." *ArXiv-Computation and Language* (Cornell University). Accessed February 10, 2022. <https://arxiv.org/pdf/2109.04870>.
- Bengoetxea, Kepa, Itziar Gonzalez-Dios, and Amaia Aguirregoitia. 2020. "AzterTest: Open Source Linguistic and Stylistic Analysis Tool." *Procesamiento del Lenguaje Natural* (6): 61–68.
- Bercuci, Loredana, and Mădălina Chitez. 2019. "A corpus analysis of argumentative structures in ESP writing." *International Online Journal of Education and Teaching (IOJET)* 6 (4): 733–747.
- Bérešová, Jana. 2011. "The Impact of the Common European Framework of Reference on Teaching and Testing in Central and Eastern European Context." *Synergies Europe* (6): 177–190.
2021. *CEFR Checker*. Cathoven AI. Accessed February 10, 2022. <https://www.cathoven.com/en/cefr-checker/>.
- Chitez, Mădălina. 2018. "Corpus Linguistics Meets Academic Writing: Examples of Applications in the Romanian EFL Context." In *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, edited by Mădălina Chitez, Claudia Ioana Doroholschi, Otto Kruse, Lukasz Salski and Dumitru Tucan, 133-148. Cham: Springer.
- Chitez, Mădălina, Loredana Bercuci, Andreea Dincă, Roxana Rogobete, and Karla Csürös. 2021. *Corpus of Romanian Academic Genres (ROGER)*. Centre for Corpus-Related Digital Approaches to Humanities. <https://roger-corpus.org/>.
- Crașovan, Elena, and Roxana Rogobete. 2020. "Analiza de text literar: o perspectivă cantitativă asupra textelor studențești." *The Annals of the West University of Timișoara. Humanities Series* 58: 93–116.
- Doroholschi, Claudia Ioana, Dumitru Tucan, Mădălina Chitez, and Otto Kruse. 2018. "Introduction: Understanding Academic Writing in the Context of Central and Eastern European Higher Education." In *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, edited by Mădălina Chitez, Claudia Ioana Doroholschi, Otto Kruse, Salski Łukasz and Dumitru Tucan, 1–14. Cham: Springer.
2015. *English Vocabulary Profile Online*. Cambridge University Press. Accessed February 10, 2022. <https://www.englishprofile.org/wordlists/evp>.

- Facultatea de Litere, Istorie și Teologie, UVT. 2021. “Plan de învățământ anul universitar 2021-2022 valabil pentru anii II–III de studiu.” Undergraduate Studies Programme - Curriculum, Department of Modern Languages and Literature, Faculty of Letters, History and Theology, West University of Timișoara, Timișoara. Accessed November 1, 2021. https://drive.google.com/file/d/10q_3NMzoGsOHvoTbItCXuwmnoxj-S2rwZ/view?usp=sharing.
- Facultatea de Litere, Istorie și Teologie, UVT. 2021. “The Regulation for the Organizing and Carrying Out the Graduation Exams of the Bachelor and Master Degree Studies at the Faculty of Letters, History and Theology.” Regulation, Faculty of Letters, History and Theology, West University of Timisoara, Timisoara. <https://drive.google.com/drive/folders/1v7kfdsWa7Z1ik6mNbzzFuZNRh1yKsGkf>.
- Little, David. 2011. “The Common European Framework of Reference for Languages, the European Language Portfolio, and language learning in higher education.” *Cercles* (DeGruyter Mouton) 1 (1): 1–21. doi:10.1515/cercles-2011-0001.
- Nesi, Hilary, Sheena Gardner, Paul Thompson, and Paul Wickens. 2008. *British Academic Written English Corpus*. Oxford Text Archive. Accessed February 10, 2022. <http://hdl.handle.net/20.500.12024/2539>.
- Schmalz, Veronica Juliana, and Alessio Brutti. 2021. “Automatic Assessment of English CEFR Levels Using BERT Embeddings.” *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)* (CEUR Workshop Proceedings) 3033. <http://ceur-ws.org/Vol-3033/paper14.pdf>.
2020. *Self-assess your language skills with our CEFR tool*. 26 June. Accessed February 10, 2022. <https://www.pipplet.com/self-assessment-scale>.
- Wang, Y, M.J.F Gales, K.M Knill, K Kyriakopoulos, A Malinin, R.C van Dalen, and M Rashid. 2018. “Towards automatic assessment of spontaneous spoken English.” *Speech Communication* 104: 47-56. doi:10.1016/j.specom.2018.09.002.
- Zheng, Ying, Yanyan Zhang, and Youyang Yan. 2016. “Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China.” *ELT Research Papers* (British Council) 16 (1): 1–20.