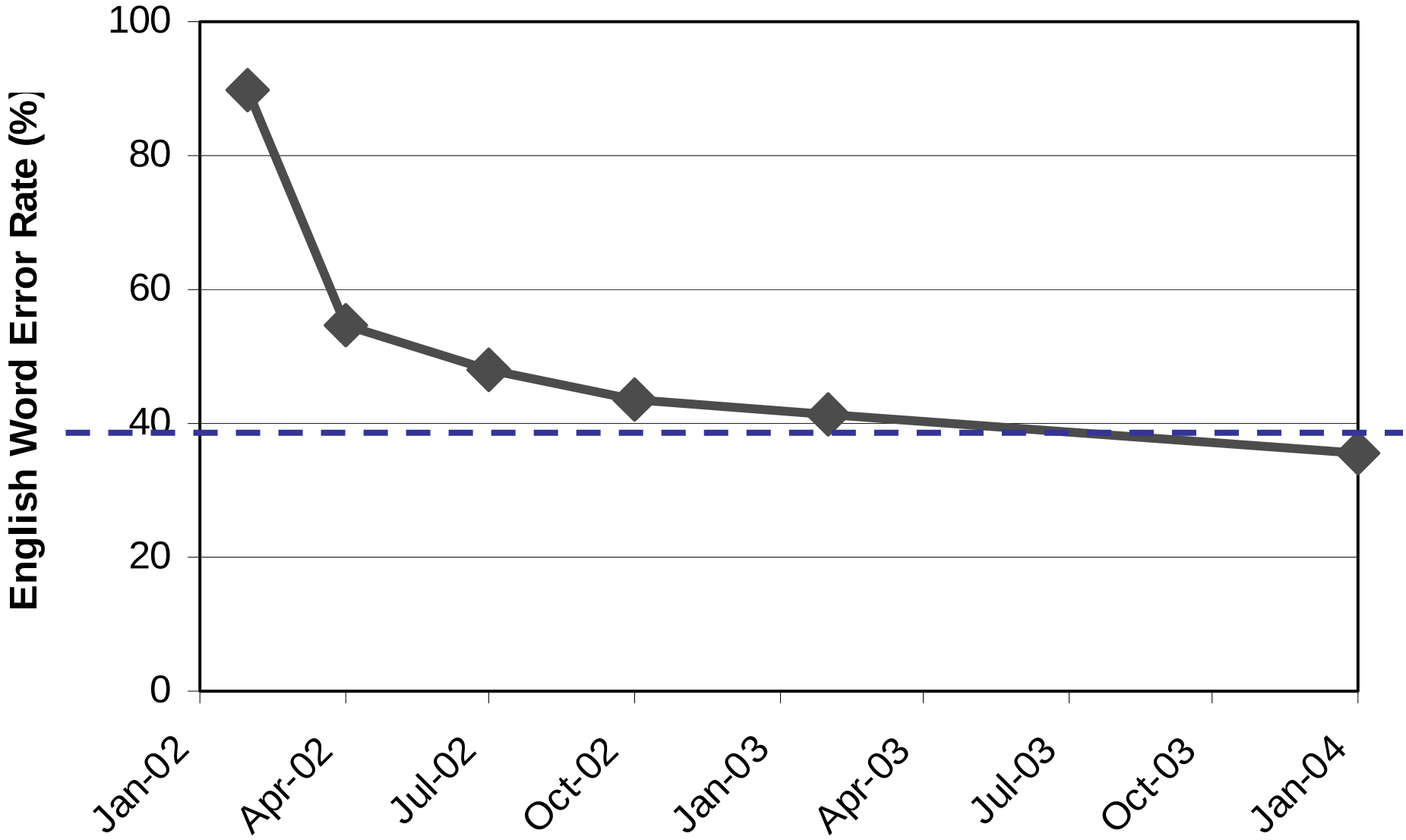
CLEF-2005 CL-SDR: Proposing an IR Test Collection for Spontaneous Conversational Speech

Gareth Jones (Dublin City University, Ireland)
Douglas W. Oard (University of Maryland, USA),
Also here:
Bill Byrne (University of Cambridge, UK)
Pavel Ircing (University of West Bohemia, Czech Republic)
Dagobert Soergel (University of Maryland, USA)

Looking Beyond Broadcast News

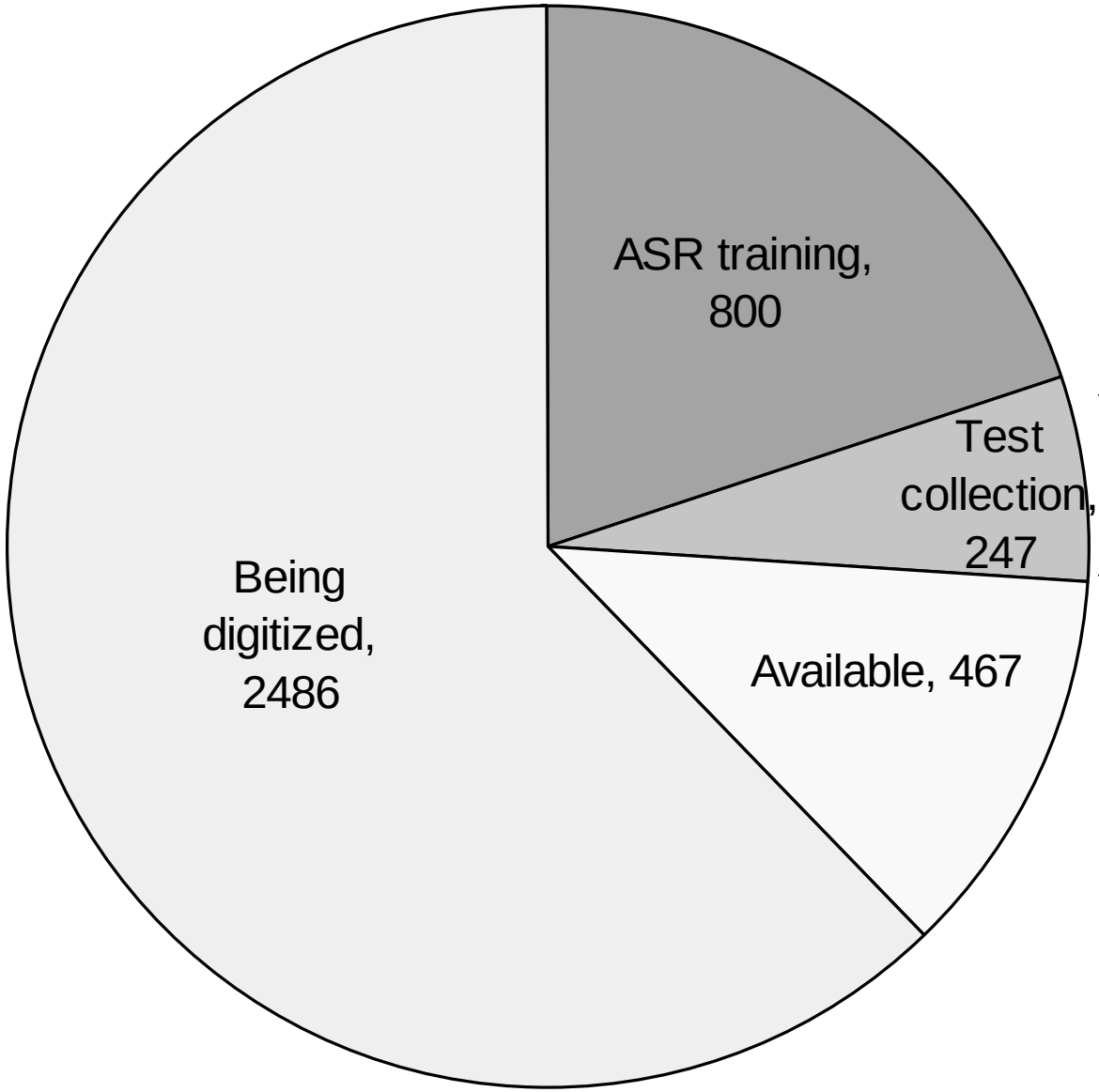
- Broadcast programming
  - News, interview, talk radio, sports, entertainment
- Scripted stories
  - Books on tape, poetry reading, theater
- Spontaneous storytelling
  - Oral history, folklore
- Incidental recording
  - Speeches, oral arguments, meetings, phone calls

English ASR Accuracy



Training: 200 hours from 800 speakers

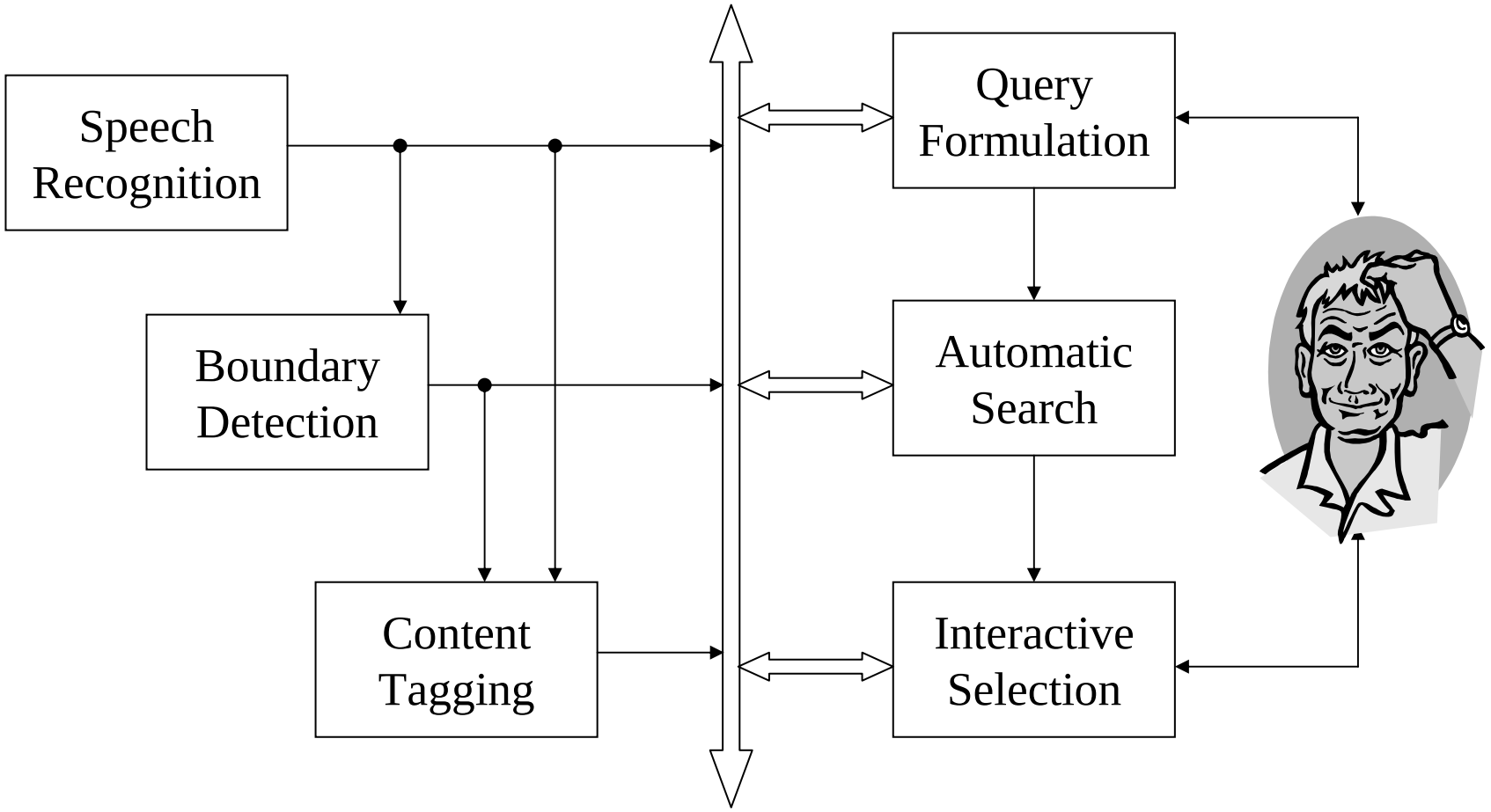
English Interviews (10,000 hours)



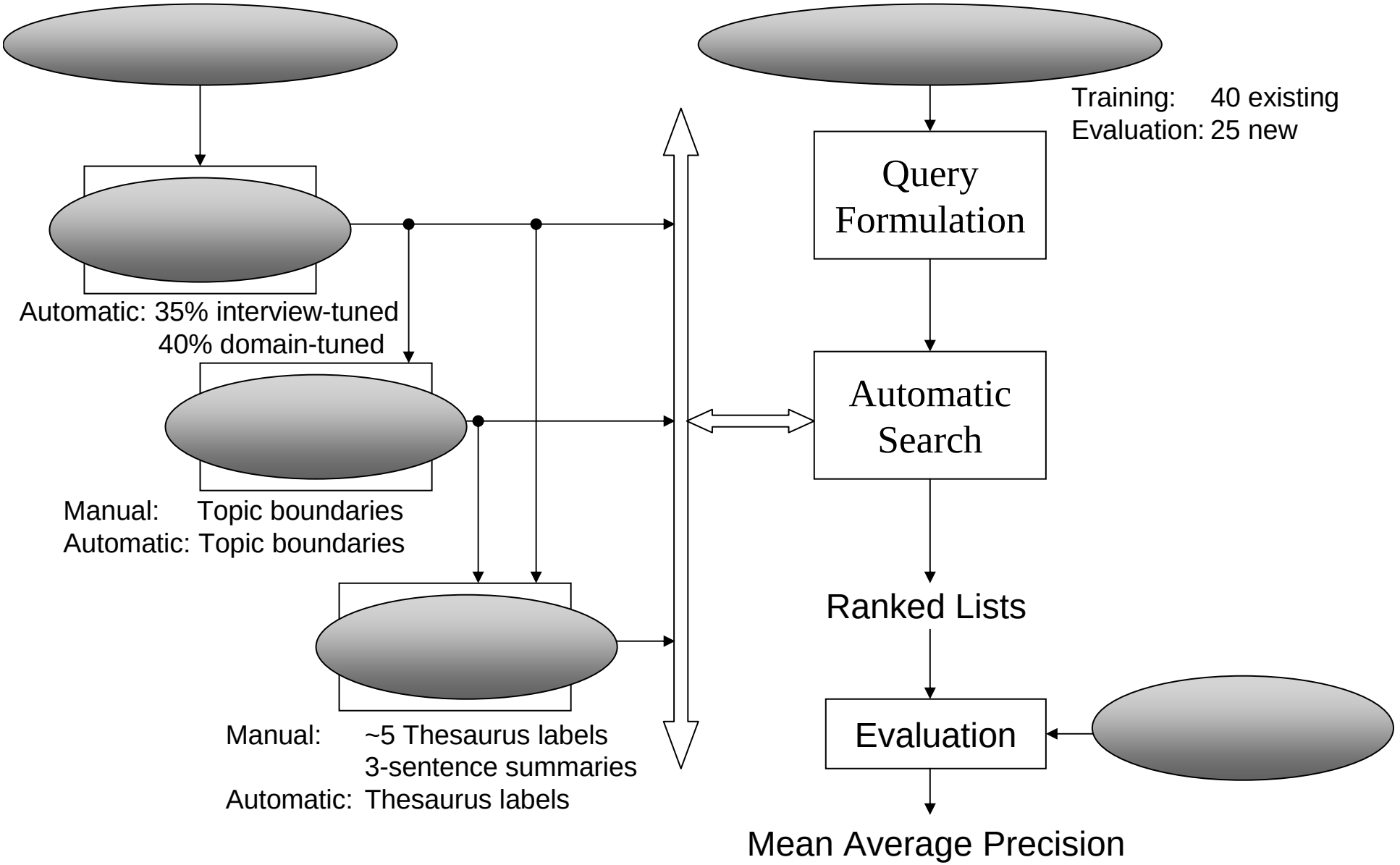
9.947 segments
~400 words each
(total: 625 hours)

“Old Indexing”

Test Collection Design



Test Collection Design



Topic Construction

- 280 topical requests, in folders
  - From scholars, teachers, broadcasters, …
- 50 selected to build an initial test collection
  - Recast in TREC topic format
  - Some needed to be “broadened”
- All assessed during Summer 2003/2004
  - At least 40 with ≥5 relevant segments

An Example 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

- **Training Topics: Search-guided**
  - Interactive search system, based on metadata
  - Iterate topic research/search/assessment
  - Augment with review, adjudication, reassessment
  - Will complete 50 topics in ~2,000 hours
- **Evaluation Topics: Pooled**
  - Requires a diverse set of ASR and IR systems
  - Augmented with search-guided assessment

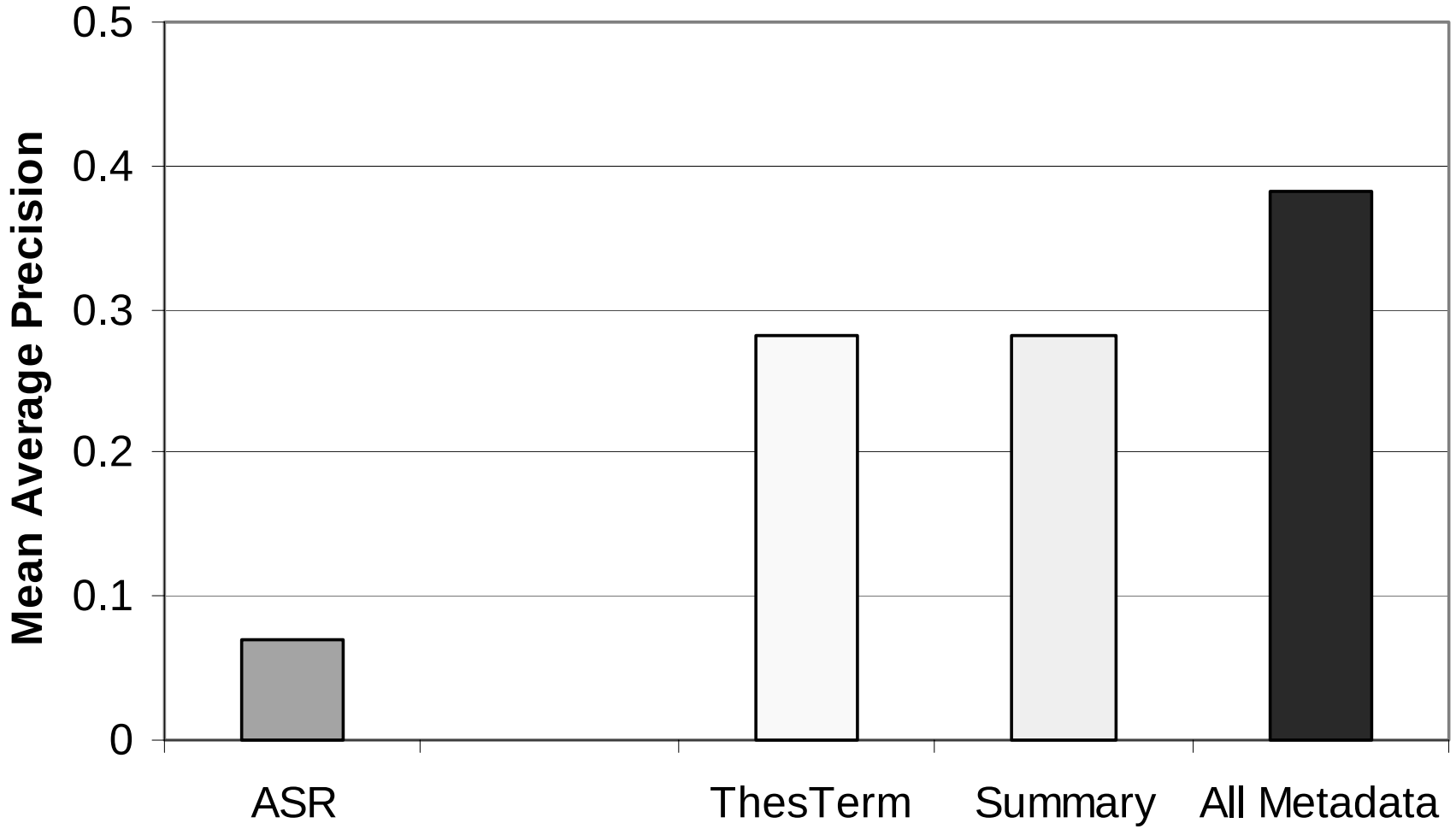
5-level Relevance Judgments

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

Some Baseline Results



Title queries, adjudicated judgments, ~40% WER

May be Available by Arrangement

- Thesaurus
  - ~3,000 core concepts
    - Plus alternate vocabulary + standard combinations
  - ~30,000 location-time pairs, with lat/long
  - Both “is-a” and “part-whole” relationships
- In-domain expansion collection
  - 186,000 3-sentence summaries
- Word lattice
  - For the 35% WER interview-tuned system
- Indexer’s scratchpad notes
- Digitized speech (.mp2 or .mp3)

Some Options for Future Years

- CLEF-2006 CL-SR:
  - Add a Czech (or Russian) collection
    - Czech ASR exists now, Russian ASR later this year
  - Larger English collection (~1,000 hours)
    - Adding word lattice as standard data
  - No-boundary evaluation design
  - ASR training data by special arrangement
    - Transcripts, pronunciation lexicon, language model
- CLEF 2007 CL-SR:
  - Add Russian, Polish or Hungarian interviews
  - Much larger English collection (~5,000 hours)

Next Steps

- Breakout session at 9.00 tomorrow
  - Decide on standard tasks, query languages, …
- Finalize data release schedule
  - Present commitment: February 1, 2005
- Recruit participants from beyond CLEF
  - IR, Speech, NLP

Key Questions

- Who is likely to participate?
  - MALACH teams, IR teams, speech teams
- Who will create the relevance judgments?
  - University of Maryland (Dagobert Soergel)
- How will we pay for this?
  - NSF (USA) funding available through 2006
- Why is this exciting?
  - Conversational speech recognition now possible
  - More conversational speech than anything else!
    - 100,000 words per coffee break!

The NSF MALACH Team

USA

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

Europe

- U. Cambridge (UK)
  - Bill Byrne
- Charles University (CZ)
  - Jan Hajic
- U. West Bohemia (CZ)
  - Josef Psutka
  - Pavel Ircing

CL-SR Track Update

- 4-7 likely participants:
  - Firm: UNED, West Bohemia, Chicago, Maryland
  - Known possible: DCU, Johns Hopkins APL, IBM
- Topic languages:
  - EN, SP, CZ, RU, DE, FR, + others by request
- Supported conditions (known boundaries)
  - ASR/TD topics/CLEF ad hoc format (required run)
  - ASR+summary+keywords/EN TDN topics (pooling)
  - Contrastive: kNN keywords+ASR confidence+…
- Data available: February 1, 2005
- Tentative plans: 2006 Czech, 2007 Russian