Cite as: Sommer-Farias, B., Vinokurova, V., Gorlova, A., Centanin-Bertho, M. (2023). *The FLTMAG* (November, 2023). https://fltmag.com/teaching-learner-corpus-data/

Teaching with Learner Corpus Data

By Bruna Sommer-Farias (Michigan State University), Valentina Vinokurova (University of Arizona), Asya Gorlova (University of Arizona), Mariana Centanin-Bertho (University of Arizona)

Introduction

Using texts to help students notice and use grammatical and lexical choices appropriately in particular communicative situations is a great way to plan classes that focus on language in context. However, available teaching materials may not address language in context that is accessible and relevant to students' ages and proficiency levels, thus providing limited or overly targeted content that leaves out important contextual or cultural information relevant to communication in particular situations. The scarcity of high-quality resources is especially common in certain settings, such as learning a less-commonly-taught-language, e.g., Russian or Portuguese in the United States.

One type of resource that can facilitate teachers' access to texts is the language corpus (plural corpora), which is defined as an electronically searchable collection of texts organized by text types and context representation, such as news articles, blog posts, dialogues, and others. While access to these texts provides teachers with an array of possible topics to inform both grammar and writing classes, it may be difficult to select texts that are appropriate to course content and students' proficiency levels. In other words, texts from native-speaker corpora may lack examples of text types relevant to learners' proficiency levels as they are full of vocabulary and grammar that is unfamiliar to students, especially at the novice and intermediate levels. For this reason, we turn to learner corpora as an alternative.

The use of learner corpus data can be highly beneficial for teachers because they offer access to a large collection of authentic texts produced by learners in multiple settings, including proficiency tests and classroom instruction. Depending on the target language, learners' proficiency level, and learning objectives, learner corpora may be more effective than native-speaker corpora. In particular, learner corpora 1) contain texts on relevant topics that tend to be covered in the language classroom; 2) are typically organized by course level or level of proficiency, making it easy to find level-appropriate texts; and 3) include samples of various genres that students are asked to produce in their language courses. When teaching a unit on the topic of travel, the teacher may more easily find level-appropriate texts about travel, including learner-produced travel narratives, touristic landmark descriptions, and restaurant reviews. Based on their specific learning objectives, the teacher can select relevant sample texts and create activities based on those texts, directing students' attention to focal communicative functions, grammatical constructions, or genre characteristics. For students, texts produced by other learners can be both a source of inspiration - an example upon which to build one's own text - and competition - an example to surpass.

There are two main ways of using corpus data to teach: the hands-off approach, where teachers select examples from the corpus to create materials for students, and the hands-on approach, where students use the corpus themselves to search for patterns guided by prompts created by the teacher. The hands-off approach is usually advised for first-time users since it eliminates the need to train students to use the corpus and its interface and provides scaffolding for those unaccustomed to corpus-based noticing activities. With more senior and/or advanced students, the hands-on approach can be a powerful way to facilitate self-directed learning, especially if students are asked to focus on features of their choice when exploring the corpus.

One concern teachers may have before using a learner corpus in the classroom for the first time is how to handle errors that may be present in texts produced by students. We have a few suggestions that we hope will encourage you to use learner texts. First, it is important to start by introducing students to the corpus, explaining that it is made up of learner texts, and preparing them to expect occasional errors. Overall, students tend to find it motivating to engage with texts produced by fellow language learners and are especially keen to discover and fix errors in these texts. We feel that it is important for learners to know that making mistakes is a natural part of language learning and to see that other learners make mistakes, too. Exposure to learner texts can be a great boost to learner confidence. However, depending on their philosophy and approach, instructors can purposefully select lines without errors when creating their own activities.

In this article, we discuss how to create corpus-informed activities using learner corpus data to teach language in context, facilitating analysis of grammar and writing in an integrated manner. First, we briefly discuss the basic elements of corpus search and their pedagogical applications. Then, we share a sample hands-off lesson created using the corpus Multilingual Academic Corpus of Assignments Writing and Speech (MACAWS), which contains written and oral assignments produced by students in the University of Arizona's Portuguese and Russian foreign language programs (Staples et al., 2019-, available at macaws.corporaproject.org). We hope that teachers use these examples as a springboard to expand this approach to other languages. To do so, we conclude by giving directions on where to find similar corpora resources and how to adapt these principles to their own contexts.

What are Learner Corpora and How Can Corpora Help?

As mentioned earlier, learner corpora (sing. corpus) are digital searchable collections of texts produced by learners and organized by specific principles, such as text type, course or proficiency level, etc. Corpora can be used to guide students to notice linguistic patterns by displaying focal features via concordance lines, or, in other words, a list of lines of text that display a search term in its surrounding context. In Picture 1, you can see an example of concordance lines displaying the results of a corpus search for the word "ficar" (‘to stay’ in Portuguese). All the sentences in Picture 1 come from texts that were written by students responding to the same prompt, in this case, giving touristic suggestions of a place they have been to. Therefore, learners analyzing those concordance lines have the chance to infer the meaning and usage of the verb "ficar" in the context they are studying, i.e. traveling. The

concordance line format allows teachers to guide students to notice words that precede and follow the search term in the center and infer lexicogrammatical rules depending on the frequency of patterns. For example, "ficar", a verb, can be preceded by the place adverb "onde" (‘where’), forming the phrase "onde ficar" (‘where to stay’). Similarly, students may notice other expressions such as "é bom ficar" (‘it is good to stay there’), and "se você quiser ficar" (‘if you want to stay there’). The fact that the repeated words are close to each other are meant to focus their attention and facilitate pattern noticing.

Picture 1 - *Concordance lines centering the word "ficar" (“to stay”) from the corpus MACAWS*

Such inductive tasks can guide students to interpret language in the format of concordance lines to serve various learning goals. For example, students can explore the form and function of the verb in travel recommendations, analyze linguistic choices such as conjunctions and pronouns, and understand the communicative purpose behind describing a travel experience and offering tips about where to stay for fellow travelers.

Corpora offer different types of resources and categorize learner texts in several ways, so it is important to explore what affordances are offered by each corpus. The MACAWS Corpus offers an online platform with tools to assist instructors and students in using corpus data for teaching (Staples et al., 2019). One of these tools is interactive Data-Driven Learning (iDDL), which allows the embedding of concordance lines in digital language learning materials. Apart from concordance lines, the MACAWS Corpus also allows teachers to select full sentences, text excerpts, or entire texts as input for activities. These different types of input cater to diverse learning outcomes. Full sentences and text excerpts are suitable for studying sentence structures, while whole texts provide additional context, such as transition words and text

markers, for comprehensive analysis as well as the whole text structure. Furthermore, teachers can filter the corpus by assignment name (e.g., food blog, culture essay, etc.), assignment topic (e.g., art, family, food, etc.), macrogenre (e.g., narration, description, etc.), mode (written or oral), and course level (for a full corpus report, see Sommer-Farias et al., 2022). The filter option is designed to make it easier to find texts that are more suitable to teachers' contexts, relevant to their curricula, and suited to their students' needs.

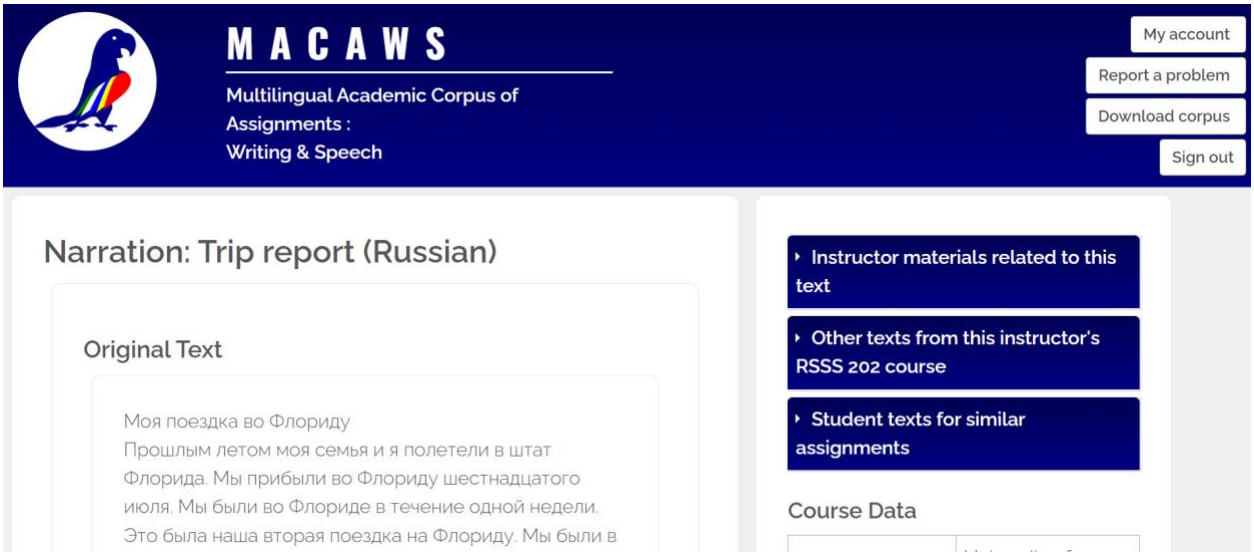
The following section will exemplify how the MACAWS Corpus can provide input for integrated teaching of grammar and writing. We hope that this step-by-step process can inspire teachers to create corpus-based tasks using corpora in other languages too.

Sample Lesson from a Beginner Russian Language Course

The sample hands-off corpus-informed lesson presented in this section was designed for a college-level first-year Russian language course. This lesson can complement any module on the topic of travel. In our case, this lesson was implemented in a second-semester course with students at a Novice High level of proficiency (ACTFL). The activities in this lesson focus on specific functions of two Russian cases, Accusative and Prepositional, namely, specifying destination and location. Students were already familiar with these cases and some of their functions from previous coursework. This lesson had the following learning objectives: to 1) associate case forms with describing motion to a destination and being at a location in trip reports; 2) practice using Accusative case and Prepositional case; and 3) create travel narratives using appropriate constructions to express destination and location. Following the principles of Data-Driven Learning, this lesson features activities aimed at illustration, interaction, and induction of focal constructions (Carter & McCarthy, 1995).

Part 1: Illustration

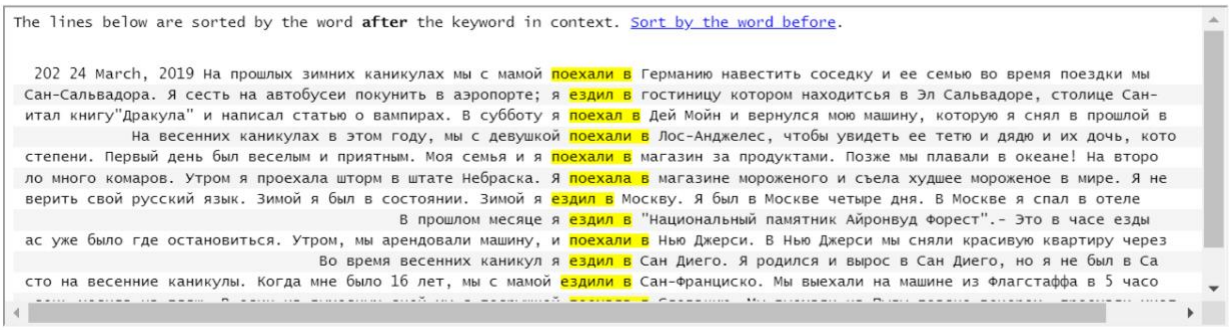
Students begin by reading a full text from the MACAWS learner corpus selected and provided by the instructor. The main goal of the illustration stage is to provide an example of a travel narrative as well as the use of the target grammatical features (Prepositional and Accusative cases) in context. Students can be provided with reading comprehension questions to engage with the text. They can also be asked to comment on the organization of the text: how effectively the text is organized, and what suggestions can be made for improving the organization of the narrative. To transition to the next stage - interaction - the teacher should ask students to note lexicogrammatical features that are characteristic of travel narratives. In other words, what do we absolutely need to write in a travel narrative? Students can be guided to notice motion verbs and expressions that show destination and location.



Picture 2 - *Example of a Learner Travel Narrative from the MACAWS Corpus*

Part 2: Interaction

At this stage, students are asked to interact closely with samples of writing that display the targeted lexicogrammatical features - in this case, Accusative and Prepositional cases. An effective way to present multiple examples of a grammatical feature is concordance lines, as mentioned above. This and the following concordance boxes, for example, problematize the expression of destination and location in Russian and highlight that Accusative and Prepositional cases are used with different verbs to express these meanings. To obtain relevant examples of the Accusative case for destination (Picture 3), the instructor searched the corpus for the phrases 'поехал/и в' and 'ездил/и в' ('went to'). The instructor took a screenshot of the results and included them in a student handout.

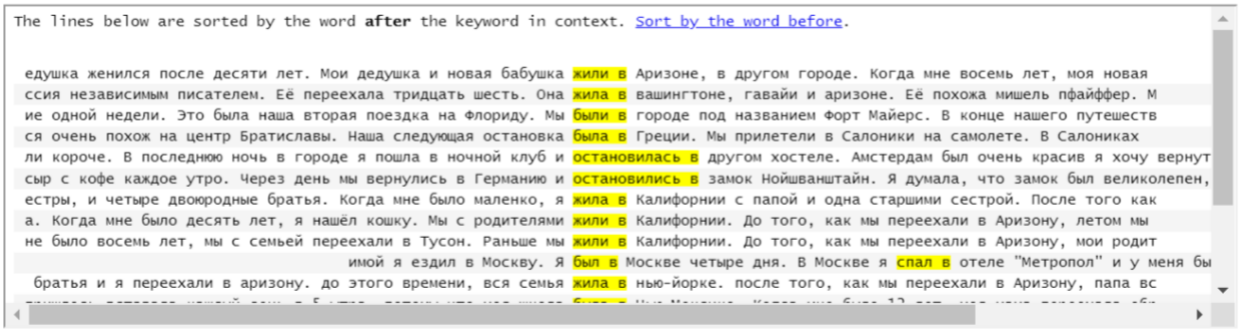


Picture 3 - *Concordance Box: Accusative Case*

The concordance box in Picture 3 is presented to the students with several questions that we suggest they answer in pairs or groups. While working in pairs or groups is not absolutely necessary, collaborative brainstorming may be helpful for students who are not accustomed to looking for patterns. First, students are directed to look at the highlighted verbs.

Since Russian uses a variety of verbs to say “to go,” students are asked to think about the meaning of these different verbs: How can they be translated into English? What is the meaning they share? Next, students are asked to consider the words to the right of the highlighted segments: What case are they in? Why? What does this case signify in this context? At this point, students are expected to notice the Accusative case forms that are already familiar to them. Given the occurrence of the Accusative case with verbs of motion, students will soon be able to conclude that the Accusative case here expresses the destination of motion. To confirm or correct student hypotheses, the teacher will explain that with verbs with the meaning ‘to go’, we need to use Accusative case to signify destination. Those of our readers who speak Russian might notice that the concordance box above contains one example with an error in case usage. In our experience, students tend to notice this error as well, because they are already familiar with other functions of the Accusative case by the time this activity is used. To validate and reward this finding, we conclude this part of the activity by asking students if they were able to find one sentence in the concordance box that violates the “motion verb + Accusative case” rule.

Now the focus of the activity shifts toward the use of the Prepositional case for location. To obtain relevant examples of Prepositional case (Figure 4), the corpus was searched for expressions ‘жил/а в’ (‘lived in’), ‘был/а в’ (‘was in’), ‘остановились/лась в’ (’stayed at’).



Picture 4 - *Concordance Box: Prepositional Case*

Similarly to the previous activity, students are directed to work in groups and analyze the concordance box. The initial questions pertain to the highlighted verbs: What do they mean? How are these verbs relevant to the topic of travel? Then, students move on to analyzing verb complements, their case forms, and significance. Lastly, they are asked to find an error in the concordance lines and explain it.

After each group task, students share their hypotheses with the entire class. As this activity is designed for a beginner-level course, this conversation takes place in English. The teacher moderates this discussion and concludes this part of the lesson by confirming the grammatical rules and functions of Accusative and Prepositional cases. Students are now ready to move to the next part of the lesson: applying the rules in context.

Part 3: Induction

The induction stage is an opportunity for students to apply the rules that they have learned in new situations. Inductive exercises can take different forms, from choosing the correct option to filling in the gaps or even writing a narrative using the target construction. In the exercise provided below, students are asked to choose the correct verb to complement the case form of the noun in bold.

Multiple-Choice Exercise

Take a look at the nouns in bold and, based on their case forms, circle the appropriate verbs:

1. Прошлым летом мы (ездили / были) в **Канаду**.
   Last summer we (went / were) ***to Canada****.*
2. В этом году Марина (поехала / жила) в **Москве**.
   This year Marina (went / lived) ***in Moscow****.*
3. Мы с семьёй (остановились / ездили) в **гостинице**.
   My family and I (stayed / went) ***in a hotel****.*
4. В августе наши студенты (ездили / были) в **Вашингтон**.
   In August our students (went / were) ***to Washington****.*

After the exercise, the instructor can choose to restate the rules, provide more detailed explanations for each sentence, or give additional examples. When the instructor is confident that students understand the rules, they can move on to productive activities. In our sample lesson, multiple-choice exercises are followed by a short writing activity, which helps to gradually move from practice to production without overwhelming the students. As such, it prepares them for a longer writing assignment that will follow.

Short Writing Activity

What about you? Write down your answers to the following questions and share them with your partner:

1. **Куда** ты ездил(а) в этом году?
   Where *did you go this year?*
2. **Где** ты остановился (остановилась)?

***Where** did you stay?*

3. **Куда** ты ходил(а) в этом городе?
   ***Where** did you go in the city?*

4. **Где** ты обедал(а), когда ты была в этом городе?
   ***Where** did you have lunch when you were in this city?*

After multiple-choice and short writing practice, students are ready to write their own travel narratives. They are instructed to refer to the constructions they learned while analyzing the concordance lines (Part 2) and to use their review and critique of the sample narrative (Part 1) to inform their writing. In-class work therefore serves as a scaffold: it provides students with a sample narrative as a jumping-off point and gives them a solid understanding of relevant grammar. Students are not asked to imitate the sample text but rather to use it as inspiration, nor are they given a list of grammatical instructions to include their narratives - the aim of the lesson is to offer support to learners rather than instill a golden standard. In their writing assignment, our students had to describe a trip they recently took. Within this narrative, they were asked to talk about their route, how they got to different places, and where they stayed. These instructions provided a basic structure for students to follow when writing their narratives and helped elicit focal grammatical constructions for expressing location and destination. This narrative was assessed using an analytic rubric with the following criteria: content and organization, accuracy, and word count.

Applying These Principles to Your Own Context and Language

Learner corpora such as MACAWS provide valuable resources and tools for instructors and students. By incorporating authentic texts and utilizing concordance lines, sentences, text excerpts, or full texts, instructors can enhance their teaching of grammar and writing and help students develop their proficiency in the target language. Concordance lines and learner texts can be used for various other purposes, such as reading comprehension, analysis of textual organization, genre analysis, and others. The sample lesson shared here was an example of hands-off use of the corpus, where students learn with texts and concordance lines selected by the teacher. Another option is to teach your students how to use the corpus interface themselves so they conduct their own searches.

To explore MACAWS further, fill out the request form available here: https://macaws-api.corporaproject.org/user/register. The form asks for a short description of how the user intends to use the corpus so that the researchers have a better idea of how the corpus is useful for teachers and researchers. For instance, if you want to create activities with concordance lines or short text excerpts, you can choose to have standard access to the interface, where only 500 characters are shown from each text. If you would like to conduct genre analysis and

compare draft and final versions of assignments, you should request access to the full access version, which displays full student texts. Additionally, you can request access to download the corpus for research purposes. Some of the projects that have been conducted based on the MACAWS corpus are 1) analysis of complexity features in L2 Russian development across program levels (Novikov, 2021); 2) the study of the influence of Spanish and English in Portuguese L3 use of copula constructions (Picoral, 2020); and 3) the analysis of acquisition paths of preposition+article contractions in L3 Portuguese among different L1- speaking learners (Picoral & Carvalho, 2020).

We also invite you to explore the resources provided below which contain links to corpora for other world languages and sample teaching materials. Keep in mind though that the choice of format of the corpus-informed task is going to depend on how each corpus presents and categorizes the data, so we suggest that the first step should be to explore how representative the available texts are for your learner population and teaching context.

Ready-to-use Russian and Portuguese activities on the MACAWS website: https://sites.google.com/email.arizona.edu/macawswebinar/activities

Webinars on how to use corpus data to teach languages: https://sites.google.com/email.arizona.edu/macawswebinar/contributions/webinars

Lists of corpora in other languages:

https://sites.google.com/email.arizona.edu/macawswebinar/other-corpora

https://corpus.cal.msu.edu/resources/

References

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*(2), 141-158.

Novikov, A. (2021). *Syntactic and morphological complexity measures as markers of L2 development in Russian* [Doctoral dissertation, University of Arizona].

Picoral, A. (2020). *L3 Portuguese by Spanish-English bilinguals: Copula construction use and acquisition in corpus data*. [Doctoral dissertation, University of Arizona].

Picoral, A., & Carvalho, A. (2020). The acquisition of preposition+article contractions in L3 Portuguese among different L1- speaking learners: A variationist approach. *Languages*, *5*(4), 45-62.

Sommer-Farias, B., Novikov, A., Picoral, A., Bertho, M., Staples, S. (2022). Multilingual learner corpus for less commonly taught languages. *International Journal of Learner Corpus Research*, *8*(2), 261-282.

Staples, S.,Picoral, A., Novikov, A., Sommer-Farias, B. (2019). *Multilingual Academic Corpus of Assignments – Writing and Speech*. Available at https://macaws.corporaproject.org/.