Гуманитарии Политеха приняли участие в семинаре «Цифровые технологии в лингвистике: современные подходы к исследованию сложности текста»

13 октября в Гуманитарном институте состоялось важное событие: научно-образовательный семинар «**Цифровые технологии в лингвистике: современные подходы к исследованию сложности текста**».

Основную аудиторию слушателей составили студенты 1 курса магистратуры и бакалавриата Высшей школы лингводидактики и перевода, поступившие в 2022 году на новые программы подготовки магистратуры «Цифровая лингвистика и бакалавриата «Цифровые технологии и иностранные языки».

Основными докладчиками и содокладчиками стали преподаватели и студенты филологического факультета СПбГУ.

Блинова Ольга Владимировна, кандидат филологических наук, доцент кафедры общего языкознания СПбГУ, руководитель гранта РНФ “Понятность официального русского языка: юридическая и лингвистическая проблематика”.

Митрофанова Ольга Александровна, кандидат филологических наук, доцент

кафедры математической лингвистики филологического факультета СПбГУ, исполнитель грантов РНФ и НИП СПбГУ вместе со своими студентами: третьекурсником **Андреем Белым** и магистрантом 2 курса **Марком Атугодаге**.

В группу «поддержки» гостей также входили **Никита Тарасов**, аспирант кафедры технологии программирования СПбГУ и **Екатерина Владимировна Трощенкова**, доктор филологических наук, профессор кафедры английской филологии и лингвокультурологии СПбГУ.

После короткого приветствия собравшихся директором Высшей школы лингводидактики и перевода, профессором А.В. Рубцовой и награждения лучших первокурсников, студентов 1 курса бакалавриата, набравших по итогам сдачи ЕГЭ больше 270 баллов (таких на программу поступило 6 человек), модератор семинара, доцент Высшей школы лингводидактики и перевода М.С. Коган предоставила слово основному докладчику О.В. Блиновой.

В своем выступлении, «Невыносимая сложность юридических текстов и методы её измерения» (аллюзия на известный роман М. Кундеры и его экранизацию Ф. Кауфманом не случайна), Ольга Владимировна очень увлеченно и живо рассказала о том, что понимают под «сложностью текста», какие типы сложности текста выделяют, зачем ее нужно определять, как это делают современные исследователи, с какими

трудностями сталкиваются, как ищут решение возникающих лингвистических проблем.

Сложность текста – объективная характеристика, измеримая для любого связного текста на естественном языке. Она влияет на трудность его восприятия конкретным читателем или категорией читателей. Существует множество текстометрических ресурсов, созданных для оценки читабельности и сложности текстов на разных языках. У анализа сложности целый ряд практических приложений: это и оценка учебных текстов для изучающих язык как второй или носителей языка как родного, и оценка разнообразной документации, и оценка языкового контента веб-сайтов.

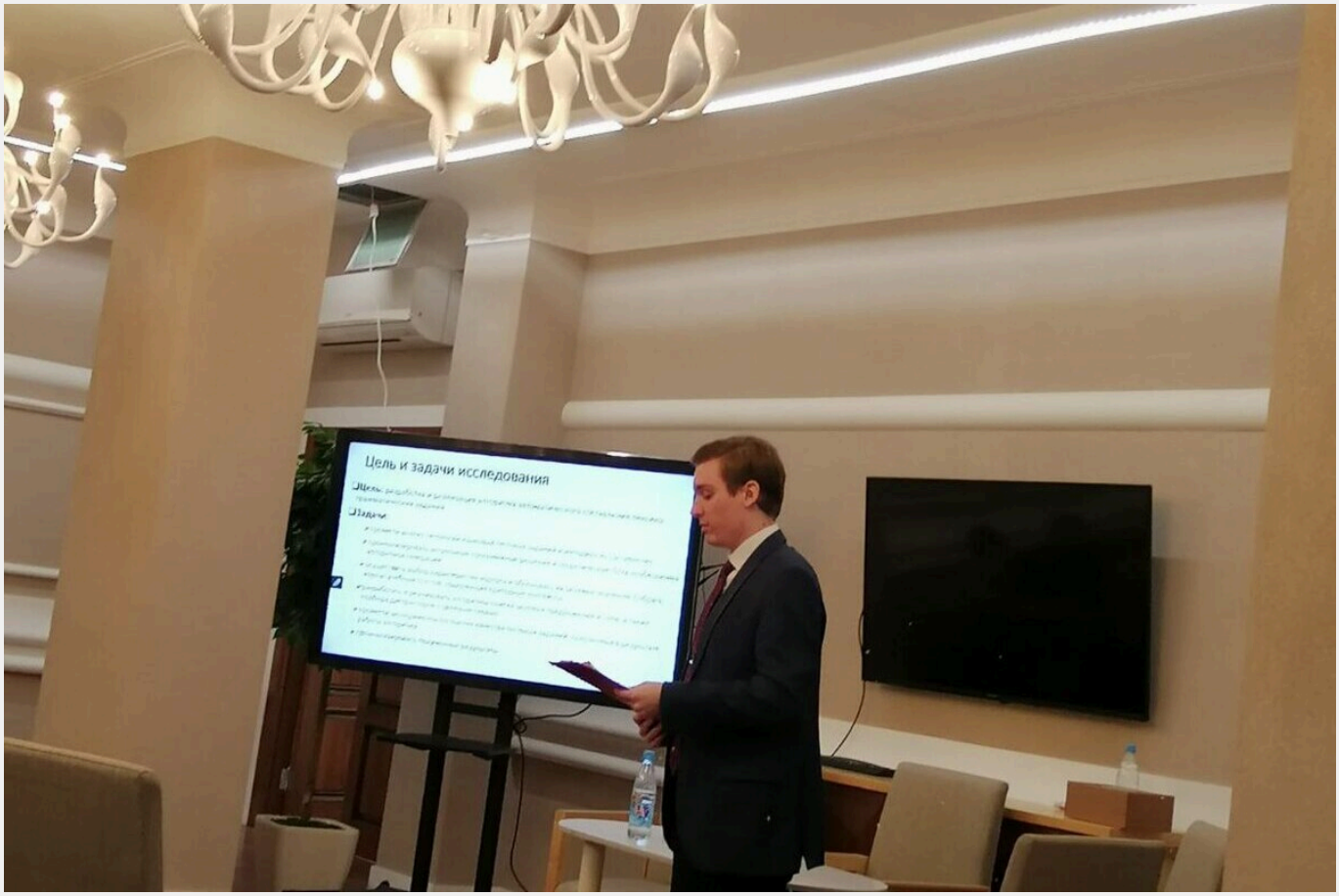
Исследовательские команды, интересующиеся сложностью текста, часто концентрируются на каком-то определённом типе текстов и задач, например, академических текстах — школьных учебниках (их анализом занимается группа из Казанского университета), учебниках по русскому языку как иностранному (такие тексты изучает группа из Института русского языка им. Пушкина); юридических текстах (Европейский университет в Санкт-Петербурге, СПбГУ) и пр. Изучение сложности текста актуально и в контексте международного «Движения к простому языку» (Plain Language Movement). Смысл движения в том, что в идеале официальный текст на языке должен быть понятен любому грамотному носителю языка. Трудно соблюдать законы, не понимая текстов законов. Между тем, юридический язык издавна критикуют за синтаксическую переусложнённость, многословие, избыточность, неоправданные повторы, архаичную лексику и т. п.

После введения слушателей в проблему Ольга Владимировна рассказала о результатах исследования группы ученых СПбГУ под ее руководством, поддержанного грантом РНФ, в рамках которого создается автоматическая модель оценки сложности русских юридических текстов. Технология включает создание представительного корпуса современных юридических документов, его морфо-синтаксичекую разметку, анализ и отбор лингвистических признаков (и основанных на них метриках), которые в модели используются для предсказания сложности (среди признаков есть и общеязыковые, и стилеспецифичные, то есть свойственные официально-деловым текстам). Наиболее эффективные для диагностики сложности текста метрики выявляются в результате тестирования модели. По мнению исследователей лучшие результаты дает Гибридный подход к оценке сложности (на данный момент модель протестирована на трех наборах юридических документов, результаты доступны на https://www.plaindocument.org/corpora).

Следует заметить, что обсуждение доклада проходило оживленно. Вопросы первокурсников, ошеломленных количеством и разнообразием метрик сложности

текста (для русского языка наибольшее количество метрик по данным литературы – 279, в работе О.В. Блиновой использовалось 130), касались этого аспекта; преподаватели Высшей школы лингводидактики и перевода задавали уточняющие вопросы по тренировке модели (при ответах на некоторые из них Ольга Владимировна воспользовалась «помощью друга» и коллеги, программиста Н. Тарасова); был даже задан вопрос, почему, по ее мнению, одну из Шнобелевских премий этого года (по литературе) вручили серьезной и актуальной работе исследователей из MIT, посвященной природе сложности юридических текстов (https://nplus1.ru/material/2022/09/16/ignobel-2022).

Опыт коллег из СПбГУ, описанный ими в ряде публикаций, несомненно, представляет собой научную ценность в плане решения исследовательских задач проекта «Цифровые технологии в лингвистике: модель автоматической оценки речевого воздействия мультимодального электронного текста», реализуемого в рамках проекта Приоритет 2030, над которыми совместно работают преподаватели Высшей школы лингводидактики и перевода и Лаборатории «Промышленные системы потоковой обработки данных» Центра НТИ СПбПУ.



2-й доклад представила О.А. Митрофанова совместно со студентами кафедры математической лингвистики СПбГУ. Он был посвящен вычислительным моделям упрощения текстов на основе опыта разработки для русскоязычных корпусов текстов.

Это своеобразный ответ на/решение проблемы использования сложных оригинальных текстов, например, в обучении иностранных студентов.

Существует несколько вариантов постановки задачи упрощения текстов. Одним из наиболее реалистичных подходов является упрощение текстов на уровне предложений, предполагающее снижение синтаксической сложности, фильтрацию словаря текста по низкочастотной, стилистически окрашенной, полисемичной лексике и т.д. Докладчик рассмотрел связь задачи упрощения и стандартных процедур автоматической обработки текстов, таких как суммаризация и перифразирование. Стандартные процедуры автоматической обработки текста основаны на использовании алгоритмов упрощения текстов, которые оцениваются с помощью существующих наборов данных; в последнее время активно и успешно используются нейросетевые алгоритмы обучения моделей упрощения. О формальном упрощении русскоязычных юридических текстов на основе машинного обучения рассказал Марк Атугодаге, занимающийся этой проблемой в рамках магистерской диссертации.

Последним на семинаре выступил А. Белый, который уже получил практически полезные результаты для преподавателей русского как иностранного (РКИ). Его работа направлена на генерацию лексико-грамматических заданий для ТРКИ с помощью предсказывающих языковых моделей. Андрей разработал и протестировал алгоритм автоматической генерации заданий со множественным выбором для ТРКИ, использующий предсказывающие статической модели распределенных векторов при генерации дистракторов. Генерируемые тестовые задания должны соответствовать определенным уровням владения русским языком как иностранным, поэтому обучение моделей производится с предварительной фильтрацией по лексическим минимумам. Проведенные эксперименты по оценке качества и пригодности генерируемых заданий доказывают высокую эффективность предложенных алгоритмов, хотя на данном этапе еще предполагают участие преподавателей по отбору лучших дистракторов и исключению наименее удачных или неверных. Поэтому задания, выдаваемые системой содержат избыточное количество дистракторов (5-6 вместо стандартных трех). Второй доклад также вызвал интерес у слушателей и привел к активному обсуждению, в ходе которого, задав ряд уточняющих вопросов, молодой сотрудник Высшей школы лингводидактики и перевода, недавняя выпускница кафедры математической лингвистики СПбГУ Д. А. Гаврилик предложила свои пути решения задач, над которыми сейчас работают начинающие исследователи ее родной кафедры.

В заключительном слове А.В. Рубцова поблагодарила гостей из СПбГУ за интересные доклады и выразила надежду на то, что они вдохновят наших студентов активно заниматься научно-исследовательской работой, начиная с первых месяцев обучения в СПбПУ. Предложение о продолжении подобных семинаров было встречено с

энтузиазмом и поддержано аплодисментами.

Материал подготовлен М.С. Коган, доцентом ВШЛиП,

руководителем образовательных программ Цифровые технологии и иностранные языки и Цифровая лингвистика (международная образовательная программа)