Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 57–69.
Terra Linguistica. 2023, Том 14, № 1. С. 57–69.

Review article
UDC 81'32
DOI: https://doi.org/10.18721/JHSS.14106

LEARNER CORPORA: RELEVANT INFORMATION AND AN OVERVIEW OF THE EXISTING FRAMEWORKS

M.V. Khokhlova ✉

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ m.khokhlova@spbu.ru

Abstract. In the modern world, there is a constant interest in foreign languages. Therefore, the question of learning about the language used by non-native speakers of a certain language, as well as describing their mistakes is a highly relevant matter. Learner corpora differ not only according to the languages they focus on, but also in relation to a number of their properties. The purpose of the study is to present a review the learner corpora available for different languages, as well as to compare the approaches that exist for their annotation. The paper considers the origins of learner corpus research, focuses on the main the stages of a project, types of learner corpora (which may differ in their tasks, students' mother tongue, language proficiency, text genre, data type, etc.), linguistic and metatextual information that accompany texts and provides a classification of errors. The paper gives a brief overview of annotation tools and corpus platforms that can be used for building a learner corpus.

Keywords: learner corpora, typology, errors, annotation, second language acquisition.

Acknowledgements: The study was carried out with the financial support of St. Petersburg State University (project No. 92563238).

Citation: M.V. Khokhlova, Learner corpora: relevant information and an overview of the existing frameworks, Terra Linguistica, 14 (1) (2023) 57–69. DOI: 10.18721/JHSS.14106

Обзорная статья
DOI: https://doi.org/10.18721/JHSS.14106

КОРПУСА УЧЕБНЫХ ТЕКСТОВ: ДАННЫЕ И ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

М.В. Хохлова ✉

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ m.khokhlova@spbu.ru

Аннотация. В современном мире не угасает интерес к иностранным языкам. Поэтому вопрос их изучения в качестве неродного, а также описание ошибок, которые допускают обучающиеся, не теряет своей актуальности. Учебные корпуса различаются не только в зависимости от языкового материала, но и по ряду своих характеристик. Целью статьи является обзор корпусов учебных текстов разных языков, а также сравнение подходов, которые существуют для их разметки (прежде всего, метатекстовой). В работе рассматриваются основные этапы разработки проектов, типы учебных корпусов (которые могут отличаться по своим задачам, по родному языку студентов, уровню владения языком, жанру текстов, типу данных и т.д.), лингвистическая и метатекстовая информация, которая сопровождает тексты, а также приводится классификация ошибок. В статье дается краткий обзор инструментов для разметки и платформ, которые можно использовать для создания учебного корпуса.

Ключевые слова: корпусы учебных текстов, типология, ошибки, разметка, усвоение второго языка.

Финансирование: Исследование выполнено при финансовой поддержке Санкт-Петербургского государственного университета (проект № 92563238).

Для цитирования: Хохлова М.В. Корпуса учебных текстов: данные и обзор существующих подходов // Terra Linguistica. 2023. Т. 14. № 1. С. 57–69. DOI: 10.18721/JHSS.14106

Introduction

Now in the 21st century, we can observe the processes of population migration. Open borders allow people to travel, study and work in different countries. The number of non-native speakers living in various countries has increased significantly over the past few years. The language of a non-native speaker has specific unusual characteristics both in vocabulary and grammar, and hence such a system deserves to be studied. The target audience of such corpora can be not only teachers or students, but also linguists who analyze second language acquisition through corpus data. This can help to create specialized tutorials, develop methods, and also to describe mechanisms of error production. Empirical linguistic evidence from learner corpora is hence the most valuable source of examples and can contribute to the understanding of the processes emerging during foreign or second language acquisition.

Overview

The origins of learner research projects can be traced back to the 1980s when computer technologies facilitated the processes of storage, retrieval and processing of large amounts of texts. This resulted in the launch of electronic collections of written and spoken linguistic data that represented the language of foreign and second language students. The work by Rosen et al. [1] proved to be one of the most profound and up-to-date monographs that are totally focused on learner corpora. Let us turn to some projects that focus on texts produced by learners. An excellent review of the projects can be found on the website of CLARIN initiative [2].

English was the first target language to have a learner corpus. The project entitled the *International Corpus of Learner English* (ICLE) [3] is often referred to as the first learner corpus that was based upon the principles of corpus linguistics [4–5]. Granger emphasizes that "the release of a learner corpus such as the ICLE marks the beginning of a new stage in the evolution of learner corpus research" [6, p. 544]. However, there were other projects focused on the analysis of learner language that foreshadowed ICLE in 1980s and 1990s. ICLE was a large-scale corpus with data collected from respondents with various L1 backgrounds. The initial corpus was about 2 mln words, while the second version was 3.7 mln words. It comprised essays written by advanced learners of English, which were university students. This corpus had a tremendous influence and launched the development of similar projects, namely, the LOCNESS (Louvain Corpus of Native English Essays), LINDSEI (Louvain International Database of Spoken English Interlanguage) LOCNEC (Louvain Corpus of Native English Conversation).

Publishing houses and testing organizations are also interested in such resources and build their own corpora. Cambridge University Press built the *Cambridge English Corpus* (CEC) [7] that includes two parts: the first subcorpus (1.8 bln words) compiles texts by native speakers (British and American), while the second subcorpus (55 mln words) focuses on how non-native speakers use English. The latter is also known as the Cambridge Learner Corpus (CLC) and comprises written data produced by more than 200,000 L2 English learners from 173 countries in such language exams as Cambridge English (all levels), CELS, IELTS and others. The corpus data is error-annotated which makes it possible to compute frequencies of different types of errors, to see the contextualized usage of the word and possible mistakes, to see difficult cases, and to compile dictionary entries and other material for language learners. Longman collected several corpora combined in the Longman Corpus Network. The Longman Learner Corpus (10 mln words), being its part, was compiled for the production of the Longman Active Study Dictionary.

Speaking about Russian learner corpora, we should name the *Russian Learner Corpus of Academic Writing* (RULEC) that was the earliest project in this direction for Russian. Its detailed description is made in [8–9]. Nowadays it is a subcorpus within the *Russian Learner Corpus* [10] that was created at the HSE Laboratory for Corpus Research (Moscow). The corpus represents the so-called "non-standard Russian". It contains samples of oral and written speech of two groups of Russian speakers: those who study Russian as a foreign language and heritage speakers. The latter includes people for whom Russian is not the main language, while they began to learn it as their first language in childhood (for example, emigrants). Along with lexical and grammatical features, error annotation is also available in the corpus. It includes spelling, morphological, syntactic, and lexical errors, as well as errors in constructions. Search results contain not only original versions but also the corrected ones. Metatextual features indicate the author's dominant language (American English, German, French, Italian, Norwegian, Dutch, Finnish, Swedish, etc.) and the level of Russian language proficiency according to the CEFR and ACTFL scales. When searching, one can specify a subcorpus taking into account the required characteristics. The corpus can be used to study the assimilation of the Russian language and in the teaching of Russian as a foreign language.

Typology

Nowadays, there are 190 learner corpora registered by the Centre for English Corpus Linguistics at Catholic University of Louvain [11]. Beyond these resources, there are also various own learner corpora, which were developed by universities or research groups. All these corpora can differ according to their properties and their volume. The issue of corpus volume is extremely important (see, for example, the discussion about the dependency of collocations on corpus volume [12] while a small amount of data does not provide sufficient evidence for frequencies and hence hinders the use of statistical tests. Nevertheless, it is difficult to say if a corpus is big or small because there is no agreement about how much data is enough. The only thing that matters, in this case, is the quality of a corpus. If it is poor, the size of a corpus will have no importance. Moreover, if a corpus was collected according to perfect and strict design criteria, even a small corpus will be of great value.

In second language acquisition (SLA), one can distinguish between second and foreign language. The former means that the language is acquired in its natural environment, for example, English as a Second Language (ESL) implies learning English in English-speaking countries (such as the United Kingdom). Foreign language acquisition (FLA) deals with studying languages in a context where it is not generally spoken. If we apply this paradigm to learner corpora, hence there can be corpora focusing on SLA and FLA tasks. However, some authors use these terms as synonyms to describe learning a second (nonnative) language.

Adopting classification schemes used by Tono [13], Granger [14–15], Rosen et al. [1], we can define learner corpora thoroughly by the following characteristics:

1. Language-related criteria

Medium. Learner corpora can cover written or oral data. The former is the dominant source, while spoken learner corpora are still rare. Nowadays, ambitious projects contain audio and video fragments known as multimedia (multimodal) learner corpora. Among them, we can name the Multimedia Adult ESL Learner Corpus [16], the PAROLE corpus [17], and the TAITO corpus that contains videos of partially transcribed discussions [18]. Granger [14] mentions the Telekorp project [18], "which results from five years of computer-mediated communication between learners of German in the US and learners of English in Germany" [14, p. 261].

Genre. Learner corpora tend to represent one kind of texts (mostly, essays). It is time and labor-consuming to collect many genres. Nevertheless, the existing projects try to overcome this drawback. For example, the BELC (Barcelona English Language Corpus) contains speech recordings across four tasks (written composition, oral narrative, oral interview, and role play) [20]. The ICLFI (International Corpus of Learner Finnish) comprises both non-fictional (e.g., essays, argumentative texts) and fictional texts (e.g., narratives, letters) [21–22].

2. Task-related criteria

Time of collection. Texts can be collected ad hoc and only once or over a period of time and thus, we can speak about cross-sectional vs. longitudinal data. Cross-sectional corpora represent data from different types of learners at a single point in time, while longitudinal corpora focus on the same learners during certain time periods. A particular mixture between them is quasi-longitudinal corpora that represent data collected simultaneously from learners with different proficiency levels. The *Longitudinal Database of Learner English* focuses on collecting longitudinal learner data from the same students over several years [23].

Task. Here we can differentiate between spontaneous and prepared texts, i.e., the ones generated in the classroom and those written at home. The use of references can also be limited.

Pedagogical use. The majority of learner corpora are corpora for delayed pedagogical use. It means that they are built on texts from a given sample of students and then will be processed and used for other (next) groups of learners and not for the ones who produce the texts. The opposite example is corpora for immediate pedagogical use (the same students can benefit from their "own" corpora).

3. Learner-related criteria

First language (L1, mother tongue). Learner corpora can contain data from learners with the same mother tongue or with several different L1 backgrounds. The ESF (European Science Foundation Second Language) Database comprises data collected by research groups from the Netherlands, Great Britain, France, Germany and Sweden [24–25]. The target languages are Dutch, English, French, German and Swedish. For each target language, two source languages were selected: Punjabi and Italian for English, Italian and Turkish for German, Turkish and Arabic for Dutch, Arabic and Spanish for French, Spanish and Finnish for Swedish. SweLL (Swedish Learner Language Corpus) presents data collected from learners who speak 64 languages from different language families [26]. The ten most frequent mother tongue backgrounds are English, Persian, German, Chinese, Russian, Arabic, Spanish, Thai, Somali and Vietnamese. The ICLE covers 11 different native languages.

Target language (L2). Usually, learner corpora focus on one language. The majority of them deal with English, but there are also corpora for other languages such as German, French, Dutch, Czech etc. The bilingual part of the CHILDES database represents data collected from children learning two or more languages [27]. The Multilingual Learner Corpus represents data from speakers of Brazilian Portuguese who learn different languages (English, German and Spanish) [28].

Proficiency in the target language. According to this feature, we can distinguish between corpora with texts collected from students at the same level of language knowledge and those with texts from speakers at various levels.

4. Data-related criteria

Owner. Collection of data for corpora can be initiated by publishing houses, companies or universities. Hence we can speak about commercial (the Longman Learners' Corpus or the Cambridge Learner Corpus) or academic corpora (the Louvain International Database of Spoken English Interlanguage (LINDSEI)). The former "tend to be much larger and have a wider range of mother tongue backgrounds" [14, p. 260].

Accessibility. This distinction is related to the one mentioned above. Depending on their funding, corpora can be freely accessible online or aimed for limited (in-house or commercial) use.

Annotation. As a rule, corpora should be annotated (at least POS-tagging), but some learner corpora represent only raw data without added linguistic information.

Levels of learner corpus design

Text collection is the first step in building corpora. The collection of texts can be organized in different ways, and it reflects the specifics of learner corpora and differs from the procedure for standard corpora. Many written learner corpora deal with one genre, e.g., essays produced by students, as it is relatively easy to collect them after exams or exercises. The compilers of a corpus should elaborate particular guidelines that include instructions for text collectors, the choice of text topics and their types, metadata description, and consent for learners about the use of texts. The next question that should be addressed with attention is the initial form of texts, i.e. whether they are electronic or written on paper, or where they are produced, i.e., at home or in a class. This is not as straightforward issue as it may seem. Electronic texts can be influenced by spell-checkers and texts written at home can have fewer errors than the ones produced in a class. The processing of hand-written texts raises the question of their OCR recognition or transcription and is time-consuming. This step usually involves an orthographic form but in case of spoken corpora texts can be supplied with phonemic or phonetic transcriptions.

Once the text is preprocessed and converted into an appropriate format, it can be annotated. The next step of corpus building deals with adding special data to texts, i.e., their mark-up. Any corpus usually has an annotation, which can be either textual or linguistic. In the case of learner corpora, their building process resembles one for standard corpora, but there are some peculiarities. Learner language abounds with errors and hence differs from the standard language, so it is necessary to take into account these discrepancies. Accordingly, learner corpora require a special type of markup, i.e., error annotation. Next, we will look at some of the principles that underlie various types of mark-up.

The first type of annotation implies specifying textual characteristics that describe texts. Standard textual annotation identifies sentences, paragraphs, sections, headings, and other features dealing with the structure of a document. It can also be helpful for learner corpora. In terms of learner corpora, one can pay attention to such unusual features as corrections (insertions or deletions) made by learners in handwritten texts. The next step involves grammatical annotation, which results in tokenization, lemmatization, POS-tagging, and determining other grammatical features. The most elaborate strategy implies syntactic parsing, semantic or discourse annotation. However, learner corpora need another type of specific annotation, namely, error annotation, that can be performed in most cases manually. This layer of annotation deals with the detection of errors, their description (categorization) and correction.

Metatextual annotation

Metatextual annotation is highly important for second language analysis as it helps to build subcorpora based on relevant features and hence to investigate linguistic phenomena inherent to the students of a particular proficiency level, age, education or social class. Metadata deal with texts as a whole and imply information about texts themselves (title, year of publication, medium, register etc) or their authors (in this case it describes sociolinguistic features), being one of the most important types of annotation in case of a learner corpus [13–15]. The usefulness of a learner corpus depends on this kind of annotation as non-properly described data fail to contribute to confirmation or rejection of linguistic hypotheses. Metatextual markup enables a user to define and select subcorpora, i.e. find those texts that meet the specified parameters.

There are different approaches to textual annotation that can focus on describing authors or texts. Granger rightly points out that "extra care has to be taken in collecting the data for learner corpora given the large number of variables affecting the learning/acquisition process" [6, p. 538]. We can proceed from the idea that the following positions can be reflected in metadata that describes the authors of texts for learner corpora: 1) the learner's first language; 2) education; 3) gender; 4) age; 5) other languages; 6) the duration of language learning. Below we will dwell on a number of projects (this list was inspired by [1]) and metadata description used by them taking into account that they are a few of many.

ICLE, on the one hand, follows the design criteria introduced by [4] and, on the other hand, tries to describe characteristics of learner texts. Granger [14] distinguishes between learner variables which concern a student and task variables that characterize the language situation. In its turn, each type can be described in terms of general variables (can apply to any corpus) and L2-specific variables. General learner features are age, gender, region and mother tongue, while L2-specific are learning context, proficiency level, amount of L2 exposure and knowledge of other foreign languages. General task variables are represented by medium, field, genre (text type), whereas task type (activities that learners are involved in: conversation, role-play, interview, essays etc.) and conditions belong to L2-specific task characteristics that can influence leaner's text generation (time limit, topic, mother tongue of interviewer etc.). The list of features used in ICLE can describe a text quite exhaustively, nevertheless the authors name one variable that plays a crucial role for learner corpora but is difficult to be recorded, that is "the teaching methodology and pedagogical materials to which the learners have been exposed" [3, p. 4].

CzeSL corpora have 15 items about the author of the text and 15 items about the text itself [1, p. 54]. While building a learner corpus for Russian, the authors [8] described 8 metadata items that were grouped into two categories, namely, author- and text-related features. The former included six subcategories, for example, language background (L2 learner or heritage speaker), dominant language (American English, German, Korean, etc.), and proficiency level according to CEFR (ACTFL). The latter contains three subcategories: mode (written or oral), genre (answers, essays, blogs, letters, stories, descriptions, etc.), and time limit (limited or unlimited).

The whole range of codes for describing texts implies the following positions [1, p. 56]:

- text id;
- date of the text collection;
- medium of the text (manuscript or electronic);
- time limit in minutes;
- permitted resources (yes or none, dictionary, textbook, other);
- part of exam (yes or n/a, interim, final);
- size limit in words;
- title of the essay;
- type of the topic (general or specific);
- activity before writing the text (exercise, discussion, visual, vocabulary, other or none);
- assigned topic (multiple choice, specified, free, or other);

- assigned genre (free or specified);
- predominant genre in the text (informative, descriptive, argumentative, or narrative);
- text length in words;
- range of text length in words.

RULEC/ RLC uses the following metadata categories: name (pseudonym), gender, language background and language experience of the student (either for L2 or HL), language proficiency level, time stamp (week and academic year when the paper was written), time limit under which the paper was written (timed or non-timed), text type (one paragraph or a long research paper), text function (e.g. narration, argumentation), and indication if the paper was written individually or in a group.

The author metadata used in CLC includes the following characteristics: age, gender, first language, nationality, exam, CEFR and ALTE levels, year, educational level, and years of English study. CLC focuses on language exams and hence here we find a range of features describing textual characteristics (exam level, date, format, style, register). It pays much attention to exam scripts and metadata and includes the binary feature of whether the exam was passed. Annotation differentiates between CEFR level student performance and CEFR level exam.

Next, we also examined four other corpora, an overview of which is given in [1], namely: the ASK corpus of Norwegian [29], "Learner corpus for Portuguese" (*COPLE2*) [30], "Croatian Learner Text Corpus (*CroLTec*) [31] and "Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache" (*Falko*) [32].

Table 1 shows the combined metafeatures that are used for describing authors in different corpora (learner-related criteria). Some metafeatures are obvious and easy to collect (such as gender or age that can be given in the questionnaire), while others are not as clear. For example, proficiency level should be additionally estimated by data collectors or teachers via language tests or the same length for the duration of study can lead to different language knowledge (due to study breaks). As we can see, with regard to meta-features, there is a core that is the same for all the corpora we have considered. Something may be different or may be implemented in a slightly different way. For instance, the "Nationality" field in some corpora can be denoted as "Country of origin" in others. Falko has the largest list of metafeatures dealing with mother tongues. The same holds true for ICLE: it indicates up to three foreign languages and the same number of mother tongues. Some corpora evaluate period spent in the country of the studied language in years, whereas ICLE counts it in months.

Some characteristics indicated as metafeatures and presented in the table serve as bases that can be chosen for dividing corpora according to their types (see section on learner corpora typology).

Error annotation

Error annotation constitutes an important part for a learner corpus and is far from straightforward as it can be difficult to make a guess about the author's intention. This stage implies the following three steps: error detection, classification and correction. Error classification usually involves a taxonomy, which predefines certain types of errors (e.g., lexical, morphological, syntactic ones). An error can be corrected on the next step and it means that the annotation will hence mark two variants for the given item (erroneous and correct ones). So we can speak about two levels of error annotation, the first one implies labeling according error categories, while the second type deals with error correction.

The question of whether it is possible to be sure that the error has been correctly detected and (if necessary) corrected deserves special attention. On the one hand, there are obvious cases, on the other hand, some examples are ambiguous and can suggest several possible correct versions or the annotator even fail to understand the author's intention.

Errors may vary depending on the level of language proficiency and include a wide range from orthographic and to stylistic ones. Rosen et al. [1] propose an elaborate error annotation scheme including two levels. The first one uses two-tier approach and allows the annotators to make their corrections without classifying the errors. The second one is based on standard linguistic categorization. The error tagset

Table 1. Metadata about authors

	id/name	sex	age	age category	first language	first language group	other languages	ACTFL	ALTE	CEF	language in the family	years in the country	program of study	duration of study	hours of study per week	textbooks	bilingual	language background	stay in the country	origin / nationality	education	Exam
CzeSL[1]	+	+	+	+	+	+	+	–		+	+	+	+	+	+	+	+	*				
RULEC /RLC	+	+			+			+		+			+					+				
ASK	+	+	+		+							+			+				+	+	+	
CLC		+	+		+				+	+				+							+	+
COPLE2	+	+	+		+		+			+	+	+		+						+	+	
CroLTec		+	+		+		+			+	+	+	+				+			+		
Falko	+	+	+		+		+					+	+	+					+		+	+
ICLE		+	+		+		+						+	+						+		

[1] CzeSL are aimed only at L2 learners.

includes: 1) incorrect words; 2) foreign words; 3) tags for incorrect inflections; 4) wrong word boundaries; 5) stylistic errors; 6) miscellaneous errors. The second type of classification involves errors in the following categories:

1) agreement;
2) valency;
3) pronominal reference;
4) analytical verb forms or compound predicates;
5) reflexive expressions;
6) negation;
7) redundant or missing items;
8) wrong word order;
9) lexicon or phraseology;
10) grammar category;
11) style.

RLC adopted the following error classification [10]:

1) orthographic errors;
2) morphologic errors;
3) syntactic errors;
4) errors in constructions;
5) lexical errors;
6) additional features.

The last group is used for tagging miscellaneous errors such as calques from other languages (interference), missed words, word, morpheme or letter substitution or other types of errors, especially of those that cannot be identified properly.

Since error correction and classification is the core part of learner corpora, it is then crucial to do it with minimal errors. The question of how many annotators are sufficient for this task is discussed in [33]. Obviously, we need to have more than one or at least two experts; however their agreement can be not so high. The evaluation of the manual annotation and its consistency deserves special attention. Many authors use the metric κ (kappa) from [34] that varies within the interval $[-1, 1]$: $\kappa = -1$ means perfect disagreement, $\kappa = 1$ shows perfect agreement, and $\kappa = 0$ suggests that the agreement is equal to chance.

Linguistic annotation

Nowadays linguistic annotation is mostly done automatically. Tools for automated analysis are integrated into corpus systems or available as separate programs. The results yields high accuracy but should be treated with caution as there are errors in lemmatization or tags. Nevertheless the importance of linguistic annotation for corpora cannot be overestimated. Linguistic data of various types makes it possible to search grammatical categories, parts of speech, sentence structure, etc.

Automatic linguistic processing usually includes sentence splitting, tokenization and morphological analysis. At this stage, the system will mark words that can potentially contain an error, since they are absent in the morphological dictionary of the system. Of course, it should be remembered that the system may not contain some word, which at the same time exists in the language and which was used by the student. Thus, taggers can help annotators to find errors.

There is a large number of different systems that can be either language-specific or not. Below we will dwell on software available for Russian. Russian is an inflectional language and thus morphological forms play a key role. There are a number of well-recommended analyzers that can be used for annotating Russian texts. The morphological annotation of the RLC was carried out with the MyStem [35–36]. Below we show the example of this annotation performed by the program for the sentence *Segodnja my pishem sochineniye o semje* 'Today we are writing an essay about a family'.

Сегодня{сегодня=ADV=}
мы{мы=SPRO,pl,1p=nom}
пишем{писать=V,ipf,tran=inpraes,pl,indic,1p}
сочинение{сочинение=S,n,inan=(acc,sg|nom,sg)}
о{о=PR=}
семье{семья=S,f,inan=(abl,sg|dat,sg)}

The sentence was split into the tokens, each of them being on a separate line: *Segodnja*, *my*, *pishem*, *sochineniye*, *o* and *semje*. Lemmata, parts of speech and grammatical information are then shown in braces. In case of ambiguity, the system generates several options for parsing, which are separated by vertical bars |. For example, *sochineniye* are treated automatically either as an accusative or a nominative form.

UDPipe is yet another example of a system that annotates texts according the following levels: tokenization, lemmatization, and morphological and syntactic parsing [37–38]. The UDPipe output is prodeuced in the CoNLL-U format [39]:

Along with tokens, lemmata, parts of speech and morphological features, this format indicates a syntactic head for the current token, which is either a value of ID or zero (if the token is the root of the sentence, for example, *pishem*), and a dependency relation to the head (for example, "object" for *sochineniye*). Additional dependency relation can be applied when sentences involve coordinate structures. The last field stores miscellaneous information that is not given in other columns, such as position in the sentence with respect to punctuation or any specific annotation.

Another example of a morphological analyzer for Russian is pymorphy2 [40]. The output format shares the same features with the above described examples. The annotation provides tokens, morphological tags, grammemes and lemmata (defined as "normal_form"). The field "score" shows the probability that tags are assigned correctly (1.0 correspond to a perfect result). The analyzer can predict lemmata and morphological features for words that are absent in the dictionary.

As it has been already mentioned, the language of learner texts is specific; therefore, automatic morphological annotation, although being important, nevertheless requires manual or semi-automatic verification and further correction.

Corpus platforms

Once texts are collected and processed, we should answer the following question: how can they be stored and accessed? Most projects deal with the XML format, which is the most suitable for describing corpus data, and use TEI guidelines to represent metadata. It is also crucial to choose a suitable platform or corpus manager that allows one to work with annotated texts and search in them. On the one hand, there are well-known systems, and on the other hand, it is possible to build a new system for a project. RLC uses the platform powered by Django [8]. It keeps texts in a MySQL database that has separate tables for each of the text layers (metadata, sentence, morphological and error annotation ones). The system enables online upload of external new texts and then automatically processes them by the MyStem tagger.

Sketch Engine [41] is widely used by many corpora, among them CLC, Arabic Learner Corpus, Estonian Corpus for Learners and Guangwai-Lancaster Chinese Learner Corpus. Fig. 1 shows an example of text types available for the Open Cambridge Learner Corpus (an uncoded subset of CLC) in Sketch Engine. Based on these attribute, a user can build an appropriate subcorpus for his (her) tasks.

Among other platforms used for corpus building and analytics we can name KonText [42], IMS Open Corpus Workbench (CWB) [43], WordSmith [44], AntConc [45] and LancsBox [46].

Conclusion

In our work, we tried to provide an overview of some learner corpora. As one can see, these resources are diverse and the language of non-native speakers deserves to be studied more profoundly. We also

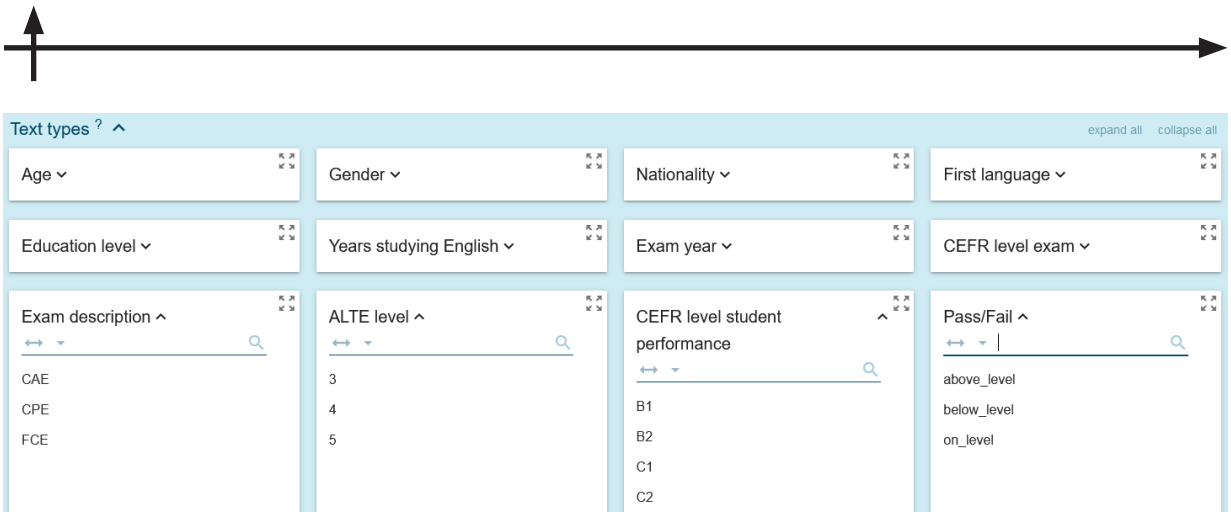

Fig. 1. Example of text metafeatures in Sketch Engine

sketched the pipeline that can be used for building a learner corpus and focused on issues related to its annotation. Error annotation requires special attention as it is the main part of such a corpus and provides empirical evidence of learner performance helping to reveal real problems the language learner encounter and not the ones that are described in dictionaries and grammars.

REFERENCES

[1] **A. Rosen, J. Hana, B. Vidová Hladká, T. Jelínek, S. Škodová, B. Štindlová,** Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech. Praha, 2021.

[2] CLARIN. Available at: https://www.clarin.eu/resource-families/L2-corpora (accessed 10.02.2023).

[3] **S. Granger, M. Dupont, F. Meunier, H. Naets, M. Paquot,** The International Corpus of Learner English. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain, 2020.

[4] **S. Atkins, J. Clear, N. Ostler,** Corpus Design Criteria. Literary & Linguistic Computing, 7 (1) (1992) 1–16.

[5] **A. McEnery, V. Brezina, D. Gablasova, J.V. Banerjee,** Corpus Linguistics, learner corpora and SLA: employing technology to analyze language use. In: Annual Review of Applied Linguistics. 39 (2019) 74–92.

[6] **S. Granger,** The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. TESOL Quarterly. 37 (3) (2003) 538–546.

[7] **D. Nicholls,** The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT, in Archer, D, Rayson, P, Wilson, A and McEnery, T (Eds), Proceedings of the Corpus Linguistics 2003 Conference, Lancaster University, 2003, pp. 572–581.

[8] **E. Rakhilina, A. Vyrenkova, E. Mustakimova, A. Ladygina, I. Smirnov,** Building a learner corpus for Russian. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, 2016. Available at: http://aclweb.org/anthology/W16-65 (accessed 10.02.2023).

[9] **O. Kisselev,** Russian Learner Corpora Research: State of the Art and Call for Action. In Bakhtiniana, São Paulo, 18 (1) (2023) 8–29.

[10] RLC. Available at: http://web-corpora.net/RLC (accessed 10.02.2023).

[11] CECL. Learner corpora around the world. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (accessed 10.02.2023).

[12] **M. Khokhlova, V. Benko,** Size of corpora and collocations: the case of Russian. In Slovenščina 2.0, 8 (2) (2020) 58–77.

[13] **Y. Tono,** Learner corpora: Design, development and applications. In Proceedings of the 2003 Corpus Linguistics Conference, D. Archer, P. Rayson, A. Wilson & T. McEnery (eds). UCREL: Lancaster University, 2003, pp. 800–809.

[14] **S. Granger,** Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) Corpus Linguistics. An International Handbook. Volume 1. Berlin & New York: Walter de 5 Gruyter (2008a) pp. 259–275.

[15] **S. Granger,** Learner Corpora in Foreign Language Education. In Van Deusen-Scholl N. and Hornberger N.H. (ed.) Encyclopedia of Language and Education. Volume 4. Second and Foreign Language Education. Springer (2008b) pp. 337–351.

[16] **S. Reder, K. Harris, K. Setzler,** The Multimedia Adult Learner Corpus. In TESOL Quarterly (2003), 37 (3) (2003) 546–557. Available at: https://www.researchgate.net/publication/251678834_The_Multimedia_Adult_Learner_Corpus (accessed: 10.02.2023).

[17] **H. Hilton,** Annotation and analyses of temporal aspects of spoken fluency. CALICO Journal, 26 (2009) 644–661.

[18] TAITO. Available at: http://urn.fi/urn:nbn:fi:lb-2014073035 (accessed: 10.02.2023).

[19] **J.A. Belz, N. Vyatkina,** Learner Corpus Research and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles, Canadian Modern Language Review/Revue canadienne des langues vivantes 62.1 (2005) 17–48.

[20] **C. Muñoz,** (ed.) Age and the Rate of Foreign Language Learning. Clevedon: Multilingual Matters. (2006).

[21] **J.H. Jantunen, S. Brunni,** Morphology, lexical priming and second language acquisition: A corpus-study on learner Finnish. In S. Granger, G. Gilquin and F. Meunier (eds.) Twenty Years of Learner Corpus Research. Louvain-la-Neuve. Presses universitaires de Louvain, 2013, pp. 235–245.

[22] **S. Brunni, L.-M. Lehto, J.H. Jantunen, V. Airaksinen,** How to annotate morphologically rich learner language: principles, problems and solutions. Bergen Language and Linguistics Studies, 6 (2015) 133–152.

[23] **F. Meunier,** Introduction to the LONGDALE project. In E. Castello K. Ackerley, & F. Coccetta (Eds.), Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment Berlin: Peter Lang Publishing, 2016, pp. 123–126.

[24] ESF. In: The ESF Database. Available at: https://www.mpi.nl/world/tg/lapp/esf/esf.html (accessed: 10.02.2023).

[25] **H. Feldweg,** The European Science Foundation Second Language Database. Nijmegen: Max-Planck-Institute for Psycholinguistics, 1991.

[26] **E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, M. Sandell,** SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In LREC Proceedings 2016, 2016, pp. 206–212. Available at: https://aclanthology.org/L16-1031 (accessed: 10.02.2023).

[27] CHILDES. Available at: https://childes.talkbank.org/ (accessed: 10.02.2023).

[28] **S. Tagnin,** A multilingual learner corpus in Brazil. In: Archer, D., Rayson, P., Wilson A., McEnery T. (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL, Lancaster University, 2003, pp. 940–945.

[29] **K. Tenfjord, J.E. Hagen, H. Johansen,** Norsk andresprkåskorpus (ASK) – design og metodiske forutsetninger. NOA norsk som andrespråk 25 (1) (2009) 52–81. Available: https://w2.uib.no/filearchive/tenfjord-hagen-og-johansen.-2009.-noa..pdf (accessed: 10.02.2023).

[30] **A. Mendes, S. Antunes, M. Janssen, A. Gonçalves,** The COPLE2 Corpus: A Learner Corpus for Portuguese. In: Proceedings of the Tenth Language Resources and Evaluation Conference – LREC'16, 23-28 May 2016, Portoroz, Slovenia, 2016, pp. 3207–3214.

[31] **N. Mikelić Preradović, M. Berać, D. Boras,** Learner Corpus of Croatian as a Second and Foreign Language. Multidisciplinary Approaches to Multilingualism. Ur. Cergol Kovačević, Kristina i Udier, Sanda Lucija. Peter Lang. Frankfurt am Main, Njemačka, 2015, pp. 107–126.

[32] **M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K. Schmidt, H. Hirschmann, T. Andreas,** Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01, 2012. Available at: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko (accessed: 10.02.2023).

[33] **R. Snow, B. O'connor, D. Jurafsky, A.Y. Ng,** Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 254–263.

[34] **J. Cohen,** A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1) (1960) 37–46.

[35] **I. Segalovich,** A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Proceedings of the MLMTA 2003. USA, 2003, pp. 273–280.

[36] **I. Segalovich, V. Titov,** Mystem, 1997. Available at: https://yandex.ru/dev/mystem/ (accessed: 10.02.2023).

[37] **M. Straka, J. Straková,** Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2017, pp. 88–99.

[38] UDPipe. Available at: https://ufal.mff.cuni.cz/udpipe/1 (accessed: 10.02.2023).

[39] CoNLL-U Format. Available at: http://universaldependencies.org/docs/format.html (accessed: 10.02.2023).

[40] **M. Korobov,** Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, 2015, pp. 320–332.

[41] **A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, M. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel,** The Sketch Engine: ten years on. Lexicography, 1 (1) (2014) 7–36.

[42] **T. Machálek,** KonText – a modern, customizable corpus query interface. Abstract of a talk presented at the conference Corpus Linguistics 2017, Birmingham, 2017. Available at: https://www.birmingham.ac.uk/Documents/collegeartslaw/corpus/conference-archives/2017/general/paper341.pdf (accessed: 10.02.2023).

[43] **S. Evert, A. Hardie,** Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK, 2011.

[44] **M. Scott,** WordSmith Tools version 8, Stroud: Lexical Analysis Software. Available at: https://lexically.net/wordsmith/ (accessed: 10.02.2023).

[45] **L. Anthony,** AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University, 2020. Available at: https://www.laurenceanthony.net/software (accessed: 10.02.2023).

[46] **V. Brezina, P. Weill-Tessier, A. McEnery,** #LancsBox v. 5.x. [software], 2020. Available at: http://corpora.lancs.ac.uk/lancsbox (accessed: 10.02.2023).

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Maria V. Khokhlova
Хохлова Мария Владимировна
E-mail: m.khokhlova@spbu.ru
ORCID: https://orcid.org/0000-0001-9085-0284

Submitted: 20.01.2023; Approved: 10.03.2023; Accepted: 17.03.2023.
Поступила: 20.01.2023; Одобрена: 10.03.2023; Принята: 17.03.2023.