Overview

• CollateX default matching
• Why you may want to override it
• How to override it

CollateX default matching

• Exact string matching
  – Near matching
• Tokenize by splitting on white space
• Punctuation marks are individual tokens
• No case normalization
• No Unicode normalization

Sample normalization overrides

• Case folding
• Unicode normalization (precomposed characters)
• Strip punctuation
• Strip markup

Soundex

• English-language surnames, 1918
• Algorithm (simplified)
  – Retain first letter
  – Delete other vowels; degenerate
  – Conflate other letters according to phonetic similarity
    (e.g., t/d = 3; m/n = 5)
  – Truncate or zero-pad to four characters
• Examples
  – Birnbaum B-651 (also ✓ Barenboim; also X Brumble)

Soundex assumptions

• More nuanced than generic edit distance
• Character differences are not all equivalent with respect to information load
  – Consonants carry more information than vowels
• Information load may be sensitive to position
  – Beginning of word carries more information than end
  – Especially true for lexical (not morphological) searching in inflected languages
Adapting Soundex to Church Slavonic

- Neutralize variant spellings of initial vowel
  - χ, γ, ξ = γ
  - ω, ια, ιε, ιη = ι
- Case fold, neutralize consonantal variants
  - Not always one-to-one, e.g., υ – υ
- Degeminate, delete other vowels, delete diacritics
  - Keep two letters of two-letter words
    - Higher information load
- Other conflations?
  - Knowledge based vs machine learning
- Expand abbreviations?
  - Ира, сф, Иа = сера (И)
- Truncate or zero-pad (to what length?)

Soundex sample

- Ch397 и възвратитъ дщерьше своєе.
- Ch384 и възвратитъ дщерьшу своё.
- Nbkm298 и възвратитъ братаницы своё.
- Berlin и възвратитъ братаниць своє.
- Ch397 и взяр дштр св
- Ch384 и взяр дштр св
- Nbkm298 и взяр бртн св
- Berlin и взяр бртн св

Two types of normalization

- Collation
  - Find alignment points
  - Coarse adjustments
  - No harm in conflating, e.g., imperfect and aorist or infinitive and supine
- Evaluation
  - Alignment points are already known
  - Finer comparisons
  - Many need to distinguish on the basis of small details

Collation after Soundex

- Greatly improved results
- Utilize forced matches
  - A B C
  - A D C
- Misses
  - Gap in alignment (no forced match)
  - Imperfect match
    - фра “фрац”
  - CollateX recognizes only perfect matches
    - Unable to recognize closest match (until last week!)

3.5

<table>
<thead>
<tr>
<th>3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lav</td>
</tr>
<tr>
<td>Tro</td>
</tr>
<tr>
<td>Rad</td>
</tr>
<tr>
<td>Akk</td>
</tr>
<tr>
<td>Ipa</td>
</tr>
<tr>
<td>Xle</td>
</tr>
<tr>
<td>Byc</td>
</tr>
<tr>
<td>Bsc</td>
</tr>
<tr>
<td>Ssc</td>
</tr>
<tr>
<td>Ssc</td>
</tr>
<tr>
<td>С</td>
</tr>
</tbody>
</table>