**УДК 81`322.5**

# РАЗРАБОТКА И ОЦЕНКА СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ И ТРАНСЛИТЕРАЦИИ ТЕКСТОВ НА КЛАССИЧЕСКОМ ТИБЕТСКОМ ЯЗЫКЕ

**Мурашкина А.В.** (НГУ)
**Научный руководитель – доктор технических наук, в.н.с. ИВМиМГ СО РАН, Барахнин В.Б.** (НГУ)

**Введение.** Старопечатные документы на тибетском языке — ценное историко-культурное наследие буддизма, веками передаваемое по наследству народами Тибета. Эти рукописи и ксилографы содержат уникальные сведения о философии, религии, медицине, истории и искусстве, играя ключевую роль в изучении культурных традиций региона. Однако со временем, под воздействием природных и антропогенных факторов, бумажные носители подвержены физическому разрушению, что ведёт к утрате информации и ограничивает доступ к этим материалам [1].

Одним из наиболее надёжных способов сохранения и систематизации исторических документов является их оцифровка [2]. Процесс включает не только создание цифровых изображений, но и разработку методов автоматического распознавания и транслитерации тибетского текста, что позволит упростить доступ к этим данным, повысить эффективность их анализа и интеграции в современные информационные системы.

**Основная часть.** В Тибетском фонде рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН находится около 70000 единиц хранения, требующих оцифровки. В рамках сотрудничества Института вычислительных технологий СО РАН и ИМБТ СО РАН поставлена цель создания программного комплекса, способного автоматически обрабатывать тексты этих документов.

**Выводы.** Настоящая работа посвящена исследованию, разработке и практической реализации программного комплекса по распознаванию и транслитерации текста со старопечатных тибетских документов. Для этого были проанализированы существующие open-source решения, включая методы оптического распознавания символов (OCR), модели на основе глубокого обучения и алгоритмы транслитерации. Для объективной оценки их качества разработан набор размеченных данных, включающий изображения оригинальных текстов и их корректные текстовые представления. Важной частью исследования является разработка системы оценки точности распознавания. Опираясь на прецеденты постановки подобных задач [3], предложен набор метрик, адаптированных к специфике работы с языковым материалом.

**Список использованных источников**:

1. Lamu Y. Protection status of domestic ancient Tibetan manuscripts and literatures // J. Ethnol. 2012. Vol. 3. N. 6. P. 54–58.
2. Ma L., Long C., Duan L. et al. Segmentation and recognition for historical Tibetan document images // IEEE Access. 2020. Vol. 8. P. 52641–52651.
3. Beshirov A. и др. Post-OCR Text Correction for Bulgarian Historical Documents // 2024.