Working with Russian Queries for the GIRT, Bilingual and Multilingual Tasks of CLEF-2001

- Fredric Gey, Hailing Jiang and Natalia Perelman, University of California, Berkeley
- (with special thanks to Aitao Chen for system support)

- Working with Russian
  - Cyrillic character set and representation
  - MT resources available
  - Transliteration
  - Finding untranslated words
- Results for bilingual and multilingual task
- GIRT task and special resources
- GIRT results
- Summary and future directions

FEATURES OF WORKING WITH RUSSIAN

- Unfamiliar alphabet (Cyrillic)
  - But phonetically European in nature
- Multiple representations (Windows, KOI-8, Unicode)
- Fewer resources than with Western European languages
  - Reasonable MT from Russian to English
  - Relatively poor MT from Russian to German
- Can work with transliterated (Romanized) representation
  - (U.S. Library of Congress standard)
  - Sometimes gives 1-1 to English for same word
  - We transliterated all queries as well as the GIRT Thesaurus

Berkeley CLEF Bilingual and Multilingual Participation

- **Bilingual experiments:** Russian → English, with German → English as a baseline
  - Russian → English using PROMPT Translator (http://www.translate.ru/) BKBIREA1, then with manual augmentation using transliteration of untranslated words (BKBIREM1)
  - German → English using L&H Power Translator, with manual augmentation using special association dictionary between German and English (BKBIGEM1)
- **Multilingual experiments:** take English translations from Bilingual experiments and retranslate to French, German, Italian, Spanish using L&H Power Translator (BKMUGEM1, BKMUREA1). Use English → all languages as a baseline (BKMUEAA1/2)

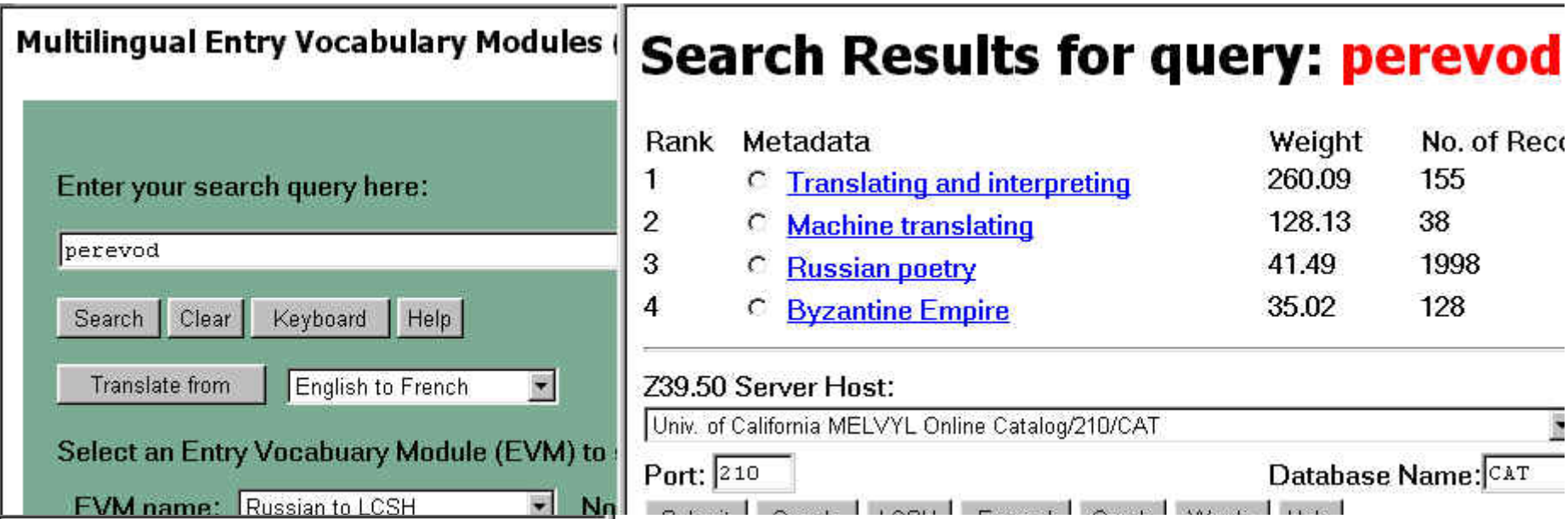
CLEF-Bilingual Russian Transliteration of Topic 50

- <num> C050 </num>
- <RU-title> Восстание в Чиапас </RU-title>
- <TR-title> Vosstanie v Chiapas </TR-title>
- <RU-desc> Найти статьи о восстании индейских крестьян в Чиапас (Мексика) </RU-desc>
- <TR-desc> Naiti stat'i o vosstanii indeiskikh krest'ian v Chiapas (Meksika) </TR-desc>
- <RU-narr> Документы излагают причины и развитие восстания индейского населения в Чиапас. Они также могут прокомментировать реакцию мексиканского правительства </RU-narr>
- <TR-narr> Dokumenty izlagaiut prichiny i razvitie vosstaniia indeiskogo naseleniia v Chiapas. Oni takzhe mogut prokommentirovat' reaktsiiu meksikanskogo pravitel'stva </TR-narr>
- Precision from 0.465 to 0.690 by adding this untranslated word!

BERKELEY STATISTICAL ASSOCIATION DICTIONARY PROTOTYPE

- Created from the University of California digital library catalog
  - Private copy, 10 million+ records (5 million non-English)
- Records in over 100 languages
- Obtained in MARC database standard format
- Foreign language titles use Library of Congress transliteration (Romanization) standard
- Search software maps between English Subject headings and
  - Arabic, Chinese, French, German
  - Italian, Japanese, Russian, Spanish
- Available at http://otlet.sims.berkeley.edu/mulevm2.html

Foreign search words can be mapped to English subject headings



Russian-German bilingual dictionary entries:

Perevod – Ubersetzung | Ubersetzung -- Perevod

German Manual Augmentation for Untranslated Words

- **Take untranslated words and submit to Berkeley statistical association dictionary prototype**
- **Add highest ranking associated English subjects to query**
- **Examples:** ***<num>C088</num>***
- ***<DE-title> beef craziness in Europe </DE-title>***
- ***<DE-desc> search documents, the cases of bovine Spongiformer,***
- ***Enzephalopathie (encephalopathy) (the beef craziness) in Europe describe. </DE-desc>***
- ***<num>C050</num>***
- ***<DE-title> rebellion in Chiapas (indians of mexico) </DE-title>***
- ***<DE-desc> reports about the revolt by Indios (indians, treatment of) in Chiapas (Mexico) is sought. </DE-desc>***

CLEF- Bilingual and Multilingual Results

- **BILINGUAL**
- BKBIGEM1 0.5088
- BKBIREA1 0.4077
- BKBIREM1 0.4204
- CLEF BL avg 0.2423

- **MULTILINGUAL**
- BKBIEAA1 0.2674
- BKBIEAA2 0.3101
- BKMUGAM1 0.2902
- BKMURAA1 0.1838
- CLEF MU avg 0.2749
- CLEF2000 avg 0.1843

CLEF- GIRT subtask Russian-German retrieval

- The GIRT collection consists of abstracts of reports and papers (grey literature) in the social science domain from Information Sozialwissenschaften, Bonn/Berlin
- 76,128 German Documents, most with English titles as well as German, some with English text – we indexed German text only
- A multilingual thesaurus available: English-German-Russian
- Almost all the documents in the collection have been manually assigned thesaurus terms. On average, there are about 10 terms assigned to each document. Keywords from documents were not used because rules would make all such use of run-type "manual"
- All Russian queries were transliterated for matching against the transliterated GIRT thesaurus

Example GIRT Russian Query

- <num> GIRT027 </num>
- <RU-title> Европейская миграция с востока на запад </RU-title>
- <TR-title> Evropeiskaia migratsiia s vostoka na zapad </TR-title>
- <RU-desc> Поиск документов о миграции с востока на запад в Европе </RU-desc>
- <TR-desc> Poisk dokumentov o migratsii s vostoka na zapad v Evrope </TR-desc>
- <RU-narr> Документы описывают существующую и ожидаемую миграцию из восточной в западную Европу, а также проблемы, которые она уже вызвала или которые ожидаются как ее последствия. Исторические описания и сообщения не являются релевантными. </RU-narr>
- <TR-narr> Dokumenty opisyvaiut sushchestvuiushchuiu i ozhidaemuiu migratsiiu iz vostochnoi v zapadnuiu Evropu, a takzhe problemy, kotorye ona uzhe vyzvala ili kotorye ozhidaiutsia kak ee posledstviia. Istoricheskie opisaniia i soobshcheniia ne iavliaiutsia relevantnymi. </TR-narr>
- <EN-title> European East-West Migration </EN-title>
- <EN-desc> Find information on East-West migration trends in Europe. </EN-desc>

GIRT Russian→German Query translation

- GIRT thesaurus transformed into a transfer dictionary
  - Used Russian-German part of the GIRT thesaurus
- PROMT web translator (Russian German version)

GIRT Thesaurus Entry and Transfer Dictionary Entry

- <entry>
- <german>Wirtschaftspolitik</german>
- <russian>экономическая политика </russian>
- <translit>ekonomicheskaia politika </translit>
- </entry>

- экономический строй Wirtschaftsordnung
- экономическая педагогика Wirtschaftspädagogik
- планирование экономики Wirtschaftsplanung
- экономическая политика Wirtschaftspolitik
- налоговый инспектор Wirtschaftsprüfer

GIRT Thesaurus lookup

- Exact matching: only 50 out of 1300 terms directly found in the thesaurus
- Why so low?
  - Different word forms

    e.g. “evropa” is in the thesaurus, but "evrope" and "evropu" are not.
- How to improve?
  - Morphological analyzer: not available
  - Fuzzy matching

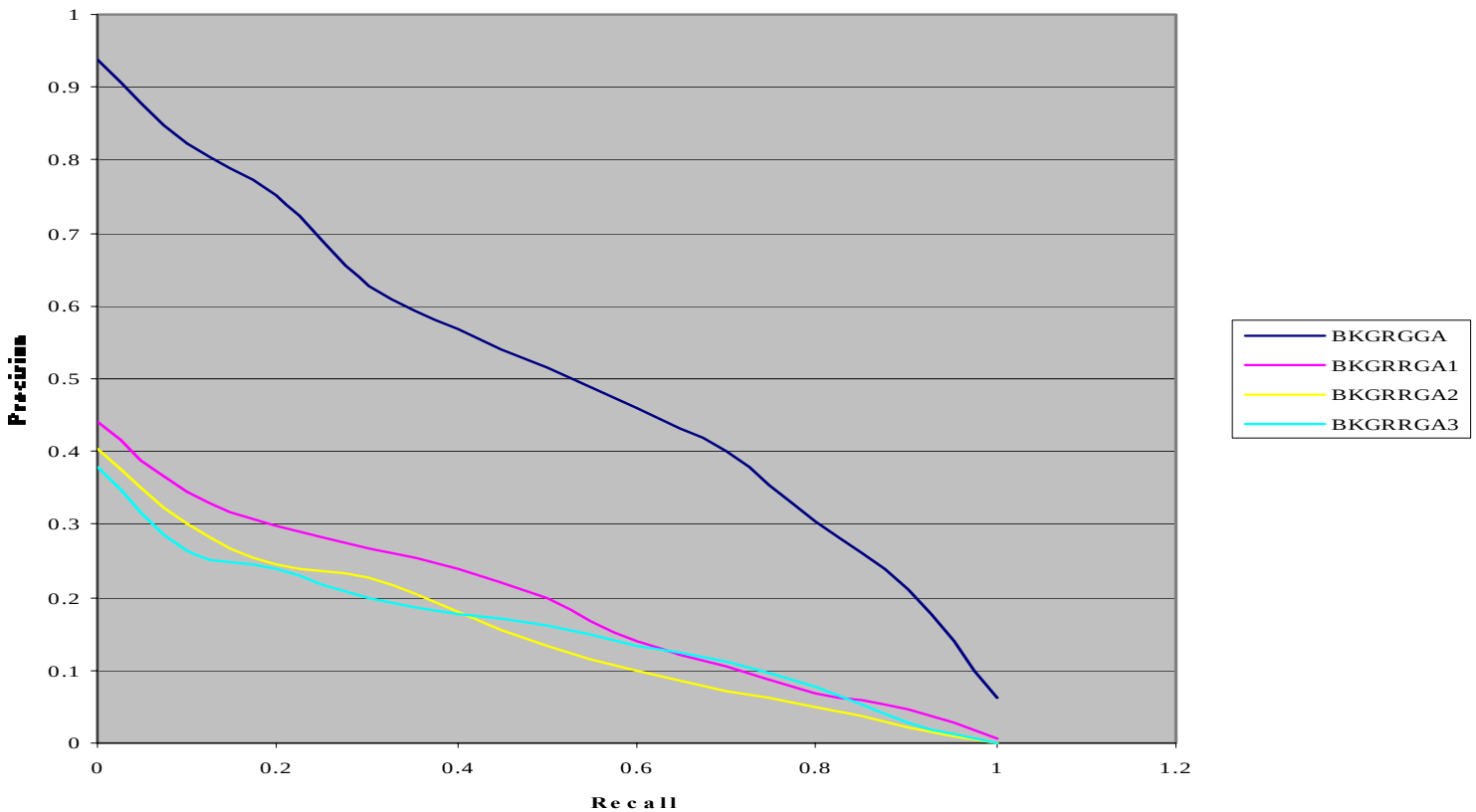
Examples of words and phrases found by fuzzy matching

Russian query terms	Russian Thesaurus terms	German Thesaurus Translation
migratsiiu	migratsiia	wanderung
Bezrabotitsei	bezrabotitsa	arbeitslosigkeit
televideniia	televidenie	fernsehen
kul'turu	kul'tura	kultur
Tekhnologicheskogo razvitiia	tekhnologicheskoe razvitie	technologische entwicklung
razvitie i organizatsiia	organizatsionnoe razvitie	organisation-sentwicklung

Berkeley's Official GIRT runs

- **BKGRGGA: German monolingual run (baseline)**
- **BKGRRGA1: Russian-German bilingual run using MT system for query translation**
- **BKGRRGA2: Russian-German bilingual run using thesaurus lookup and fuzzy matching.**
- **BKGRRGA3: identical to BKGRRGA2 except that only title and desc sections used**

Results of GIRT official runs



Lessons learned and recommendations

- Russian was a challenging query language for both the CLEF main tasks and the GIRT domain specific task
- For Russian→German our best performance was only about 36 percent of German → German monolingual – a lot of work remains to be done
- External statistical association dictionaries (from library catalogs) show promise for words not translated by traditional sources
- Transliteration and fuzzy matching also show promise for words not translated
- The GIRT domain specific task should be released of the restriction not to use index keywords from documents