AI for Repair of Multilingual Records incorporating compound transliteration and diacritic restoration

Presenter: Jennifer Proctor
The Problem:

In Micro

THE MUSIC IN THE BOX

Language complicates search

43. Юмореска

В. ГЕВИКСМАН

Не очень быстро, шутливо, грациозно
*But wouldn’t the Cyrillic work if the record was just complete?
A Problem with Many Roots

Human Factors

- Different priorities - ex: Russian libraries vs cataloguers who know no Cyrillic
- Different institutional and national norms leading to inconsistent data entry

Technology

- Inconsistent retention of diacritics, non-Roman characters, and stress marks
- Programs misinterpret characters like stress marks
- Transliteration is not 1-to-1
Pre-processing, data transformation, Machine Learning, and Statistical Analysis were combined so that data (& search results) will be more predictable in future.
4.5 million records labeled (OpenRefine, Python)
Text into Numbers
Computers only speak Math

- **HASHING**
  Letters and words become numbers

- **WORD VECTORS**
  What specific number is influenced by relationships between words

- **NOMINALIZATION**
  Encoding Text Categories as Unique Numbers with a code book
While dozens of ML algorithms tested and scalability options were explored (b/c cannot allocate 33 TB RAM errors), ultimately the horizontal scalability issue made
Error Catching And Correcting!

**NGRAMS**

Breaking apart transliterated text into ngrams was way more complicated than normal because 1 Cyrillic letter could equal as much as 6 or as few as 1 transliterated character (which may or may not be a letter)

**FREQUENCIES**

Presence of common typos can help QCers hone in on error locations

**SUBSTITUTIONS**

By matching some common typos to their best match in the actual Russian language frequencies, some replacements can be made automatically and generally once this process is done, a good context-aware Russian spellchecking tool is able to sort out the rest
Final Project Pipeline

Data File Formatting (BIG Data) → Text into Hashing Vectorizer → Nominals to numeric via Codebook → Apply Model → Statistical Analysis on flagged

TRAIN: random forest calibrated with sigmoid linear regression

Scripting n-gram splits w/ diacritics, compounds, etc

Finding bad n-grams, replacing w/ good
A Shadow Problem: Where did the of Diacritics go?

Hidden deep inside the software we use there is a set of assumptions - valuations - about what languages, and therefore what letters, should be included. This goes far beyond ensuring you can understand the GUI - it impacts what data is saved and how, and what happens to data that is NOT saved.
Today, the Alabama Language is spoken by less than 100 people, largely on the Alabama-Coushatta Reservation in Texas.

This Alabama-English dictionary contains meanings and pronunciations of tens of thousands of words and phrases digitized to make them more accessible to Indigenous Language Learners. Unfortunately, export from the transcription software erases many of the diacritic characters.