Сбор образцов живой русской речи для корпуса

Для чего это нужно?
Записи спонтанных разговоров являются ценным материалом для изучения разговорной речи – со всеми ее запинками, оговорками, нестандартным порядком слов, особенностями произношения, интонации и проч. Если корпуса литературного русского языка довольно обширны и разнообразны, то данных устной спонтанной речи, доступных исследователям, кот наплакал. В проекте **LiveCorpus** мы каждый год, силами лингвистов-первокурсников, собираем и аннотируем новые записи. Мы стараемся, чтобы в корпусе были представлены говорящие разного возраста, занятий и места жительства, хотя, конечно, в силу задачи, у нас больше говорящих-студентов.

Что делать?

Вариант 1 (Тема 2017 года: Комментарий видео)
Заранее подготовьте ролик с трейлером фильма, записью музыкальной группы, стримом компьютерной игры и т. п. Попросите говорящего прокомментировать то, что он видит. Говорящий не должен знать заранее, о чем будет этот ролик, чтобы эффект спонтанности сохранялся. Во время показа ролика вы можете задавать говорящему вопросы и вообще перевести разговор в режим диалога.

Чтобы не было наложения двух звуковых дорожек, для прослушивания ролика используйте наушники.

Вариант 2 (Стандартный-2)

Попросите своих старших родственников или знакомых рассказать о чем-то из своей жизни (например, рассказ бабушки о том, как вы были маленьким, как она ездила в колхоз на картошку и т.п.). Обсуждение современных тем (про политику, музыку, спорт) тоже возможно.

Частые ошибки

Не старайтесь записывать "гладкие" монологи людей с хорошо подвешенным языком – для наших целей это неинтересно. С другой стороны, диалоги вроде

```
"А чего говорить-то?" (пауза)  — "Ну скажи что-нибудь" (пауза)
```

тоже будут неудачны -- постарайтесь вначале разговорить людей, чтобы они перестали стесняться камеры и в идеале забыли про нее.

Качество записи
Видео обычно записывается с помощью камеры на телефоне или планшете, если есть видеокамера или профессиональный диктофон, их тоже можно использовать (диктофон – как дополнительное записывающее устройство).
Время звучания ролика - 5 минут (примерно).
Старайтесь, чтобы размер получившегося файла был не более 400-500 Мб (файлы объемом больше гигабайта трудно закачивать и обрабатывать).
Формат файла - .wav, .avi, .mpg (в принципе, любой формат, поддерживаемый вашим устройством, годится, но потом файлы придется конвертировать).

Старайтесь, чтобы запись была сделана не в шумном месте. Например, очень неудачны записи в кафе, ночном клубе, при сильном ветре – вы просто не сможете разобрать, о чем говорят собеседники. В кадре могут быть все собеседники, допустимо также, чтобы один из собеседников (снимающий ролик) был за кадром. Идеально, чтобы у говорящего были видны руки (жестикуляция), но портретная съемка тоже годится.

Сведения о говорящих

Соберите сведения о говорящих - пол, примерный возраст, образование (среднее, высшее, ниже среднего), профессия, регион проживания, регион, где родился. Вся информация должна быть занесена в эту таблицу:
https://docs.google.com/spreadsheets/d/1pyBAUsgsRbaNsgzV4eNY1P1sL2fS0Vh8y2_pVUWudzk/edit?usp=sharing

Использование материалов

Чтобы мультимедийные материалы могли быть помещены в учебный корпус, вы должны соблюдать правила научно-исследовательской этики:

• говорящий должен быть поставлен в известность, что его записывали
• говорящий должен дать согласие на то, что его видео появится в корпусе (на анонимных условиях, т. е. без указания имени и т.д.).
Если в разговоре участвуют двое или больше говорящих, то согласие должен дать каждый из них.

• любая информация, которая может нанести вред говорящему или другим, шифруется как в видеозаписи, так и в транскрипте (таковой могут быть: номера паспортов, резкие оценки других людей с называнием их имени и др.)
Расшифровывать видео будет тот, кто его записал. Обработанные материалы будут общедоступны.

Если говорящий не согласен на размещение видео, обсудите компромиссные возможности использования аудио- или даже просто письменного транскрипта речи в корпусе.
Пожалуйста, сразу пометьте себе "уровень согласия" говорящих там же, где вы записали сведения о них.

Я записал(а) ролик, что дальше?
Выложите файл с записью в облачное хранилище (дропбокс, Яндекс.Диск, Гугл-драйв и т.д.). Лучше заранее вырезать лишние фрагменты из записи, приведя ее к желательному объему в 5 минут.

Имя файла с записью может быть любым (в латинице), но желательно, чтобы оно было коротким (8-11 символов) и отражало тему разговора. В дальнейшем все файлы, которые вы будете создавать по проекту, должны начинаться так же, как имя файла с записью.

Вся работа по письменной расшифровке и улучшению ее качества, лингвистической аннотации будет проводиться в ходе дальнейших домашних заданий.

Если вы показывали говорящему другой видеоролик (рисунки или другие стимулы), также дайте ссылку на него в таблице (он должен быть доступен в вашем репозитории или где-то еще онлайн. Имейте в виду, что если файл размещали не вы, в один прекрасный день он может исчезнуть из интернета, поэтому лучше его скачать к себе, по возможности).

Если хотите, вы можете сделать первичную расшифровку видеозаписи. Полностью расшифровка потребуется в следующем домашнем задании по корпусу живой разговорной речи.

Примеры роликов прошлых лет

https://yadi.sk/d/DP7ZV5oIAgnrV

https://yadi.sk/d/15bpqjtIAgn9u

Чеклист:
***** Видео-файл, выложенный в облако.

* Заполненная таблица со сведениями о говорящих и ссылкой на репозиторий с записью.