

Петербургские учёные разработали первую в России систему распознавания карельской речи

21.04.2025

Специалисты Санкт-Петербургского Федерального исследовательского центра РАН (СПб ФИЦ РАН) создали программный комплекс распознавания речи для карельского языка (ливвиковского наречия) по аудиоматериалам. Предложенная система направлена на документирование и сохранение карельского языка, но также может стать важным инструментом в области автоматизации лингвистических исследований культур коренных народов Российской Федерации. Результаты исследования опубликованы в научном журнале [«Информатика и автоматизация»](#).



Сегодня на карельском говорят около 30 тысяч человек по всему миру, порядка 25 тысяч из которых проживают в России. В повседневной жизни этот язык используют максимум 5–7 тысяч человек, что ставит его под угрозу исчезновения, отмечают исследователи.

«Изучение карельского языка имеет огромное значение не только для сохранения культурного наследия, но и для поддержки языковой традиции его носителей. Создание системы распознавания карельской речи — один из способов если не возродить, то хотя бы задокументировать этот язык», — рассказывает старший научный сотрудник лаборатории речевых и многомодальных интерфейсов СПб ФИЦ РАН **Ирина Кипяткова**.

Распознавание речи на таких малоресурсных языках — это особое направление компьютерной лингвистики, в рамках которого разрабатываются методы, позволяющие добиться хорошей точности при сравнительно небольшом количестве обучающих ресурсов (текстов, аудиозаписей, словарей).

Обучение нейросетей проводилось на базе данных карельского языка, в основу которой легли материалы карельских периодических изданий, текстов на ливвиковском наречии из открытого корпуса вепсского и карельского языков, а также ряда других открытых письменных источников. Кроме того, учёные собрали два набора данных разговорной речи. Первый — на основе материалов радиопередач «Родной берег» ГТРК «Карелия». Второй был специально собран в ходе записи карельской речи, содержащей включения русскоязычных слов. Материал обработан участниками проекта, владеющими ливвиковским наречием карельского языка.

«Кроме нашей системы, в мире есть лишь одна модель, поддерживающая карельский язык. Однако она является многоязычной, её обучение базировалось на очень небольшом объёме данных, что существенно снижает точность распознавания именно карельской речи. Мы же учитываем реальные условия использования карельского языка, где часто встречается переключение между карельским и русским. Многие носители карельского в повседневной речи смешивают два языка, и мы обучаем нашу систему корректно обрабатывать такие случаи», — отмечает Ирина Кипяткова.

Разработка в перспективе может применяться для автоматического стенографирования речи на карельском языке, например, для расшифровки аудиоархивов и устной речи, в том числе для лингвистических исследований, что должно способствовать сохранению карельского языка и поможет упростить создание лингвистических корпусов для его дальнейшего исследования. Результаты в том числе могут быть использованы для разработки ассистивных систем, систем машинного перевода, анализа медиаконтента.

Проект поддержан грантом РНФ ([№ 24-21-00276](#)). В состав участников проекта вошли сотрудники Карельского научного центра РАН.

Источник: пресс-служба СПб ФИЦ РАН.

[Новости Российской академии наук в Telegram ➔](#)