l'analisi linguistica e letteraria xxxi (2023) 25–38
issn: 1122-1917

Проект устного учебного корпуса русского языка

Tatsiana Maiko
Università degli Studi di Milano
tatsiana.maiko@unimi.it

Received May 2023; accepted August 2023; published online October 2023

This article introduces the *Russian Spoken Learner Corpus*, a new resource for language learning research. The corpus was created by a research team at the Department of Languages, Literatures, Cultures and Mediations of the University of Milan. It comprises longitudinal and quasi-longitudinal oral data produced by Italian learners of Russian across different proficiency levels, from A0→1 to C1. In the longitudinal part of the project, data collection is conducted twice a year within the same group of students throughout their three/five-year study program. The quasi-longitudinal subcorpus includes data produced by students from the first to the fifth year of study. In addition to learner data, the corpus also includes two reference subcorpora. One subcorpus contains interviews with native speakers of Russian, while the other one consists of interviews with bilingual (Italian-Russian) speakers. The interviews are transcribed following explicit conventions. The database contains audio files, their transcripts, and detailed metadata about the interviewee, the interviewer, and the tasks.

Keywords: Language Learning Research, Spoken Learner Corpus, Longitudinal Corpus, L2 Russian

1. *Введение*

Наряду со стремительным развитием корпусной лингвистики и корпусных подходов к исследованию языка, в последние десятилетия изучение речи неносителей языка получило новый импульс благодаря созданию и анализу учебных корпусов (англ. *learner corpora*). Собрания образцов речи студентов, изучающих иностранный язык, позволяют проводить количественные и качественные исследования различных аспектов интеръязыка, в том числе в рамках сравнительного интеръязыкового анализа (Granger 1996), предполагающего сопоставление продукции носителей языка и студентов с разными родными языками. На сегодняшний день большинство учебных корпусов состоит из письменных текстов (среди наиболее крупных международных проектов[1] упомянем *International Corpus of Learner English*[2] [5,5 миллионов слов] и *Corpus and repository of writing*[3] [16 миллионов слов]). Тем не менее, существует ряд

[1] Обновляемый список учебных корпусов, составленный *Centre for English Corpus Linguistics* Католического университета в Лувен-ла-Нев, см. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (последнее обращение 5 августа 2023).
[2] https://uclouvain.be/en/research-institutes/ilc/cecl/iclev2.html (последнее обращение 5 августа 2023).
[3] https://crow.corporaproject.org (последнее обращение 5 августа 2023).

проектов по созданию учебных корпусов, содержащих образцы устной речи (например, *Trinity Lancaster Corpus*[4] [4,2 миллиона слов], *Louvain International Database of Spoken English Interlanguage*[5] [*LINDSEI*, более 1 миллиона слов])[6]. В целом объем доступных письменных учебных данных остается выше, чем устных[7].

Самым крупным учебным корпусом русского языка является *Русский учебный корпус*[8] (общий объем 2.258.869 слов), содержащий образцы письменной (2.165.488 слов) и устной (93.381 слово) речи студентов с 48 доминантными языками[9]. В итальянский подкорпус *Русского учебного корпуса* (60.188 слов) входят образцы в основном письменной (55.789 слов) и частично устной (4.399 слов) речи студентов с доминантным итальянским языком и различным уровнем владения русским языком. Как видно, устные данные студентов, изучающих русский язык, представлены в *Русском учебном корпусе* намного меньше, по сравнению с письменными текстами. Нам неизвестно о других проектах по созданию учебного корпуса русского языка на основе устной продукции, кроме коллекций образцов устной речи, собранных исследователями для решения определенных задач. Например, для описания особенностей речи информантов-билингвов Е. Протасова (Protassova 2016) записала 41 рассказ-описание по картинкам из книги М. Мейер "*Frog, where are you?*"[10]; для изучения дискурсивных элементов в речи студентов Н. Стоянова (Stoyanova 2020) собрала образцы устной речи итальянских студентов (82 рассказа), используя задания проекта "*Истории о подарках и катании на лыжах*"[11].

Что касается лонгитюдных корпусов[12], целью которых является сбор образцов речи определенной группы студентов на протяжении нескольких лет (не менее одного раза в год), то единственным известным нам проектом для русского языка как иностранного является лонгитюдный корпус письменной академической речи *RULEC* (*The Russian Learner Corpus of Academic Writing*)[13] (579.011 слов), состоящий из текстов, написанных

[4] http://cass.lancs.ac.uk/trinity-lancaster-corpus/ (последнее обращение 5 августа 2023).

[5] https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html (последнее обращение 5 августа 2023).

[6] Кроме того, некоторые корпуса содержат сбалансированные письменные и устные данные, например *Guangwai-Lancaster Chinese Learner Corpus* (672.328 токенов в письменном подкорпусе и 621.900 токенов в устном).

[7] Данная тенденция характерна и для корпусов, содержащих образцы речи носителей языка. Так, к примеру, письменная часть *British National Corpus* составляет 86,3 миллионов слов, в то время как устная часть – 10 миллионов слов; основной корпус *Национального корпуса русского языка*, состоящий из письменных текстов различных жанров, содержит 375 миллионов слов, а устный корпус представлен 13,4 миллионами слов.

[8] *Russian Learner Corpus*, *RLC*, http://web-corpora.net/RLC (последнее обращение 5 августа 2023).

[9] О концепции создания корпуса *RLC* и работе над ним см. (Rakhilina et al. 2016).

[10] Об идее и методе сбора данных с использованием данной книги см. (Berman, Slobin 1994).

[11] http://www.spokencorpora.ru/showcorpus.py?dir=03pands/rus (последнее обращение 5 августа 2023).

[12] Приведем несколько примеров существующих лонгитюдных учебных корпусов: *LONGDALE* (*Longitudinal Database of Learner English*) (письменная и устная речь), *Barcelona English Language Corpus* (*BELC*) (письменная и устная речь), 5 подкорпусов *FLLOC* (*French Learner Language Oral Corpora*) (устная речь), некоторые подкорпуса *Corpus of Learner German* (*CLEG13*) (письменная речь).

[13] Данный корпус является отдельной частью корпуса *RLC*.

36 англоговорящими американскими студентами на протяжении четырехлетней программы обучения. На данный момент не существует лонгитюдных учебных корпусов устной речи студентов, изучающих русский язык. Кроме того, насколько нам известно, попыток составления лонгитюдного корпуса данных (как устной, так и письменной речи) студентов с родным итальянским языком не предпринималось.

В этой статье мы хотим представить проект учебного корпуса устной речи итальянских студентов, изучающих русский язык, создаваемый группой преподавателей и исследователей департамента иностранных языков, литературы и межкультурной коммуникации Миланского университета (О. Беженарь, П. Котта Рамузино, К.Г. Маканьо, Т. Майко). Создание такого ресурса, являющегося базой речевых образцов разных уровней языковой компетенции, позволит изучать постепенное развитие интеръязыка студентов на различных стадиях процесса овладения языком.

В статье представлены концепция и принципы создания корпуса (§ 2), приводятся сведения о структуре интервью (§ 2.1), метаданных (§ 2.2) и процессе транскрибирования записей (§ 2.3), а также описываются первые результаты и дальнейшие перспективы работы над проектом (§ 3). Статья завершается обсуждением возможных областей применения создаваемого корпуса (§ 4).

2. *Концепция и принципы создания корпуса*

Учебные корпуса могут представлять собой лонгитюдные или нелонгитюдные собрания письменных или устных речевых образцов, изучающих иностранный язык. В случае лонгитюдных корпусов их создатели собирают продукцию определенной группы учеников в течение нескольких лет (обычно рекомендуется проводить повторный сбор данных хотя бы один раз в год и в целом не менее трех раз [Ployhart, Vandenberg 2010, 9]). Стоит отметить, что количество создаваемых лонгитюдных учебных корпусов намного меньше, чем нелонгитюдных, что связано с дополнительными сложностями создания таких корпусов: в первую очередь, с более длительным процессом сбора данных и вероятностью того, что студенты не смогут по каким-либо причинам продолжать участвовать в проекте. В условиях отсутствия лонгитюдных корпусов исследователи могут прибегнуть к псевдолонгитюдным (или квазилонгитюдным) корпусам (Granger 2002, 11; Gass, Selinker 2008, 56–57). При создании таких корпусов вместо того, чтобы отслеживать прогресс одной группы учащихся на каждом этапе обучения языку, исследователи собирают данные нескольких групп учащихся с разным уровнем владения языком. Для того чтобы гарантировать некоторую однородность данных, группы информантов обычно обладают рядом общих характеристик (например, один и тот же родной язык или одинаковая среда обучения).

Настоящий проект предусматривает создание как лонгитюдного, так и квазилонгитюдного подкорпусов. В рамках лонгитюдной части проекта сбор данных (чекпоинт) проводится 2 раза в год с интервалом 6 месяцев у определенной группы студентов, начиная со второго семестра первого года обучения, на протяжении трех-/пятилетней программы обучения. В квазилонгитюдный подкорпус входят данные,

произведенные студентами с уровнем владения русского языка A0→1 – C1. Таким образом, создаваемый нами корпус позволит изучать динамику овладения русским языком итальянскими студентами на протяжении пяти лет обучения, начиная с уровня A0→1 и до уровня C1.

Те же задания предлагаются контрольной группе студентов-носителей русского языка для создания референтного корпуса, содержащего сопоставимые с учебным корпусом данные, отличающиеся минимальным количеством переменных. Это позволяет проводить контрастивные исследования в рамках сравнительного интеръязыкового анализа (о важности сравнения сопоставимых данных см. Granger, Dagneaux, Meunier 2002, 40). Кроме того, в отдельный подкорпус собираются образцы устной речи студентов-билингвов, родными языками которых являются итальянский и русский языки (и другие языки, например румынский или украинский), для изучения особенностей речи так называемых эритажных носителей[14].

2.1 Структура интервью

Опираясь на опыт международных проектов (в первую очередь *Louvain International Database of Spoken English Interlanguage* [Gilquin, De Cock, Granger 2010] и *Trinity Lancaster Corpus* [Gablasova, Brezina, McEnery 2019]), мы разработали трехчастную структуру для проведения неформальных интервью. Структура интервью одинакова для всех подкорпусов, что обеспечивает сопоставимость собираемых данных. Средняя длительность составляет примерно 30 минут. Вначале студенты получают краткое разъяснение относительно специфики интервью (целью его проведения является сбор образцов устной речи студентов с разным уровнем языковой компетенции; исправление ошибок не предусматривается) и его структуры.

Первая часть интервью представляет собой свободный монолог на одну из четырех предложенных тем, касающихся личного пространства учащихся и их опыта (“Путешествия”, “Свободное время”, “Семья”, “Моя учеба. Университет”). Студенту дается минута, чтобы обдумать свой ответ[15]. Монолог студента не ограничен по времени и длится в среднем 5-10 минут на каждую тему. Далее следует обсуждение выбранной темы с интервьюером в форме вопрос-ответ. Данная часть интервью не предполагает времени на размышление и максимально приближается по своим характеристикам к незапланированной диалогической речи. Кроме того, для того чтобы разнообразить собираемые речевые образцы, студентам предлагаются утверждения, стимулирующие выразить согласие/несогласие и высказать собственное мнение

[14] Эритажными (херитажными) носителями русского языка называют людей, эмигрировавших из русскоязычной страны в детском возрасте или родившихся за пределами русскоязычных стран, в той или иной степени усвоивших русский язык в домашних условиях, но использующих в качестве доминантного языка другой язык (об эритажных носителях см. Polinsky, Kagan 2007; Выренкова, Полинская, Рахилина 2014; об эритажных носителях русского языка в Италии см. Перотто 2013; Perotto 2019).

[15] Ряд исследований (Skehan 1998; Ortega 1999; Ahangari, Abdi 2011) показал положительное влияние времени на обдумывание ответа на объем и качество образцов устной речи, а также на снижение эмоциональной и когнитивной нагрузки на интервьюируемых.

(например, “Некоторые думают, что у мужчины и женщины должны быть разные роли в семье”; “Сегодня у людей больше свободного времени, чем в прошлом” и др.).

Вторая часть интервью организована таким же образом, но студент выбирает одну из четырех тем, не касающихся непосредственно личной информации и опыта учащихся (“Праздники”, “Итальянская и русская кухни”, “Изобразительное искусство. Поход в музей” и “Проблемы современного общества”).

Несмотря на то, что темы монолога заданы, такую продукцию можно считать свободной (или полусвободной), так как во время интервью студент не получает никакого предварительного инпута и самостоятельно выбирает языковые средства для выражения своих мыслей.

В третьей части интервью студенту предлагается посмотреть небольшое видео, пересказать и прокомментировать увиденное, а затем ответить на вопросы интервьюера. Для этого задания используются видеоматериалы, доступные онлайн: короткометражный фильм “Все равно”[16] и отрывок короткометражного фильма “С днём рождения”[17]. Речевые образцы, собранные в рамках этой части интервью, представляют собой в определенной степени контролируемую продукцию, поскольку участники пересказывают одно и то же видео. Однако задание включает в себя и элементы сторителлинга, т.к. студентам предлагается предположить, что произошло до событий, показанных в видео, или произойдет после.

Безусловно, интервью не представляет собой аутентичную коммуникативную ситуацию, однако мы стремились максимально приблизить условия проведения интервью к условиям спонтанного речепорождения: в случае предложенной нами модели интервью проводятся в учебном контексте, с интервьюером-преподавателем и записываются на специальное устройство[18], но в отличие от тех случаев, когда исследователи используют образцы устной речи, собранные во время экзамена, студент не готовит свою речь заранее и знает, что его продукция не будет оцениваться.

В рамках лонгитюдной части проекта интервью проводятся каждые 6 месяцев (апрель-май и октябрь-ноябрь) с определенной группой студентов. Во время первого чекпоинта студент выбирает одну из четырех предложенных тем в первом задании и одну из четырех тем во втором задании. Во время второго чекпоинта студенты выбирают две другие темы. Во время третьего чекпоинта им снова предлагаются первые две выбранные ими темы, во время четвертого чекпоинта – выбранные во второй раз темы и т.д. Такая же схема предусмотрена и для задания с видео: студентам предлагается видеоролик А во время первого, третьего, пятого и т.д. чекпоинтов и видеоролик Б во время второго, четвертого, шестого и т.д. чекпоинтов. Такой подход (сбор речевых образцов одних и тех же информантов на повторяющиеся темы) обусловлен стремлением создать максимально благоприятные условия для сбора сопоставимых

[16] https://www.youtube.com/watch?v=0-B6jmK8SZA&list=LL (последнее обращение 5 августа 2023).

[17] https://www.youtube.com/watch?v=vs6f1U6GLI0 (последнее обращение 5 августа 2023).

[18] Запись производится при помощи программы *Microsoft Teams* при проведении интервью в удаленном формате и при помощи диктофона в случае очного интервью.

данных, сведя к минимуму переменные. В то же время студентам не предлагается одна тема и одно видео, чтобы избежать эффекта запоминания задания.

2.2 Метаданные

Важным элементом процесса создания учебных корпусов является сбор метаданных об информантах, учитывать которые необходимо из-за их возможного влияния на интеръязык (о важности метаданных см. Gilquin 2015, 9–34; Myles 2021, 260 и след.). Кроме того, наличие подробных метаданных, описывающих структуру учебного корпуса, типы заданий и специфику представленных в нем образцов речи, повышает его доступность, вероятность использования широким кругом исследователей и сопоставимость с другими корпусами[19].

В рамках нашего проекта для каждой группы информантов была разработана специальная анкета, предложенная затем студентам в формате *Google Forms*.

Италоязычные студенты, участвующие в проекте, должны заполнить анкету, указав следующую информацию:

- возраст;
- гендер;
- курс университета;
- страну происхождения;
- родной(-ые) язык(-и);
- домашний(-ие) язык(и);
- владение другими языками;
- продолжительность изучения русского языка;
- опыт изучения русского языка до университета (в школе / частные уроки / самостоятельно) с указанием продолжительности, учебных материалов, являлись ли учителя носителями или неносителями русского языка и др.;
- изучение русского языка в университете с указанием продолжительности, учебных материалов, являются ли учителя носителями или неносителями русского языка и др.;
- наличие международных сертификатов владения русским языком;
- опыт пребывания и обучения в русскоязычной стране;
- вид и частоту языковой экспозиции[20] (общение в устной или письменной форме с носителями и неносителями русского языка; чтение книг, Интернет-источников и др.; просмотр фильмов, телевидения и др.; использование социальных сетей; прослушивание аудиокниг, подкастов, песен; учебный контент (приложения/социальные медиа/подкасты, посвященные русскому языку);

[19] В связи с этим см. предложение о стандартизации метаданных, включаемых в учебные корпуса (König et al. 2022).

[20] Некоторые переменные, играющие ключевую роль в процессе усвоения второго языка, такие как языковая экспозиция и мотивация, очень редко включаются в метаданные учебных корпусов (исключение составляют, например, *SCooLE* [*Secondary-Level Corpus of Learner English*] и *ICNALE* [*The International Corpus Network of Asian Learners of English*]).

– мотивацию к изучению русского языка (интерес к культуре, языку; личные отношения; желание участвовать в волонтерских программах; профессиональные причины; желание переехать в русскоязычную страну и др.).

Уровень владения русским языком информантов устанавливается их преподавателями на основании работы на занятиях и результатов на экзаменах.

Студенты-билингвы должны указать:

– возраст;
– гендер;
– курс университета;
– родной(-ые) язык(-и);
– домашний(-ие) язык(-и);
– страну происхождения (если страна происхождения не Италия, то сколько лет проживают в Италии и сколько лет и в каком формате изучают итальянский язык);
– язык обучения в школе / университете;
– владение другими языками;
– опыт изучения русского языка (дома / частные уроки / в русской школе выходного дня / в школе в русскоязычной стране / в школе в Италии / в университете) с указанием продолжительности, учебных материалов и др.;
– наличие международных сертификатов владения русским языком;
– опыт пребывания и обучения в русскоязычной стране;
– вид и частоту языковой экспозиции;
– мотивацию к изучению русского языка.

Анкета, разработанная для студентов-носителей русского языка, включает в себя такие вопросы, как:

– возраст;
– гендер;
– курс университета;
– родной(-ые) язык(-и);
– домашний(-ие) язык(-и);
– страна происхождения;
– формат и продолжительность изучения русского языка (в школе / университете);
– продолжительность проживания в Италии;
– длительность изучения итальянского языка;
– владение другими языками.

Все участники должны дать согласие на использование своих данных для исследовательских целей.

Кроме того, каждый аудиофайл и его расшифровка сопровождаются метаданными, описывающими интервьюера (возраст, гендер, родной(-ые) язык(-и), иностранные языки) и задание.

2.3 Транскрибирование

Среди корпусов устной речи можно выделить два подтипа: корпуса, включающие в себя текстовые файлы и файлы со звучащей речью (англ. *speech corpus*), и те, материал в которых представлен в виде транскриптов, но соответствующий звучащий текст недоступен (англ. *mute spoken corpus*).

Транскрипция звучащей речи может производиться с разным уровнем детализации. Большинство корпусов устной речи содержит только орфографическую расшифровку, при которой слова за редкими исключениями записываются в стандартной орфографической форме, а не в фонетической транскрипции (например, устный[21] и мультимедийный[22] подкорпуса *Национального корпуса русского языка* (НКРЯ), см. подробнее Гришина 2005). В некоторых корпусах используется более детальная транскрипция: например, в проекте *Рассказы о сновидениях и другие корпуса звучащей речи*[23] указываются просодические особенности текстов и такие нюансы произнесения, как губные смычки, придыхание, ускоренное произнесение и др. Лишь немногие корпуса (например, часть корпуса *Один речевой день*[24]) снабжены детальной акустико-фонетической транскрипцией звуковых файлов. Это связано, в первую очередь, с тем, что транскрибирование устной речи – трудоемкий процесс, предполагающий квалифицированную ручную обработку и требующий значительных затрат времени.

В случае устных учебных корпусов к сложностям, характерным для процесса транскрибирования в целом (выявление в аудиозаписи и передача в транскрипте пауз, хезитаций, речевых сбоев и др.), добавляются специфические трудности (распознание единиц звучащей речи при индивидуальных особенностях произношения студента, проблема отражения в транскрипте нестандартных форм слов, ударения и др.). Согласно подсчетам, приведенным в (Gilquin 2015), на транскрибирование одной минуты записи в рамках проекта по созданию устного учебного корпуса *LINDSEI* уходит 20-30 минут (включая финальную вычитку)[25].

Для того чтобы ускорить транскрибирование устных корпусов, производятся попытки частичной автоматизации процесса с помощью программ автоматического распознания звучащей речи. В силу специфики данных, о которой упоминалось выше, полностью автоматизировать транскрибирование образцов устной речи студентов-инофонов на данный момент не представляется возможным, но ведется работа по созданию специальных программ распознания речи неносителей языка (Wang, Schultz 2003). По причине повышенных затрат времени и усилий большинство устных учебных корпусов содержат только орфографическую расшифровку аудиозаписей.

---

[21] https://ruscorpora.ru/new/search-spoken.html (последнее обращение 5 августа 2023).

[22] https://ruscorpora.ru/new/search-murco.html (последнее обращение 5 августа 2023).

[23] http://www.spokencorpora.ru/ (о создании корпуса см. Кибрик, Подлесская 2009) (последнее обращение 5 августа 2023).

[24] https://ord.spbu.ru/ (последнее обращение 5 августа 2023).

[25] Jendryczka-Wierszycka (2009) отмечает, что на расшифровку одного интервью (средней продолжительностью 15 минут) при создании польского подкорпуса *LINDSEI* потребовалось в среднем около 5 часов.

Что касается процесса транскрибирования в нашем проекте, то начальной целью является минимальная транскрипция, правила которой были разработаны нами, опираясь на опыт российских и международных проектов, в частности устного подкорпуса *НКРЯ*, устных корпусов *Рассказы о сновидениях*, *Один речевой день*, устных учебных корпусов *LINDSEI* и *The Trinity Lancaster Corpus*. Поскольку первоочередными исследовательскими задачами, которые можно будет решать с применением создаваемого нами корпуса, мы видим изучение явлений лексики, синтаксиса, фразеологии и прагматики в интеръязыке итальянских студентов, изучающих русский язык как иностранный, то решение создать на данном этапе менее детальную и аналитически менее глубокую транскрипцию представляется оправданным.

Для упрощения процесса транскрибирования на начальном этапе была использована программа *speechpad.ru*, применение которой, однако, показало удовлетворительные результаты только при обработке речевых образцов студентов с высоким уровнем языковой компетенции, а также студентов-билингвов и носителей русского языка. Индивидуальные особенности произношения и обилие нестандартных форм слов и нестандартного построения предложений, характерные обычно для начальных уровней владения иностранным языком, не позволили успешно использовать программу автоматического аннотирования в случае обработки записи речи студентов уровня A1-A2 (и в некоторых случаях B1)[26].

Для орфографической расшифровки речевого материала был разработан набор правил и принят ряд обозначений, основные принципы которых представлены ниже.

Общепринятая пунктуация в транскрипте не используется. После отрывка с интонацией завершения ставится знак ‘/’. Для передачи вопросительных и восклицательных реплик используются знаки ‘?’ и ‘!’. Заглавная буква используется только для передачи имен собственных, но не ставится в начале реплики.

Хезитационные паузы, широко представленные в речи учеников, оформляются с помощью круглых скобок и многоточия ‘(...)’. Поскольку хезитации могут быть заполнены разными звуками, точная передача которых иногда вызывает затруднения, все возможные заполненные паузы были сведены к следующим вариантам: (э), (эм), (а), (ам), (и), (м), (ну), (но).

Так как на данном этапе было принято решение не включать фонетическую транскрипцию в орфографическую расшифровку, индивидуальные особенности произношения интервьюируемых не фиксируются. При неправильной постановке ударения ударная гласная выделяется с помощью большой буквы: *вОкзал*, *бЫла*. В случае неразборчивого произнесения слова или фрагмента реплики, используется ремарка ‘(нрзб.)’. Если интервьюируемый и интервьюер произносят что-либо одновременно, используется помета ‘(вместе)’.

[26] Об особенностях процесса транскрибирования речи студентов на начальных уровнях обучения см. Saturno 2014.

Нестандартные формы слова, словосочетания или конструкции записываются так, как их произнес студент[27]: *в лесе*, *некоторые информации*. Если студент использует 'инновацию' (например, *иностранций* вместо *иностранец*, *африканин* вместо *африканец*, *совестический* вместо *советский*), 'неправильное' слово помечается с помощью '*', а в скобках указывается искомое слово: **африканин* (*африканец*). В случае самоисправления, когда студент, испытывая трудности в процессе порождения речи, не заканчивает слово и заменяет его на другое, в транскрипте используется знак '=' в конце незавершенного слова: *я приле= приехала*. Если интервьюируемый употребляет иностранное слово, то используется помета '(ин.)', после которой записывается, если это возможно, употребленное студентом иностранное слово.

Некоторые междометия условно отображаются следующим образом: *угу* (произносимое с закрытым ртом утвердительное междометие), *не-у* (произносимое с закрытым ртом междометие отрицания), *м*? (переспрос с закрытым ртом). Числа записываются словами.

Каждый транскрипт затем вычитывается другим участником проекта, чтобы повысить качество расшифровки и снизить риск разной интерпретации услышанного. Как показал ряд исследований (Detey 2012, 234; Zechner 2009), двойная проверка транскриптов особенно важна при составлении устных учебных корпусов, т.к. при транскрибировании ученической речи количество спорных моментов и несовпадений в расшифровке разными исследователями выше, чем при транскрибировании речи носителей языка.

3. *Первые результаты и перспективы развития*

Перед началом процесса сбора данных важно протестировать проект и выявить потенциальные проблемы (о важности пилотного проекта при создании учебных корпусов см. Bell, Payant 2020). В рамках работы над устным учебным корпусом изначально разработанные анкеты и двухчастная структура, предполагающая обсуждение одной темы на выбор и видео, были протестированы с небольшой группой участников. После чего анкета была значительно расширена и разработана версия для студентов-билингвов. В структуру интервью было решено добавить блок со второй темой на выбор, чтобы расширить круг обсуждаемых тем. Кроме того, чтобы увеличить сопоставимость данных и снизить вероятность прайминга, была разработана инструкция для интервьюеров, включающая примеры вопросов и реплик-стимулов. На данный момент в лонгитюдном проекте участвует 14 студентов. В целом было проведено 99 интервью (см. таблицу 1[28]).

[27] Несмотря на то, что автоматическое извлечение из корпуса нестандартных форм затруднительно, а иногда невозможно, решение отображать в транскрипте произведенную студентом форму объясняется желанием максимально сохранить аутентичность данных.

[28] В таблице указано приблизительное количество слов, так как работа над расшифровкой и вычиткой продолжается.

Таблица 1 - *Состав корпуса*

Группа информантов	*Количество интервью*	*Количество слов*
A1	14	≈12000
A2	21	≈20000
B1	22	≈23000
B2	20	≈27000
C1	3	≈5000
Билингвы	9	≈14000
Носители русского языка	11	≈22000

Работа над корпусом продолжается в настоящее время по двум направлениям: пополнение его новыми записями от участников лонгитюдного проекта и от новых информантов (стремясь сбалансировать состав корпуса по уровню языковой компетенции информантов), а также расшифровка собранного материала.

Последующие этапы обработки материала будут представлять собой его описание на разных уровнях в соответствии с различными исследовательскими задачами (изучение морфологических, синтаксических, лексических, фразеологических, а также прагматических единиц) и создание специализированной базы данных, позволяющей добавить металингвистическую разметку и разметку ошибок[29]. На данный момент корпус загружен на платформу *Sketch Engine*[30], служащую для создания и исследования корпусов и предоставляющую такие возможности обработки данных, как отображение лексической и грамматической сочетаемости лексических единиц (англ. *word sketches*), создание списков частотности употребления, извлечение *n-grams* и ключевых слов и др. Кроме того, платформа *Sketch Engine* позволяет внести метаданные о интервьюируемом, интервьюере и задании, которые сопровождают каждый файл[31].

Для того чтобы сделать корпус доступным для широкого круга пользователей, ведется работа по созданию базы данных на платформе *Unimi Dataverse*[32]. База данных разделена на четыре подкорпуса. Для каждого интервью создается датасет, включающий в себя три аудиофайла (один на каждое задание), их транскрипты и анкету интервьюируемого.

[29] Для разметки ошибок планируется использовать специальную программу – *UCLEE* (*Université Catholique de Louvain Error Tagging Editor*), позволяющую ускорить процесс разметки ошибок и внесения исправлений в тексты учащихся. В настоящее время в разработке находится новая версия программы (*UCLEEv3*). Инструкцию по использованию программы и подробное описание тагсета см. Granger, Swallow, Thewissen 2022.

[30] https://www.sketchengine.eu (последнее обращение 5 августа 2023).

[31] О добавлении метаданных и других типов аннотации см. https://www.sketchengine.eu/guide/document-annotation-tool/ (последнее обращение 5 августа 2023).

[32] https://dataverse.unimi.it/dataverse/RuSLC/ (последнее обращение 5 августа 2023).

4. *Заключение*

Создаваемый нами устный учебный корпус разделен на четыре подкорпуса: лонгитюдный и квазилонгитюдный подкорпуса образцов устной речи студентов-носителей итальянского языка, подкорпус образцов устной речи студентов-билингвов и контрольный подкорпус продукции носителей русского языка. Полная реализация проекта в виде репрезентативной базы образцов устной речи всех представленных в корпусе категорий говорящих будет иметь важное значение для решения теоретических научных задач, таких как всестороннее изучение спонтанной речи студентов русского как иностранного с родным итальянским языком, а также сравнительный интеръязыковой анализ с речью студентов с другими родными языками, студентов-билингвов, чьими родными языками являются итальянский и русский, и носителей русского языка. Кроме того, создаваемый нами корпус сможет быть использован для решения актуальных прикладных задач в области преподавания русского как иностранного благодаря доступу к аутентичному репрезентативному материалу, который может быть использован, например, для диагностики более и менее проблемных тем, для иллюстрации удачных или неудачных высказываний, для определения последовательности введения материала и выработки преподавательских стратегий и т.д. Материал корпуса позволит устанавливать этапы усвоения определенных языковых явлений и определять случаи трансфера с родного и иностранных языков.

Литература

Ahangari, Saeideh, Morteza Abdi. 2011. "The Effect of Pre-Task Planning on the Accuracy and Complexity of Iranian EFL Learners' Oral Performance." *Procedia – Social and Behavioral Sciences* 29: 1950–1959.

Bell, Philippa, Caroline Payant. 2020. "Designing Learner Corpora." In *The Routledge Handbook of Second Language Acquisition and Corpora Routledge*, edited by Nicole Tracy-Ventura, Magali Paquot, 53–67. New York: Routledge. https://doi.org/10.4324/9781351137904.

Berman, Ruth A., Dan I. Slobin. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Detey, Sylvain. 2012. "Coding an L2 Phonological Corpus: from Perceptual Assessment to Non-Native Speech Models: An Illustration with French Nasal Vowels." In *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, edited by Yukio Tono, Yuji Kawaguchi, Makoto Minegishi, 229–250. Amsterdam/Philadelphia: John Benjamins.

Gablasova, Dana, Vaclav Brezina, Tony McEnery. 2019. "The Trinity Lancaster Corpus: Development, Description and Application." *International Journal of Learner Corpus Research* 5(2): 126–158.

Gass, Susan M., Larry Selinker. 2008[3]. *Second Language Acquisition*: *An Introductory Course*. New York: Routledge.

Gilquin, Gaëtanelle, Sylvie De Cock, Sylviane Granger. 2010. *Louvain International Database of Spoken English Interlanguage*. *Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, Gaëtanelle. 2015. "From Design to Collection of Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger, Gaëtanelle Gilquin, Fanny Meunier, 9–34. Cambridge: Cambridge University Press.

Granger, Sylviane. 1996. "From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora." In *Languages in Contrast. Text-based Cross-linguistic Studies*, edited by Karin Aijmer, Bengh Altenberg, Mats Johansson, 37–51. Lund: Lund University Press.

Granger, Sylviane. 2002. "A Bird's Eye View of Learner Corpus Research." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, edited by Sylviane Granger, Joseph Hung, Stephanie Petch-Tyson, 3–33. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.6.04gra.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier. 2002. *International Corpus of Learner English: Handbook and CD-ROM*, Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Jendryczka-Wierszycka, Joanna. 2009. "Collecting Spoken Learner Data: Challenges and Benefits. A Polish L1 Perspective." In *Proceedings of the Corpus Linguistics Conference, University of Liverpool, UK, 20–23 July 2009*, edited by Michaela Mahlberg, Victorina González-Díaz, Catherine Smith. Available at http://ucrel.lancs.ac.uk/publications/cl2009/230_FullPaper.doc (last accessed May 9, 2023).

Granger, Sylviane, Helen Swallow, Jennifer Thewissen. 2022. *The Louvain Error tagging Manual. Version 2.0. CECL Papers 4*. Louvain-la-Neuve: Université catholique de Louvain. https://cdn.uclouvain.be/groups/cmseditorscecl/Granger%20et%20al._Error%20tagging%20manual_v2.0_2022.pdf.

König, Alexander, Jennifer C. Frey, Egon W. Stemle, Aivars Glaznieks, Magali Paquot. 2022. "Towards standardizing LCR metadata." The 6th International Conference for Learner Corpus Research, Padova, 22.–24.9.2022.

Mayer, Mercer. 1969. *Frog, Where Are You*? New York: Dial Press.

Myles, Florence. 2021. "Commentary: An SLA Perspective on Learner Corpus Research." In *Learner Corpus Research Meets Second Language Acquisition*, edited by Bert Le Bruyn, Magali Paquot, 258–273. Cambridge: Cambridge University Press.

Ortega, Lourdes. 1999. "Planning and Focus on Form in L2 Oral Performance." *Studies in Second Language Acquisition* 21(1): 109–148. https://doi.org/10.1017/S0272263199001047.

Perotto, Monica. 2019. "Acquisizione e apprendimento linguistico degli *heritage speakers* russofoni della scuola N. Gogol' di Roma: ultimi sviluppi dell'indagine." In *Studi di linguistica slava. Nuove prospettive e metodologie di ricerca*, a cura di Iliyana Krapova, Svetlana Nistratova, Luisa Ruvoletto, 425–438. Venezia: Edizioni Ca' Foscari.

Ployhart, Robert E., Robert J. Vandenberg. 2010. "Longitudinal Research: The Theory, Design, and Analysis of Change." *Journal of Management* 36(1): 94–120. https://doi.org/10.1177/0149206309352110.

Polinsky, Maria, Olga Kagan. 2007. "Heritage Languages: In the 'Wild' and in the Classroom." *Language and Linguistics Compass* 1(5): 368–395.

Protassova, Ekaterina. 2016. *Narrative. Frog Stories in Russian: 41 Transcripts – Ages 5, 6, 7, 8, 9, 10, and Adult*. Software http://childes.talkbank.org/access/Frogs/Russian-Protassova.html (last accessed on May 9, 2023). https://doi.org/10.21415/T5SG69.

Rakhilina, Ekaterina V., Anastasia S. Vyrenkova, Elmira Mustakimova, Alina Ladygina, Ivan Smirnov. 2016. "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*, edited by Ele-

na Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam, Lars Borin, 66–76. Linköping: LiU Electronic Press.

Saturno, Jacopo. 2014. "Issues in the transcription of initial learner varieties." 47th Annual Meeting of the Societas Linguistica Europaea, Adam Mickiewicz University, Poznań, 11–14.09.2014.

Skehan, Peter. 1998. "Task-Based Instruction." *Annual Review of Applied Linguistics* 18: 268–286. https://doi.org/10.1017/S0267190500003585.

Stoyanova, Nataliya. 2020. "Diskursivnye elementy v russkoj reči ital'jancev: nekotorye zakonomernosti usvoenija." In *Systèmes linguistiques et textes en contraste*: *Études de linguistique slavo-romane*, dir. par Olga Inkova, Małgorzata Nowakowska, Sebastiano Scarpel, 357–374. Kraków: Wydawnictwo Naukowe Uniwersytetu Pedagogicznego.

Wang, Zhirong, Tanja Schultz. 2003. "Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization." In *Proceedings of the 8th European Conference on Speech Communication and Technology*, *1–4 September 2003*, 1449–1452. Geneva. Available at www.cs.cmu.edu/~tanja/Papers/Euro03-WangSchultz.pdf (last accessed May 9, 2023).

Zechner, Klaus. 2009. "What Did They Actually Say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous Speech Responses in an English Proficiency Test." In *Proceedings of the International Speech Communication Association International Workshop on Speech and Language Technology in Education* (*SLaTE*), *3–5 September 2009*. Warwickshire. Available at http://www.eee.bham.ac.uk/SLaTE2009/papers/SLaTE2009-09-v2.pdf (last accessed May 9, 2023).

Выренкова, Анастасия С., Мария С. Полинская, Екатерина В. Рахилина. 2014. "Грамматика ошибок и грамматика конструкций: эритажный (унаследованный) русский язык." *Вопросы языкознания* 3: 3–19.

Гришина, Елена А. 2005. "Устная речь в Национальном корпусе русского языка." В *Национальный корпус русского языка*: 2003–2005. *Результаты и перспективы*, под ред. Владимира А. Плунгян, 94–110. Москва: Индрик.

Кибрик, Андрей А., Вера И. Подлесская. 2009. *Рассказы о сновидениях. Корпусное исследование устного русского дискурса*. Москва: Языки славянских культур.

Перотто, Моника. 2013. "Два поколения русскоязычных в Италии: условия сохранения и утраты языка." В *Русский язык зарубежья*, под ред. Марии М. Ровинской, 237–257. Санкт-Петербург: Златоуст.