Оцифровка и обработка печатных материалов на малоресурсном языке: проблемы и решения

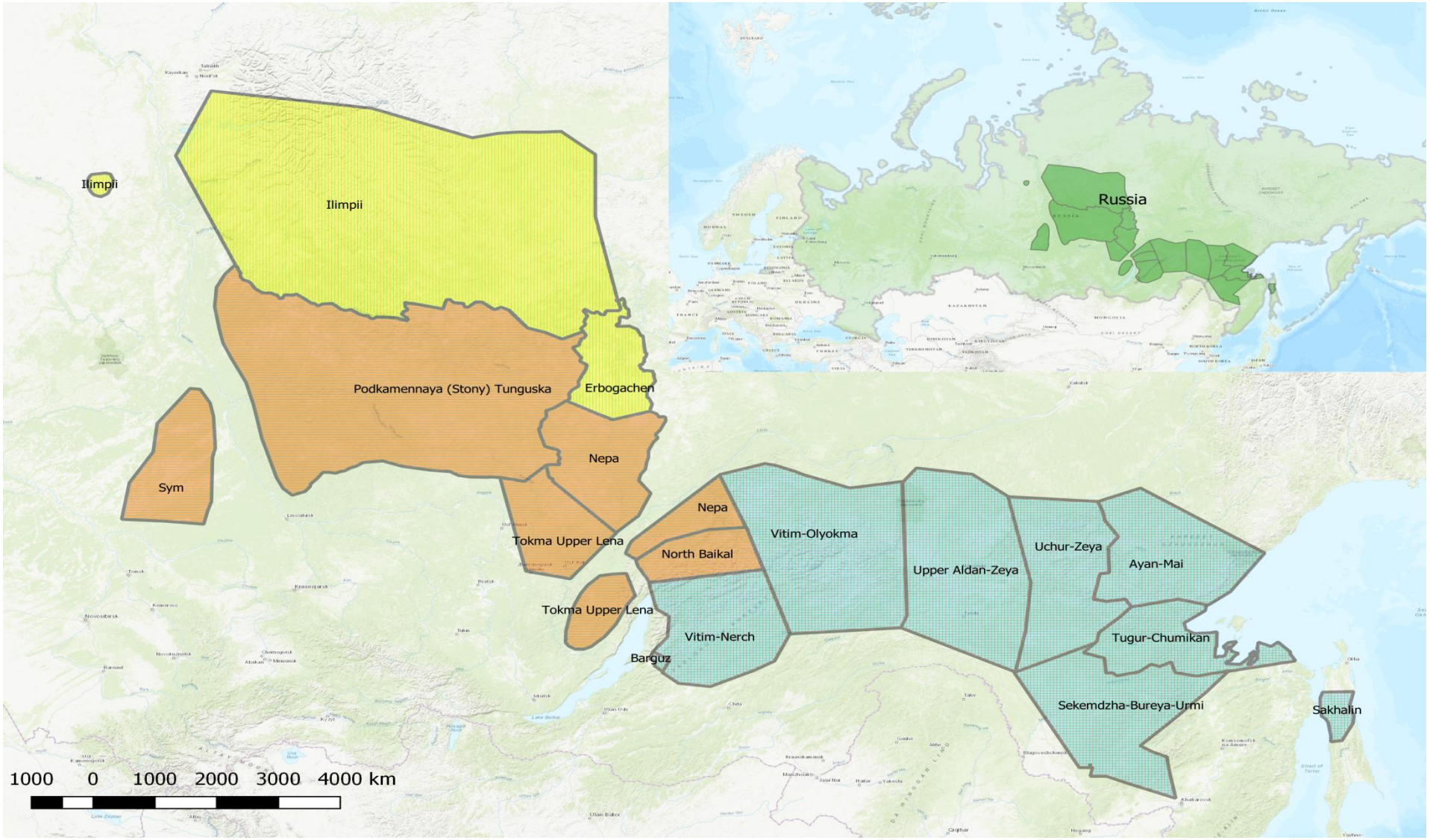
«От языковых машинных фондов к лингвистическим корпусам: памяти В.М. Андрющенко».

Москва, 2018
Е. Клячко

Содержание

- Малоресурсные языки
  - Эвенкийский язык
- Что даст оцифровка?
- Издания на эвенкийском языке
- Проблемы и решения

Эвенкийский язык



Г. М. Василевич «Очерки диалектов эвенкийского языка», 1948 г.,

Переработка Н. А. Мамонтовой

Ресурсы на эвенкийском

- http://www.evengus.ru/ и http://evenkiteka.ru/

Цифровые словари, книги

- Корпуса ИЭА РАН

http://corpora.iea.ras.ru/corpora

(полевые записи различных наречий + тексты из «Библии для детей» и газеты «Эвэды ин» («Эвенкийская жизнь»))

- Корпус ЛАЛС НИВЦ МГУ

http://siberian-lang.srcc.msu.ru/ru/textspage

(полевые записи различных наречий, в основном северного, мультимедийные записи)

Что даст оцифровка?

- Большие словари
- Пополнение корпуса
- Архивные данные, в том числе и по исчезающим диалектам
- Переиздание литературы
- Использование в учебных пособиях

Диалекты эвенкийского языка и литературный язык

- Северное, южное и восточное наречие
- Литературный язык на базе южных говоров
- Неоднозначное отношение носителей эвенкийского и лингвистов к литературному языку

- Язык изданий:
  - Диалекты
  - Литературный

Издания на эвенкийском языке


NAUMOW
DUKUCAN, OJOCON
BI ON MOTIWA WAM
OGIZ MOLODAJA GWARDIJA 193
МУННГИ ОТРЯДВУН
A. SALATKIN
AGI
ƏWIƷƏRƏN
ГАРПАНИНДЯ
ИЗДАТЕЛЬСТВО ГЛАВСЕВМОРПУТИ
СЕВЕРНАИЛ НГИНАКИР
САМАСЭЛВЭ ХУСКИ
И. Сталин
Советский Союз
войнами
РАНЬШЕ и
№ 1
K. P. SCUKIN
On Rossija hawammilin, krestjanisalin mərduwer tiamukitwə ƨawacatin
LENPARTIZDAT 1933

Проблемы

- Графическая вариативность:
  - Алфавиты (латиница до 1937 г., кириллица)
  - Орфографические различия
  - Выбор шрифтов
- Отсутствие / малое количество готовых инструментов:
  - OCR
  - словари

Инструменты OCR

	Abbyy Fine Reader	Cuneiform	tesseract
Поддержка эвенкийского языка	(кириллический) алфавит, без словарной поддержки	Нет	нет
Возможности обучения	Свой словарь, ручное обучение	? (не документированы)	Обучение модели
Цена	Коммерческий продукт	Open source	Open source
Порог вхождения	Низкий	?	Высокий

Попытка работы с tesseract

- https://github.com/lalsnivts/evenki_ocr_texts
- Сложности в подготовке модели

Подготовка словаря для распознавания (Abbyy)

Список словоформ

- словари http://www.evengus.ru/slovari/ + морфоанализатор в режиме генерации
- скачивание текстов газеты «Эвенкийская жизнь» на эвенкийском языке:
  - обход web-страниц
  - исключение текстов на русском языке
  - токенизация

➔более 80 тысяч словоформ

➔качество распознавания улучшилось

Морфологический анализ

- foma: finite-state transducer
- Словари evengus
- правила на основе грамматик эвенкийского языка

https://github.com/gisly/evenkiMorph

➔

Малый объем словаря

Плохое качество на «ненормализованных» текстах

Результаты

- Большая доля ручной работы
- С каждым новым оцифрованным словарем/текстом ручной работы становится меньше