ISSN 2559-3390 RST VII, 2023
DOI:10.62229/rst/7.1/3

Metadata design for the first electronic learner corpus of Romanian

Carmen Mîrzea Vasile
University of Bucharest / "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics, Romanian Academy
carmen.vasile@unibuc.ro

Elena Irimia
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy/ University of Bucharest
elena@racai.ro

To cite this article: Mîrzea Vasile, C., E. Irimia, 2023. „Metadata design for the first electronic learner corpus of Romanian". *Romanian Studies Today*. VII, 2023, p. 31-47.

Abstract: This paper introduces an ongoing work of collecting, annotating and documenting the first digital Romanian Learner Corpus (LECOR), focusing on its metadata. We shortly describe the institutional context of the project, the current state of the art in the field, the objectives in terms of structure, dimensions and annotations and what work has already been done at this stage of the project. Then we present the modular structure of the metadata scheme and a detailed account of all the metadata fields and their possible values, from general metadata concerning the whole corpus (Section 3.1), to metadata organised around the student/learner (Section 3.2) and text/composition (Section 3.3). We will also give some examples of how metadata has been dealt with in various researches (including based on LECOR corpus).

Keywords: corpus metadata; individual differences; variables in language learning; Romanian as L2/FL; learner corpus of Romanian.

1. The LECOR project

The article describes the metadata used for the LECOR corpus, which is currently under construction. Metadata are collections of information used to describe in a standardised manner some specific data: in our case, the described data is a collection of foreign students' productions in Romanian, processed, annotated, sorted and integrated in a corpus; in other words, metadata are data about data and they are essential for documenting both the work done in the project and the content produced in this process.

LECOR corpus is developed through a two-year project (*Learner Corpus of Romanian (LECOR). Collection, Annotation and Applications*), a grant aided by the Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI), as part of the sub-programme dedicated to research projects to stimulate young independent teams (TE). Its host institution is the University of Bucharest (Faculty of Letters, "Solomon Marcus" Center for Computational Linguistics) and the deadline is May 14th, 2024 (for other administrative details, see Barbu *et al.* 2023).[1]

The main goal of the LECOR project is to build the first digitalized learner corpus for Romanian, covering an essential gap in the field of both Second Language Acquisition (SLA) and Natural Language Processing (NLP) research. There are two recent learner corpora for Romanian: CORLS (*Corpus Oral de Limba Română ca Limbă Străină*, part of the volume *Achiziția limbii române ca L2. Interlimba la nivelul A1*; it contains 70,000 words, for details see Vasiu 2020) and the corpus compiled by Constantinescu & Stoica (*Româna ca limbă străină. Corpus*; it contains 125,000 words, for details see Constantinescu & Stoica 2020). Apart from their rather small size, both corpora are only in print format and therefore not accessible for digital processing. The metadata annotation is minimal: e.g., in terms of variables describing the learners, Vasiu (2020) encodes only the country of origin, the gender and the native language(s) (L1), while Constantinescu & Stoica (2020) encode the country of origin, age, L1 (and a second native language, when the case), level of formal instruction (high school, college, etc.) and previous knowledge of Romanian.

The LECOR corpus will be thus the first electronic learner corpus for Romanian, available in open-access format, scalable and with rich metadata annotation. This learner corpus is designed to have the following features: (a) it is a monolingual corpus; (b) it is a multi-L1 learner corpus (samples are collected from learners with more than 20 different native languages: Arabic, Chinese, Korean, Albanian, Greek, Armenian, Turkmen, Turkish, Persian, Bulgarian, Serbian, Ukrainian, Belarusian, Russian, etc.); (c) it contains written (80%) and

[1] https://aclanthology.org/2023.ranlp-1.16.pdf

oral (20%) learners' samples, in total, a number of 4,000 samples (aiming for over 600,000 words); (d) it is a general corpus (it does not contain samples of language for specific purposes, for example); (e) the sample collection is mainly controlled (i.e. it consists of essays, dialogues, etc., created in the classroom, as homework or during exams); (f) it is both synchronic/ cross-sectional and diachronic/ longitudinal corpus; (g) it is automatically annotated at the morpho-syntactic level and manually annotated by linguist specialists for learner morphological and syntactical errors.

At this moment, the current state of progress in collection, storage and annotation of the LECOR corpus is as follows: 1) most of the corpus is already collected and systematically ordered (over 3500 samples of raw texts – compositions in word format, photos of handwritten compositions, recordings in mp4 format and their transcriptions in .txt format); 2) the texts and all the other additional materials are safely stored on a server at the University of Bucharest; 3) the texts are partially anonymized, they are not annotated at all at the level of learner error (see Barbu, Irimia, Mîrzea Vasile and Păiș 2023), but the list of error types that will be manually annotated in a small part of the corpus (only 3,500 sentences) is almost definitive; 4) the metadata, which we will present in this article, is fully collected; 5) the interface, which will be in the platform Sketch Engine (see Kilgarriff *et al.* 2014), is not fully ready.

The corpus samples are produced by learners enrolled in the one-year, intensive academic programme called "The Preparatory Year", at the Faculty of Letters (Centre for Romanian Studies), University of Bucharest. This programme is mandatory for those students who wish to pursue their studies in Romanian language in Romania and need a language certificate for B1 or B2 level to fulfil their aim.

2. Variables/metadata in learner corpora. The general situation of LECOR

Díaz-Negrillo and Thompson (2013: 15) note that most learner corpora are not well-documented and do not contain enough information about the learners and the samples they produced. Information such as the learner's native language, whether he knows other foreign languages, whether it is a sample from an exam or done as homework, etc., are essential in quantitative and qualitative accurate analyses based on non-native samples. A thorough understanding of the influence of different learner- and task-related variables on the L2 use has important theoretical and pedagogical implications, therefore metadata availability and richness are the first thing researchers look at to determine the suitability of the corpus to their research objectives.

An overview of the situation of variables/metadata availability for learner corpora is made by Granger and Paquot (2017); they underline the idea that this additional information must be collected with great care and in such a way as to ensure the interoperability, re-usability and sustainability of learner corpus/corpora (see also Lange 2022).

The relevance of documenting learner and L2 learning variables, as well as task-related variables has been discussed in detail in a number of volumes and monographic chapters (see, for example, individual differences in second language acquisition, in general, in Li, Hiver and Papi 2022, Kerz and Wiechmann 2020, Ädel 2015, Saville-Troike 2012; the affective individual difference variables, e.g., emotions, motivation, linguistic self-confidence, self-esteem, etc., in Albert 2022, Dörnyei and Ryan 2015; language aptitude, language learning strategies, personality, in Dörnyei 2005, Dörnyei and Ryan 2015; the task variability in language learning research, in Tracy-Ventura and Myles 2015, the situational variables, in Gablasova 2020), as well as in more in-depth studies (some of which will be cited below).

LECOR metadata scheme follows the core metadata scheme for learner corpora developed by König *et al.* (2022), a revised and standardised version of the Granger and Paquot (2017) schema proposal. Besides a documentation about the error annotation, text transcription, anonymization, and POS annotation, LECOR corpus will be accompanied by a manual to define and describe all the variables encoded in the metadata.

The recording of LECOR variables is completed with full regard to all ethical regulations in all the stages of compiling the corpus (collection, storage, annotation) and in its exploitation in research; both learners and researchers handling the data must comply with national laws regarding the processing of personal information and the EU General Data Protection Regulation; learners, teachers and project members sign a consent form.

The recording of LECOR metadata was made with the end purpose of making them available to the corpus users (learners, researchers, etc.). They will be accessible for searching, together with the actual data, by indexing within the NoSketchEngine (Kilgarriff *et al.* 2014) open-source corpus query platform. In such a platform, any number of metadata variables can be used as searching criteria to extract specific usage examples from the corpus.

3. LECOR metadata description

The metadata schema we designed, based on other well-known learner corpora descriptions (synthesized and standardized in LCR König *et al*. (2022) metadata scheme), is intended to balance the richness of metainformation available for such a resource and the degree of interest prompted by this information in the corpora

studies field with the cost of systematically recording and structuring the metadata, especially in terms of human work. Therefore, we came with a three-module architecture for our metadata: the general corpus metadata (see Section 3.1), the metadata associated with each student that provided compositions for LECOR (see Section 3.2) and the metadata associated with each composition (), where by composition (including transcription of an audio file) we understand a specific answer of a specific student to a specific learning task (see Section 3.3). The task description is registered for each composition: see fields like TASK_TYPE, TASK_REQUIREMENTS, TIME_LIMIT and LENGTH_REQUIREMENT in Section 3.3.

The Corpus Metadata module is independent in the architecture, while the Student/Learner module and the Work/Composition Metadata module are connected through the STUDENT_ID field.

Most of the fields in the metadata have predefined values, which, together with the guidelines that we will provide, will facilitate the user experience of searching the corpus. Compositions in the corpus having specific attributes can be searched only if the specific attribute is clearly encoded in the corpus indexing process. E.g., if we want to select only written samples as opposed to those provided in audio form, we need to access the TASK_FORM metadata and chose only the documents for which TASK_FORM has the specific value: "written". In this case, the only possible predefined values are "written" and "oral" (selected from a drop-down list) and the restriction on the values is necessary to avoid inconsistencies in data retrieval caused by misspelling or heterogeneity of the values. Whenever a new value occurs (for example, a leaner declares a L1/L2 that was not present in the predefined value list), the list is updated by including this new value.

Very few fields have descriptive values and are designed only to provide supplementary information, not to be searched on: valid URL links of the original scanned (for written) or recording (for oral) compositions, detailed requirements of the tasks executed in the compositions, student and composition/work ID, student name, age, email and group (name, e-mail and group are metadata that will not be publicly available), etc.

Some information about the student/learner can vary along the academic year, therefore we provide two different fields for the same property, to be recorded successively in the first and second semester. A student's evolution is followed only along two semesters, because the educational setting for collecting the samples is one full-time intensive university year (in total, 28 weeks) dedicated to L2 language training; the curriculum of the Preparatory year of Romanian language comprises 28 hours per week in the first semester and 30 in the second one. In Section 3.2, we specified only fields associated with the first semester (see fields whose names start with 1st_SEM), to avoid tedious repetition, but in the schema, all this information is accompanied by second semester equivalents.

Most student metadata are self-reported, being collected based on a socio-linguistic survey created for LECOR project's purposes and self-administered by the students in class, with teacher assistance. Metadata about general language aptitude/abilities, regularity of class attendance, degree of motivation for studying and degree of creativity are mostly subjective variables evaluated intuitively for the students by the teachers who accompany them in a one year long learning process. All metadata about the task requirements and the resulting compositions are also produced by the teachers.

3.1. Corpus metadata

The fields selected to describe the corpus are a reduced collection of fields from the administrative and corpus design LCR metadata scheme (König *et al.* 2022). General identification corpus metadata like CORPUS_NAME (corpus full name in English), CORPUS_ACRONYM, LANGUAGE (language targeted by the corpus; Romanian, in our case), VERSION (states the particular version of a corpus; the first digit represents a major update, e.g. adding new data or new annotations, while the second digit represents minor update corrections of annotations; the value will be 1.0 in our case); URL (a URL of the interface where the corpus will be available for searching), REF_ARTICLE (the article that the authors would like to be used when referring to the learner corpus), DESCRIPTION (summary description of the corpus to help the user establish the resource's fitness to the user's purpose) are necessary when advertising the corpus on different online platforms.

For further details concerning the context of the corpus creation we introduced variables like CORPUS_DOCUMENTATION (links to extensive documentation of the corpus, including metadata description, transcription guidelines/decisions; correction guidelines/decisions), CORPUS_AUTHORS (names of the people responsible for the corpus creation and annotation), CONTACT_EMAIL: e-mail address of a reachable contact (institution or person); RELATED_RESEARCH_PROJECT (name of the financing research project and link to the project webpage), PUBLISHER (name of the institution responsible for the distribution of the corpus, in our case, the Faculty of Letters of the University of Bucharest), DATE_OF_PUBLICATION (publication date of the current corpus version), OTHER_VERSIONS (name of versions and links to other versions of this dataset), AVAILABILITY and LICENCE. Regarding the licence, given that LECOR will be a corpus in free access (open for searching through the provided interface) but it will not be downloadable, we need to document what the Romanian legislative regulations are for this specific situation (whether or not a licence is needed).

General metadata describing the content and format of the corpus that we employed are: CHARACTER_ENCODING (the character encoding for the text

files: UTF-8, the character encoding standard used for electronic communication), FILE_FORMAT (the file formats for all types of files (text, audio, images) in the corpus), L1_LANGUAGE (list of all the L1 languages of the students that contributed to the corpus), CORPUS_MODE (multimodal, containing written, spoken and image documents; in our case, written and spoken-based corpus, with broad orthographic transcription), LONGITUDINAL (value: yes; longitudinal corpora follow the same learners/students over some time − 2 semesters in our case − and thus include more than one text per learner), TIME_OF_DATA_ COLLECTION (the period when the data has been collected; 2019-2024 in our case), PLACE_OF_DATA_COLLECTION (the city and country of data collection: Bucharest, Romania), DATA_PRODUCTION_SETTING (the setting in which the data were collected; in our case: language instruction setting), EDUCATIONAL_ SETTING (specification of the setting for the students' language acquisition; in our case, in general, the value is "undergraduate"). Elementary statistics of the data are expressed by variables like CORPUS_SIZE_TOKENS (Number of tokens in the corpus; numbering is done on the students' original texts or transcriptions), CORPUS_SIZE_TEXTS (the number of compositions recorded by the corpus), CORPUS_SIZE_STUDENTS (the number of learners/students that contributed to the corpus).

Corpus metadata act like a calling card for corpora advertising: this type of information is either presented as such on the specific resource homepage or indexed in various online resource repositories.

3.2. Student/Learner metadata

Learner variables, which refer to corpus design criteria based on the students who contributed to the corpus, can be classified as general versus L2-specific. General learner metadata fields that we introduced in our schema are self-reported in a detailed learner profile questionnaire: NAME (the complete name of the student, with the family name(s) followed by given name(s)), EMAIL (e-mail address provided by the student), GROUP (the number of the group the student belongs to in the series of students), TEACHER (teacher's / teachers' name field is a complete name of the teacher, with the family name(s) followed by given name(s)), ACADEMIC_ YEAR (the academic year the student was enrolled in, e.g. 2019-2020), AGE (a number that represents the student's age), GENDER (student's gender, with possible values: female, male, rather not say or non-binary), NATIVE_LANG ((1st) native language of the student), BILINGUAL/TRILINGUAL (values: yes/no; specifies if the learner is bilingual/trilingual, that is they have two or three native languages), 2nd LANGUAGE/3rd LANGUAGE (the second/third native language of the learner, correlated with options for the BILINGUAL/TRILINGUAL field).

L2-specific self-reported variables are: LEARN_REGION (the region where the learner is learning Romanian, with possible values: Romania/Other country; currently, the LECOR corpus includes only productions by students studying Romanian in Romania (in the speaking country), but, in the future, the hope is that the scalable LECOR corpus will incorporate samples produced by learners from outside Romania (in the non-speaking environment), in which case, the general LECOR metadata will be updated); 1st_SEM_MOTIVATION (motivation for studying Romanian; the predefined list of possible values is: Studies, Business, Citizenship, Personal interest, Other), OTHER_FOREIGN_LANG (knowledge of other foreign language(s) and proficiency level (beginner, intermediate or advanced); the values are free combinations of languages and levels (ex: English-intermediate, French-advanced), separated by comma), 1st_SEM_PARALLEL_STUDYING (is the learner studying another foreign language in parallel with Romanian, in the first semester? the values are yes or no), 1st_SEM_PARALLEL_LANG (specification of the language studied in parallel in the first semester, with values selected from a multiple values drop down list; a "none" option has to be correlated with a "no" option for the previous field), 1st_SEM_RO_STUDY_MODE (mode of study of L2 Romanian in the first semester; the predefined list of possible values is: in an academic context (at school/university); on a private course at some private school private classes with a teacher; informally (immersion); using online platforms and mobile applications listening and watching Romanian TV programme; watching movies; listening to music; listening to audiobooks/podcasts; self-instruction (grammar/vocabulary/etc.); talking with native speakers in Romanian; others), 1st_SEM_CLASS_HOURS_ PER_WEEK (in the case of study at an institution, the number of hours spent in class each week in the first semester; the value is a number), 1st_SEM_ COURSE_TYPE (in the case of study at an institution, what sort of a course was it in the first semester? the predefined list of possible values is: general language, Romanian literature, language for specific purposes, Romanian culture), 1st_SEM_RO_PREVIOUS_STUDY (whether, prior to the beginning of the process of collecting samples from him/her in the first semester, the learner had already studied Romanian; the values are yes or no), 1st_SEM_RO_PREVIOUS_ STUDY_DETAILS (if previous study was undergone before the first semester, this field provides more details like in what kind of programme they had studied, how many lessons a week, etc.;), 1st_SEM_HOME_LEARNING (does the learner speak Romanian in their family/home context in the first semester? Possible values: yes or no), 1st_SEM_USAGE_FREQ (how often does the student use Romanian in their interactions with native speakers in the first semester? Possible values: not at all; very little; a little; quite a lot; a lot; most of the time; the whole time).

Variables describing the linguistic context of the corpus collection are contributing to an image of the learning environment, which highly impacts language production performance. In our case, the students are immersed in a second language context and, therefore, exposed to Romanian in many day-to-day social activities, sometimes both in and outside the home. They also have access to online and media produced oral and written materials. Combining different such variables as parameters that influence proficiency can facilitate very interesting and useful research studies.

The learner's native language (L1) recording is one of the most important metadata, without which learner corpus research (LCR) would be quite irrelevant. Firstly, there has been and is a growing interest in the errors that a particular population of learners makes, for the improvement of teaching/learning of that language[2]. Secondly, without recording the native languages of the learners, it would not be possible to do the Contrastive Interlanguage Analysis (CIA, see Granger 2015) applied to samples produced by learners with different mother tongue backgrounds (e.g., the Romanian produced by learners with Arabic as an L1 and the Romanian produced by learners with Turkish or Bulgarian as an L1).

Studies also showed that differences between L1 and L2 affect L2 acquisition of morphosyntax (see, e.g., McManus 2015).

Based on a LECOR subcorpus, Șinca (2023) analyses the communicative strategies used by four populations of learners of Romanian: with L1 Turkmen, Arabic, Albanian and Greek. She finds that Turkmen learners use communicative strategies much more frequently (of the examples identified in her subcorpus, 40% belong to these learners), which might indicate that they place more emphasis on getting the message across than on the correct form of the message (as native speakers of Arabic and Albanian seem to do). As regards the types of strategies preferred, the author remarks that Greek, Arabic, and Albanian learners prefer the strategy of approximation, while Turkmen learners prefer the strategy of code-switching (with English, which seems to be a principal second language for them).

Studies in L3/Lx acquisition state that all previous language learning experiences have an impact on the interlanguage development (Jessner, Megens, and Graus 2016). Therefore, not only the mother and home tongue(s) have to be documented, but also all additional languages and living-abroad experiences of the students (see also immediately above the mention of code-switching with a second language, not the native language, Șinca 2023).

---

[2] See, for example, the textbooks authored by Cerneva (2007), *Limba română pentru bulgari* [Romanian language for Bulgarian learners] or by Yacob (2007), *Limba română pentru arabi* [Romanian language for Arabic learners]. Such textbooks would be more effective for the Bulgarians, Chinese, etc. learning Romanian if, besides the learners' native metalanguage and contrastive grammar explanations, they insisted on the specific errors of these natives.

L2-specific learner variables reported by the teachers are: 1st_SEM_ LANG_ABILITIES (general language aptitude/abilities in the first semester, expressed in grades and rating/qualifications; possible values: 10 – excellent; 9 – very good; 8 – good; 6-7 – satisfactory; 5 – sufficient; 1-4 – unsatisfactory[3]), 1st_SEM_CLASS_ATT_REGULARITY (the regularity of attendance at class in the first semester, expressed in percentages and rating/qualifications; possible values: 100% – excellent, 90% – very good; 80% – good; 60-70% – satisfactory; 50% – sufficient; 10-40% – unsatisfactory), 1st_SEM_MOT_DEGREE (degree of motivation for studying in the first semester; possible values: 1; 2; 3; 4; 5, with 5 standing for the highest degree of motivation), CREATIVITY (degree of creativity; possible values: 1; 2; 3; 4; 5, with 5 standing for the highest creativity), OTHER_OBS: general remarks about the learner's group (the general proficiency level of the group; extra-classroom activities, such as: visits to museums, excursions, intercultural events, etc.); the role of the student in the group (a student who helps their colleagues a lot, who is a leader or one who does not display these qualities, etc.).

The list of fields is completed with a STUDENT_ID, which is a natural number (in the range (0,n), where n is the total number of students participating with compositions for the corpus) that uniquely identifies a student. Since all personal data have to be anonymised, to assure privacy rights for the learners, the STUDENT_ID will be the only publicly available reference to the student.

Specialists in the field insist that a learner proficiency measure (in our case, 1st_SEM_LANG_ABILITIES, but see also Section 3.3) should always be included in the metadata, acknowledging that it is the most time-consuming assessment to be obtained, since it is not sufficient to encode years of exposure to language or institutional status of the learner. Standardized placement tests corroborated with self-proficiency ratings at specific points in time are recommended proficiency approximations (Tono and Díez-Bedmar 2014, Lozano and Mendikoetxea 2013, a.o.).

Internal-affective measures, like the motivation for studying or the creativity, are also metadata of interest for investigations on their correlation with language usage performance. They are also correlated with the corpus collection setting, since voluntary contributions are typically associated with higher degree of motivation and confidence, while contributions collected in a formal educational institution show a wider range on the motivation scale (Gilquin 2015), as is the case in our corpus.

Existing corpus-based studies dedicated to the effects of learner variables on the variation of L2 use have focuses mostly on the two most commonly

[3] In the Romanian university system, the maximum mark is 10 and the minimum is 1; 5 is the mark for passing an exam.

available variables, L2 proficiency and L1 background (Gablasova *et al.* 2019, who looked at environment variables, is a notable exception).

Lu (2011) was one of the earliest learner corpus analysis studies on the variation of syntactic complexity across proficiency levels, using a set of 422 timed argumentative essays from a single college. Römer and Garner (2019) were interested in how the use of verb-argument constructions varies in the L2 usage of spoken English across proficiency levels and across different L1 backgrounds (Italian and Spanish), using Trinity Lancaster Corpus Sample (TLCS). Gilquin (2019) focused on the study of cross-proficiency variation in the use of light verb constructions in spoken English among English as a foreign language and English as a second language learners. Other studies examined variation in the use of lexical bundles (Staples *et al.* 2013)) or multi-word sequences (Garner 2016) across different three and five proficiency levels, respectively.

Stormbom (2018) investigated differences in the use of epicene pronouns ("he", "he or she", and "they") among L1 English writers and L2 English writers with different L1 backgrounds (Bulgarian, Czech, Dutch, Finnish, German, Italian, Norwegian, Polish, Russian, Spanish, Swedish, and Turkish), using the pronoun choice as indicator for the degree of non-sexist language use. Castello and Gesuato (2019) were interested in the variation of the use of lexical backchannels[4] (as signals of active listener-ship and discourse competence) in the oral discourse of L2 English learners from different L1 backgrounds.

Some studies were interested in the correlation between L2 proficiency level and L1 background concerning their effect on L2 acquisition. E.g., we have seen before that Römer and Garner (2019) were working with both L1 Italian and Spanish learners of English at different proficiency levels. Lu and Ai (2015) investigated variation in the syntactic complexity of L2 English writing among L2 English learners with different L1 backgrounds, while Shatz (2019) looked into how the learners' capitalization abilities were affected by L2 English learners' L1 backgrounds in interaction with L2 proficiency level[5].

3.3. Work/sample metadata

Researchers proved that completing learner metadata with more self/teacher-reported data associated to each specific work, like the place of the text production (school vs. home), the using of supplementary materials, feedback opportunities,

[4] Two categories of lexical backchannels: *Convergence* (25 items, such as "absolutely", "probably", "I know", "that's right") and *Request for Confirmation* (1 item, "really").

[5] We have found it useful to list several studies that have considered different variables, in the belief that their relevance, as described in detail in the literature (see the references in Section 2 above), is self-evident. Reporting the results of these studies, as suggested by an anonymous reviewer, would in fact imply a rather detailed presentation of these studies and would unbalance the content of the present article, which is mainly aimed at a detailed presentation of LECOR metadata.

time dedicated to the task, etc. is even more instructive for L2 language usage performance analysis: through these variables, the naturalness of the task solving or the learners inclination on producing meaningful vs. grammatically correct sentences (Bell and Payant 2020) can be appreciated.

In our corpus, the metadata associated with each composition will describe conditions of collection, information about the task being accomplished in the composition and general level of proficiency at the moment of the task assigned, as appreciated by the teacher.

DATE, which encodes the date the composition was provided, it is very important for diachronic studies, concerned with the evolution of the leaner performance over time. Yuldashev, Fernandez and Thorne (2013), for example, conducted an examination on the production of multi-word units over 38 weeks of online Spanish training, in weekly blogs and instant messages of the learners. Designing a longitudinal study is not an easy endeavour: academic constraints, like long-term funding or the pressure for constant production of publications, together with the necessity for sustainability and reliability of different context elements (like students' and teachers' availability, hardware and software used for storing and processing the data etc.), are only some of the many details that had to be taken into account in LECOR when designing a long term corpus collection process.

The INSTITUTION variable, with possible values "University of Bucharest" and "Others", was encoded as we envisage the possibility of future extension of the corpus from collaboration with other institutions.

Access to the original forms of the compositions, either scanned images or audio/video recordings, is provided through HAND_WRITTEN_COMP and AUDIO_VIDEO_COMP variables, whose values are links to png/mp4 files on the project server.

TASK_FORM (is the task written or oral? Possible values: oral/written), TASK_REQUIREMENTS (the requirements laid down for completion of the task, expressed by free form string values), TIME_LIMIT (possible values: timed/untimed as attributes of the task completion), LENGTH_REQUIREMENT (the completion of the task has a length requirement; possible values: yes/no), and LENGTH_REQUIREMENT_DETAILS (number of words, lines, minutes, number of replies for dialogues; recommended format: words: 350, lines: 12, approx. 5 min; the field value "none" has to be correlated with a "no" value for the previous field) are variables describing the task that has to be accomplished, and therefore can be common to more compositions from different students accomplishing the same task. Although the number of spoken learner corpora (comprised entirely of oral compositions) and the number of oral compositions in mixed/multimodal corpora is continuously increasing, they are outnumbered by written corpora (Gilquin 2015) due to the significantly greater costs in terms of human labour and time that recording and transcribing compositions entails, as compared to computer-written and even

hand-written compositions collection. These facts also reflect in our decision to include only 20% oral compositions in LECOR.

Conditions of composition production are important since they are indicators of the actual proficiency level of the learner at the moment, beyond the level expressed in the composition. E.g., SPONTANEITY, with the possible values Spontaneous/Prepared, can indicate that the production is either made on the spot, and therefore is a good indication of proficiency, or prepared, most probably at home and using reference materials, situation that will bias the proficiency evaluation. SPONTANEITY can be corroborated with a) REF_SOURCE_ USAGE, that indicates the possibility of use of external sources of reference (dictionaries, internet, etc.) in producing the composition, as estimated by the teacher (possible values: unknown, most probably no, most probably yes); b) TASK_TYPE, with possible values "an exercise that has been completed in class", "something set as homework", "work done in an exam", since exam and class produced compositions are clearly more reliable than homework in terms of proficiency evaluation; c) WRITING_TYPE, indicating if the task is written by hand or in word-processing applications like MS Word or LibreOffice, which is important because word-processing applications have word proofing functions. GENERAL_ PROFICIENCY, encoding general proficiency level of the group when the task was assigned (possible values: A1, A1-A2, A2, A2-B1, B1, B1-B2, B2, B2-C1, C1, C1+) is also to be taken into account when analysing composition proficiency.

Some examples of research that have taken into account this very important variable have been provided in Section 3.2 above. We add to those the results of the analysis of a LECOR sub-corpus (samples produced by 160 A2-B1 learners of Romanian, see Mîrzea Vasile 2024) showing that in the development of interlanguage progress is not linear and certain specific features may not correlate with the proficiency level. The study examines the acquisition of adjective gradation, focusing mainly on the preposition required by the graded adjective; the findings are that there is a slightly higher error rate especially with the relative superlative among B1 students (13.19%) than among the A2 students (5%).

DIACRITICS (with possible values: not applicable, yes, mostly yes, mostly no), specifying the usage of diacritics in the composition production, are important beyond the indication of the learner correct acquisition of the Romanian alphabet and vocabulary. The usage of diacritics in Romanian can seriously affect Natural Language Processing tools such as the POS taggers, which are trained on diacritically correct documents, and therefore the POS tagging annotation is less reliable on texts with incomplete or incorrect diacritic usage.

Other composition variables describe the type of content expressed in the texts: TEXT_STYLE (the literary style of the composition: narration, description, argumentation, etc.;), TEXT_REGISTER (the registers of the text: essay, summary, translation, letter or email, diary, creative writing, scientific writing, social media,

oral dialogue, oral monologue, etc.), TOPIC (one or more topics referred to in the composition, like leisure, transportation, famous people, daily routines, etc.). All these content related aspects may correlate with different types of linguistic features (like grammatical structures, vocabulary, etc.) and impact research conclusions (Tracy-Ventura and Myles 2015).

Finally, the list of composition metadata includes a WORK_ID (a number identifying the work/composition contained by the document) and is associated with the STUDENT_ID described in the learner metadata section, creating a mapping from the student to all their associated compositions.

4. Conclusions and future perspectives

As it could be seen, the LECOR corpus contains a very rich metadata schema, which records the variables that influence RL2/FL learning. These metadata will allow accurate studies based on the corpus, and their results can be explained in a relevant way taking into account these variables. For example, we will be able to check which errors are common in Romanian interlanguage development (regardless of the learners' native language) and which errors are more likely to be produced by native language influence, we will be able to check whether the learners' gender has or does not have an influence on the learning process, which hierarchy of factors leads to a successful learning, etc.

Apart from designing the corpus with rich metadata (thus conforming to a quality standard of the reference literature), using the corpus properly also requires explaining how this metadata was collected, how values were assigned, etc., which is the very idea behind this article.

Although LECOR is still under construction, it has already allowed some analysis on various aspects of the RL2/RFL acquisition. Many other studies on interlanguage development and many applications are possible in the future (wordlists for each level, grammars, tailored materials for specific groups of learners, etc.).

When the electronic version of the corpus is fully ready, a discussion of the challenges of fully implementing metadata and its use will be useful.

ACKNOWLEDGEMENTS

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS – UEFISCDI, project number PN-III-P1-1-1.1-TE-2019-1066 (*Learner Corpus of Romanian (LECOR). Collection, Annotation and Applications / Corpus de română ca limbă străină (LECOR). Colectare, adnotare și aplicații*).

References

Albert, Á., 2022, *Investigating the Role of Affective Factors in Second Language Learning Tasks*. Springer.

Ädel, A., 2015, "Variability in learner corpora", in S. Granger, G. Gilquin and F. Meunier (eds), *The Cambridge handbook of learner corpus research*. Cambridge, Cambridge University Press, 401-420 (*https://doi.org/10.1017/CBO9781139649414.002*).

Barbu, A.M., E. Irimia, C. Mîrzea Vasile and V. Păiș, 2023, "Designing the LECOR Learner Corpus for Romanian", in G. Angelova, M. Kunilovskaya and R. Mitkov (eds), *Deep Learning for Natural Language Processing Methods and Applications (Proceedings of International Conference Recent Advances in Natural Language Processing, RANLP 2023, Varna, 4–6 September, 2023)*. Bulgaria, INCOMA Ltd. Shoumen, 143-152 (*www.acl-bg.org*).

Bell, P., and C. Payant, 2020, "Designing learner corpora: Collection, transcription, and annotation", in N. Tracy-Ventura and M. Paquot, *The Routledge handbook of second language acquisition and corpora*. London / New York, Routledge, 53-67.

Castello, E., and S. Gesuato, 2019, "Holding up one's end of the conversation in spoken English: Lexical backchannels in L2 examination discourse", *International Journal of Learner Corpus Research* 5(2), 231-252 (*https://doi.org/10.1075/ ijlcr.17020.cas*).

Cerneva, M., 2007, *Limba română pentru bulgari / Ruminski ezik samoucimel b gualozu*, Editura Gramma.

Constantinescu, M.-V., and G. Stoica, 2020, *Româna ca limbă străină. Corpus*. București, Editura Universității din București.

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

Dörnyei, Z., 2005, *The Psychology of the Language Learner. Individual Differences in Second Language Acquisition*. Mahwah, NJ, Lawrence Erlbaum.

Dörnyei, Z., and S. Ryan, 2015, *The psychology of the language learner revisited*. New York / London, Routledge.

Gablasova, D., 2020, "Variability", in N. Tracy-Ventura and M. Paquot, *The Routledge handbook of second language acquisition and corpora*. London / New York, Routledge, 358-369.

Gablasova, D., V. Brezina, and A. McEnery, 2019, "The Trinity Lancaster Corpus: Development, Description and Application", *International Journal of Learner Corpus Research*, 5 (2), 126-158 (https://doi.org/10.1075/ijlcr.19001.gab).

Garner, J.R., 2016, "A phrase-frame approach to investigating phraseology in learner writing across proficiency levels", *International Journal of Learner Corpus Research* 2 (1), 31-68 (https://doi.org/10.1075/ijlcr.2.1.02gar).

Gilquin, G., 2019, "Light verb constructions in spoken L2 English: An exploratory crosssectional study", *International Journal of Learner Corpus Research* 5 (2), 181-206 (https://doi.org/10.1075/ijlcr.18003.gil).

Gilquin, G., 2015, "From design to collection of learner corpora", in S. Granger, G. Gilquin and F. Meunier (eds), *The Cambridge handbook of learner corpus research.*

Cambridge, Cambridge University Press, 9-34 (*https://doi.org/10.1017/CBO9781139649414.002*).

Granger, S., 2015, "Contrastive interlanguage analysis: A reappraisal", *International Journal of Learner Corpus Research* 1(1), 7-24.

Granger, S., and M. Paquot, 2017, "Towards standardization of metadata for L2 corpora" (invited talk), *The CLARIN workshop on Interoperability of Second Language Resources and Tools*, 6-8 December 2017, University of Gothenburg, Sweden.
(*https://sweclarin.se/sites/sweclarin.se/files/event_atachements/Granger_Paquot_Metadata_G%C3%B6teborg_final.pdf*)

Jessner, U., M. Megens and S. Graus, 2016, "Crosslinguistic influence in third language acquisition", in R. A. Alonso (ed.), *Crosslinguistic influence in second language acquisition*. Bristol: Multilingual Matters Clevedon, 193-214.

Kerz, E., and D. Wiechmann, 2020, "Individual differences", in N. Tracy-Ventura and M. Paquot, *The Routledge handbook of second language acquisition and corpora*. London / New York, Routledge, 394-405.

Kilgarriff, A. *et al.*, 2014, "The Sketch Engine: ten years on", *Lexicography* 1 (1), 7-36.

König, A., J.C. Frey, E.W. Stemle, A. Glaznieks and M. Paquot, 2022, "Towards standardizing LCR metadata", Book of abstracts of *The 6th International Conference for Learner Corpus Research (LCR 2022*, Padova, 22.9.20220), 78 (*http://www.maldura.unipd.it/lcr-2022/docs/LCR2022_BoA.pdf*).

Lange, H., 2022, "Metadata Formats for Learner Corpora: Case Study and Discussion", in *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*. Louvain-la-Neuve, Linköping University Press, 108-113 (*https://aclanthology.org/2022.nlp4call-1.11.pdf*).

Li, S., P. Hiver and M. Papi, 2022, *The Routledge handbook of second language acquisition and individual differences*. New York, Routledge.

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

Lu, X., 2011, "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development", *TESOL Quarterly* 45 (1), 36-62 (*https://doi.org/10.5054/tq.2011.240859*).

Lu, X., and H. Ai, 2015, "Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds", *Journal of Second Language Writing* 29, 16-27 (*https://doi.org/10.1016/j.jslw.2015.06.003*).

McManus, K., 2015, "L1–L2 differences in the acquisition of form–meaning pairings in a second language", *The Canadian Modern Language Review* 71 (2), 155-181 (*https://doi.org/10.3138/cmlr.2070.51*).

Mîrzea Vasile, C., 2024, "Aspecte privind gradarea adjectivelor și complementul comparative la nenativi. Studiu de corpus" (unpublished).

Römer, U., and J. R. Garner, 2019, "The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency", *International Journal of Learner Corpus Research* 5 (2), 207-230 (*https://doi.org/10.1075/ijlcr.17015.rom*).

Saville-Troike, M., 2012, *Introducing second language acquisition*, 2nd ed. Cambridge, Cambridge University Press.

Shatz, I., 2019, "How native language and L2 proficiency affect EFL learners' capitalisation abilities: A large-scale corpus study", *Corpora* 14 (2), 173-202 (*https://doi.org/10.3366/cor.2019.0168*).

Staples, S., J. Egbert, D. Biber and A. McClair, 2013, "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section", *Journal of English for Academic Purposes* 12 (3), 214-225 (*https://doi.org/10.1016/ j.jeap.2013.05.002*).

Stormbom, C., 2018, "Epicene pronouns in intermediate to advanced EFL writing", *International Journal of Learner Corpus Research* 4 (1), 1-22 (*https://doi.org/10.1 075/ijlcr .16016.sto*).

Șinca, I., 2023, "Communicative strategies in a1 written Romanian. A case study", poster presentation at the workshop *Corpus based and experimental research in SLA. The state-of-the-art and future directions* (8th Bucharest Colloquium of Language Acquisition, 16–18.11.2023.2023).

Tono, Y., 2016, "What is missing in learner corpus design?", in M. Alonso-Ramos (ed.), *Spanish learner corpus research: Current trends and future perspectives.* Amsterdam, Johh Benjamins Publishing Company, 33-52.

Tono, Y., and M. B. Díez-Bedmar, 2014, "Focus on learner writing at the beginning and intermediate stages: The ICCI corpus", *International Journal of Corpus Linguistics* 19 (2), 163-177.

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

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

Yacob, N., 2007, *Limba română pentru arabi*. Cluj-Napoca, Casa Cărții de Știință.

Yuldashev, A., J. Fernandez and S.L. Thorne, 2013, "Second language learners'contiguous and discontiguous multi-word unit use over time", *The Modern Language Journal*, 97 (S1), 31-45.