ASPECTS REGARDING THE USE OF THE LEARNER CORPUS OF ROMANIAN (LECOR)

Carmen MÎRZEA VASILE[1]
Ana-Maria BARBU[2]
Valentina COJOCARU[3]
Mihaela CRISTESCU[4]
Elena IRIMIA[5]
Simona NEAGU[6]
Vasile PĂIȘ[7]
Isabella ȘINCA[8]
Monica VASILEANU[9]

Abstract
This article presents some ways in which the Learner Corpus of Romanian (LECOR) can be used. The first section describes what data and metadata LECOR contains and how it can be accessed through the query interface. The second section presents types of applications based on language facts extracted from the corpus. For instance, case studies on the correct use of the imperfect and the indirect object in dative are treated, as well as applications on communicative strategies at A1 level and the construction and querying of metadata-based subcorpora.

Keywords: learner corpus; query interface; applications; L2 Romanian; case study.
DOI: 10.24818/SYN/2025/21/1.11

[1] Carmen Mîrzea Vasile, "Solomon Marcus" Center for Computational Linguistics, Department of Linguistics, Faculty of Letters, University of Bucharest; "Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Romanian Academy, carmen.vasile@unibuc.ro.
[2] Ana-Maria Barbu, "Solomon Marcus" Center for Computational Linguistics, Faculty of Letters, University of Bucharest; "Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Romanian Academy, anamaria.barbu@g.unibuc.ro.
[3] Valentina, Cojocaru, Center for Romanian Studies, Faculty of Letters, University of Bucharest; "Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Romanian Academy, valentina.cojocaru@litere.
[4] Mihaela Cristescu, Center for Romanian Studies, Faculty of Letters, University of Bucharest, mihaela.cristescu@litere.unibuc.ro.
[5] Elena Irimia, Research Institute for Artificial Intelligence, Romanian Academy, elena@racai.ro.
[6] Simona Neagu, Școala Doctorală Litere, Faculty of Letters, University of Bucharest, simonaneagu94@yahoo.com.
[7] Vasile Păiș, Research Institute for Artificial Intelligence, Romanian Academy, vasile@ racai.ro.
[8] Isabella Șinca, Școala Doctorală Litere, Faculty of Letters, University of Bucharest, isabella.sinca@s.unibuc.ro.
[9] Monica Vasileanu, "Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Romanian Academy; Department of Linguistics, Faculty of Letters, University of Bucharest, monica.vasileanu@litere.unibuc.ro.

1. LECOR: Brief description and use

Learner Corpus of Romanian (LECOR)[10] is the result of a government project[11] and represents a large corpus of texts in Romanian produced by foreign students, in written or oral format. The texts are collected in the period 2019-2024, from international students attending the preparatory year, i.e., a year-long intensive program of Romanian language, organized by the University of Bucharest. Most texts are at beginner and intermediate proficiency levels, as assessed by classroom teachers who were also project members. Specifically, over 55.6% of texts are at beginner level (A1: 21.7%; A1-A2: 9%; A2: 24.9%), 12.4% are at A2-B1 level, 28.3% are at intermediate level (B1: 11.4; B1-B2: 8.2; B2: 8.7), and 3.4% are at C1 level (plus 0.2% at B2-C1 level). The students have different native languages (see Fig. 1) and are going to study various subjects in Romania, such as medicine, mathematics, computers, etc.

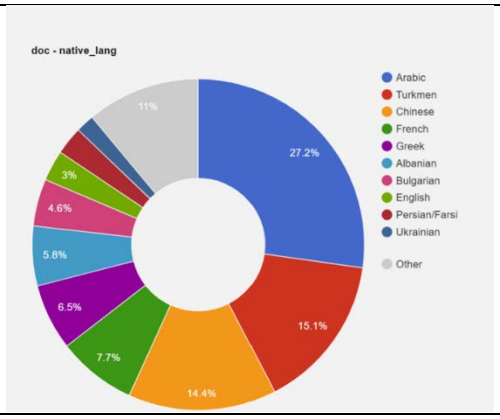

Figure 1. Distribution of texts by students' native languages

Figure 2. LECOR variants

The current open-access version of LECOR, accessible at http://lecor.unibuc.ro/crystal/#open, uses the NoSketch Engine query interface, benefiting from numerous functions made available by it, which we will not dwell

[10] This section focuses on the practical usage guidance for the corpus, including essential descriptive information without which usage directions would lack relevance. Information previously included in earlier papers and presentations (Barbu et al. 2023, Mîrzea Vasile and Irimia 2023, Mîrzea Vasile 2024, Mîrzea Vasile et al. 2024) has been supplemented with data available only after the completion of the electronic corpus development, and some usage guidelines have been elaborated in greater detail. The second section of the article presents several concrete applications. The term 'use' in the title refers to both practical guidelines and concrete research applications enabled by LECOR.

[11] This work was supported by CNCS - UEFISCDI, project number PN-III-P1-1-1.1-TE-2019-1066: "Learner Corpus of Romanian (LECOR). Collection, Annotation and Applications".

on in this article (see instead section Concordance: About in the LECOR platform). Also, we will not go into details regarding the corpus building and technical solutions adopted, for which we refer to Barbu et al. (2023). Instead, we will focus on what is made available to users in this corpus. The LECOR home page (Fig. 2) proposes first of all the selection of the corpus type from four available variants – note that *Long LECOR* is not intended for public use. Two contain simple texts, such as the students' initial texts, with errors: *LECOR student texts*, and corresponding minimally corrected versions: *LECOR teacher texts*. The first can be of interest for research, for example, on the individual evolution of students or the word count of the learners' texts. The second is useful for research on terms (forms and grammatical annotations) or topics covered in, which might otherwise become inaccessible due to errors. The other two available variants, namely *LECOR* and *LECOR All texts 2*, are based on word-aligned pairs of student text and teacher text. This alignment provides the correct variants for student errors. They differ only in the visualization of these corrections, thus:

- *LECOR* – displays, for pairs that do not match, the correct (teacher) variant in green (except for the searched word, which is bold red), separated by a vertical bar from the student variant in red, see Table 1a. When searching for a keyword in context (KWIC), the interface provides matches for both student and teacher variants without distinction. By convention, the symbol ++ indicates the deletion of the word in the corrected version, and -0- indicates the insertion of a word in the corrected version. For example, "sp|++ -0-|să" can indicate the replacement of *sp* with *să* (by deleting *sp* and inserting *să*).
- *LECOR All texts 2* – displays the student's text, whereas the teacher's version can only be viewed if the view option *teacher* is selected, see Table 1b. This eliminates the redundancy and text overload that characterizes the LECOR variant, and the corrected variant is treated as an annotation to the student text, alongside other annotations such as lemma, morphosyntactic tag, etc. related to each word in the student's text, see Table 1c. Note that -0- indicates words that exist only in the corrected version and RED, words that exist only in the student version.

Table 1. Ways of viewing the text and its annotations

a.	LECOR	Dimineața| **Dimineață** trebuie sp|++ -0-|să -0-|se trezește|trezească devreme ,
b.	LECOR All texts 2	**Dimineața** trebuie sp -0- -0- trezește devreme , Dimineață RED să se trezească
c.	Annotations	**Dimineața** trebuie sp -0- -0- trezește Dimineață/dimineață/Rgp /trebui/Vmip3 RED/sp/Ncms-n să/să/Qs se/sine/Px3--a------w trezească/trezi/Vmsp3

After choosing the corpus variant, the searches will be done through the *Concordance* section resulting in examples of keyword use in context (see Table 1). It is important to know that linguistic analyses can benefit from searches focused on any of the information included in the LECOR platform. These are shown in Fig. 3.

In addition to the KWIC corresponding to the *word* attribute, searches can also be made by the following attributes:
- *lemma* – all inflectional forms of the respective lemma are found in the text;
- *tagupos* – corpus tags (CTAG) indicating the part of speech, for example ADP (preposition), AUX (auxiliary), ADJ (adjective), ADV (adverb), etc.;
- *tagxpos* – morphosyntactic descriptions (MSD), which contain more grammatical categories than the part of speech, for example Ncmrs (common noun masculine singular right case)[12];
- *error* – presence of an error automatically annotated by comparing the student's text with the teacher's: WRONG (wrong use or form); -0- (missing word); OK (correct word);

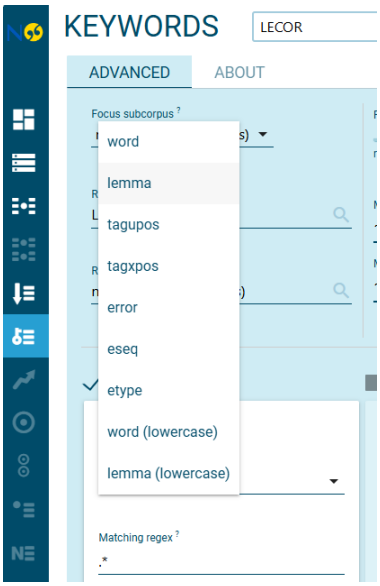

Figure 3. Searchable attributes

- *eseq* – sequences of corrected words marked at the ends with B (begin) and E (end), respectively, which usually include mistakes regarding the topic or lexical phrases;
- *etype* – type of error annotated by the teacher, much more informative than error, for example GVM (grammar, verb, morphology), FS (form spelling), LP (lexical phrase errors), etc. (cf. Granger, Swallow & Thewissen, 2022).

The values of these attributes can be filtered by regular expressions (see *Matching regex* in Fig. 3), for example, for the attribute *tagxpos* the expression Afpf.* will result in a list of all MSDs specific to feminine adjectives found in the corpus.

LECOR also allows combining criteria through complex queries. For example, if one wants to study the error of the type "encliticly articulated noun + genitive article", he/she can use the CQL expression: [tagxpos="Nc...y"] [lemma="al" & tagxpos="Ts.*"] as in Fig. 4.

LECOR has a very rich metadata module, described in Mîrzea Vasile and Irimia (2023), which allows results to be filtered both by student data, such as native language, gender, age, proficiency level, etc., and by text characteristics, such as handwritten, MS Word or oral format, topic, circumstances of elaboration, etc. This metadata is accessible in the Concordance section, Text Types, see Fig. 5.

[12] For the MSD tagset see https://www.sketchengine.co.uk/romanian-tagset/ or http://nl.ijs.si/ME/Vault/V3/msd/html/ (for Romanian).

Figure 4. Complex query and its result

Figure 5. Metadata filters

With this rich palette of search possibilities available, the analyses that can be done on the basis of LECOR are also very diverse: sociolinguistic analyses, synchronic or longitudinal linguistic analyses, manuals textbooks and possibly individualized exercises. The following section provides just a few examples of applications regarding the use of the corpus.

2. Types of applications

This corpus can be used both directly, through direct consultation as practice in the classroom, and indirectly, as a research base for the development of didactic materials (textbooks, vocabulary books, exercises, etc.) or for case studies. Students can directly access the corpus, possibly under the guidance of the teacher, to become aware of various aspects regarding the structure of the Romanian language. They can observe an incorrect (lexical, grammatical) form in comparison with the correct one, can see the lemma (dictionary form) of a word or grammatical information about each form separately etc. Repeated mistakes can be highlighted more easily so that the learning process can focus on them. Exercises can involve, for example, reading the text to identify and correct mistakes; comparing the two text variants (with and without errors) and explaining the differences; identifying which nouns/verbs have incorrect plural forms from a list, etc. For typical applications of this type of corpora, see, for instance, Granger, Gilquin, and Meunier ((eds) 2015: 443-627), and Díaz-Negrillo and Thompson (2013: 17-23). The corpus can be used, of course, to inform L2 Romanian acquisition research on rather theoretical issues, but triangulation of methods is recommended.

It is worth mentioning that LECOR does not offer, through the corrected version (in fact, a normalized variant in terms of spelling, grammar, and vocabulary), samples of native Romanian language and of good linguistic practice, especially at the lower proficiency levels, as in the case of a corpus of genuine Romanian texts. This is due to the minimalist way of correction that leads to the outline of an interlanguage stage, where the student does not use the second language like a native, but does not make lexical, grammatical or pragmatic mistakes either.

We can talk about an indirect use of LECOR when the corpus is used by teachers, researchers, or materials developers (rather than by learners themselves). Several results based on qualitative and quantitative corpus analysis have been published and/or presented over the course of the project, for instance, expression of inalienable possession, dative nominals, the comparative and superlative, use of the definite article, the imperfect, etc. In what follows, we focus on some of these case studies.

2.1 The imperfect

In Romanian, the imperfect is an anaphoric tense, as it encodes a past action or state which is (partially) simultaneous with a specific point in the past (GR 60). It is part of a complex system of past tenses, among which the compound perfect, which refers to an action or state prior to the utterance, is the most frequently used (GR 57). This is why Romanian learners often struggle with the use of the imperfect, and replace it with the compound past in narrative contexts, where the imperfect typically occurs. However, it is important to note that there are no strict rules mandating the exclusive use of the Romanian imperfect tense in these situations, but rather quantitative differences.

A first attempt to tackle the problem was undertaken by Mîrzea Vasile and Preda Cincora (2023), who carried out a contrastive interlanguage analysis (for CIA, see Granger 2015) on Romanian L1 and L2 samples. The authors aimed at examining the variation between the imperfect and compound past in native and non-native Romanian. To achieve this, they chose a context in which the imperfect typically occurs, i.e., the expression of age, and queried two corpora for two exemplary phrases, formulated with both tenses: *când* 'when' + *a fi* 'be' + *copil* 'child' (e.g. "când eram copil" vs. "când am fost copil", 'when I was a kid') and *când* 'when' + *a avea* 'have' + numeral + *ani* 'years' (e.g., "când aveam 7 ani" vs. "când am avut 7 ani", 'when I was 7 years old'). They performed their queries on a subcorpus of LECOR (1,406 texts, 235,391 words, 275,443 tokens, A1-B2) and RoTenTen21, a web corpus of L1 Romanian (7,876,464 texts, 2,763,173,824 words, 3,324,975,990 tokens). The authors observed that the compound past occurred only in a minority of cases, whereas the proportion was higher in non-native Romanian. Re-running the analysis on the complete version of LECOR 2024, the same trends were observed (Table 2): the imperfect dominates both in native and non-native Romanian, but the

compound perfect appears in a higher proportion in the learners' samples (Mîrzea Vasile et al. 2024).

Table 2. The use of imperfect and compound past in expressing age in L1 and L2 Romanian (Mîrzea Vasile et al. 2024).

The construction in the query	L1 (roTenTen21)	L2 (raw) LECOR subcorpus (275.443 tokens)	L2 LECOR 2024 (593.252 tokens)
când ***eram*** *copil / copii*	3.669 occurrences (97,79 %)	9 occurrences (75 %)	34 ocurrences (80,95%)
când ***am fost*** *copil / copii*	83 occurrences (2,21%)	3 occurrences (25%)	8 occurrences (19,05%)
când ***aveam*** *... ani / an*	3.506 occurrences (98,79%)	3 occurrences (100%)	11 (*a avea* "to have") + 4 occurrences (*a fi* "to be") = **15** imperf. **(71,43%)**
când ***am avut*** *... ani / an*	43 occurrences (1,21%)	0 occurrences	3 (*a avea* "to have") + 3 occurrences (*a fi* "to have") = **6** C.P. **(28,57%)**

Moreover, our analysis highlights the grammatical variation in the expression of age, as may be observed from the examples below, with (1) and (3) containing the imperfect, (2) and (4), the compound past:

(1) *Am fost la Cilieni, când* ***eram*** *copil, dar nu am prins niciodată pește.* (RoTenTen21)
'I went to Cilieni when I was a kid, but I never caught any fish'[13]
(2) *Când* ***am fost*** *copil, tatăl meu mi-a arătat cum să fac un cal să zboare.* (RoTenTen21)
'When I was a kid, my dad showed me how to make a horse fly'
(3) *Când* ***eram*** *copil, în vacanță, îmi plăcea să mă joc cu prietenii*. (LECOR, m., Turkmen, A2-B1, *Childhood vacations*, written)
'When I was a kid, during the holidays, I liked to play with my friends.'
(4) *Prietenii mei sunt foarte veseli. și se joacă cu mine de când* ***am fost*** *copil.* (LECOR, f., Arabic, A2-B1, *A funny incident*, written)
'My friends are very cheerful. And they play with me since I was a kid.'

Thus, even in native Romanian, there is notable variation between the imperfect and compound past. In the analyzed contexts where imperfect is much more frequent in the L1 standard Romanian, compound past specifically occurs when: relating to other perfective aspect tenses, appearing as a regional variant (particularly in northern dialects), or conveying meanings of 'to turn, fulfill, reach (a number of years)'. This variation should be considered when assessing, correcting, and explaining the use of these two tenses in L2 samples.

[13] The English translation captures what we believe is the most likely meaning the learner intended to convey.

To ensure reliability, the data were triangulated with a questionnaire-based investigation, completed by 151 B1 Romanian learners (Mîrzea Vasile et al. 2024). The questionnaire comprised sentences where the imperfect was expected, as in (5), and the students were asked to choose the correct verb forms. For example (5) below, the options were a) *am fost* ('be' in the compound past) b) *eram* ('be' in the imperfect), c) *am avut* ('have' in the compound past) and d) *aveam* ('have' in the imperfect), thus assessing not only the acquisition of tenses, but also the lexical selection. In Romanian, the verb *a avea* 'have' is used to express age, whereas English and other languages use the verb *be*. All the options had previously been attested in LECOR.

(5) *Când* _____ *7 ani, am primit cadou o bicicletă.*
'When _____ 7 years old, I received a bike as a gift'

While the expected answer was the most frequently selected one by students, the proportion of responses containing the imperfect is lower than in the corpus counts (Table 3). However, the findings display the same trend, although the proportions differ: the imperfect is preferred, but the compound past is also used (28.58% of the contexts identified in the corpus vs. 39.1% in the questionnaire).

Table 3. Corpus and questionnaire findings for the expression of age (Mîrzea Vasile et al. 2024)

când aveam 7 ani "when I was 7 years old"		**L2 LECOR 2024 (593.252 tokens)**	**Questionnaire responses**
a avea "to have" IMPERF.	*când aveam X ani*	11 **(52.38%)**	51 **(33,8%)**
a fi "to be" IMPERF.	*când eram X ani*	4 (19.05%)	41 (27,2%)
a avea "to have" C.P.	*când am avut X ani*	3 (14.29%)	27 (17,9%)
a fi "to be" C.P.	*când am fost X ani*	3 (14.29%)	32 (21,2%)
Total		21	151 responses

The questionnaire results indicate a lower acquisition rate of the lexicogrammatical structure compared to the corpus analysis. This discrepancy can be attributed to several factors: the item framing, where some of the provided sentences contained a verb in the compound past, potentially priming responses; the timing of the assessment, as the questionnaire was not administered immediately following explicit instruction on the imperfect, unlike many corpus compositions; cross-linguistic influence, with potential transfer from L1 and/or other known languages (e.g., English); and L1 usage patterns, where the compound past can occur in similar contexts.

2.2 Communicative strategies at the A1 level

Learners of Romanian as L2, like any other L2 learners, are often faced with the need to express a concept or idea in L2, while lacking the linguistic resources to do so. In

this case, they resort to a communication strategy (see Dörnyei and Scott 1997) to fulfil their aim. Strategies should not be seen as mere mistakes; they are an important part of the L2 acquisition learning process. Șinca (2024a,b) used the taxonomy of communication strategies proposed by Dörnyei and Scott (1997) to analyze A1 Romanian learners' written productions, and contrasted her findings with previous results by Vasiu (2020), who focused on A1 oral interactions.

The analysis of 800 written samples by native speakers of Arabic, Turkmen, Albanian, French, Turkish, Bulgarian, Korean, and Greek (Șinca 2024b) showed that A1 Romanian learners employ mainly four strategies (see Fig. 6):

(a) approximation, i.e., using an alternative word that is close in meaning to the intended word (as in (6), where *lung* 'long' is used instead of *mult* 'a lot'[14]);
(b) code-switching, as in (7), where the student used Eng. *mug* instead of Ro. *cană* 'mug';
(c) word coinage, i.e., creating a new word in the target language, according to an alleged rule, as in (8), where the verb *felicitărim*, which is supposed to mean 'celebrate', is coined from Ro. *felicitări* 'congratulations';
(d) literal translations, i.e., words or structures from L1 or another foreign language translated into the target language, as in (9), where the student translated Eng. *ponytail*.

(6) *Vreau să stau mai* ***lung*** *cu familia mea*. (LECOR, f., Albanian, A1, *Timetable and the academic year*, in class)
'I want to spend more time with my family'.
(7) *Sub ceas este o masă cu un computer și un* ***mug***. (LECOR, f., French, A1, *The living room*, in class)
'Under the clock is a table with a computer and a mug'.
(8) *voi avea vacanță de iarnă pentru că* ***felicitărim*** *Crăciun* (LECOR, f., Albanian, A1, *Timetable and the academic year*, in class)
'I'll have a winter vacation because we're celebrating Christmas'.
(9) *fac o frumoasă* ***coadă de cal*** (LECOR, f., Greek, A1, *A working day*, homework)
'I do a nice ponytail'.

[14] In this example, the use of the adverb *lung* instead of *mult* can be explained by a possible literal translation from the L1 Albanian, where *gjatë* encodes both special and temporal extension, see MHAED I, s.v. *gjatë*, and/or L2 English, another foreign language known by the student, where *long* has a similar synthetic meaning.

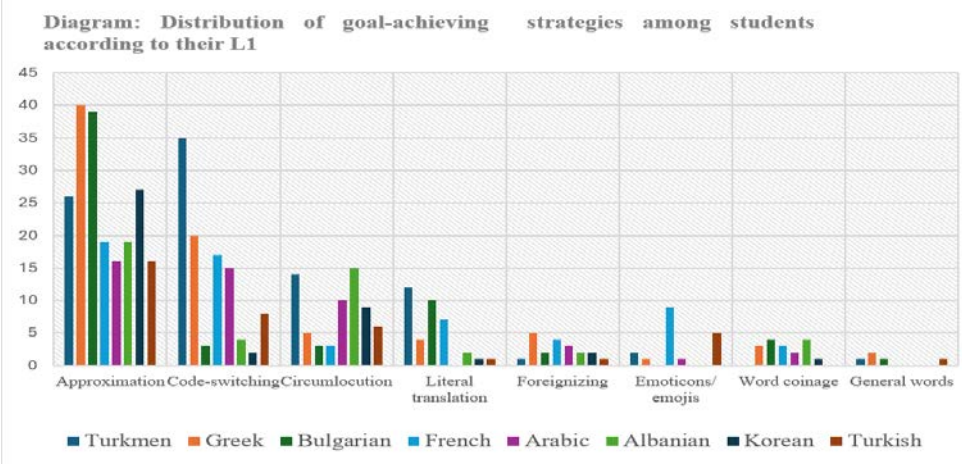

Figure 6. Communication strategies in the written productions of Turkmen, Greek, Bulgarian, French, Arabic, Albanian, Korean and Turkish learners of Romanian (Șinca 2024b)

More interestingly, the quantitative analysis in Șinca (2024b) showed that certain native groups resort to communication strategies more often than others: Turkmen and Greek students use the strategies most, while Korean and Turkish natives the least (see Fig. 6). Her findings suggest that Turkmen and Greek learners of Romanian focus more on getting the message across at all costs ("focus on meaning"), while the other groups, especially Turkish and Korean natives, focus more on communicating correctly in Romanian ("focus on form"). In addition, the less frequent use of code-switching and almost no use of literal translation by Arabic and Turkish learners of Romanian is correlated with a lower level of proficiency in English, compared to the other groups (Șinca 2024b).

Moreover, Șinca (2024b) compared her findings with the results in Vasiu (2020: 145-153), suggesting that learners use different strategies in oral and written communication. While some strategies occur only in writing (for example, emojis/emoticons) or only in oral communication (for example, fillers, e.g. *well; you know; actually*, and repetitions, Dörnyei and Scott 1997: 190), the only strategy that is frequently used by all learners in both settings is approximation (Șinca 2024b).

The results observed in this case study on communication strategies can facilitate L2 Romanian teaching in several ways. Key aspects include: differentiated approaches based on native speaker groups (e.g., varied emphasis on grammar practice for those sensitive to formal correctness versus on conversation practice for those focused on communication despite errors made) and the integration of some strategy manifestations in teaching, such as the correct use of generic words (*lucru* 'thing', *ceva* 'something', *altceva* 'something else', *faptul* 'the fact', *asta* 'this', etc.).

2.3 The indirect object in dative

A case study based on LECOR was dedicated to the acquisition of the Romanian dative structures by non-native speakers (Neagu and Mîrzea Vasile 2023). The research aimed to discuss aspects regarding the usage of the indirect object constructions and clitic doubling. For this purpose, a subcorpus of LECOR was examined, containing approximately 1000 written and 450 oral productions of 100 foreign students (A2-B2 level). The study investigated and classified the most frequent errors in the production of Romanian dative constructions, such as the replacement of the nominal or pronominal forms in dative by the corresponding forms in accusative or nominative, and also the absence of the pronominal clitic forms.

It was noticed that the foreign students enrolled in the Preparatory Year encounter many difficulties in the acquisition of the Romanian indirect object structures and the doubling rules. On the one hand, this situation is justified by the fact that Romanian, a language with rich inflection, is more difficult to be assimilated by non-native speakers, because it contains numerous affixes, and sometimes the same affix can express several grammatical categories. On the other hand, the homonymy between the genitive and the dative structures causes some confusions to the students. Furthermore, the acquisition of the Romanian dative is also problematic due to the speakers' native languages.

Starting from well-known interlanguage error classifications (Corder 1971, Dulay, Burt and Krashen 1982, Dagneaux, Sharon and Granger 1998, James 1998), several types of errors regarding the acquisition of the Romanian dative have been identified in this study, such as substitution, omission, and spelling errors. The quantitative analysis revealed that in the LECOR subcorpus, the substitution and the omission errors were the most frequent. It was noticed that students tend to use an accusative clitic instead of a dative one (10). Among the Romanian verbs that select indirect object (being, in general, ditransitive), the following ones occurred in erroneous structures: *a recomanda* 'to recommend', *a mulțumi* 'to thank', *a răspunde* 'to answer', *a telefona* 'to phone', *a zice* 'to say', *a cumpăra* 'to buy', *a face* 'to do', *a trimite* 'to send', *a da* 'to give', *a scrie* 'to write' (11). In addition, the examples investigated showed a preference for the constructions with indirect object expressed only by a noun, without the clitic doubling (12). Another frequent substitution error consists in using wrong inflected forms of nouns and indefinite, demonstrative, and relative pronouns, instead of the ones in dative (13). In addition, it was noticed (mostly among Arabic speakers) a tendency of replacing the noun in dative by a structure containing the preposition *la* 'to' and the noun in nominative (11), a pattern intensively used in spoken Romanian.

(10) *Lui Alex* ***îl*** *place un pulover*. (LECOR, f., Bulgarian, *The story*, written)
'Alex likes a sweater.'
(11) *am spus* ***la familia mea*** *că nu pot să merg*. (LECOR, m., Arabic, *Holidays*, written)
'I told my family I couldn't go.'
(12) *După somn,* ***lui Dan*** *era foarte foame*. (LECOR, m., Turkish, *The letter*, written)
'After sleeping, Dan was very hungry.'
(13) ***Toți oamenii*** *le place să călătorească in țari diferite*. (LECOR, f., Ukrainian, *An ideal voyage*, written)
'Everybody likes to travel to various countries.'

Among the Romanian structures with dative pronouns, special attention was dedicated to the verb *a-i plăcea* 'to like', which is introduced from A1 level and raises difficulties in the acquisition process (Neagu and Mîrzea Vasile 2023: 208, 212). In the analysed subcorpus, a number of 650 structures containing this verb were identified. In most of the cases (92%), the verb selects an indirect object expressed only by the clitic pronoun. A substitution error consists in using a stressed pronoun or a noun in nominative, instead of the stressed one in dative (14). Less frequently, a stressed pronoun in accusative was also used, preceded by a preposition (15).

(14) ***Iubitul meu*** *place omul care eu sunt*. (LECOR, f., Bulgarian, *Holidays*, written)
'My boyfriend likes the person I am.'
(15) ***Pe mine*** *nu îmi place deloc*. (LECOR, m., Turkmen, *Holidays*, written)
'I don't like it at all.'

Another class of Romanian structures with dative pronouns consists of fixed expressions such as *a-i fi foame/sete/somn/frig …* 'to be hungry/thirsty/sleepy/cold …'. These constructions raise problems to non-native speakers, as they tend to use stressed pronominal forms in nominative instead the correct ones in dative (16) and they omit doubling the indirect object expressed by a noun (see 12 above).

(16) *noi nu ne foame*. (LECOR, m., Turkmen, *A message for mum*, written)
'We are not hungry'

The list of the most frequent errors in the acquisition of the Romanian indirect object is helpful in the improvement of the teaching methods and the elaboration of more efficient didactic materials and courses. For example, attention can be drawn to the most frequent errors and error correction exercises can be developed using authentic examples produced by L2 learners.

3. Conclusions

The fact that LECOR is scalable will permit the development of an increasingly large annotated corpus of Romanian. This fact is very important, because the larger a digitalised corpus, the more reliable it is for a wider range of applications. So, since it is and will be a large corpus, partially annotated and partially with open access, LECOR has many possible end-uses in language teaching and in natural language processing. It has an immediate pedagogical use. It can be used in classrooms, or by learners themselves, since this kind of data are relevant for the (error) producers.

It also can be used to inform instructional materials design (such as textbooks, wordlists, dictionaries, etc.), for language teachers training and for language testing. Moreover, the metadata will allow drawing a certain 'difficulties profile' for learners with a specific mother tongue and thus will enable teachers to design more specific materials for their target groups of learners.

References and bibliography

Barbu, A.M., Irimia, E., Mîrzea Vasile, C. and Păiș, V. 2023, "Designing the LECOR Learner Corpus for Romanian", in Angelova, G., Kunilovskaya, M., R. Mitkov (eds.), *Deep Learning for Natural Language Processing Methods and Applications* (Proceedings of the 14th International Conference Recent Advances in Natural Language Processing, RANLP 2023, INCOMA Ltd., Shoumen, Varna, Bulgaria: 143-152. Retrieved from https://aclanthology.org/2023.ranlp-1.16.pdf. Accessed on 24 January 2025.

Corder, S.P. 1971. Idiosyncratic dialects and error analysis, in *IRAL: International Review of Applied Linguistics in Language Teaching*, 9 (2): 147-160.

Dagneaux, E., Sharon, D. and Granger, S. 1998. "Computer-aided Error Analysis", in *System: An international Journal of Educational Technology and Applied Linguistics*, 26 (2): 163-174.

Díaz-Negrillo, A. and P. Thompson. 2013. "Learner corpora. Looking towards the future" in Díaz-Negrillo, A., Ballier, N., P. Thompson (eds). *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam/Philadelphia: John Benjamins Publishing Company: 9-30.

Dörnyei, Z. and Lee Scott, M. 1997. "Communication strategies in a second language: Definitions and taxonomies", in *Language learning*, 47.1: 173-210.

Dulay, H., Burt, M., S. D. Krashen. 1982. *Language Two*. Rowley: Newbury House.

GR = Pană Dindelegan, G. and Maiden, M. (eds). 2013. *The Grammar of Romanian*, Oxford University Press.

Granger, S. 2015. "Contrastive interlanguage analysis: A reappraisal", in *International Journal of Learner Corpus Research*, 1.1: 7-24.

Granger, S., Gilquin, G. and Meunier, F. (eds). 2015. *The Cambridge Handbook of Learner Corpus Research*, Cambridge: Cambridge University Press.

Granger, S., SwallowH. and Thewissen, J. 2022. *The Louvain Error tagging Manual*. Version 2.0. CECL Papers 4, Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain. Retrieved from cdn.uclouvain.be/groups/cms-editors-cecl/cecl-papers/Granger et al._Error tagging manual 2.0_final_CC.pdf. Accessed on 24 January 2025.

MHAED I, II = Mann, Stuart. E., 1948, *An Historical Albanian-English Dictionary*. Part I: A-M, Part II, N-Z, London/New York/Toronto: British Council by Longmans, Green and Co. LDT.

James, C. 1998. *Errors in Language Learning and Use. Exploring Error Analysis*, Essex: Longman.

Mîrzea Vasile, C. 2024, "Ghid de utilizare. Corpus de română ca limbă străină (LECOR)", https://unibuc.ro/wp-content/uploads/2024/09/Ghid_LECOR_5.09.24_v0.pdf. Accessed on 24 January 2025.

Mîrzea Vasile, C., Barbu, A.-M., Irimia, E. and Păiș, V. 2024. *Using (new) corpora for L2 Romanian research and teaching*, presentation at the University of Lisbon, 24 October 2024.

Mîrzea Vasile, C. and Irimia, E. 2023. "Metadata design for the first electronic learner corpus of Romanian", in *Romanian Studies Today* VII, București: Editura Universității din București: 31-47.

Mîrzea Vasile, C. and Preda Cincora, E. 2023. „Imperfectul în româna ca limbă străină. Observații pe baza corpusului", in Bălășoiu, C., Ene, C., Nedelcu, I., A. Toma (eds.), *Actele celui de-al 22-lea Colocviu Internațional al Departamentului de Lingvistică: Lingvistică sincronică, diacronică și tipologică (București, 18–19 noiembrie 2022)*, București: Editura Universității din București: 387-401.

Neagu, S. and Mîrzea Vasile, C. 2023. "Complementul indirect și dublarea în româna ca interlimbă. Studiu de corpus", in Platon, E., Bocoș, C., Roman, D., L. Vasiu (eds.), *Actele Conferinței Internaționale Discurs polifonic în limba română ca limbă străină (RLS), ediția a IV-a, 6-7 octombrie 2023*, Cluj-Napoca: Editura Presa Universitară Clujeană: 204-216.

Șinca, I. 2024a. „Strategii de comunicare la nivelul A1 în producții scrise de nativi arabi, turkmeni, albanezi și francezi. Studiu de caz", in Platon, E., Bocoș, C., Roman, D., L. Vasiu (eds.), *Discurs polifonic în româna ca limbă străină (RLS). Actele Conferinței Internaționale,* Cluj-Napoca, ediția a IV-a / 2023, Nr. 4/2024: 189-203.

Șinca, I.T. 2024b. *Strategii de comunicare în româna ca limbă nenativă în texte scrise (nivelul A1). Câteva studii de caz*, dissertation, Faculty of Letters, University of Bucharest, unpublished.

Vasiu, L.-I. 2020. *Achiziția limbii române ca L2. Interlimba la nivelul A1*, Cluj-Napoca: Editura Universitară Clujeană.

The authors

Carmen Mîrzea Vasile has published two books on adverbs and has co-authored works published at Oxford University Press, De Gruyter Mouton, John Benjamins, the Publishing House of the Academy, etc. She has led a project resulting in the construction of the first electronic corpus of Romanian as a non-native language ("Learner Corpus of Romanian (LECOR). Collection, Annotation and Applications", 2020-2024). Her main interests are the (derivational) morphology and its interfaces, the syntax and pragmatics of the adverbs, language standardization, language research resources, and Romanian as foreign/second language.

Ana-Maria Barbu, having a career built on the foundation of two specializations: computer scientist and philologist, has constantly focused for decades on building digital linguistic resources, lexical databases, corpora of annotated texts, and her scientific research has addressed topics of theoretical and computational linguistics. Among the most relevant contributions for this study are the participation in the development of the Romanian tagset within MULTEXT-east project, the construction of two lexical resources of inflectional forms and forms separated into syllables, as well as the Dictionary of Verbal Valences for Romanian (https://dcv.lingv.ro/).

Valentina Cojocaru has published several papers on discourse markers, language contact, and teaching Romanian as a foreign language, in local prestigious journals such as *Revue roumaine de linguistique* and *Studii și cercetări lingvistice*. Her primary research interests include discourse analysis, bilingualism, and second language acquisition, with a focus on the pragmatic and sociolinguistic dimensions of language use. In addition to her academic work, Valentina is deeply involved in organizing and coordinating the Linguistics Olympiad in Romania.

Mihaela Cristescu an Assistant Professor at the Centre for Romanian Studies, Faculty of Letters, University of Bucharest, teaching Romanian as a Foreign Language and a seminar in General Linguistics, since 2016. She worked for 6 years as a Linguist Researcher, in the Natural Language Processing field, contributing to the formal description of Romanian. She has worked as a corpus annotator in two COST Actions, PARSEME (2017-2018) and UniDive (2022-present), classifying and analysing Romanian verbal multiword expressions. Her main research interests are negation, multiword expressions, corpus annotation, error analysis, Romanian as a FL, general linguistics.

Elena Irimia is a Senior Scientific Researcher III that participated in the development of different standardized annotated corpora for Romanian: a hand validated journalistic corpus (**RoCo_News**), a balanced corpus (ROMBAC), a reference corpus (CoRoLa, https://corola.racai.ro/), a learner corpus (LECOR corpus). She developed a treebank nucleus for Romanian and later participated in the development of a reference treebank for Romanian. She has expertise in: corpora collection; corpora annotation, both manually and automatically; metadata creation. She participated in different national and international projects concerned with (mono/multi-lingual, uni/multi-modal) corpora collection and annotation and with developing tools and platforms for processing and managing corpora.

Simona-Ștefania Neagu is a PhD student in linguistics at the University of Bucharest, since 2018. Her doctoral research focuses on the morphosyntax of Old Romanian and is based on data collected from original documents from the 17th and 18th centuries. She presented a series of papers regarding her subject at several national and international conferences. Her main research interests are morphology, diachronic syntax, linguistic typology, and Romance linguistics.

Vasile Păiș has vast experience in natural language processing, large language models, multimodal processing, and platforms for language resources and tools. He recently received the "Gh. Cartianu" award of the Romanian Academy for multiple publications and the development of the RELATE platform (used in LECOR project). He participated in the creation of numerous resources for the Romanian language, such as: the Representative Corpus of the Contemporary Romanian Language (CoRoLa), CURLICAT, MARCELL, LegalNERo (the first resource for named entity recognition in the Romanian legal domain), RoMEMEs (the first multimodal resource investigating the usage of Internet memes in the Romanian language).

Isabella Șinca has done numerous research projects, focusing on the communication strategies employed by learners of Romanian as a second language (L2). She has presented her studies at various conferences and symposiums. Some of her work has already been published, while other pieces are currently in the publication process. Additionally, she collaborated on the Lecor (Learner Corpus of Romanian) project, contributing to the development of the first electronic corpus of Romanian as a non-native language.

Monica Vasileanu is a researcher at the Institute of Linguistics and a lecturer at the University of Bucharest. She is involved in the editing of major academic Romanian dictionaries, such as *Dicționarul etimologic al limbii române (DELR)* and *Dicționarul explicativ al limbii române (DEX)*. She is the author of a monograph and more than 40 academic papers mainly dedicated to the Romanian lexis. Her most recent publication is *RLS, pls! Manual de limba română ca limbă străină pentru nivelul A1*, EUB, 2024.