Designing the LECOR Learner Corpus for Romanian

Ana-Maria Barbu[1,2], Elena Irimia[3], Carmen Mîrzea Vasile[1,2], and Vasile Păiș[3]

[1]Faculty of Letter, University of Bucharest, 5-7 Edgar Quinet, Bucharest
[2]"Iorgu Iordan - Al. Rosetti" Institute of Linguistics, 13 Calea 13 Septembrie, Bucharest, Romania
[3]Romanian Academy Research Institute for Artificial Intelligence, 13 Calea 13 Septembrie, Bucharest, Romania
anamaria.barbu@g.unibuc.ro elena@racai.ro
carmen_marzea@yahoo.fr vasile@racai.ro

Abstract

This article presents a work-in-progress project, which aims to build and utilize a corpus of Romanian texts written or spoken by non-native students of different nationalities, who learn Romanian as a foreign language in the one-year, intensive academic program organized by the University of Bucharest. This corpus, called LECOR – Learner Corpus for Romanian – is made up of pairs of texts: a version of the student and a corrected one of the teacher. Each version is automatically annotated with lemma and POS-tag, and the two versions are then compared, and the differences are marked as errors at this stage. The corpus also contains metadata file sets about students and their samples. In this article, the conceptual framework for building and utilization of the corpus is presented, including the acquisition and organization phases of the primary material, the annotation process, and the first attempts to adapt the NoSketch Engine query interface to the project's objectives. The article concludes by outlining the next steps in the development of the corpus aimed at quantitative accumulation and the development of the error correction process and the complex error annotation.

1 Introduction

The LECOR corpus is developed through the project "Learner Corpus of Romanian (LECOR). Collection, Annotation and Applications", funded by the Romanian Government as part of the subprogram dedicated to research projects to stimulate independent young teams (TE). The host institution of the project is the University of Bucharest. In this section we present the core features of the LECOR corpus and the different uses for which it has been designed. The main goal of the project is to build and make accessible the first Romanian electronic learner corpus and, at the same time, to professionalize human resources in this field of research. It will be a corpus in free access, open for searching through the provided interface. LECOR will be downloadable only by request and only for research purposes, which will be in accordance with the informed consent signed by learners.

LECOR is a monolingual general learner corpus, collected over four academic years (2019-2024). The target language is (general, standard) Romanian, learned in a native speaking environment and taught by native teachers. The learners are university students, aged generally between 18 and 25. Their native languages are various (more than 20 mother tongues): Arabic, Chinese, Korean, Albanian, Greek, Armenian, Turkmen, Turkish, Persian, Slavic languages (Bulgarian, Serbian, Ukrainian, Belarusian, Russian, etc.). Among the internationally spoken languages, Arabic and Mandarin Chinese are well represented, in contrast to English, Spanish, and Hindi, which are scarcely or not at all found, just like all Western and Northern European languages. The Romance languages are also weakly represented, especially through their non-standard varieties: Spanish and Portuguese from Latin America, and French from Africa. As a general characteristic, the learners' native languages exhibit typological distance both from Romanian and among themselves. As the corpus processing is not yet complete, an exact proportion of native languages in LECOR cannot be provided at this time.

In its first version, LECOR will contain written (80%) and oral (20%) learners' samples. The corpus is planned to be of large size, including 4,000 samples (of which, 800 audio transcriptions). The 4,000 samples in the corpus have very different sizes (ranging from samples provided by A1-level students, consisting of at least 40-50 words, to samples provided by B2+ level students, containing more than 350 words). Taking an average of 150-170 words per sample, we can approximate

Proceedings of Recent Advances in Natural Language Processing, pages 143–152
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_016

the size of LECOR at the end of the project to be over 600,000 words. At the moment, about 4,500 written productions are collected (from which a selection will be made, prioritizing exam papers and texts produced in class) and about 500 files with audio homework and exam recordings; users of the corpus will have access to original samples (handwritten texts and audio files). All the texts will be automatically annotated with lemma and POS-tag, while a small part will be annotated for errors. The LECOR corpus was designed to be scalable, so the aim is to increase its size and continue the annotation.

In each of the four academic years on the duration of the project, samples are collected, mostly in controlled contexts, from approximately 50-70 students per year, documenting their progress from A1 to B1 or B2 proficiency level. Therefore, the corpus can be used for both synchronic/cross-sectional and diachronic/ longitudinal research.

LECOR is all the more valuable as it encompasses, besides A2, B1 and B2 samples, at least one quarter of A1 samples and as it thoroughly documents the interlanguage development of several dozens of learners, who have produced approximately 60 samples throughout an entire academic year[1].

Regarding the representativeness of proficiency levels, LECOR is a relatively balanced corpus: the number of samples from A2 and B1 learners is comparable, whereas A1 learners contribute slightly fewer samples, and B2 learners produce the least number of samples.

The text types are varied and comply with the minimum proficiency level requirements (e.g. argumentative essays are not required at beginner level). In LECOR there are descriptions (of a city, of a (class)room, a house, a person, etc.), especially at A1 and A2 levels, narratives (*What I did today, What I used to do on holiday as a child, A nightmare trip*, etc.), argumentative essays (Why it's good to learn languages, Online shopping – pros and cons, Protecting the environment, etc.). Several description and story-telling tasks are based on pictures. The text genre are also diverse: e-mail (letter), long WhatsApp message, review, essay, description, procedure (recipe, health instructions), etc.

LECOR is a very well documented resource, learner variables/metadata, as well as text and task variables/metadata being carefully and thoroughly recorded, following the core metadata scheme for learner corpora (see König et al., 2022). Because it is a large annotated corpus with rich metadata and a high degree of representativeness, LECOR will have many possible end-uses.

At first, it will be used in studies about non-native Romanian acquisition and, in general, in second language acquisition research (for testing particular SLA theories, to set up the interlanguage profile at certain stages of SL / FL development, to track individual differences, etc., see also Granger et al. (2015)).

Then, the corpus can be used in language teaching and in natural language processing. Traditionally, learner corpora are used for didactic purposes (for an overview of applications, see at least McEnery et al. (2006); Díaz-Negrillo and Thompson (2013); Granger et al. (2015); Mitchell (2021)). On the one hand, it can be used to inform instructional materials design, such as course books (e.g. *Learning from common mistakes*, Brook-Hart (2009)), learner dictionaries (see, for example, Macmillan English dictionary advanced learner, Rundell (2007)), wordlists (e.g. *Focus on Vocabulary 2: Mastering the academic word list*, Schmitt and Schmitt (2011)), etc.; moreover, the metadata will allow for creating a '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. Such materials do not exist at all for Romanian and are obviously long overdue by both learners and instructors. More precisely, based on the learner corpus (and, in many cases, a contrasting, language-target corpus), numerous research questions can be addressed: *How does second language evolve across different levels of proficiency? Which errors are developmental (specific to all learners) and which are likely to be caused by transfer from the native language? What are the specific features of interlanguage at a given proficiency level for a given population of learners?* For Romanian as a target language, Vasiu (2020) tries to identify, based on an own corpus what is the specificity of interlanguage at A1 level with respect to the learners' native language. Using quantitative analysis, the author reaches conclusions such as: at A1 level, for all native language groups, preposition acquisition is the most difficult; all A1 learners tend to

[1]Corpora of beginners are in general infrequent (Tracy-Ventura et al., 2021) and corpora with truly longitudinal data are accordingly rare.

omit adverbs; there are agreement errors between nouns and adjectives, except for possessive adjectives (for 1st and 2nd person, singular), which are memorized as formulas (*mama mea* 'my mother', *profesorul meu* 'my father'); Arabic learners tend to omit the copulative verb most frequently; Romance speakers superfluously use the preposition *la* 'at', etc.

LECOR can also be used for language teachers training and for language testing (Callies and Götz, 2015). On the other hand, the corpus will have an immediate pedagogical use; it will be available for use in classrooms or by learners themselves, since this kind of data is relevant for the (error) producers.

LECOR can be used also for native language automatic identification[2], in forensic linguistics. Non-native speakers of Romanian make errors characteristic of learners with a specific mother tongue. Thus, the native language of a malicious individual can be discovered by mapping the type of errors made in his/her use of Romanian. This is an important means of identifying such individuals and it is very useful in the context of increasing social media threats.

The learners' errors identified in LECOR can be used also to improve the technology for automatic translation (McEnery et al., 2015).

LECOR can also be used for automatic grammar- and spell-checking and automated scoring of L2 written and oral performance (also Granger et al. (2015)).

2 Related work

In the last three decades, the construction of learning corpora has experienced a remarkable development, as evidence of their increasing importance, as can be seen in the list of about 200 corpora provided on the website of the Catholic University of Louvain[3] or in the CLARIN infrastructure[4].

The series of these corpora was opened at the time of the publication of the International Corpus of Learner English (Granger et al., 2009) and is by far dominated by the broad interest in learning English, but there are also corpora with written, audio or multimodal content for learning many other languages from different language families, such as Arabic, Czech, Finnish, French, German, Greek, Mandarin, Japanese, etc.

The Romance languages, of which the Romanian language is a part, are also well represented in this field, with written or spoken corpora, of which we mention the general ones, with native students of different languages and a unique target language: COPLE2 (Mendes et al., 2016) for Portuguese, the Spanish learner corpora (SLC) (Alonso-Ramos, 2016), CELI (Spina et al., 2022), LIPS (Gallina, 2017) or VALICO (Corino and Marello, 2017) for Italian, or the FLLOC platform (Marsden et al., 2002) or PAROLE (Hilton, 2009) for French. For the Romanian language, apart from small in-house bespoke corpora, there are only two printed corpora (Constantinescu and Stoica, 2020; Vasiu, 2020), which gather Romanian raw texts produced by foreign students. The corpus compiled by Constantinescu and Stoica (2020) comprises more than 450 samples (380 written samples / 65,000 words, and 79 oral transcriptions / 60,000 words); it was produced by 61 A1-B2 learners in the period 2004-2016 in various instructional contexts. Vasiu (2020) corpus contains transcriptions of oral samples produced by 172 A1 students at proficiency tests in 2014-2017; its size (70,000 words) is comparable to the oral part of the previous corpus. The digitalization and integration in our project of the two printed corpora would be a difficult endeavour, in terms of copy right issues and collaboration between independent working teams. Moreover, we had a different design in mind: large-scale corpus, richer metadata, longitudinal scope, internationally used annotation schemes, etc. In this perspective, our project comes to cover an important gap in this field.

For our project, we also benefited from the experience of building other corpora, such as the corpus

[2]For the identification of the native language (NLI), a large-sized corpus and a high-quality dataset (comparable to the International Corpus of Learner English (ICLE), used for NLI research, which comprises 6,085 essays written by speakers of 16 different L1s, see Jarvis and Paquot (2015)) are necessary, with a large number of native languages to enable comparison; uniform topics; comparable text sizes; thoroughly evaluated proficiency levels (Jarvis and Paquot, 2015). LECOR will be a medium-sized learner corpus, containing 4,000 samples, scalable (with the possibility to increase over time), covering L1 languages at least as diverse as those ones in ICLE.

[3]https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (accessed 01.08.2023).

[4]https://www.clarin.eu/resource-families/L2-corpora (accessed 01.08.2023).

for Czech, CzeSL (Rosen et al., 2020), for Latvian, LaVA (Darġis et al., 2022) or for Croatian, CroLTeC (Preradović et al., 2015) and many others. From these we took as a model the set of metadata, the text annotation with POS and error codes, the association of student texts with variants corrected by teachers or multi-level error annotation. Regarding the corpus query interface, although many of the existing corpora use the TEITOK interface (COPLE2, CzeSL, CroLTeC, etc.), we followed LaVA's example by adopting the NoSketch Engine interface, due to operating system version and configuration incompatibilities between our server and TEITOK.

3 Raw-Material Collection and Organization

The corpus is collected from foreign students coming for studies in Romania from Eastern countries (Far East, Near East, Middle East), South-eastern Europe, Latin America and, less often, Western and Central Europe; their native languages (Arabic, Chinese, Bulgarian, Albanian, Serbian, Turkish, Greek etc.) are therefore both typologically distant from Romanian and from each other. They are learning in mixed groups. The learners are generally high school graduates, but there are also masters and PhD students (aged between 18 and 25 years). The one-year program they are enrolled in is intensive, totalling 800 hours of classroom instruction. In the first part of the academic year, they have in their curriculum 28 hours per week of general course of Romanian, while in the second part, 30 hours per week (with the addition of languages for specific purposes). In general, students' interlanguage progress is documented from absolute beginner (A1) to intermediate level (B1 or B2).

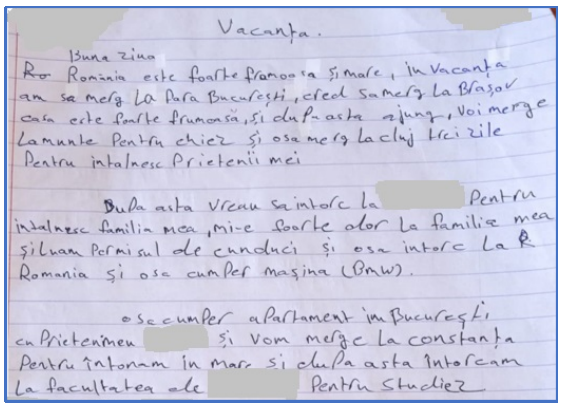
The raw material of the corpus (scans of the hand-written work samples, digital textwork samples, audio and video recordings) comes from different sources (it was collected by several teachers from the foreign students enrolled in the one-year intensive program "Preparatory year" at Faculty of Letters, University of Bucharest), in different folder and archiving structures (organized by student, by work sample or by teacher, archived or not archived) and in different file formats (.mp3/.mov/.mp4 for audio/video, .jpg/.png/.heic/.pdf for scans, Word/PDF for digital texts). Moreover, some scans cover more than one work sample and have to be split through different methods, according to their format: e.g., .pdf files are automatically page-split, .png/.jpg files are manually cropped in Paint. Then .heic/.png/.jpg files are converted to .pdf, for harmonization and because PDF format offers small size and quick loading with unsignificant loss in quality together with the possibility to concatenate photos depicting different pages of the same work sample. Video files are converted to audio files and the common chosen format was .mp3.

All scanned and audio/video files are manually transcribed following common orthographic transcription guidelines; some of the transcription principles are: (1) do not take into account strikethrough words in the scanned files or hesitations/repetitions in the audio/video files, but transcribe only the final version of a word provided by the student; (2) ignore syllabification of words at the end of the line, including the erroneous ones; (3) transcribe bracketed words; (4) keep the paragraph structure of the original but do not mark the original indenting; (5) for errors in lower/upper case spelling, only the ones concerning proper names or beginning of sentence are transcribed; other inappropriate uses of case spelling are ignored. In this manner, a transcribed .txt version of each work sample (further referred to as transcribed student form) was created, which is the basis for all preprocessing, correction, annotation and indexing steps in the corpus generation flow.

In order to protect the integrity of learners, we followed guidelines for research data management from University of Bucharest[5] and Catholic University of Leuven[6], which address various issues about anonymization, pseudo-anonymization, and encryption of sensitive data. All files are completely manually anonymized to protect student's right to privacy (see an example in Figure 1). This procedure is done after checking the transcripts, generally at the same time of samples correction. Despite the task not being performed automatically, due to the fact that not all samples contain personal data (e.g. only about 15% of audio files require anonymization), the total time required for the task is reasonable, e.g. 1-2 minutes to clear personal

[5] https://cometc.unibuc.ro/wp-content/uploads/2018/09/UB_Ghid-protectia-datelor_100918.pdf (accessed 01.08.2023).

[6] https://www.kuleuven.be/rdm/en/guidance/legal-ethical/anonymise-pseudonymise (accessed 01.08.2023).




```
Buna ziua
România este foarte fromoasa și mare, in
vacanța am sa merg La fara București,
cred sa merg La Brașov casa este foarte
frumoasă, si dupa asta ajung, voi merge
La munte Pentru chiez și osa merg La
cluj trei zile Pentru intalnesc
Prietenii mei

Dupa asta vreau sa intorc La Maroc
Pentru intalnesc familia mea, mi-e
foarte dor La familia mea si luam
permisul de cunduci si osa intorc La
Romania și osa cumper masina (BMW).

osa cumper apartament im București cu
Prietenmeu Omar si vom merge La
constanța Pentru întonam in mare si dupa
asta întorceam La facultatea de Litere
Pentru studiez
```

Figure 1: A handwritten learner text with anonymizations and its transcript version with pseudonymizations. The sample belongs to an A2 learner (male) with Arabic as mother tongue; the text topic is summer holidays.

information from an audio sample. "Coding" sensitive data is relatively easy.

The sensitive data in scanned images is covered in Paint (for PDF sources they are converted to Paint, anonymized and converted back to PDF), audio sensitive data are replaced with beep sounds in Audacity, while the personal information in text documents is pseudonymized, i.e. replaced with similar plausible data (e.g. "Mohammed" is replaced with "Ahmed", "35 years old" is replaced with "29 years old", etc.) to maintain the morpho-syntactic coherence of the sentence (full anonymization, e.g replacing "35 years old" with "xxx" will impact negatively on the POS-tagging performance in future steps). The sensitive data we are targeting in the anonymization/pseudonymization process concerned student's name, age, birth date and birth place, previous school/university/-work place, etc. (our internal list is broadly similar to that in Megyesi et al. (2018)). In case the pseudonymized word have morphological features like gender/number/case, they have to be replaced with similar values: e.g., feminine, genitive etc. For more challenging cases, where covering up sensitive data would affect the overall message of the text (e.g., replacing the name of a town with another would make its description inappropriate), we decided to keep the original place name.

Next step is producing the corrected form for each sample, further called teacher form. The correction is made by linguists/Romanian language teachers on the transcribed student sample by using commonly agreed general principles: e.g., 1. minimal corrections (if possible, we do not change the part of speech and the words order or number), 2. we do not provide more correction alternatives, but only one correct option; 3. semantic accuracy is preserved: e.g. if the work sample is made based on an image and the image depicts a "red skirt", the "yellow skirt" syntagm provided by the student is considered an error and corrected. A distinction is made between actual errors affecting form, grammar, vocabulary, punctuation, etc., and infelicitous constructions, register and stylistic inaccuracies and other awkward language (for a similar approach, see Granger et al. (2022)). The category of infelicities includes, for example: informal language, such as the short demonstrative forms *asta, ăsta* ('this (one)') instead of the standard (long) ones *această/aceasta, acest/acesta* ('this (one)'), the shortened forms of numerals (*treișpe*, instead of *treisprezece* 'thirteen'), address formulae with an inappropriate degree of politeness, text sequences with unclear meaning, etc. At the moment, it has not been decided how exactly infelicities will be annotated, but the annotation will definitely be done manually.

This preprocessed material is than organized in two steps:

1. According to the student: each student has a unique ID and a metadata file containing information associated to that specific student; the name format of a student folder is *StudentID_student* and the metadata file is in the .tsv format (see examples in Table 1).

2. Inside the student folder, the files are organized in folders dedicated to different work samples: work samples also have unique IDs (unique in a list of work samples of a specific student) and

Folder	Folders and Files
1_student	1_student.tsv
	1_1_text
	1_2_text
	1_3_text
2_student	...
1_1_text	1_1_text.tsv
	1_1_t.txt
	1_1_s.txt
	1_1_o.pdf

Table 1: Example of the corpus folder structure. File names are in italics and folder names are in normal text.

associated text metadata files containing information corresponding to that specific work sample. The name format of a work sample folder is *StudentID_WorkID_text* (see examples in Table 1) and the folder has the following structure: original student form file + transcription of the original form file + teacher corrected form file + metadata file. For the file name convention, *o* stands for original form, *s* stands for student form, and *t* stands for teacher form. The original student form file (name format: *StudentID_WorkID_o*) can be: (a) a scanned work sample in the .pdf format; (b) an audio work sample in the .mp3 format; or (c) a student digital text work sample in .docx or .pdf format. The transcription (name format: *StudentID_WorkID_s*) and the teacher corrected transcription (*StudentID_WorkID_t*) are text files. The corresponding metadata is a .tsv file (see examples).

Metadata are collected in shared online Excel files (one for students and another for work sample). The StudentID field connects work sample metadata entries with student metadata entries. Important student/learner metadata fields specify gender, age, region for learning Romanian, native language(s), (bi/tri)linguality information, languages studied in parallel with Romanian, motivation for studying Romanian, degree of motivation, frequency of interaction with Romanian native speakers, mode of study, etc. Important work sample fields refer to spontaneity, time/length limits or requirements, writing type (hand-written or digital), use of diacritics, level of proficiency of the student, etc. Some of this metadata fields will be indexed and used at searching, while others will only be displayed in the search results. Scripts were designed to automatically extract metadata from the shared files and distribute them in the proper folders in .tsv format.

4 Annotation Procedure

Once the source files are organized in the manner presented above, the annotated corpus is created based on a procedure that includes the following two stages:

1. morphosyntactic annotation (POS-tagging) of the student version and the teacher one;

2. comparing student–teacher texts, which involves the alignment of the two versions and the automatic annotation of errors/differences.

This procedure is semi-automated, requiring the corpus files to be passed through the external POS annotation platform (located on a server other than the corpus server), then the annotated files are uploaded to the LECOR server for automatic error annotation. To make working with a large volume of data more efficient, scripts were created to detect and process only files added to the LECOR corpus or modified after the last annotation.

4.1 POS-tagging

Both the student and teacher forms of the work samples were annotated automatically in the RELATE platform (Păiș et al., 2020), dedicated to processing Romanian language. For this purpose, an export script was devised to transfer the documents to the platform. Following the annotation, an import script was used to transform the annotated documents into the LECOR specific format. From the multiple text processing pipelines available in RELATE (Păiș et al., 2019; Păiș, 2020), for the purpose of the LECOR project, we used UDPipe (Straka et al., 2016) with a recent model (Păiș et al., 2021) trained on the Romanian RRT corpus version 2.7 (Barbu Mititelu et al., 2016).[7]

The resulting documents, in CoNLL-U Plus format, included the following: segmentation (sentence and token), lemma, part-of-speech (UPOS and MSD tags), see Figure 2 for the student version (on the first five columns) and the teacher version (on the following five columns).

The CoNLL-U Plus format allows for additional annotation levels to be included in the future, if

[7]The tagger performance on a general corpus (the test subcorpus of RRT) was evaluated at: 99.88 F1 for token segmentation; 97.39 F1 for sentence segmentation; 95.91 accuracy for lemmatization; 97.15 UPOS accuracy for POS tagging. Further evaluation of the tool on LECOR corpus remains to be done at the end of the project; we expect important decrease in the tagger performance, given the specificities of a learner corpus.

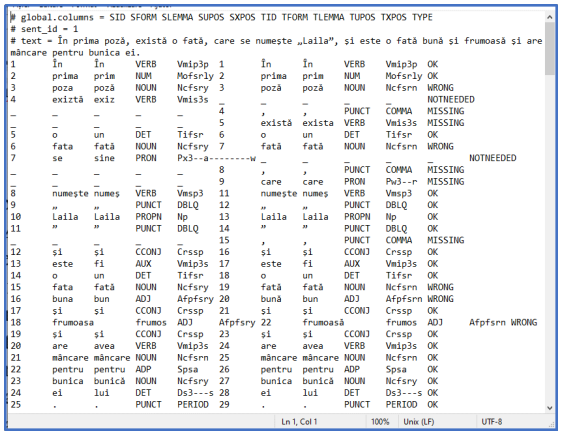


Figure 2: Differences between student and teacher variants.

needed.

4.2 Comparing Student–Teacher Texts

The student and teacher versions of the work samples were aligned at token level. First, an automatic process was used, employing a modified Dynamic Time Warping (DTW) algorithm. This allowed matching partial words or words with mistakes and marking such issues. Considering two tokens T1 and T2 (each with the attributes form and lemma), the matching formula is:

```
Match(T1,T2)=(T1.form==T2.form || T1.
lemma==T2.lemma || removeDia(T1.form)==
removeDia(T2.form) || removeDia(T1.lemma
)==removeDia(T2.lemma) || lev(removeDia(
T1.form), removeDia(T2.form))<2)
```

In this equation, removeDia is a function that removes Romanian diacritics, lev is the Levenshtein edit distance. Parts of this equation may seem redundant (such as comparing both form and lemma). However, due to possible mistakes in the student work samples, the lemmatization process may produce different results. For example, the Romanian word *copii* may have either the lemma *copil* ("child") or the lemma *copie* ("copy"/"duplicate"), depending on the context. Similarly, words with different forms may yield the same lemma, for example in the case of wrong singular/plural form. Furthermore, the equation was devised without having in mind a particular lemmatization algorithm.

Following the automatic process, a manual process was needed to confirm the differences between the teacher and student forms. The result is a CoNLL-U Plus file containing 5 columns for each of the student and teacher versions (token id, word form, lemma, UPOS, MSD) and an 11th column

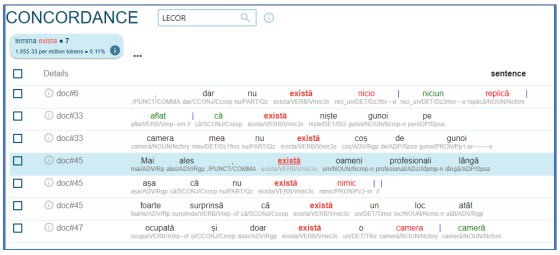


Figure 3: NoSketch Engine interface for LECOR.

with the error type as labelled by the aligning algorithm (no error, missing word, additional word that is not needed, spelling mistake), see Figure 2.

4.3 Corpus Query Interface

The LECOR corpus will employ the NoSketch Engine (Rychlỳ, 2007; Kilgarriff et al., 2014) open-source corpus query platform to allow searching access. The primary content indexed in the platform is represented by the differences file with error annotations, as described in the previous subsection. In addition, metadata about the student and the work sample will be indexed in order to allow querying sub-corpora based on different criteria. For this purpose, a dedicated script was created to convert from the CoNLL-U Plus file to the "vertical" file format used by NoSketch Engine, with the additional metadata inserted into specific file structures. A small sample of the LECOR corpus is currently available online in the NoSketch Engine installation[8]. The interface allows for both simple querying (based on words or lemmas) or complex CQL based queries. Sollutions for accessing the original annonymized scanned or audio work sample from the NoSketch Engine interface, by clicking on a query result, will be explored.

Figure 3 shows the search result for the lemma *domn* ("mister, sir") with the option to provide the lemma and MSD for KWIC only. At the bottom of the figure, the full text of the first line of concordances has been opened, containing the student's text along with the related corrections. The red words belong to the student and the green ones are their corrected forms. This mixing of student-teacher versions creates the problem of getting matches for wrong and corrected forms indiscriminately. For example, line 2 in Figure 3 matches the wrong form, and line 3 matches the corrected form. This problem needs to be corrected.

[8] http://lecor.unibuc.ro/crystal/#dashboard?corpname=lecor (accessed 01.08.2023).

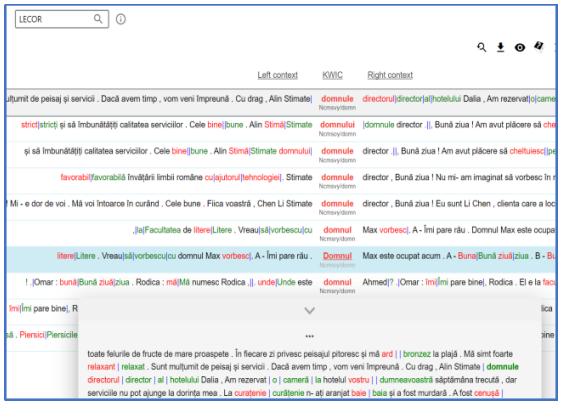


Figure 4: Access to word annotation.

In Figure 3 we have selected for display only the MSD and lemma for KWIC, but the interface has access to all the information associated with each word, represented in Figure 4, except the last column. The information in the last column concerns the type of error committed by the student and will be subject to further processing including manual annotation by the teacher and searches by error type.

As seen in Figure 4, under each word in the text, the information about the lemma as well as the part of speech (UPOS) and the morpho-syntactic descriptions (MSD) are available. For example, for the first word Ro. *dar* "but", the annotated information is *dar*/CCOMJ/Ccssp where the lemma is *dar*, UPOS is CCONJ (coordinating conjunction) and MSD is Ccssp (see MULTEXT-EAST specifications[9]).

5 Conclusions and Further Work

In this study we presented the design stage of the LECOR corpus emphasizing the conceptual framework for building and utilization of the corpus. The next phase involves the quantitative accumulation of primary material.

Regarding the text correction, we have already established the general criteria and developed a proofreading manual, and further we will validate these criteria by correcting a representative volume of texts in parallel and establishing the Inter-Annotator Agreement.

Error annotation in this phase is done at a basic level and is strongly correlated with automatically detected differences between the student/teacher versions. This phase is very useful for what we intend to do, which is a manual error annotation, on multiple levels, as is already practiced in the field. The inventory of errors is already established, it remains to build the technical annotation method, especially for errors whose correction involves changes in the word order.

The NoSketch Engine query interface will be explored further to see how well it can be adapted to the project's goals. For example, a distinction must be made between the student version and the teacher version, possibly with separate searches on each version. We will investigate whether the option of querying parallel corpora provided by NoSketch Engine solves this desideratum. Another issue concerns the use of metadata about students and work samples. We consider using the platform's option to create subcorpora, which can potentially be selected by certain metadata values.

[9] https://www.sketchengine.eu/romanian-tagset/ (accessed 01.08.2023).

Acknowledgments

This work is supported by 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".

References

Margarita Alonso-Ramos. 2016. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Benjamins, Amsterdam.

Verginica Barbu Mititelu, Radu Ion, Radu Simionescu, Elena Irimia, and Cenel-Augusto Perez. 2016. The romanian treebank annotated according to universal dependencies. In *Proceedings of The Tenth International Conference on Natural Language Processing (HrTAL2016)*.

Guy Brook-Hart. 2009. *Learning from common mistakes*. Cambridge University Press, Cambridge.

Marcus Callies and Sandra Götz. 2015. *Learner Corpora in Language Testing and Assessment*. John Benjamins Publishing Company, Amsterdam.

Mihaela-Viorica Constantinescu and Gabriela Stoica. 2020. *Româna ca limbă străină: Corpus*. Editura Universității din București, București.

Elisa Corino and Carla Marello. 2017. *Italiano di stranieri: I corpora VALICO e VINCA*. Guerra, Perugia.

Roberts Darģis, Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova, and Kristīne Pokratniece. 2022. LaVA – Latvian language learner corpus. In *Proceedings of the Thirteenth Language Resources and*

Evaluation Conference, pages 727–731, Marseille, France. European Language Resources Association.

Ana Díaz-Negrillo and Paul Thompson. 2013. *Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson (eds). Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Francesca Gallina. 2017. *Anna Gudmundson, Laura Alvarez Lopez, Camilla Bardel (eds.), Romance languages. Multilingualism and Language Acquisition*. Peter Lang, Frankfurt am Mein.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, Cambridge.

Sylviane Granger, Helen Swallow, and Jennifer Thewissen. 2022. The louvain error tagging manual. version 2.0.

Heather Hilton. 2009. Annotation and analyses of temporal aspects of spoken fluency. *Calico Journal*, 26(3):644–661.

Scott Jarvis and Magali Paquot. 2015. Native language identification. *Cambridge handbook of learner corpus research*, pages 605–628.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Emma Marsden, Florence Myles, Sarah Rule, and Rosamond Mitchell. 2002. Oral french interlanguage corpora: Tools for data management and analysis. occasional paper.

Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Routledge, London/New York.

Tony McEnery, Richard Xiao, and Yukio Tono. 2015. *The Cambridge Handbook of Learner Corpus Research*, chapter Learner corpora and natural language processing. Cambridge University Press, Cambridge.

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of gdpr: Insights from the creation of a learner corpus of swedish. In *7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018), Stockholm, Sweden, 7th November, 2018*, pages 47–56. Linköping University Electronic Press.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3207–3214, Portorož, Slovenia. European Language Resources Association (ELRA).

Rosamond Mitchell. 2021. *The Routledge Handbook of Second Language Acquisition and Corpora*, chapter Corpora and Instructed Second Language Acquisition. Routledge, Taylor & Francis, London/New York.

Vasile Păiș. 2020. Multiple annotation pipelines inside the relate platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.

Vasile Păiș, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufiș. 2021. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.

Vasile Păiș, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France. European Language Resources Association.

Vasile Păiș, Dan Tufiș, and Radu Ion. 2019. Integration of romanian nlp tools into the relate platform. In *International Conference on Linguistic Resources and Tools for Natural Language Processing*.

Nives M. Preradović, Monika Berać, and Damir Boras. 2015. *Cergol Kovačević and Sanda Lucia Udier (eds.) Multidisciplinary Approaches to Multilingualism*, chapter Learner Corpus of Croatian as a Second and Foreign Language. Peter Lang.

Alexandr Rosen, Jiří Hana, Barbora Hladká, Tomáš Jelínek, Svatava Škodová, and Barbora Štindlová. 2020. *Compiling and annotating a learner corpus for a morphologically rich language. CzeSL, a corpus of non-native Czech*. Karolinum Press, Charles University.

Michael Rundell. 2007. *Macmillan English dictionary advanced learner*. MacMillan.

Pavel Rychlỳ. 2007. Manatee/bonito-a modular corpus manager. In *RASLAN*, pages 65–70.

Diane Schmitt and Norbert Schmitt. 2011. *Focus on Vocabulary 2: Mastering the Academic Word List*. Pearson Education, White Plains, NY.

Stefania Spina, Irene Fioravanti, Luciana Forti, Valentino Santucci, Angela Scerra, and Fabio Zanda. 2022. Il corpus celi: una nuova risorsa per studiare l'acquisizione dell'italiano l2. *Italiano LinguaDue*, 14(1):116–138.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

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