# CLEF-2005 Cross-Language Speech Retrieval Track Overview

UMD: **Douglas Oard**, Dagobert Soergel, Ryen White,
Kim Braun, Xiaoli Huang, Craig Murray,
Scott Olsson, Jianqiang Wang
DCU: Gareth Jones
IBM: Bhuvana Ramabhadran, Martin Franz
VHF: Sam Gustman

# Why Cross-Language Speech?

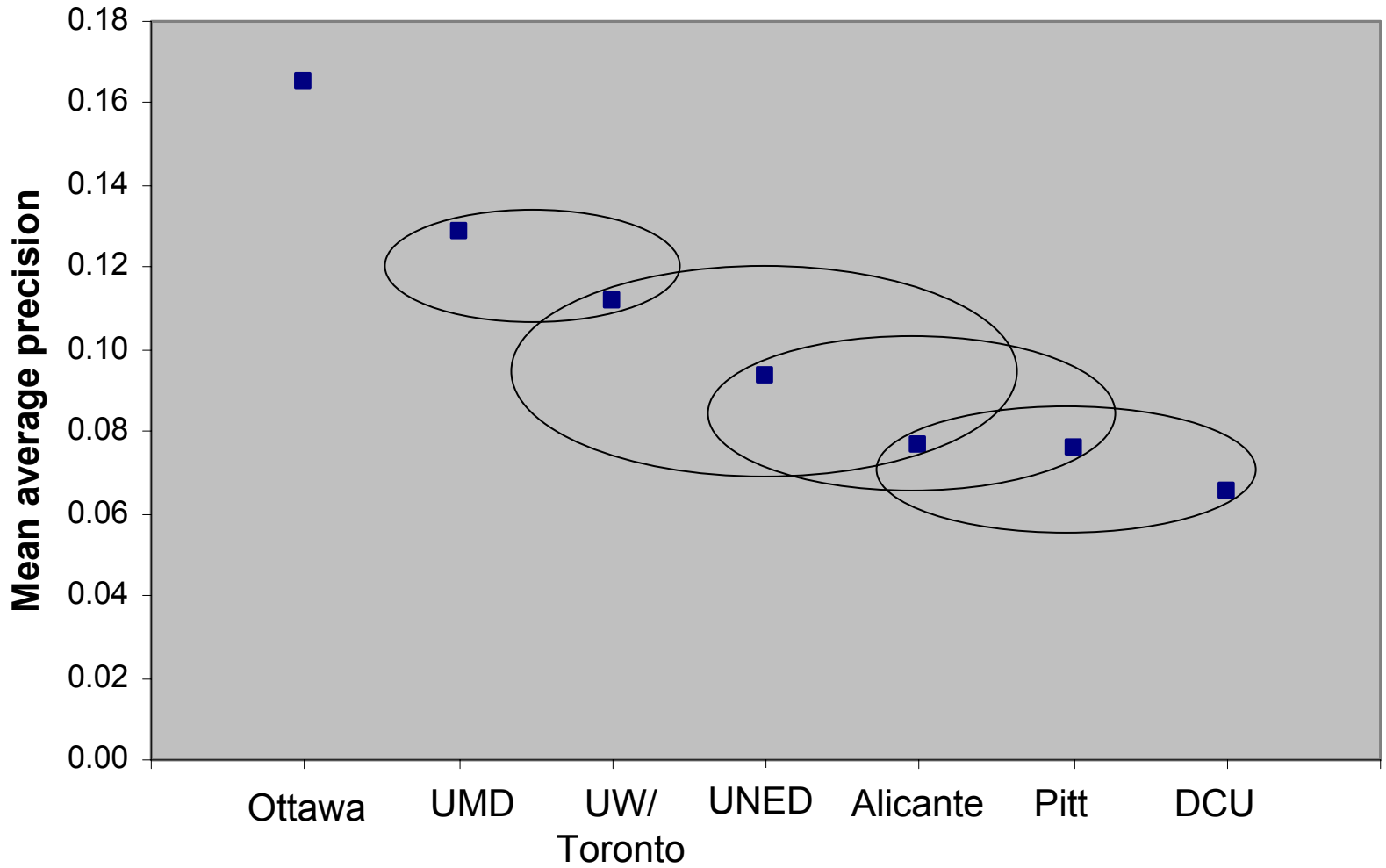
- Good News: We can search news well
- Bad News: News is a niche market
- Words spoken/day > present Google index
  - Words spoken/CLEF break > 1 CLEF paper
- Just 1 language has >10% of the speakers
  - And just 10 people in this room can speak it

# CLEF-2005 CL-SR Participants

- 7 teams / 4 countries
  - Canada: **Ottawa**, Waterloo
  - USA: Maryland, Pittsburgh
  - Spain: Alicante, **UNED**
  - Ireland: DCU

- 5 “official” runs per team (35 total)
  - Baseline required run: ASR / English TD topics

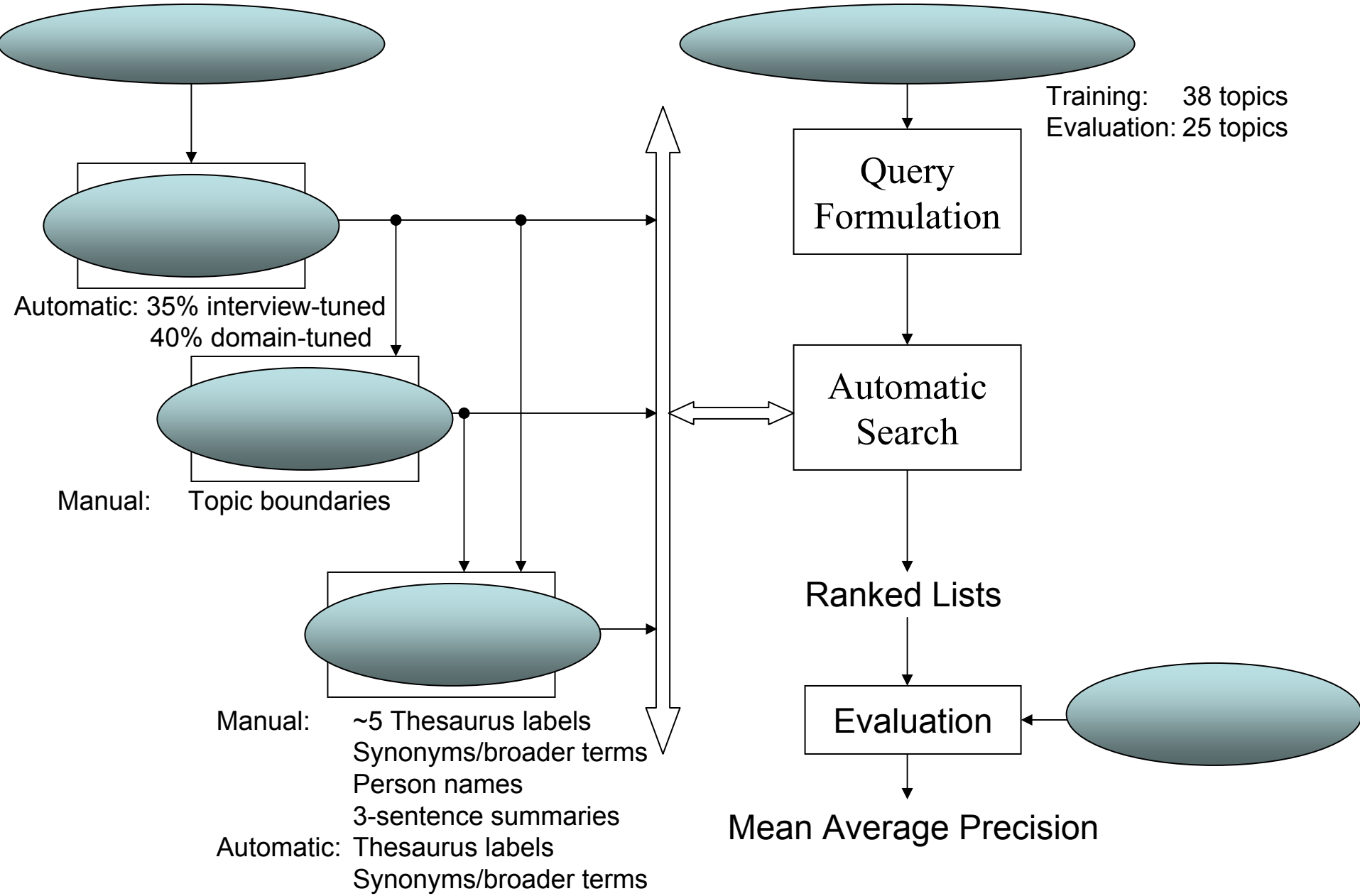
# TD Queries / English Automatic



# CLEF-2005 Contrastive Conditions

| Site (query/document) | MAP English | ΔMAP Czech | ΔMAP German | ΔMAP French | ΔMAP Spanish |
|---|---|---|---|---|---|
| Ottawa (TD/ASR04,AK1,AK2) | 0.1653 | | | + 2% | |
| Ottawa (TDN/ASR04,AK1,AK2) | 0.2176 | | - 41% | | - 14% |
| Maryland (TD/NAMES,KW,SUM) | 0.3129 | | | - 21% | |
| Waterloo (T/ASR03,ASR04) | 0.0980 | - 52% | | - 13% | |
| UNED (TD/ASR04) | 0.0934 | | | | - 60% |
| DCU (T/ASR03,ASR04,AK1,AK2) | 0.1429 | | | + 16% | |

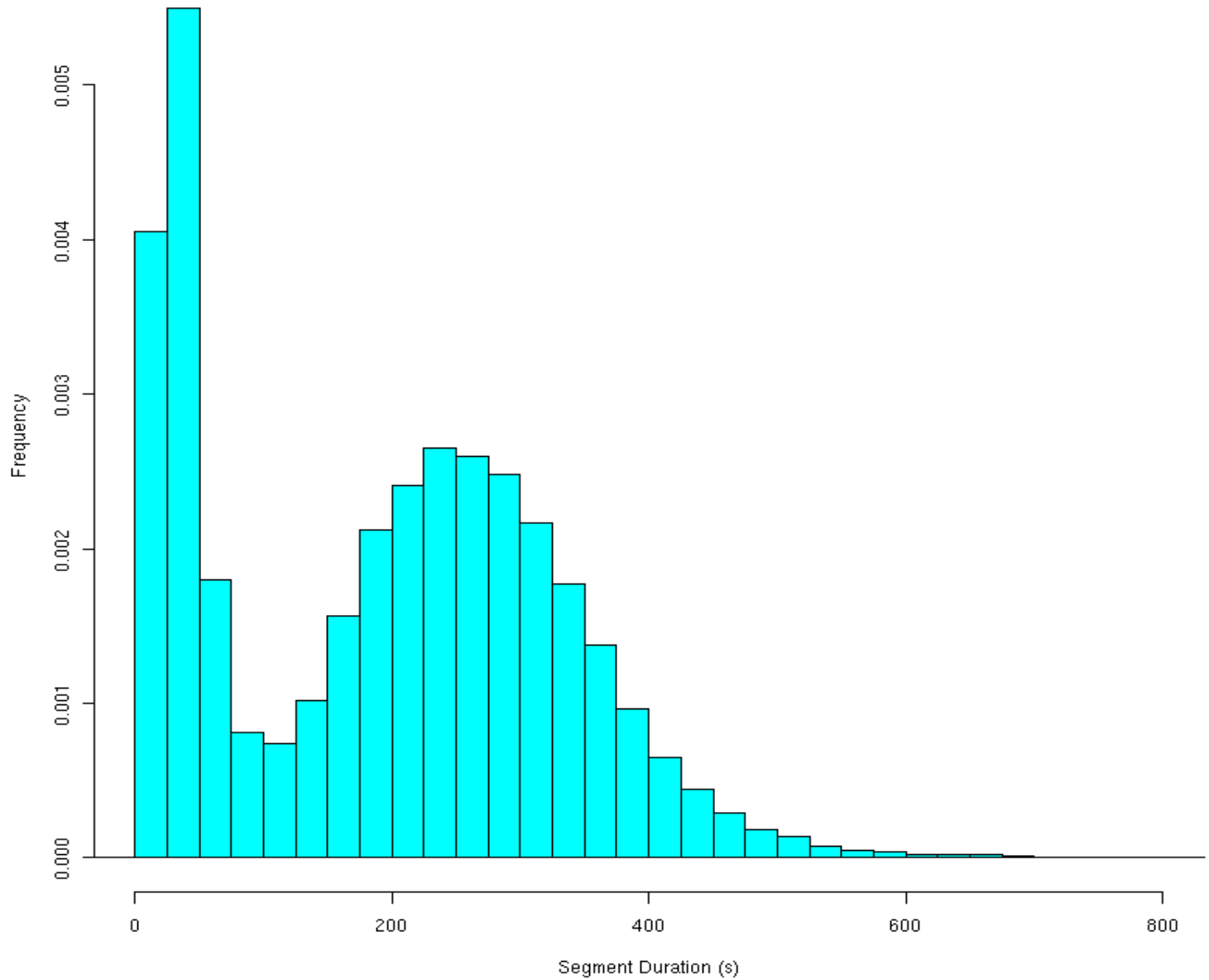
# Test Collection Design



# CLEF-2005 CL-SR Collection

- 8,104 topically-coherent segments
  - 297 English interviews
  - "Known boundary" condition (VHF Segments)
  - Average 503 words/segment
- 63 "topics" (information need statements)
  - **T**itle / **D**escription / **N**arrative
  - 38 training + 25 (blind) evaluation
  - 5 topic languages (EN, SP, CZ, DE, FR)
- 48,881 Relevance judgments (9.6%)
  - Search-guided + post-hoc judgment pools
- Distributed to track participants by ELDA

# Segment Duration



**<DOCNO>**VHF**22029**-209918.**033**

**<INTERVIEWDATA>** Z… | 1922 | Esther Malke | Ettu | R… | R… | Ettuka

**<SUMMARY>**EZ speaks of her family's difficulty to adjusting to Montréal, Canada. Tells of her work in her husband's restaurant. EZ notes the birth of her son. Speaks of her first husband's profession.

**<NAME>**Yehuda Z…, Helen L…, Sandra B…, Allan Z…, Moishe S…

**<MANUALKEYWORD>** Montréal (Quebec, Canada) | family businesses | acculturation | working life | extended family members | Canada 1945 (May 8) - 2000 (January 1) | occupations, spouse's

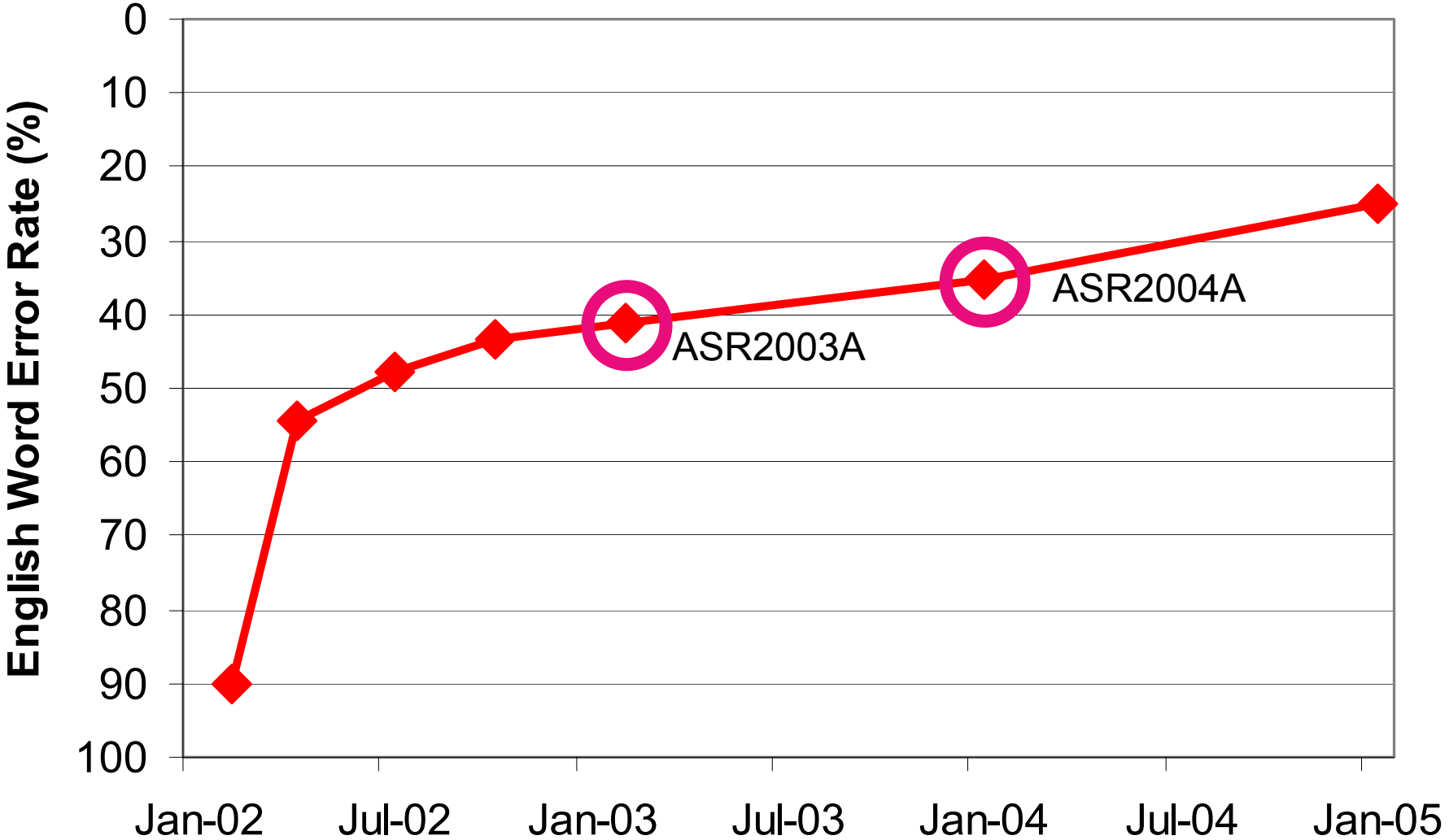
**<AUTOKEYWORD2004A1>** extended family members | friends | working life | introduction of friends and/or family members | migration to the United States | migration from Germany | schools | Jewish-gentile relations | family businesses | family life | cultural and social activities | occupations, father's | aspirations for the future | emigration and immigration policies and procedures | aid: assistance in migration | means of adaptation and survival | United States 1945 (May 8) - 1952 (December 31) | France 1918 (November 11) - 1939 (August 31) | Paris (France) | Poland 1944

**<AUTOKEYWORD2004A2>** extended family members | fate of loved ones | anti-Jewish measures and legislation | Budapest (Hungary) | Poland 1939 (September 1) - 1945 (May 7) | aid: assistance in hiding and/or evasion | working life | seizure of property | United States 1945 (May 8) - present | living conditions | Poland 1941 (June 21) - 1944 (July 21) | hiding | New York (New York, USA) | Poland 1918 (November 11) - 1939 (August 31) | occupations, interviewee's | Berlin (Prussia, Germany) | aid: provision of shelter | China 1939 (September 1) - 1945 (September 1) | separation of loved ones | contact with loved ones, renewed

**<ASRTEXT2004A>**well here the first five years we were not happen we wanted to go back to israel my husband couldn't red cross that we opened a grocery and then we lost that a lot of money and then we had a cousin he said you know what that to make a living and best for me in a restaurant so you bastard and arrest and um and after two weeks because when they came smart about a sticky he didn't want to be there anymore so we didn't want to lose the two thousand dollar so we thought we going to sell it but the themselves so fast there was another jewish area on them near takes three percent province so to make a living my husband took me and to be cooked and we could make a living and i became pregnant with holland and my husband wanted to go to a um venezuela we made already passport and her sister from pennsylvania again there might be care of the group and he was a very angry at him we should come to pennsylvania and to settle we didn't want to go to a small town so he came back and when i was and i continued and the restaurant i myself i didn't want to let go and this time was the monthly a star they took away from him the building about the start clothes and discrete guy wanted to buy my arrested because he heard them expecting a baby and we made a lot of friends and we had the store because it was a jewish tired after i make a lot this and on nothing cash and she skate and i was very very good cook to people came to me trust the muscle and paid that double the price like somebody else and it and alan was for them discrete came in and he gave me six thousand dollars for the rest of two thousand intention four thousand dollars in notes in my name there was my first time that a canadian man and he liked us he invited us for parties on monday my husband and got a terrible terrible parkland and he wanted to start to build because uh you know this was a straight would he uh he he bark wood from the forest and sold it to the sea on our way to the trains so he knew measurements that we couldn't he wanted to build the when he told our neighbors time there that he wants to build we kissed them he said we want to be a part so he became part of for five years and my first time that we built a law that houses and that time of the trial one girl yeah yeah and then me that happy after the children that could i could children my husband was sick alone so i manic i was in the office i attempted he build always then we started to the small buildings like thirty units fifty units and my rent and then he uh

**<ASRTEXT2003A>**from here the first five years we were not have been wanted to go back to israel my husband couldn't red cross and we opened a grocery and then we must have a lot of money and then we had a cousin he said you know what have to make a living in that for me in a restaurant so you bastard and the rest of the um and after two weeks because when they came smart about a sticky he didn't want to be there anymore to we didn't want to lose the two thousand dollar so we thought we going to sell it but the themselves so fast there was a jewish area on them near creeks because some products so to make a living my husband took me into the cook cook it and he could make a living and i became pregnant with holland and my husband wanted to go to um venezuela we made already passport and her sister from pennsylvania again there might be the the group and he was a very angry at him he should come to pennsylvania and to settle they didn't want to go to a small town so he came back and when i was and i continued and the rest of it and i myself i didn't want to let go and this time was them on the us us they took away from him in the building a contest our clothes and this great guy wanted to buy my arrested because he heard them expecting a baby and we made a lot of friends and we had the store because it was a to we started to i make a lot this and run out and kishka and she skate and i was very very good cook to people came to me that the muscle and they did double the price like somebody else and a half an hour was for them discrete came in and he gave me six thousand dollars for the rest of two thousand intention four thousand dollars in notes in my neighbors was my first time that factory in man and he liked us he invited us for parties on monday my husband and got a terrible time in parkland and he wanted to start to build because they were of this was this trade would he uh he he bark wood from the

# English ASR




Training: 200 hours from 800 speakers

# An English Topic

**Number:** 1148

**Title:** Jewish resistance in Europe

**Description:** Provide testimonies or describe actions of Jewish resistance in Europe before and during the war.

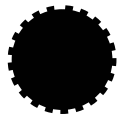
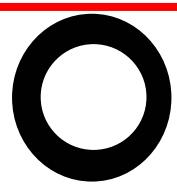
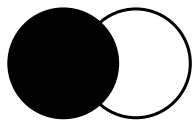
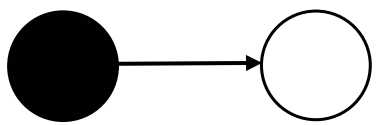
**Narrative:** The relevant material should describe actions of only- or mostly Jewish resistance in Europe. Both individual and group-based actions are relevant. Type of actions may include survival (fleeing, hiding, saving children), testifying (alerting the outside world, writing, hiding testimonies), fighting (partisans, uprising, political security) Information about undifferentiated resistance groups is not relevant.

# Relevance Judgments

- Search-guided
  - Iterate topic research/query formulation/judging
    - Done by grad students in History
    - Trained by expert librarians
  - Essential for collection reuse with future ASR
  - Done between topic release and submission

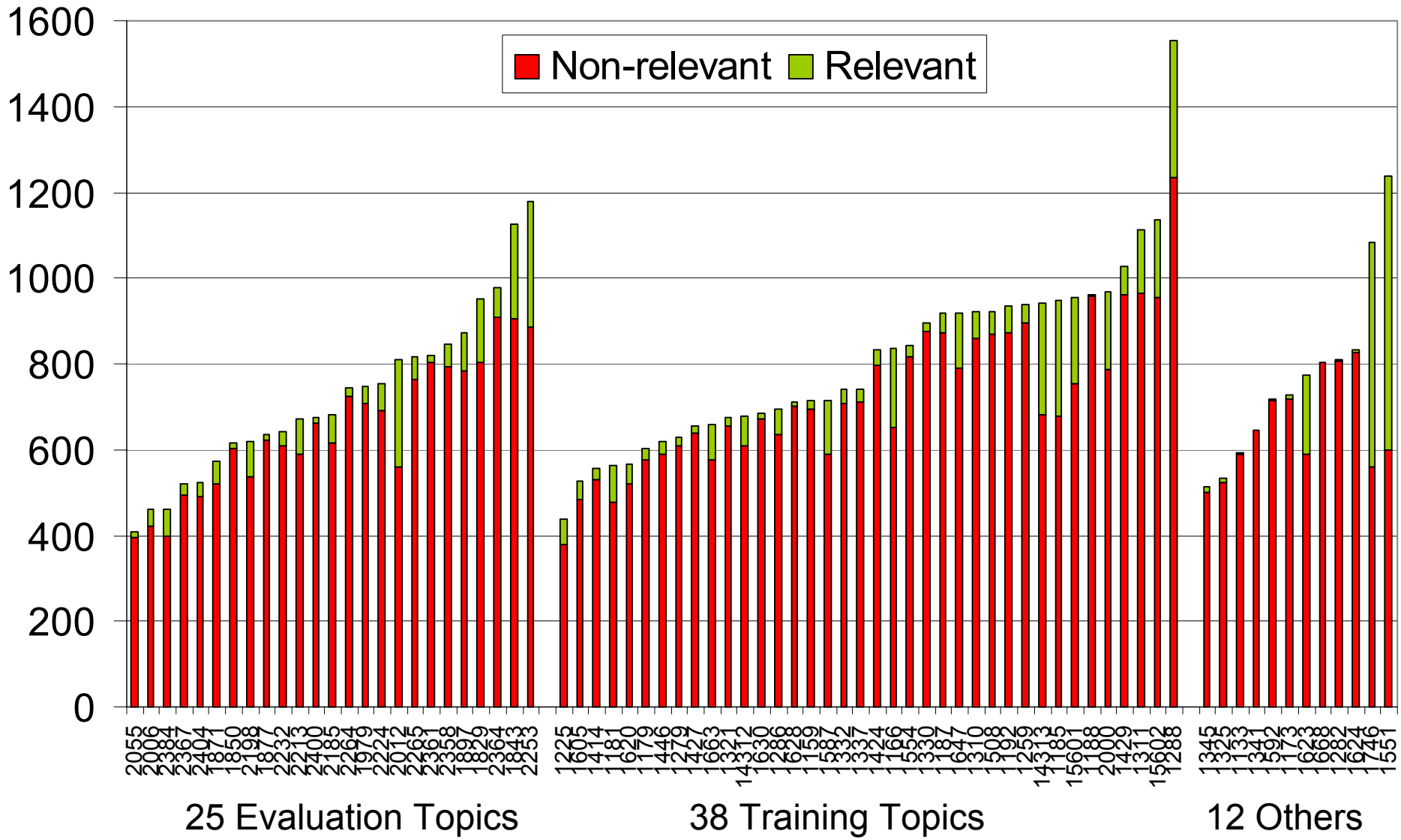
- Top-ranked (="Pooled")
  - Same assessors (usually same individual)
  - 100-deep pools from 14 systems
    - 2 per site, chosen in order recommended by sites
    - Omit segments with search-guided judgments

# 5 Types of Relevance

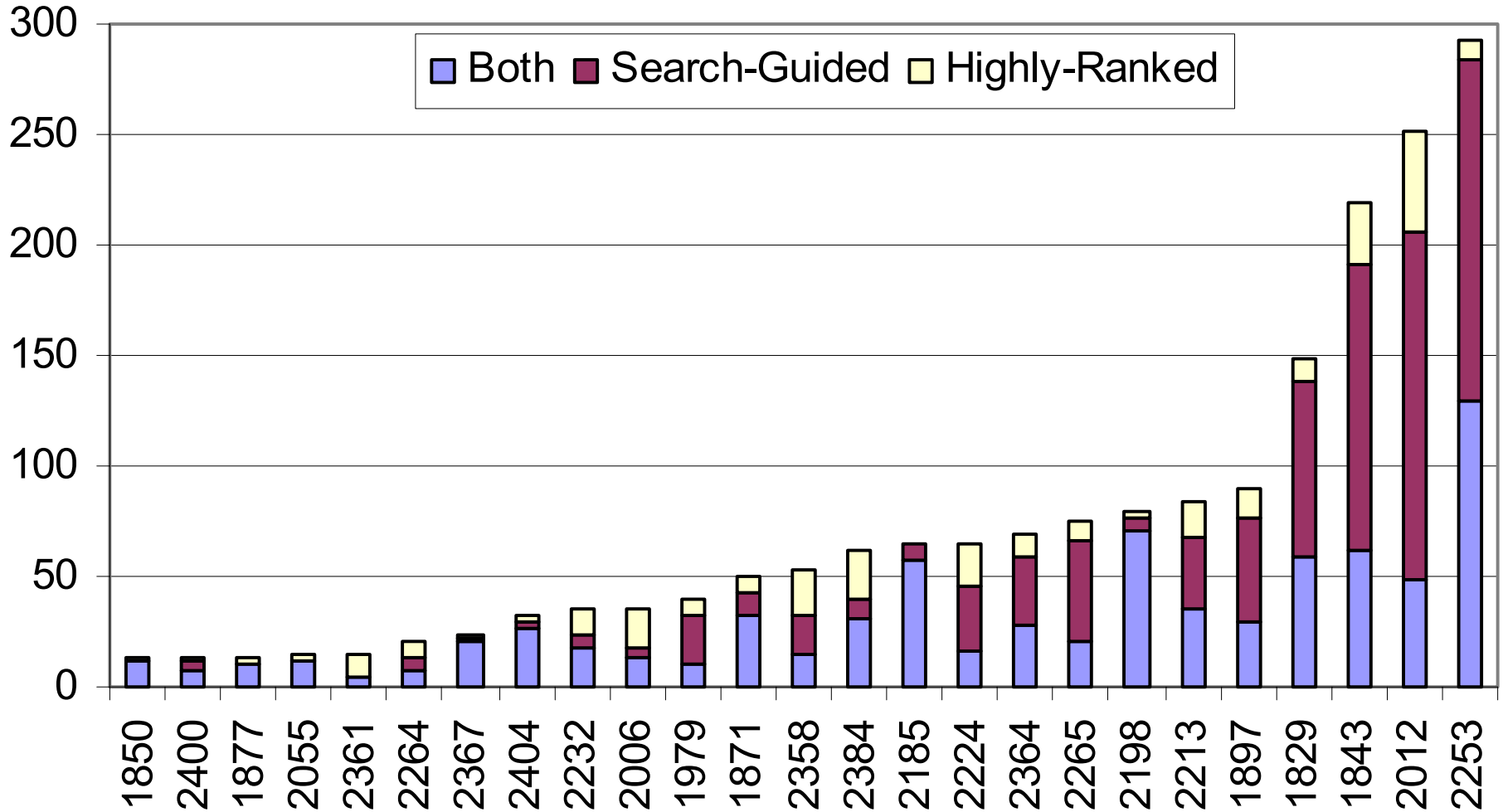
| Topical Relevance Type | On topic | Relationship between overall document topic & overall user need topic | Diagram Illustrating Relationship[1] | Relatedness to the given topic | Specificit of informati |
|---|---|---|---|---|---|
| **Direct** | Directly | Matching |  | High | High |
| **Indirect** | After inference | Matching |  | High ~ medium | High |
| **Context** | No | Surrounding |  | Low ~ medium | Low |
| **Comparison** | Partially | Comparing/ contrasting |  | Low ~ medium | High ~ medium |
| **Pointer** | No | Apart |  | Low | High |

- Five levels of relevance for each (0=none, 4=highly)
- Collapsed to binary relevance using {Direct≥2} U {Indirect≥2}
- Script provided that sites can use to explore other combinations

# Total (Binary) Judgments

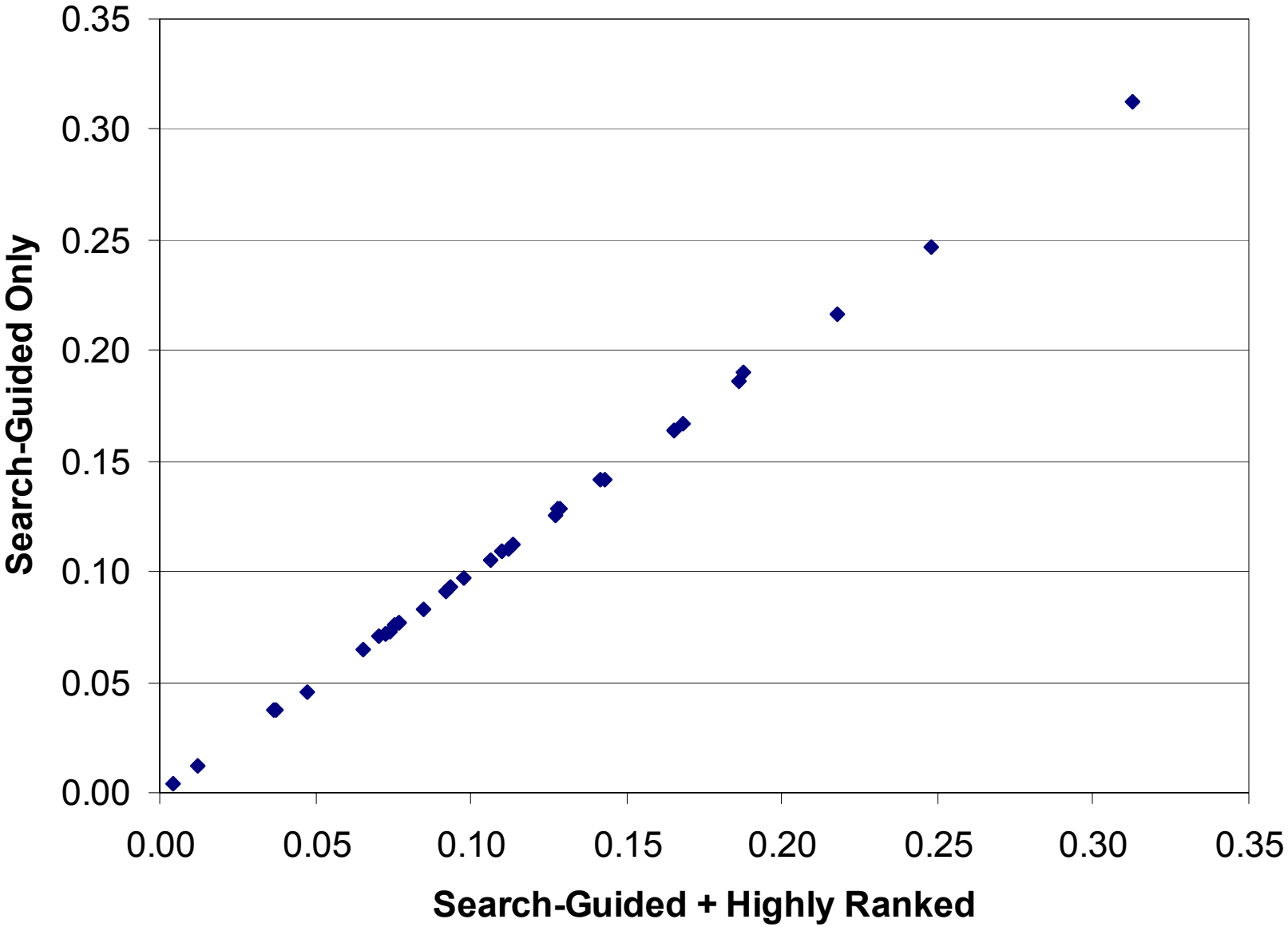


# Contributions by Assessment Type



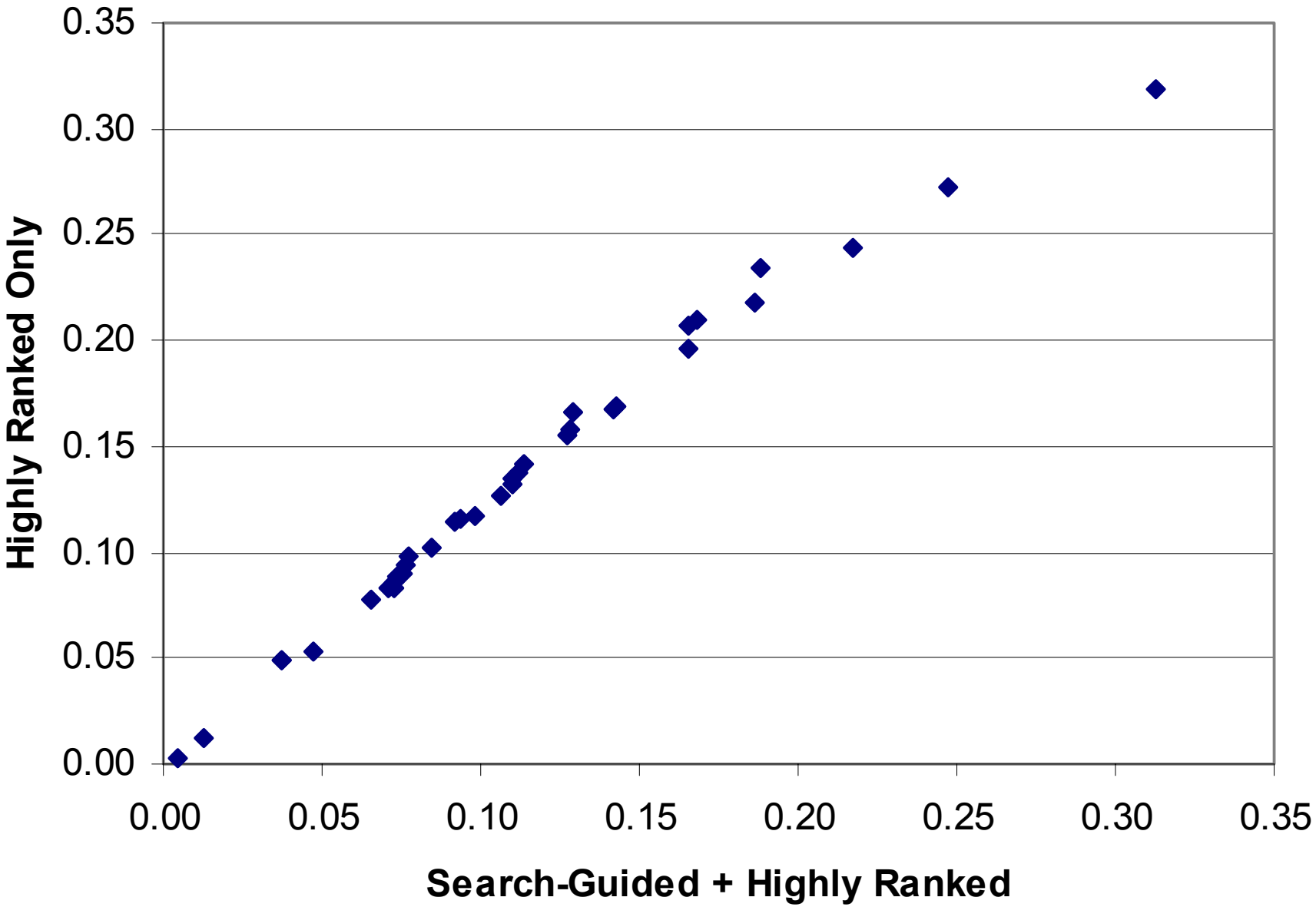

Binary relevance, 25 evaluation topics, 14-run 100-deep pools

# Search-Guided: Stable Rankings




Mean average precision, 35 official CLEF-2005 CLEF runs, 25 evaluation topics

# Highly Ranked: Less Stable




Mean average precision, 35 official CLEF-2005 CLEF runs, 25 evaluation topics

# Almost A Unified Track for Everything

- We're an ad hoc bilingual track
  - Standard distribution is a standard CLEF track
- We're a domain-specific track
  - Thesaurus, human indexing, lead-in vocabulary
- We're a Geo track
  - Every "document" is location and time tagged
- We'll be a "new language" track in 2006
  - Czech "documents"
- We could be an image retrieval track
  - Thousands of images, with captions and ASR
- And, of course, we're a speech track!

# For More Information

- CLEF Cross-Language Speech Retrieval track
  - Come see the posters!
  - http://clef-clsr.umiacs.umd.edu/
- The MALACH project
  - http://www.clsp.jhu.edu/research/malach
- NSF/DELOS Spoken Word Access Group
  - http://www.dcs.shef.ac.uk/spandh/projects/swag

# Backup Slides

# 5-level Relevance Judgments

Binary qrels

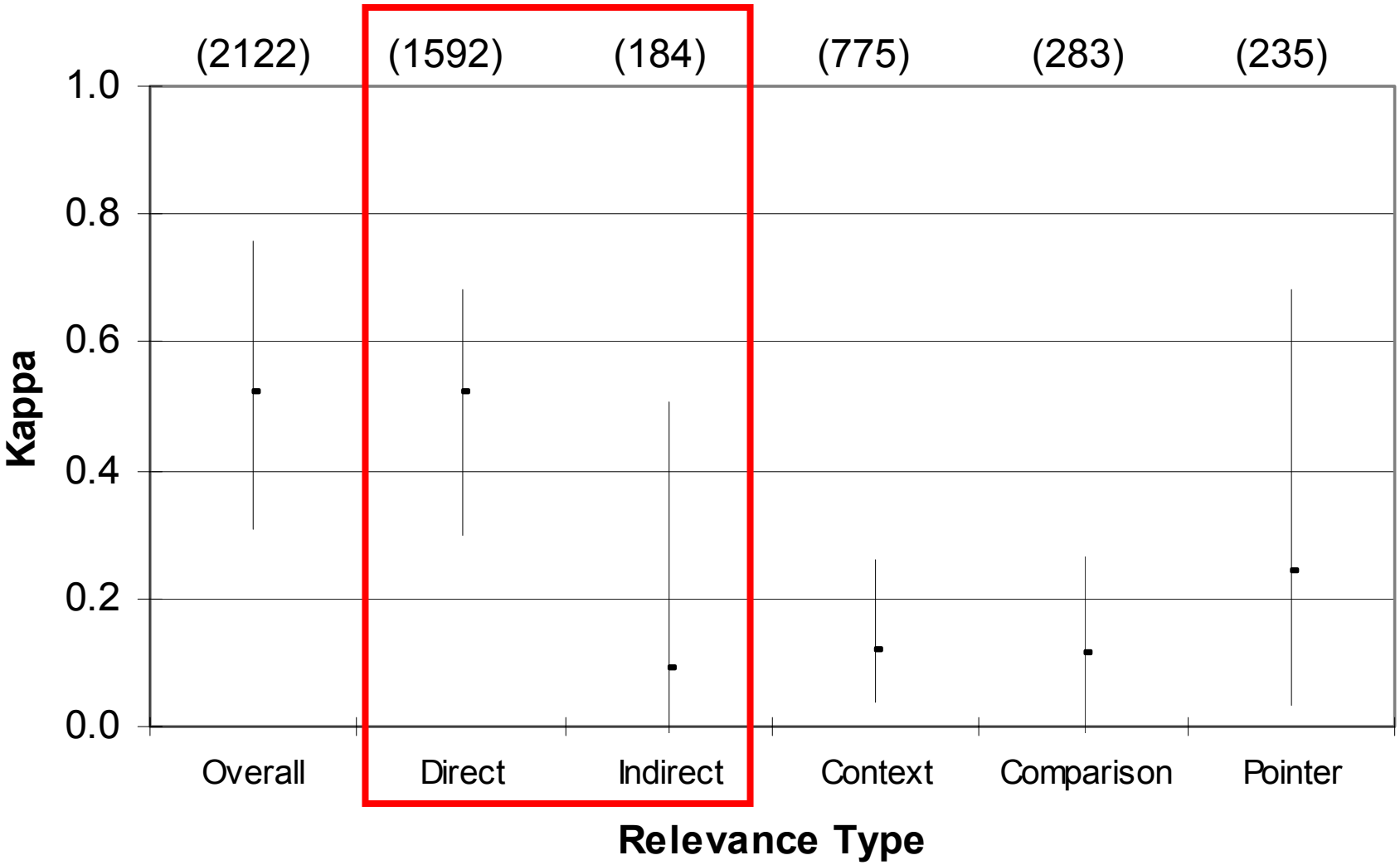
- "Classic" relevance (to "food in Auschwitz")
  - Direct — Knew food was sometimes withheld
  - Indirect — Saw undernourished people

- Additional relevance types
  - Context — Intensity of manual labor
  - Comparison — Food situation in a different camp
  - Pointer — Mention of a study on the subject

# Supplementary Resources

- Thesaurus (Included in the test collection)
  - ~3,000 core concepts
    - Plus alternate vocabulary + standard combinations
  - ~30,000 location-time pairs, with lat/long
  - Synonym, is-a and part-whole relationships
- Digitized speech (provided to Waterloo)
  - .mp2 or .mp3
- In-domain expansion collection (MALACH)
  - 186,000 scratchpad + 3-sentence summaries

# Assessor Agreement




44% topic-averaged overlap for Direct+Indirect 2/3/4 judgments

14 topics, 4 assessors in 6 pairings, 1806 judgments

# The CLEF CL-SR Team

**USA**

- Shoah Foundation
  - Sam Gustman
- IBM TJ Watson
  - Bhuvana Ramabhadran
  - Martin Franz
- U. Maryland
  - Doug Oard
  - Dagobert Soergel
- Johns Hopkins
  - Zak Schefrin

**Europe**

- U. Cambridge (UK)
  - Bill Byrne
- Charles University (CZ)
  - Jan Hajic
  - Pavel Pecina
- U. West Bohemia (CZ)
  - Josef Psutka
  - Pavel Ircing
- UNED (ES)
  - Fernando López-Ostenero

# CL-SR Track: Plans for 2006

DCU: **Gareth Jones**,

UMD: Doug Oard, Dagobert Soergel, Ryen White

CU: Jan Hajic, Pavel Pecina

UWB: Josef Psutka

JHU: Bill Byrne (Cambridge), Zak Schefrin

IBM: Bhuvana Ramabhadran

USC: Sam Gustman (Shoah Foundation)

# Tentative CLEF-2006 Plans

(Intent is to do both)

- New Czech test collection
  - ~700 hours unseen speakers (+ 700 hours seen)
    - ~35% mean Word Error Rate
  - 25 evaluation topics (+2-3 topics for training)
  - New unknown-boundary relevance assessment

- Larger English test collection
  - ~900 hours unseen speakers
    - ~25% mean Word Error Rate (2006: ~35% WER)
  - 25 evaluation topics (+63 topics for training)
  - Word lattice distributed as standard data

# CLEF-2007 Options

(probably only 1 will be possible)

- English
  - Expand to ~5,000 hours of unseen speakers
  - 25 new topics
  - Full ASR training data
- Czech
  - 25 new topics (total 50 across 2 years)
  - Full ASR training data
- Russian
  - 25 topics
  - Full ASR training data

# Breakout Session Agenda

- Lessons learned from this year
  - Evaluation design, logistics
- English lattice options for 2006
  - Word lattice, phone lattice, partial decoding
- Czech test collection design
  - 2-channel ASR, unknown boundary measures
- Resource sharing among participants
  - Czech morphology/translation, in-domain expansion
- Additional data from VHF (special arrangement)
  - ASR training, BRF expansion,
- ELDA test collection release to nonparticipants

# CL-SR Breakout Session

Doug Oard and Gareth Jones

# Topic Construction

- English topics based on mailed requests
  - 25 new topics
- Czech topics based on discussions w/users
  - 25 topics, some from existing (English) topic set
  - Don't tune Czech system to existing topics!
- All topics created in English
  - Standard translations into CZ, DE, FR, SP
  - Other languages possible if you can help

# Standard Resources for Czech

- Brain-dead automatic segmentation
  - 400-word passages, with 50% overlap
  - Passage-to-start time conversion for results
- Prague Dependency Treebank (LDC)
  - Czech/English translation lexicon
    - Precompiled morphology
  - Czech morphology
    - For both formal and colloquial

# Run Submission

- English: ranked lists of segments
- Czech: ranked lists of start times

- 5 official runs for each language?
  - Additional runs can be scored locally

- One required condition for each language
  - Monolingual TD queries automatic data only?

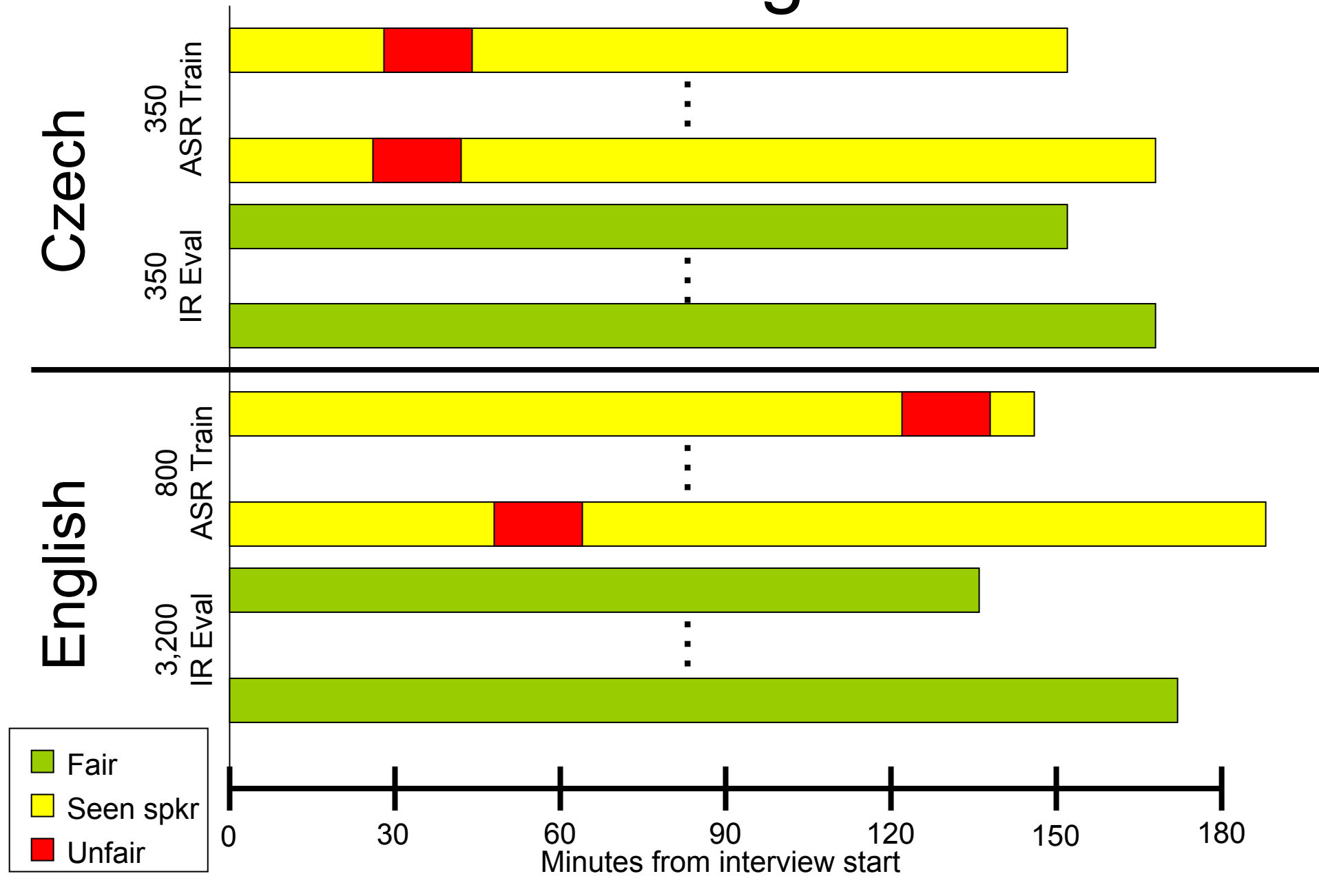
# Czech Relevance Assessment

- Integrated search
  - Thesaurus-based onset marks
  - ASR results
- Unknown-boundary evaluation measure
  - With unknown boundaries in ground truth
- Summary generation
  - Descriptor history + ASR results
- Thesaurus translation into Czech

# Potential Concerns

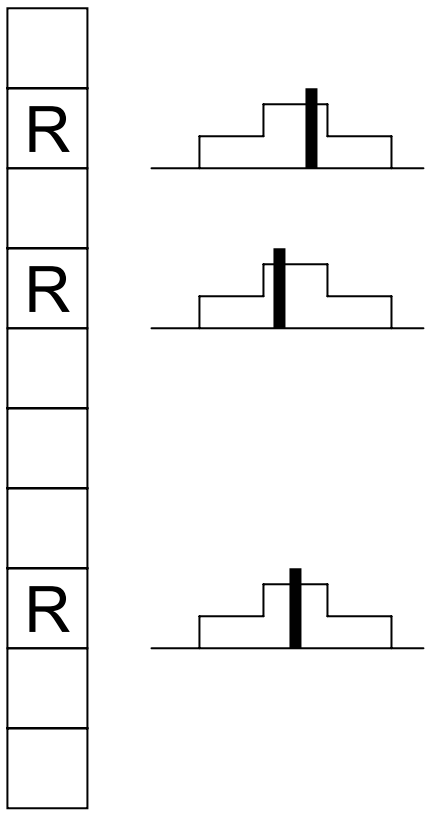
- Czech
  - # of teams contributing to highly-ranked pools
  - Clearance of interviews for ELDA distribution
  - Measure: Tuning [rank] vs. [start time error]
  - How to treat speakers seen during ASR training
  - Colloquial usage
- English
  - Cost-benefit tradeoff for highly-ranked judging
  - Partial decoding design and data format

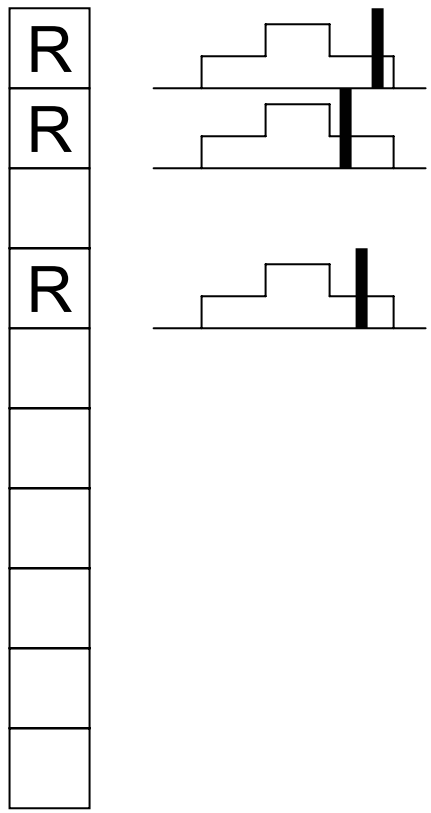
# Interview Use in English & Czech



# Ranked Start Time Evaluation

Key Idea: Weight contribution to average precision by f(absolute error)




**AP**=(1.0*1/2 + 1.0*2/4 + 1.0*3/8)/3
= **0.46**

**AP**=(0.5*1/1 + 0.5*2/2 + 0.5*3/4)/3
= **0.46**

# Eval Measure Design Issues

- What error function to use?
  - Symmetric? [yes]
  - Max allowable error? [1.5 minutes]
  - Shape? [step function, 1-minute increments]
- How to prevent double counting?
  - Guard band w/list compression? [4 minutes]
- What time granularity to use
  - For ground truth? [15 seconds]
  - For system output? [1 second]

# Known Problems

- “Synonyms” are really “lead-in vocabulary”
  - Extended family members -> aunts, uncles, …
- Missing tapes cause errors
  - For segments crossing missing-tape boundary
- Incomplete judgments for added English
  - Only a problem for the 63 existing topics
  - But questionable judgments are marked
- Some 2006 topics are in the 75 from 2005
  - Don’t look at topics at topics beyond the 63!

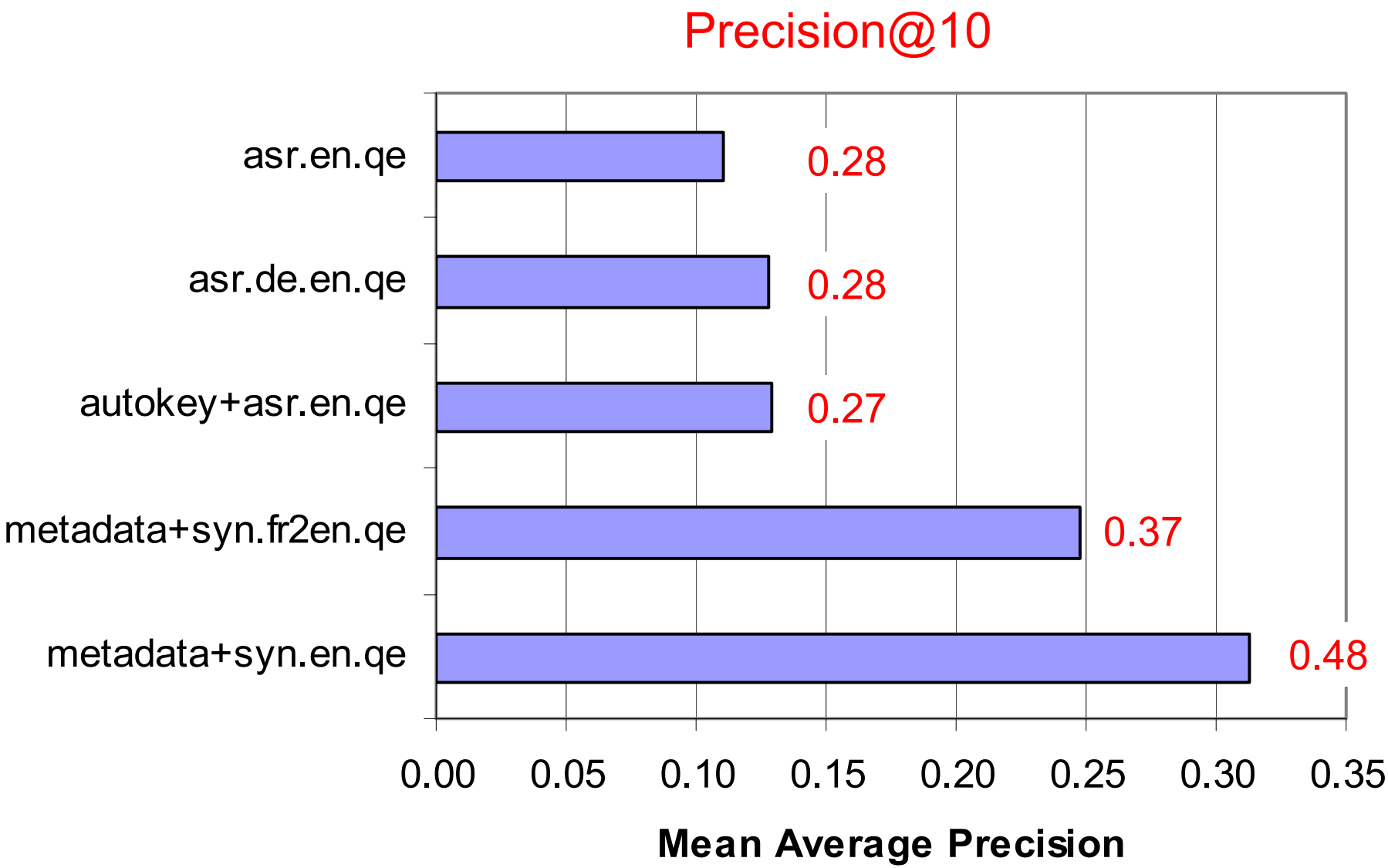
# Things to Think About

- Effective and efficient lattice search
- Failure analysis using judgment types
  - Direct, indirect, context, comparison, pointer
- Clever expansion approaches
  - Thesaurus, sequence, time, similarity
- Interactive experiments

# CLEF-2005 at Maryland: Cross-Language Speech Retrieval

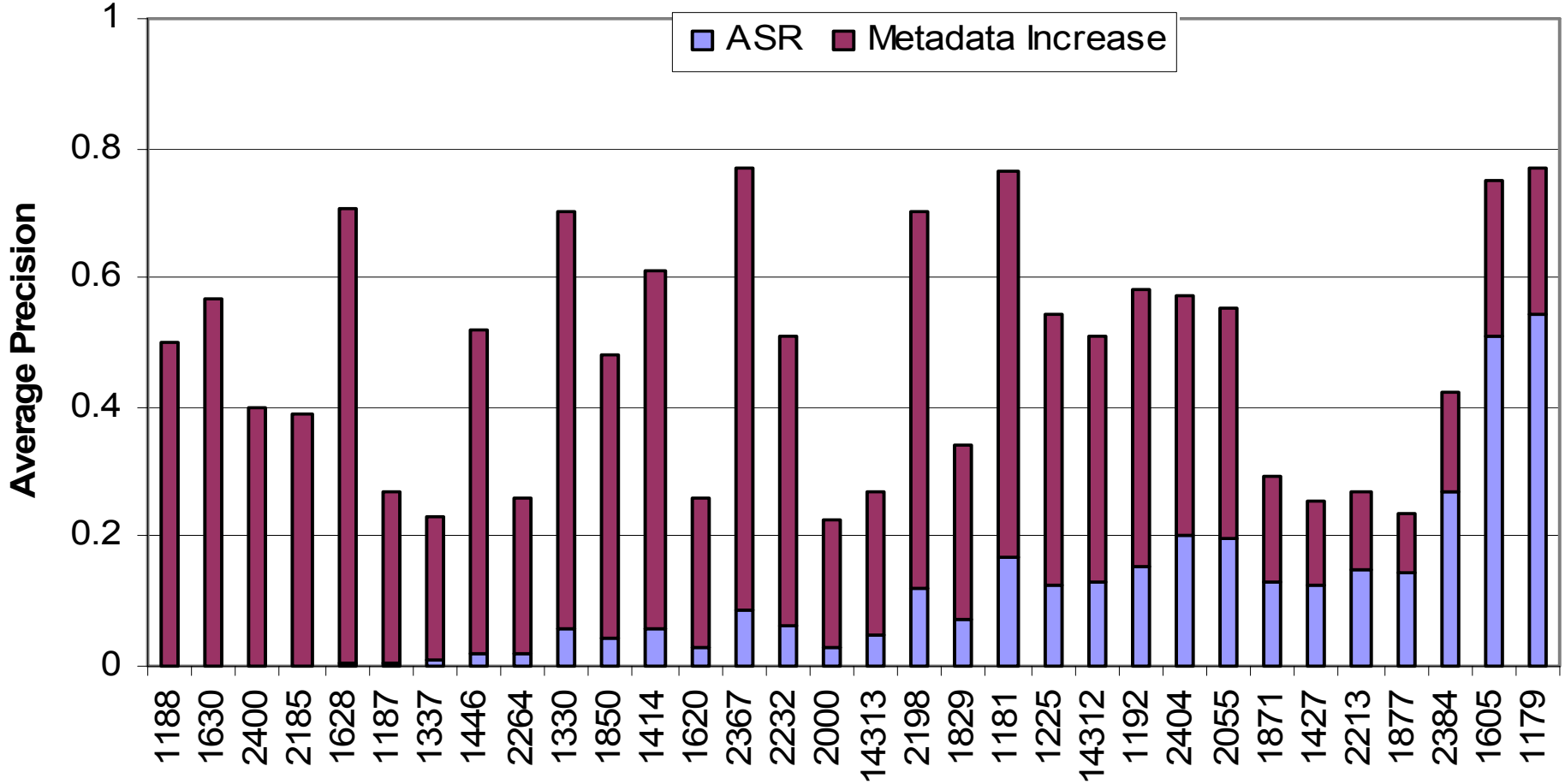
**Douglas Oard**, Jianqiang Wang, Ryen White

# Maryland CLEF-2005 Results




25 evaluation topics, English TD queries

# Comparing ASR with Metadata

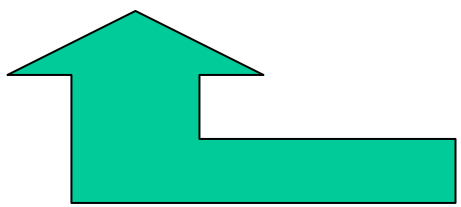


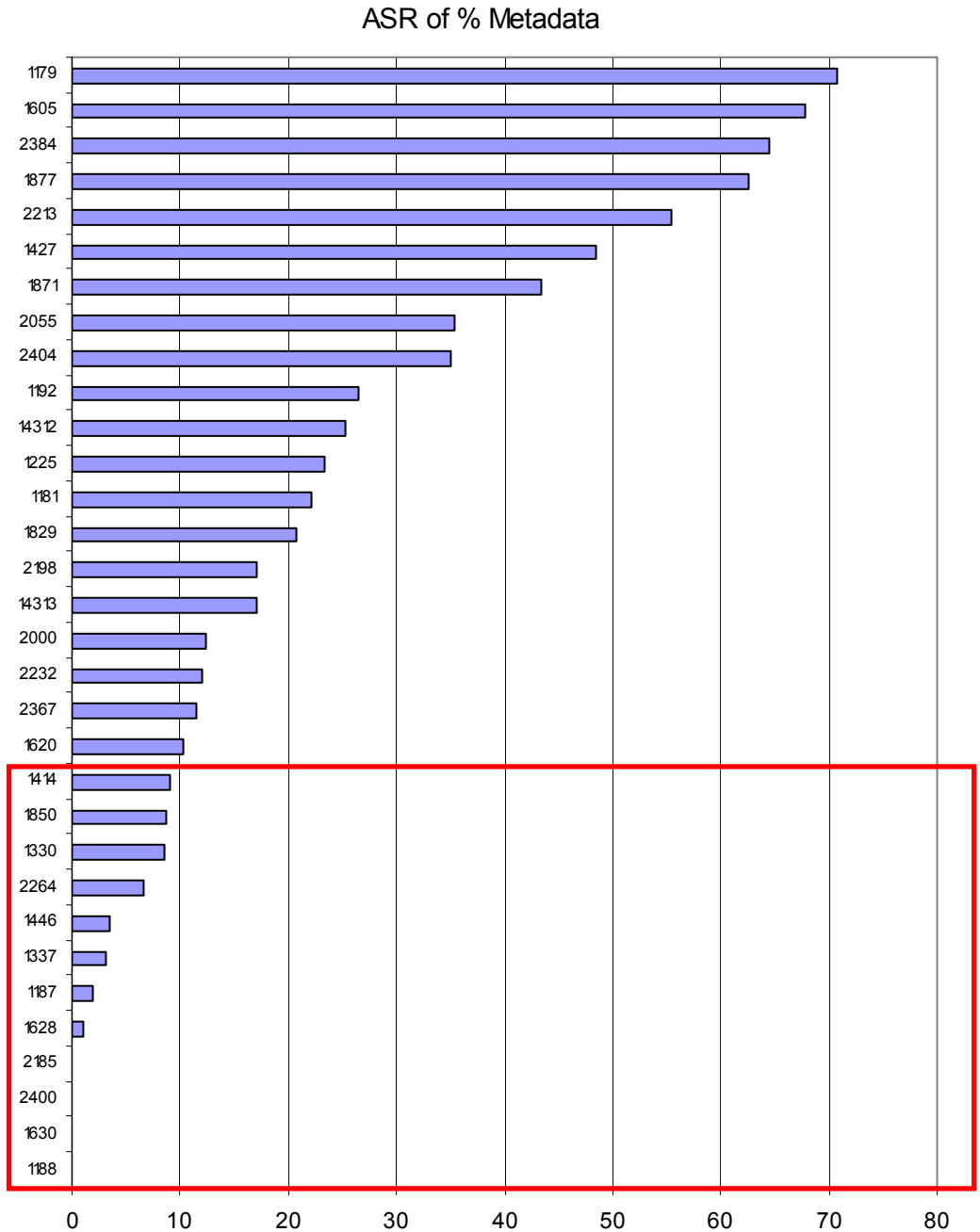

CLEF-2005 training + test – (metadata < 0.2), ASR2004A only, **Title queries**, Inquery 3.1p1

# Error Analysis

| Somewhere in ASR | Only in Metadata | |
|---|---|---|
| wallenberg (3/36)* rescue jews | | |
| wallenberg (3/36) | eichmann | |
| abusive female (8/81) personnel | | |
| minsko (21/71) ghetto underground | | |
| art auschwitz | | |
| labor camps | ig farben | |
| slave labor | telefunken | aeg |
| holocaust | sinti roma | |
| sobibor (5/13) death camp | | |
| witness | eichmann | |
| jews | volkswagen | |

* (ASR/Metadata)






CLEF-2005 training + test – (metadata < 0.2), ASR2004A only, Title queries, Inquery 3.1p1