MALACH : Multilingual Access to Large spoken ArCHives

IBM

UNIVERSITY OF MARYLAND

SURVIVORS OF THE SHOAH VISUAL HISTORY FOUNDATION

http://www.clsp.jhu.edu/research/malach

(funded under NSF ITR Award 0122466)

Sam Gustman

Survivors of the Shoah Visual History Foundation

Bhuvana Ramabhadran, Michael Picheny,
Martin Franz, Nanda Kambhatla

IBM T. J. Watson Research Center

William Byrne

CLSP, Johns Hopkins University

Josef Psutka

University of West Bohemia

Jan Hajic

Charles University

Dagobert Soergel, Douglas W.Oard,

CLIS, University of Maryland

Examples of Spoken Archives

Source	Description
Vincent Voice Library (MSU)	Speeches, Performances, Lectures, Interviews, Broadcasts, etc. 50000 recordings
Oyez! Oyez! Oyez! (NWU)	Supreme Court Proceedings 500 hours
History and Politics Out Loud (NWU)	Significant political and historical events and personalities of the twentieth century
Informedia (CMU)	2 TB of Digital Video
National Gallery of Spoken Word (MSU)	Spoken word collections from the 20th century

VHF Multimedia Data Collection

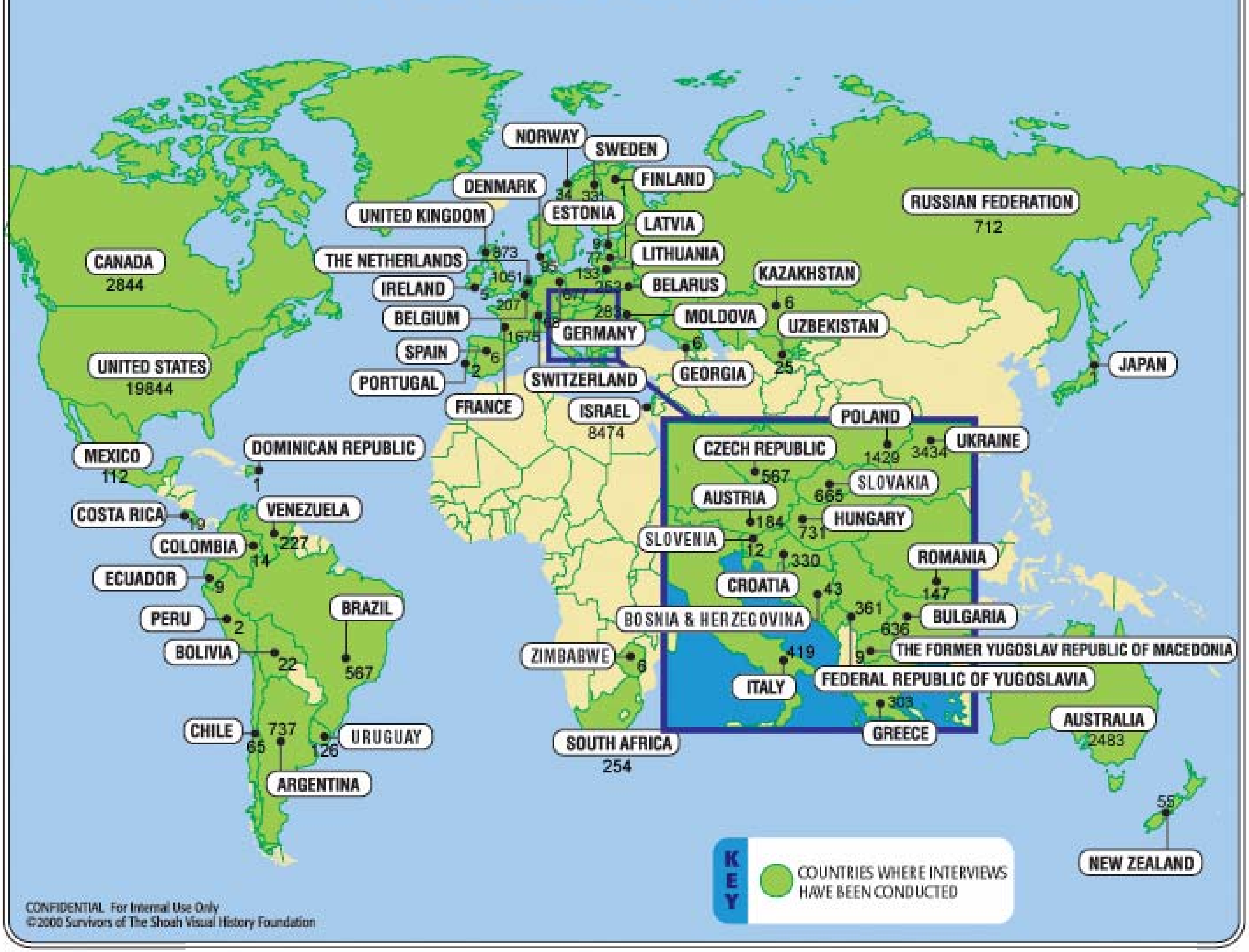
Data

- VHF has collected testimonies 52000 testimonies (2 1/2 hours each) in over 32 languages (180 TB of digital video) - the largest and most complex single topic digital video library in the world

Mini02.mov

http://www.vhf.org/archive.htm

INTERVIEW MAP



CONFIDENTIAL For Internal Use Only

©2000 Survivors of The Shoah Visual History Foundation

Number of Interviews by Country

Argentina	737
Australia	2,483
Austria	184
Belarus	253
Belgium	207
Bolivia	22
Bosnia & Herzegovina	43
Brazil	567
Bulgaria	636
Canada	2,844
Chile	65
Colombia	14
Costa Rica	19
Republic of Croatia	330
Czech Republic	567
Denmark	95
Dominican Republic	1
Ecuador	9
Estonia	9
Finland	1
France	1,675
Georgia	6
Germany	677
Greece	303
Hungary	730
Ireland	5
Israel	8,474
Italy	419
Japan	1
Kazakhstan	6
Latvia	77
Lithuania	133
Macedonia	9
Mexico	112
Moldova	283
Netherlands	1,051
New Zealand	55
Norway	34
Peru	2
Poland	1,429
Portugal	2
Romania	147
Russia	712
Slovakia	665
Slovenia	12
South Africa	254
Spain	6
Sweden	331
Switzerland	68
Ukraine	3,434
United Kingdom	873
United States	19,843
Uruguay	126
Uzbekistan	25
Venezuela	227
Yugoslavia	361
Zimbabwe	6

Total:
51,649 testimonies
57 countries

Testimony Language Statistics

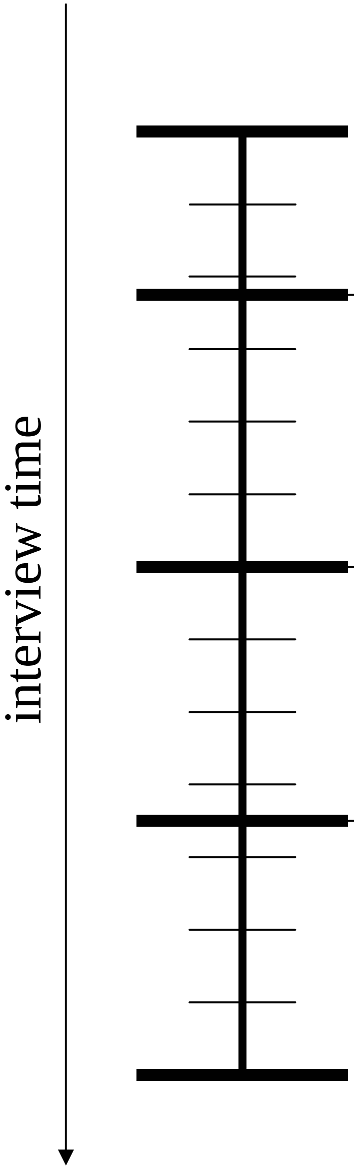
Bulgarian	622
Croatian	394
Czech	574
Danish	72
Dutch	1,080
English	24,947
Flemish	5
French	1,886
German	933
Greek	303
Hebrew	6,317
Hungarian	1,285
Italian	432
Japanese	1
Ladino	10
Latvian	6
Lithuanian	45
Macedonian	9
Norwegian	34
Polish	1,571
Portuguese	563
Romani	28
Romanian	123
Russian	7,011
Serbian	374
Sign *(3 American & 1 Hungarian)*	
Slovak	574
Slovenian	6
Spanish	1,350
Swedish	269
Ukrainian	318
Yiddish	513

Total:
51,649 testimonies
32 languages

Manual Indexing System

- Cataloguers listen to the audio data
- Divide data into large segments
- For each large segment
  - Divide into smaller segments
  - For each smaller segment, make notes on what the speaker said
  - Annotate these notes with keywords that can be used to index this data
  - Associate with video, stills, artifacts, etc.
  - Summarize these notes
  - About 4000 testimonies catalogued in this fashion
- Clearly expensive and time-consuming – depending upon the nature of the archive, cost may be prohibitive.
- Alternatively used fixed 1-minute segments

An Example



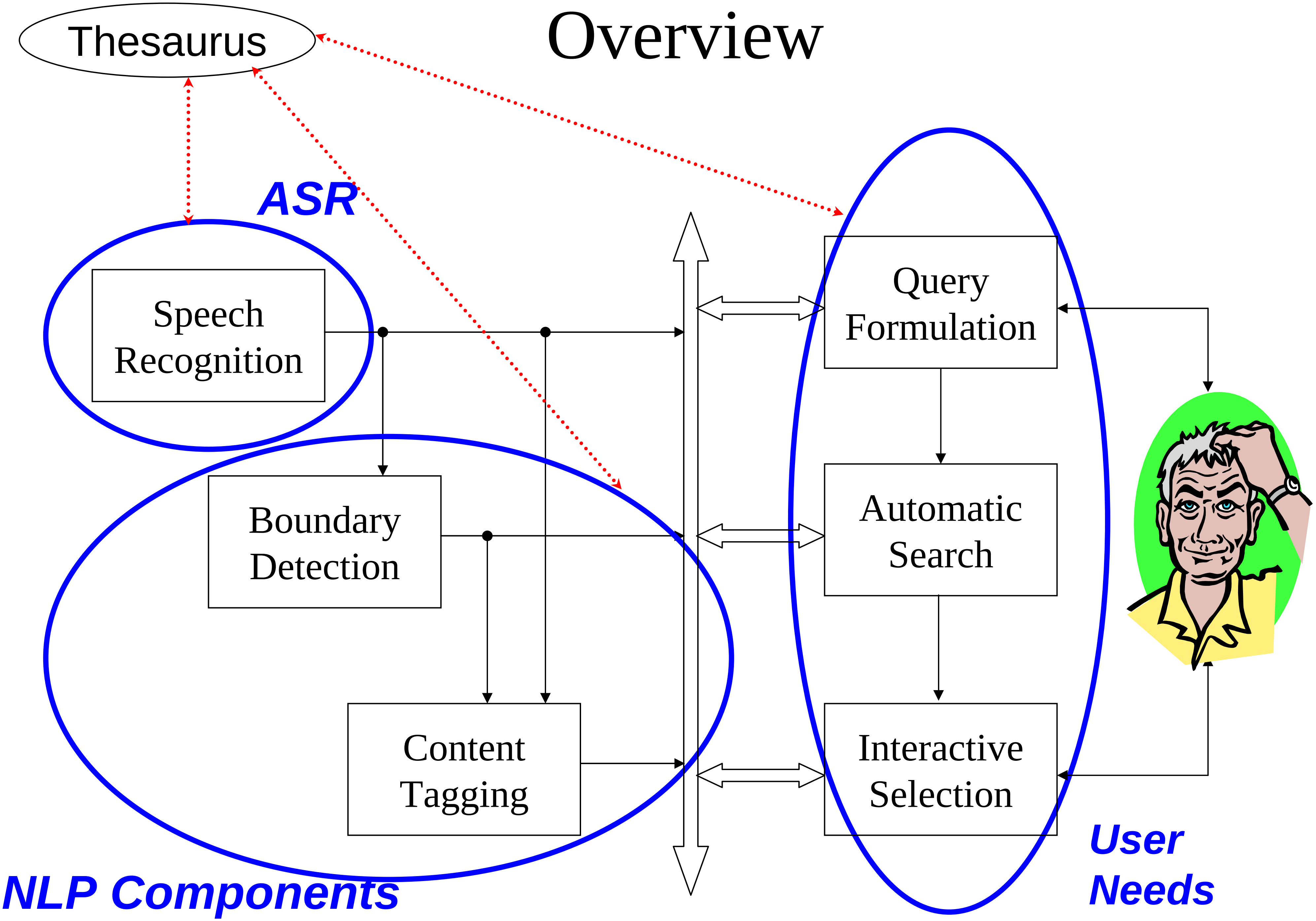
Location-Time	**Subject**	**Person**
Berlin-1939	Employment	Josef Stein
Berlin-1939	Family life	Gretchen Stein Anna Stein
Dresden-1939	Relocation Transportation-rail	
Dresden-1939	Schooling	Gunter Wendt Maria

MALACH: Multilingual Access to Large Spoken ArCHives

The objective of MALACH is to dramatically improve access to large multilingual spoken archives by capitalizing on the unique characteristics (unconstrained natural speech) of the Survivors of the Shoah Visual History Foundation's (VHF) multimedia digital archive of oral histories

- Specific goals include:
  - ➡ Advances in speech recognition technology to handle spontaneous and emotional speech with disfluencies, heavy accents, elderly speech, and dynamic switching between multiple languages
  - ➡ Advances in information retrieval technologies to provide efficient indexing, search and retrieval
  - ➡ Automated techniques for the generation of new metadata to label segments
  - ➡ Automated translation of domain-specific multilingual thesauri
  - ➡ Workshops and user studies to evaluate the social and scientific value of the technology and see how it can be applied to other large archives.

Overview



Thesaurus

ASR

Speech Recognition

Boundary Detection

Content Tagging

NLP Components

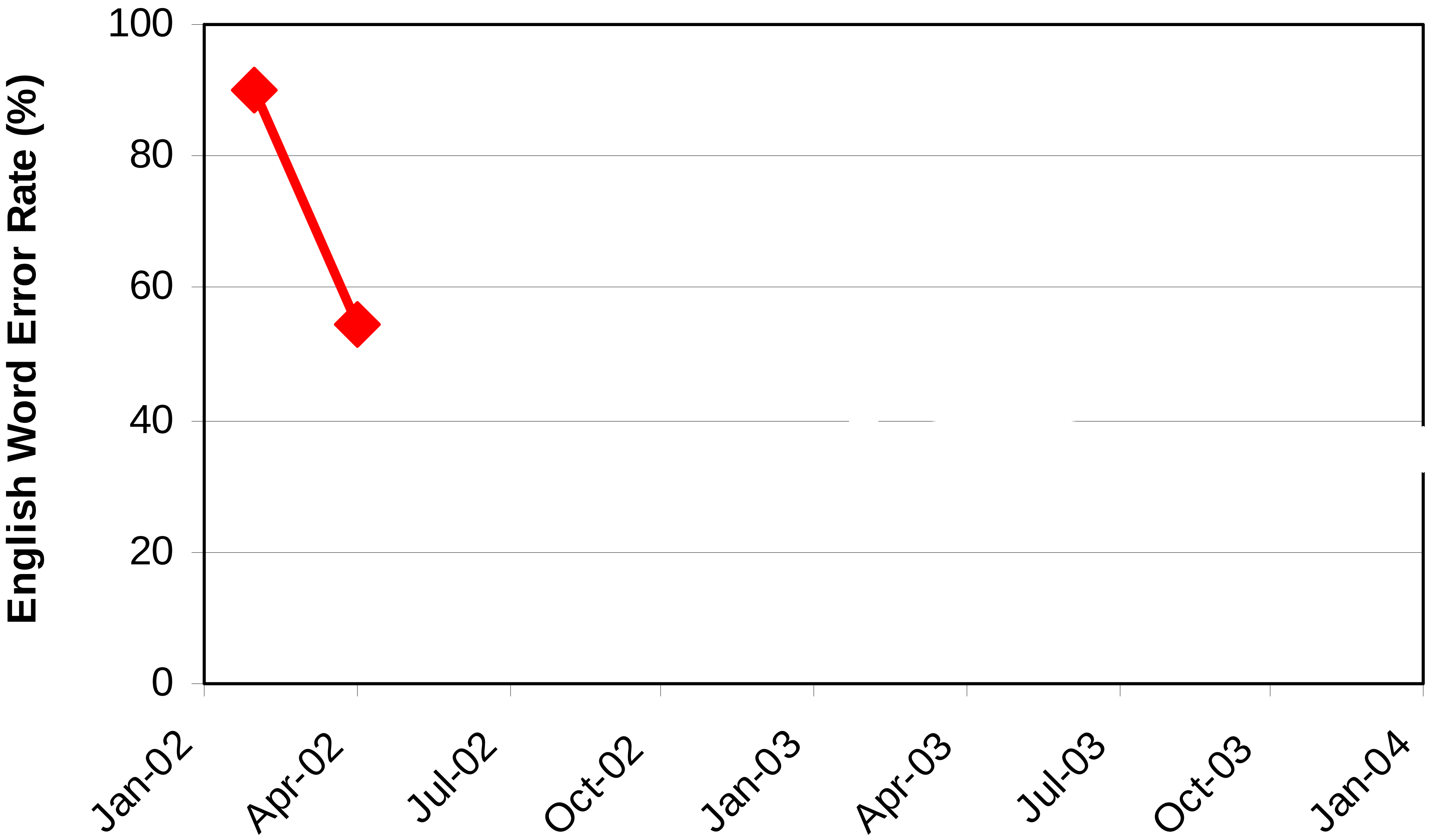
Query Formulation

Automatic Search

Interactive Selection

User Needs

English ASR Accuracy



Why is Speech Recognition Hard?

- **Unusual Words**
  - My middle name m- my my middle brother he had two names in lost- in- before the war Shloma Hasich and me, that’s Chuna Moskovitch, I was the baby at home and the sisters name was Miriam all were Mosokowiz
  - **my middle name** from my mental emitter but out the heck in the **shloma** hostage the meat and scorn are much as **I was the baby home** and desist his name rose mary an
- **Disfluencies**
  - A- a- a- a- band with on- our- on- our- arm
  - **a** hat and bend **with** the on **on our** farm
- Emotional speech
  - a young man they ripped his teeth and beard out they beat him
- Sections of frequent interruptions
  - CHURCH TWO DAYS these were the people who were to go to march TO MARCH and your brother smuggled himself SMUGGLED IN IN IN IN
  - **church** H. to data this these **people** who have to go to court each and two brothers **smuggled** some drugs and

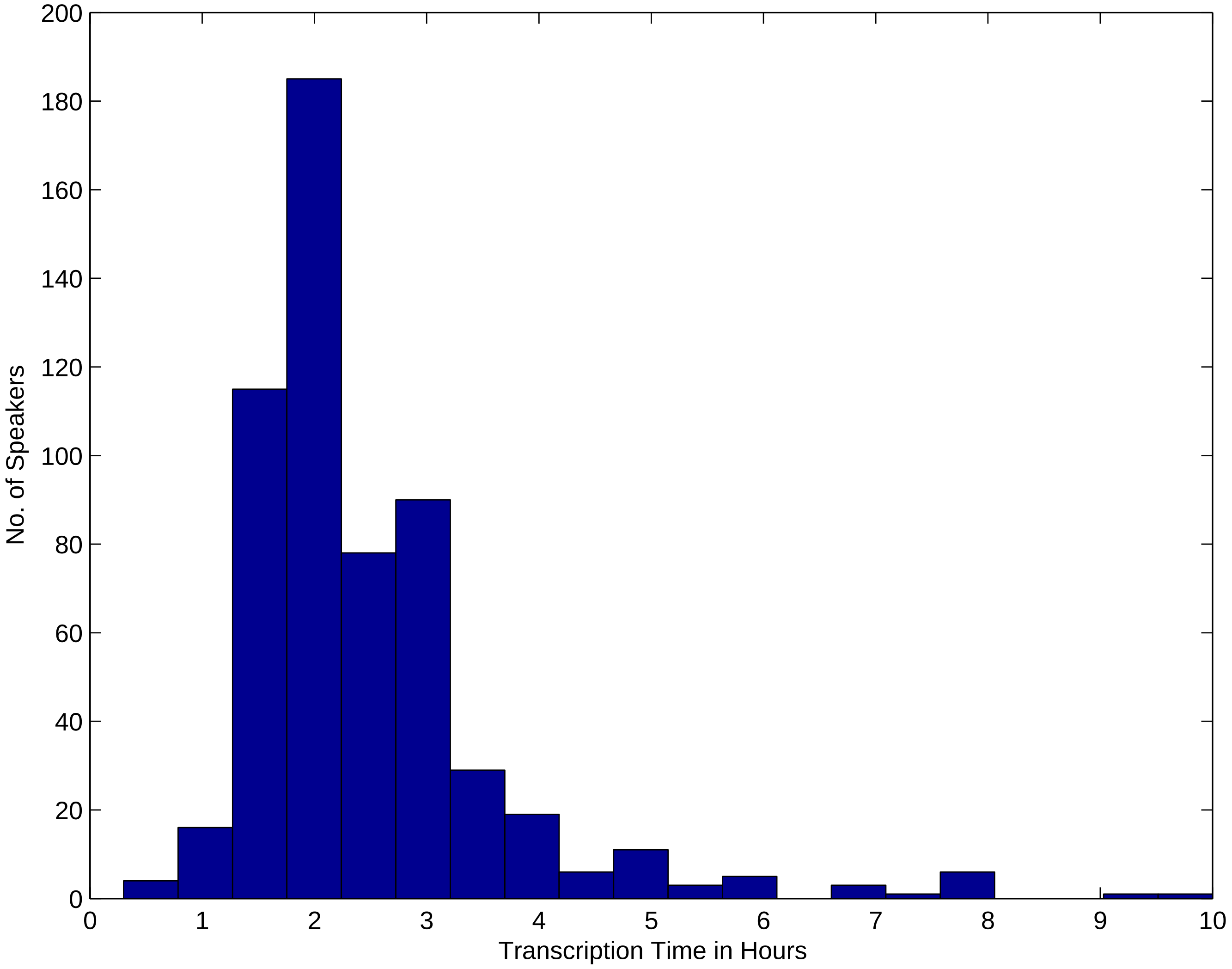
Unexpected Surprises

- Stereo format recordings with interviewee and interviewer in the same channel
- Some with low volume and some with no data in it at all
- Many, many non-English testimonies
  - There is no guarantee that a testimony is in English, even if the interviewer starts speaking in English and says that it is in English!
- As many as 9 speakers in some testimonies
- Lots of cross talk – less of this in interviewers with British and Australian accents
- Some interviewees say very little.
  - A few testimonies, interviewers did all the talking – forced yes/no type answers

Other observations

- Lots of foreign words, unsure words, names, places
- Noisy Background:
  - Static noise, Airplane noise, Buzzing Sound, Hammering noise in the background, Coughing, Laughter, Emotion (crying, screaming), Many conversations in the background, Badly placed microphone

Histogram of Transcription times



Examples of foreign words, names…

ADAKCLAUS ADDUS-YIS-HOREL

ARBEIT-MACHT-FREI ARNHEIM

ARONAFISCHSTRASSEN

BABUSHKAS CZESTOCHOWA

HA-NOR-YAT-SA-NEE HASLACH

JUDENANRAT SZMALCONIKI

VERMIETEN YANZICHITZ

YAKUBOVICH YITZKAH

YU-OV-DOV-SKY YUDENLAGER

ZWILLINGEN ZOSHA

ASR Performance

- Gender Dependent Systems
  - – Two gender dependent systems trained with about half the training data (~100h male speakers, ~78h female speakers)

		65h		200h
SI	45.5	46.6	41.0	42.3
SAT	41.9	43.3	37.6	38.2
MLLR	39.4	39.6	35.1	35.2

- Performance improvements of 1.4% absolute at the SAT level obtained with 65h of training data went away after MLLR
- Gains not seen with 200 hours of training data (0.6% overall gain with gender dependent systems)

Decoding the Test Collection

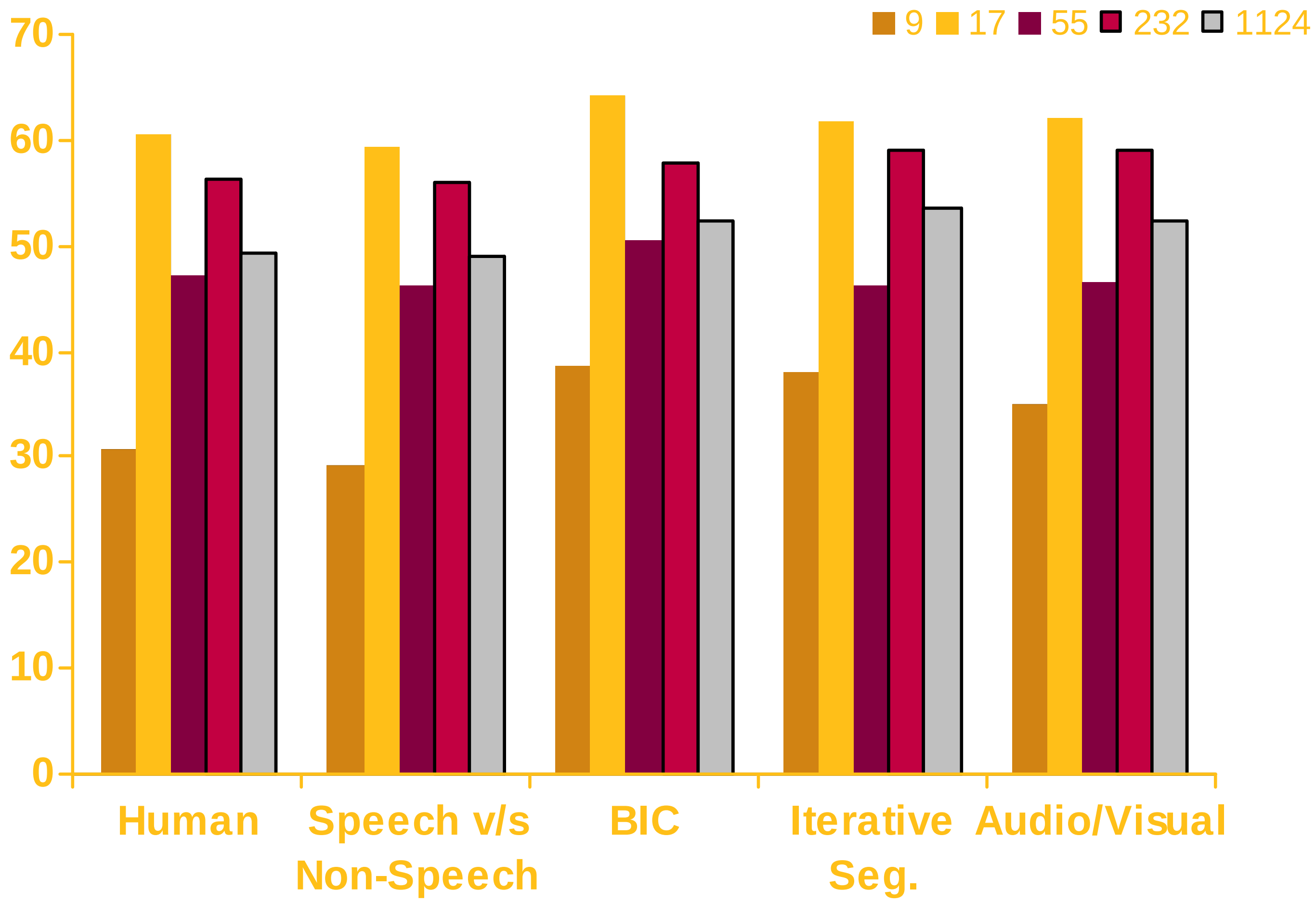
- **Why is this important?**
  - Test collection is being used in training models for automatic topic segmentation, categorization and search
- **Collection Details**
  - Compressed audio (Sampling Frequencies: 44.1 KHz and 48KHz)
  - 625 hours done (computing done ~4xRT)
    - 580 hours of speech
    - Models used had an SI WER of 46.7% and speaker-dependent word error rate of 39.6%

Total Tapes	Full Testimonies	Partial Testimonies
1294	199	47

Why is acoustic segmentation necessary? (Eurospeech 2003)

- Automatically identify and remove non-speech segments
- Reduce computational load
- Speaker labeling of segments allows adaptation to be performed on speaker-coherent clusters
- Manual process is time-consuming and expensive
- Goal is to improve recognition performance on tens of thousands of hours of spoken material

First Pass Decoding with several automatic segmentation schemes



Segment Clustering

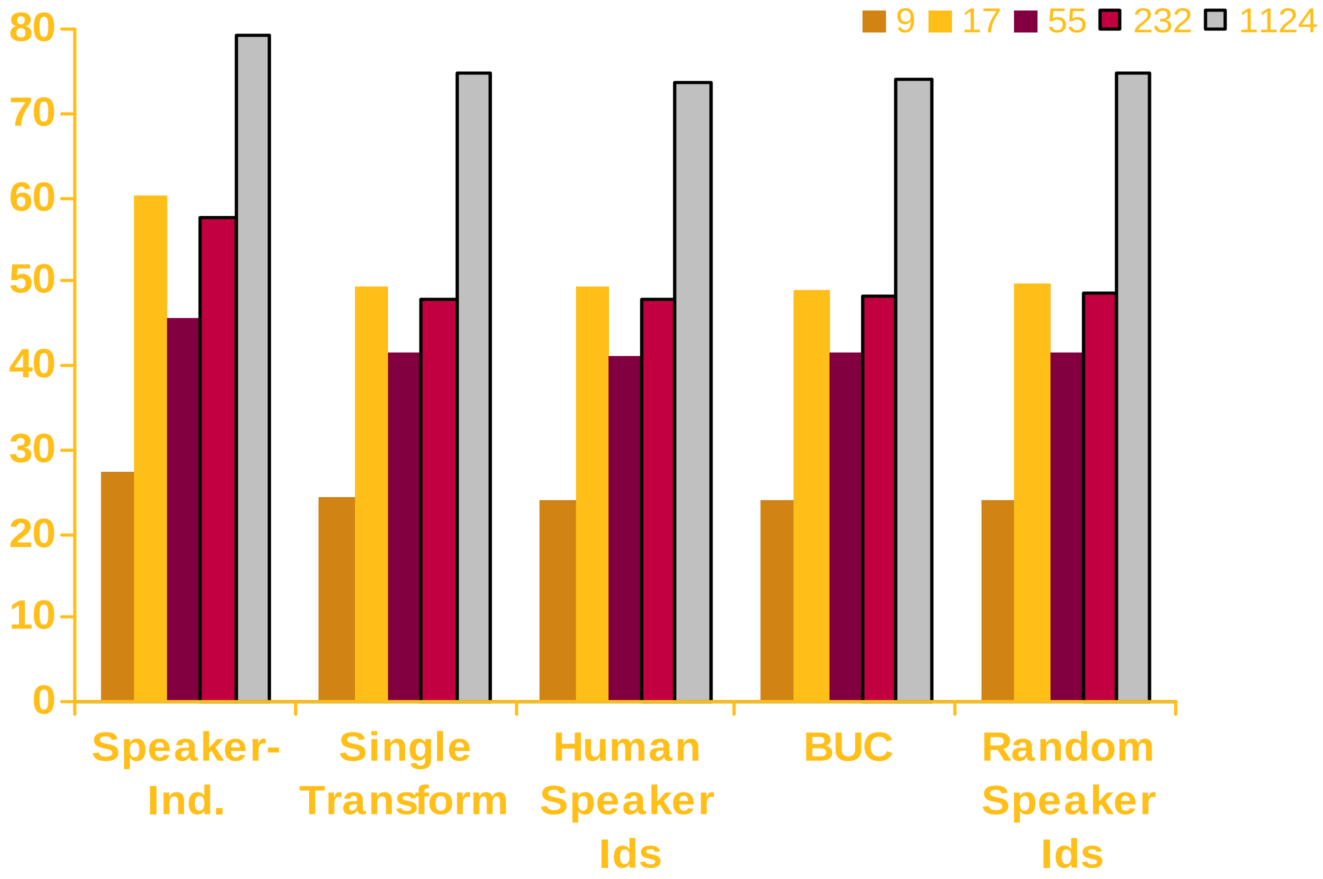
- Bottom-up clustering scheme to two clusters (interviewee and interviewer)

- Single Cluster (i.e one transform only)

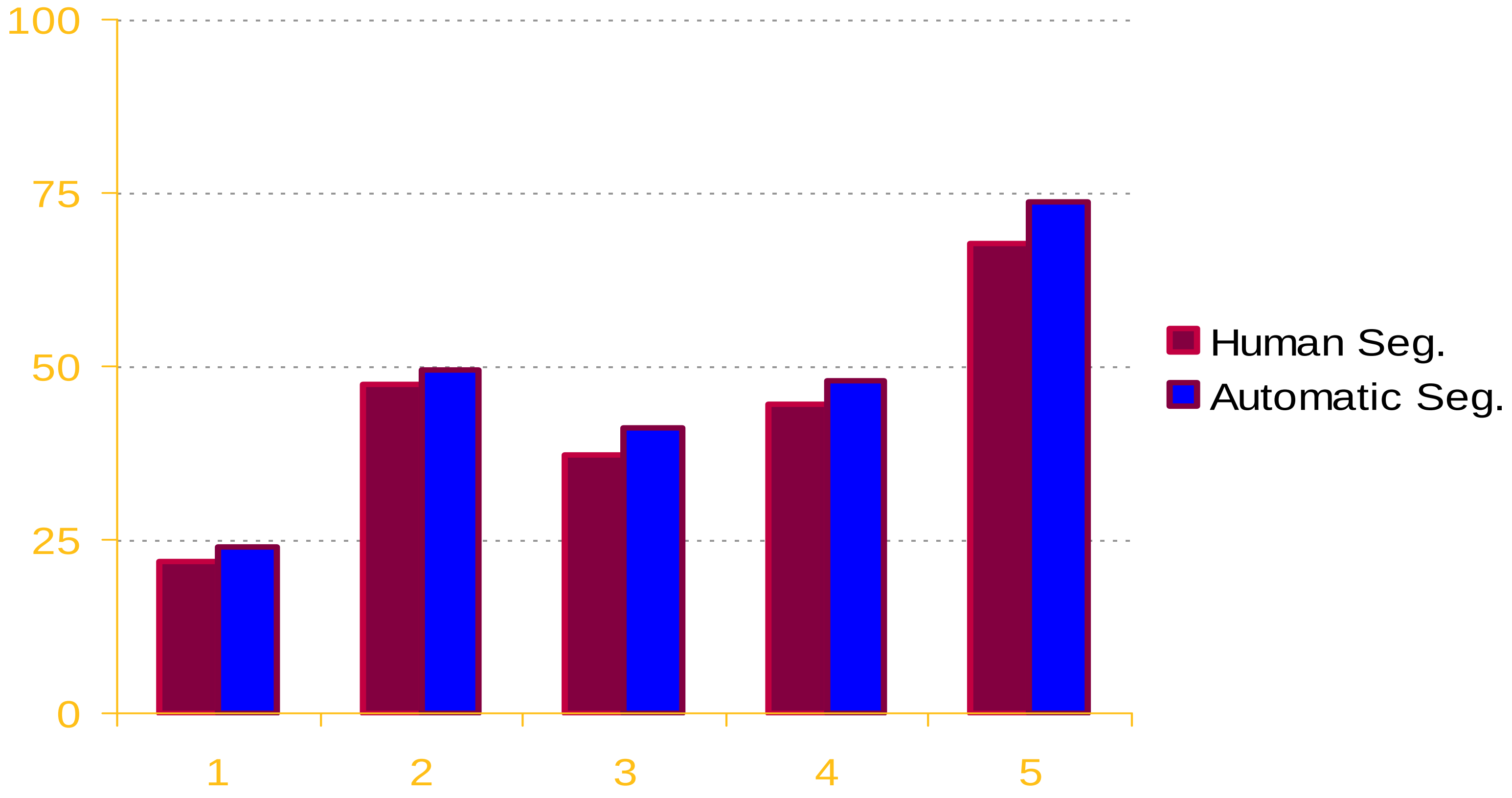
- Manually marked speaker ids

- Randomly assigned speaker ids

WER : Effect of Automatic speaker clustering on Automatic Segmentation (Speech/Non-Speech scheme)



WER after adaptation – how far are we from the best we can do?



Lessons learned

- Automatic segmentation schemes can do as good as if not better than manual segmentation
- For adaptation, best performance is obtained when the segments are speaker-coherent
- Significant impact on interviewer’s speech (less than 18%) and mostly in impure segments
- Future work to focus on deriving speaker-pure segments

ASR accuracy on names, locations and organizations *(named entities)*

- Manual Annotations on 3 ½ hours of a testimony used as reference
  - Named entities: 593
    - Person names: 118 (56 uniq names)
    - Locations: 229 (63 uniq names)
    - Organization names: 61 (17 uniq names)
    - Country names: 185 (17 uniq names)

- Overall recognition accuracy on NE : 28%

Pronunciations

- Language of origin of the words was used as a guiding principle to capture the most likely (representative) pronunciation
- German was the most frequent first rank variant language
- US English variants were added by default
- Distribution on a reasonable sample set

French	39%
Polish	20%
Hungarian	12%
Russian	11%
Italian	5%
Czech	5%
Dutch	4%
Spanish	4%

WER goes down by 1%!!

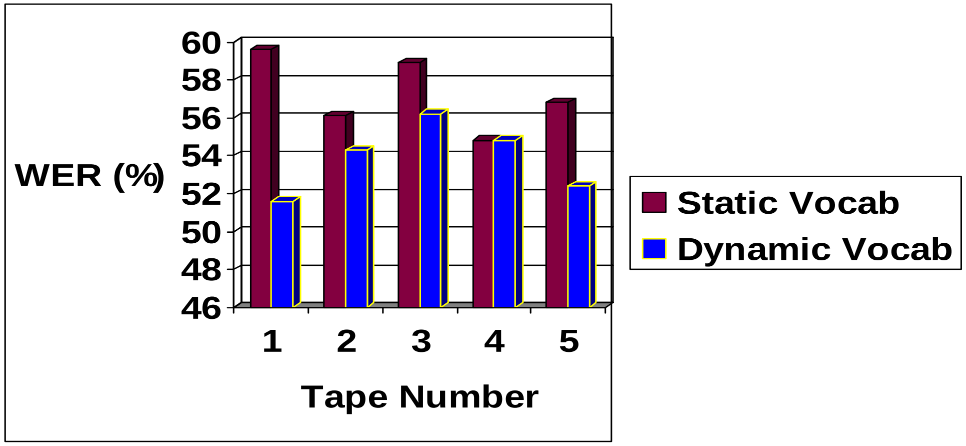
Syllable centric models (ASRU 2003)

- Insufficient coverage for many syllables in training data. Also, test data vocabulary is different and introduces new syllables. Thus we need mixed phonetic-syllable pronunciations.
  - Phonetic: B ER K AX N AW
  - Syllabic: B _ER K_ AX N _AW
  - Mixed : B ER K _AX N _AW
- 5796 distinct syllables in the MALACH vocabulary
- WER improves marginally (0.5%)

Dynamic lexicon

- Different vocabulary for different testimonies
- Built using PIQ and Segment_PIQ_Person information
  - Accuracy on Named Entities: 49%

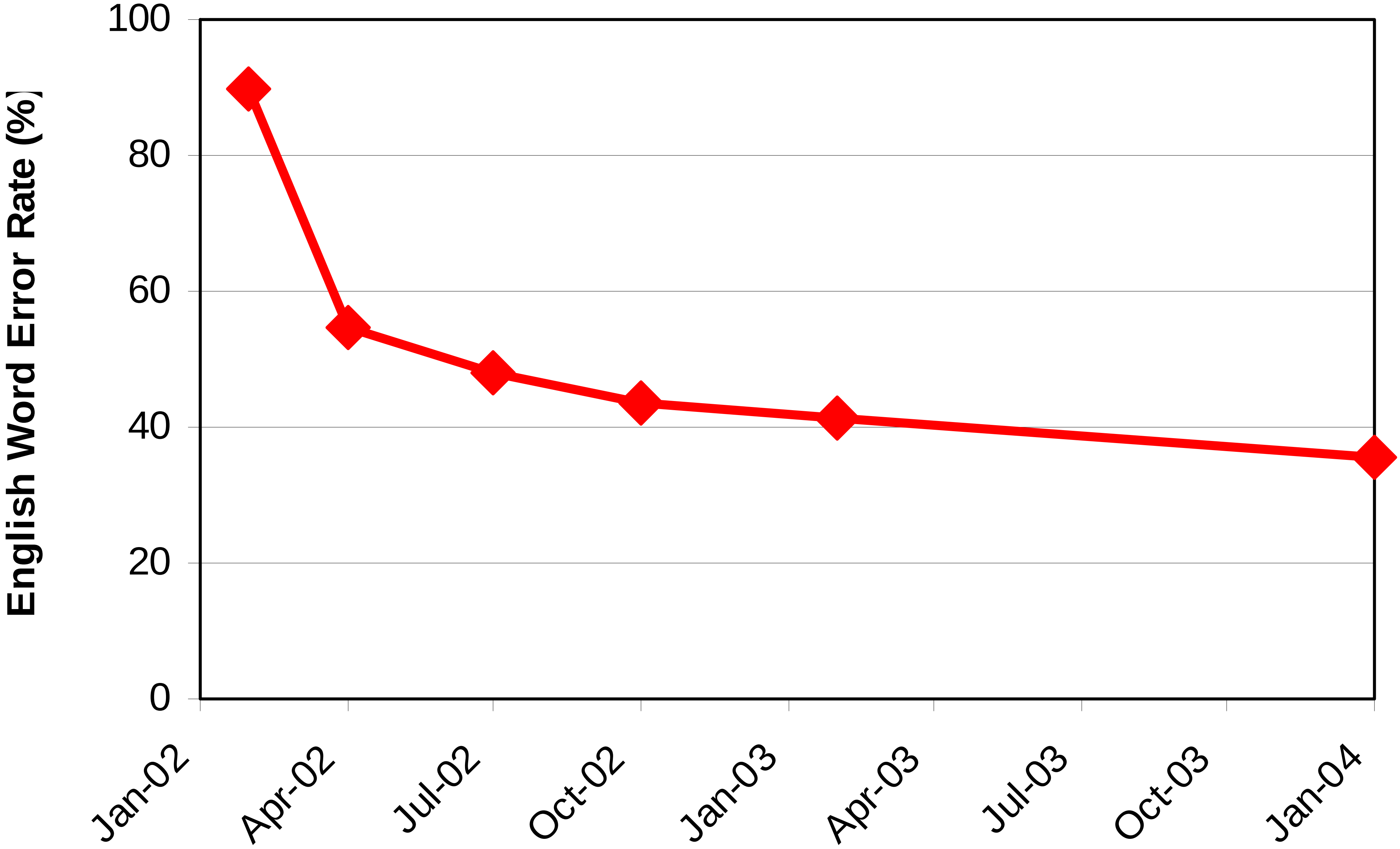
Overall WER Variation across tapes



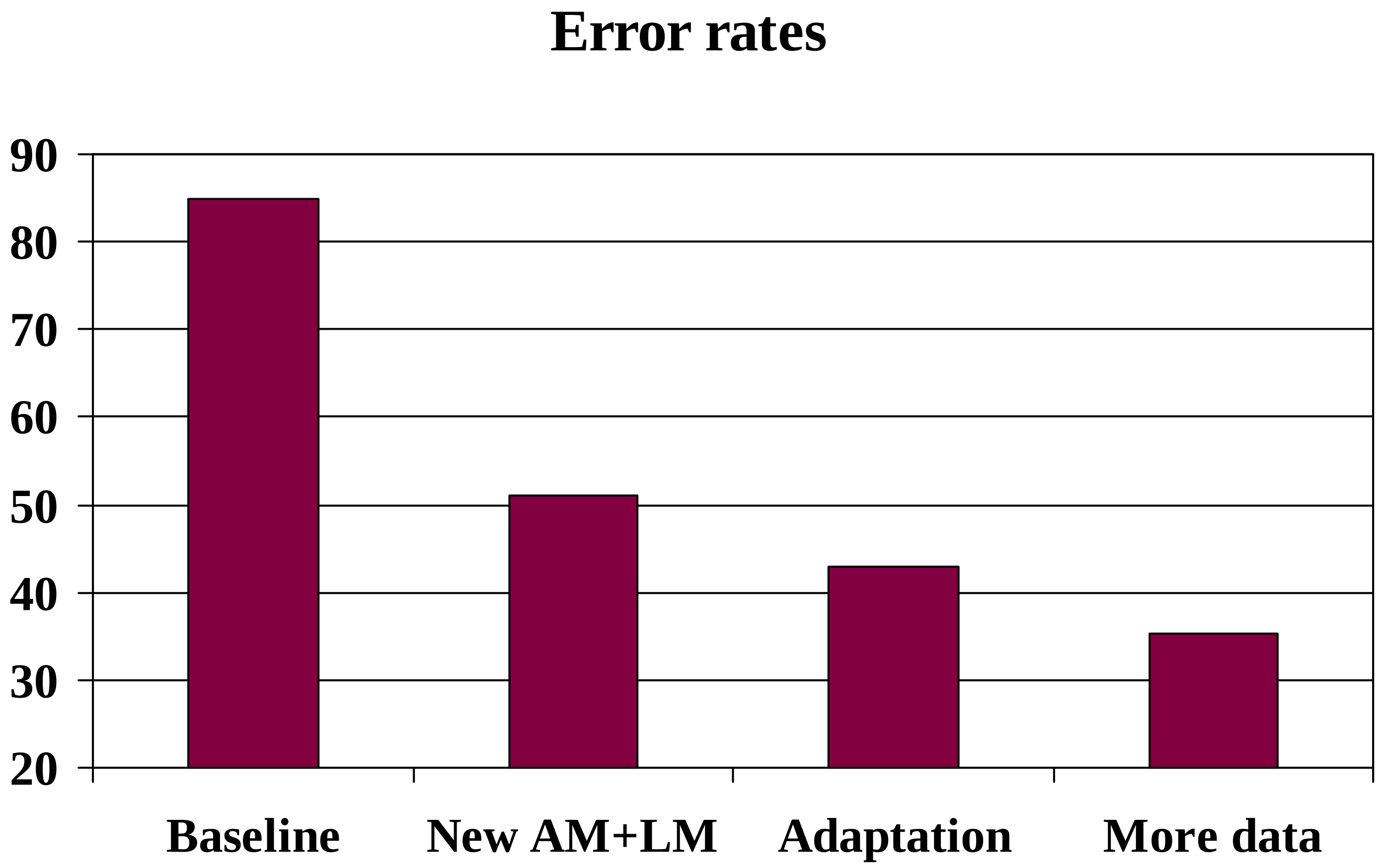
OOV on NE: 25.5%

Vocab	NE **Accuracy** (%)	Overall **WER** (%)
Static	31	47.6
Dynamic	48	43.4
Gain	54.8	8.8

English ASR Accuracy



ASR Summary



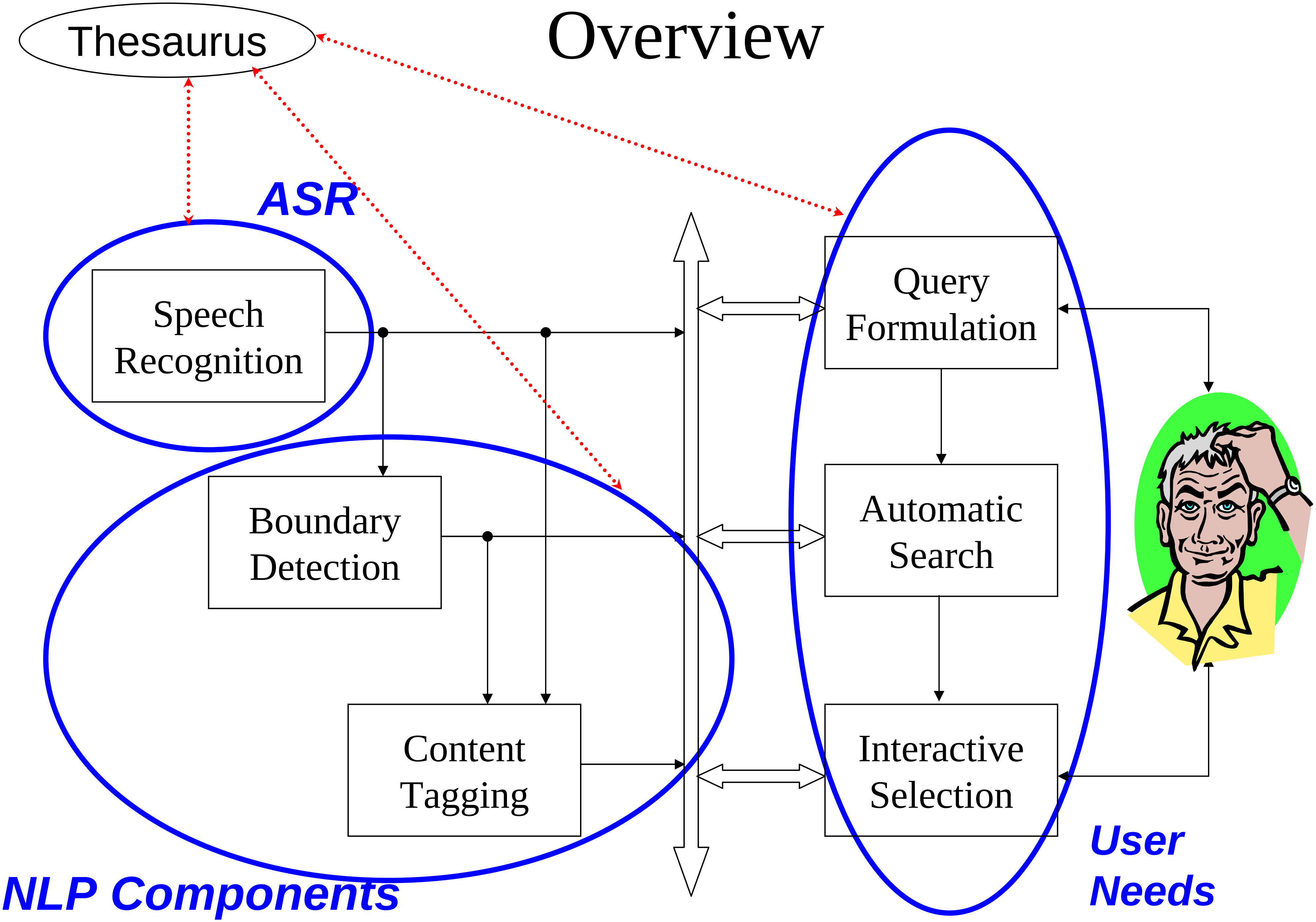
Short-term enhancements:

- System combination
- Improved vocabulary coverage
- Additional training data

Long-term enhancements:

- Accent and disfluency modeling
- Adaptation
- Robustness to background noise and speech
- Segmentation, Speaker id

Overview



Thesaurus

ASR

Speech Recognition

Boundary Detection

Content Tagging

NLP Components

Query Formulation

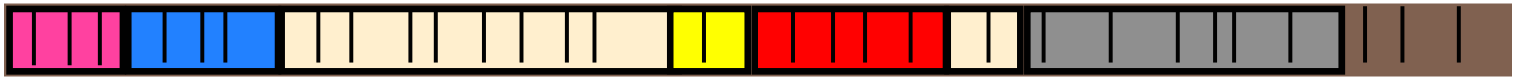
Automatic Search

Interactive Selection

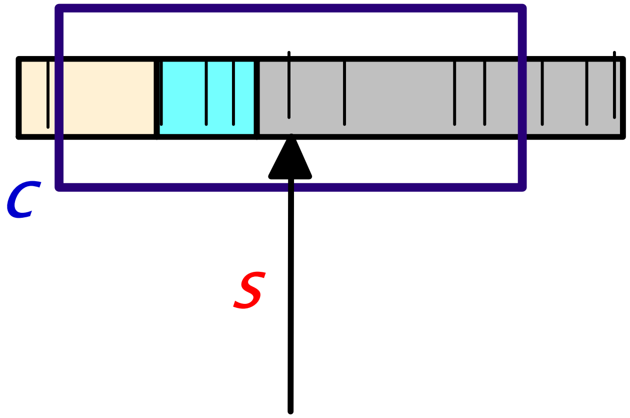
User Needs

Boundary Detection (Segmentation)

- Identify topically cohesive intervals in a stream of text
- Compute the probability of a topic boundary occurring at a given sentence boundary



Statistical Models for Segmentation



Probabilistic models for $P(s \mid c)$

- s a binary random variable denoting presence or absence of topic boundary at any given point
- c context -- text and acoustics surrounding any given point
- binary features: $\phi(s, c) \rightarrow [0\ 1]$

Combination of Decision Tree and Maximum Entropy models

Topic Segmentation – Data Sample

… because the roads were crowded with with army units going back and forth you know .. and you also were off you had to walk no on the main road because you were afraid you were going to be picked up for work .. that's what some did they came to Loetche and some people were picked up and held four weeks for work .. when they came home they told us

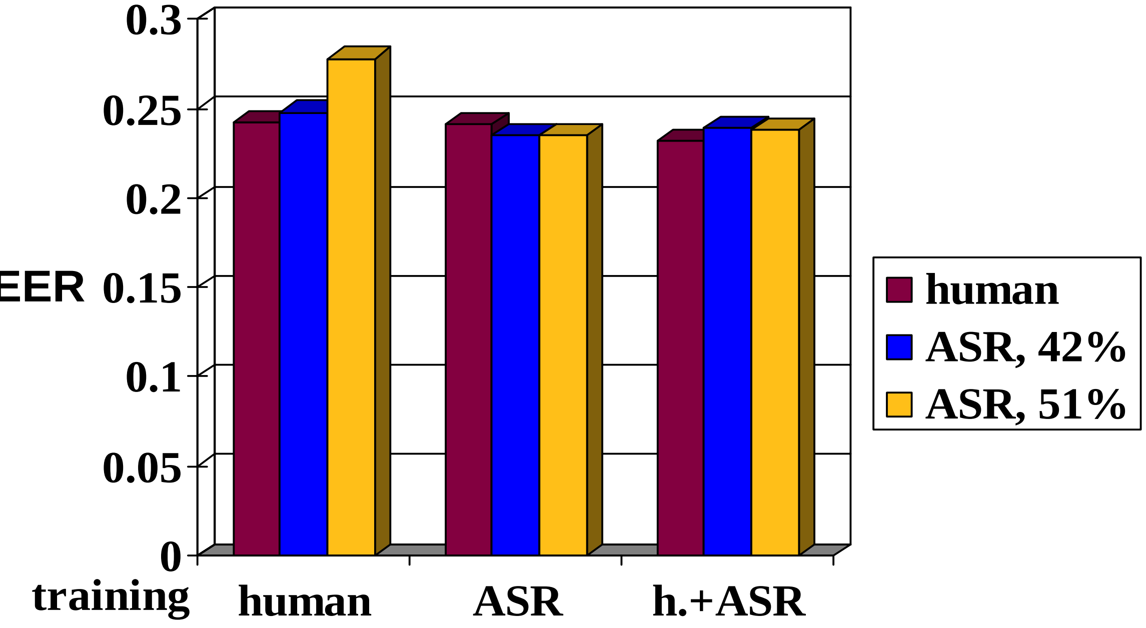
on the way

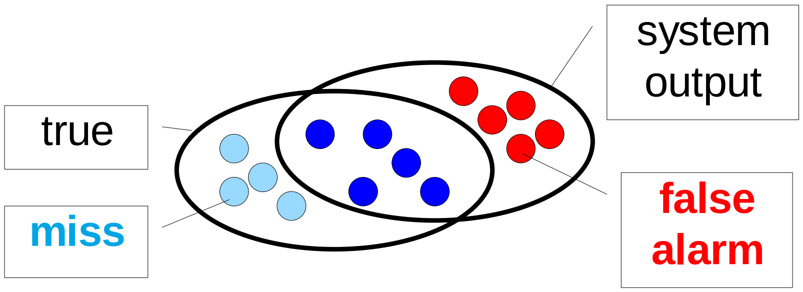
--- segment boundary ---

we came we came home was was about the time of Succoth .. you know the city was deserted there was a they were already taking people to work .. when we came home we couldn't recognize the city .. my parents first of all they confiscated everything .. they told us to get out of the orchard .. they took whatever they wanted they took over the whole ranch ...

arrival

Topic Segmentation: ASR-based Training





Equal Error Rate (Miss Rate = False Alarm Rate)

test \ training	human	ASR	human+ASR
human	0.242	0.241	0.232
ASR, 42% WER	0.248	0.235	0.239
ASR, 51% WER	0.278	0.235	0.238

Segment with Keywords

my brother my sister and I went to live with my grandmother in Billibeck Westphalia and we spend a year there and we went to school there and this little town of two thousand was Catholic

and I had a lot of good friends there

I went to the public school back into grade school because they did n't have any high school in this little town

then my parents left Moers they went to Billib- I mean they went to Berlin

so my sister and my brother and I moved with them in nineteen thirty six

we were enrolled in a private Jewish school

it took my father a very long time to find a position and he finally found one as a sales rep for a men 's wear in a

and the naturally they started to prepare us for emigration

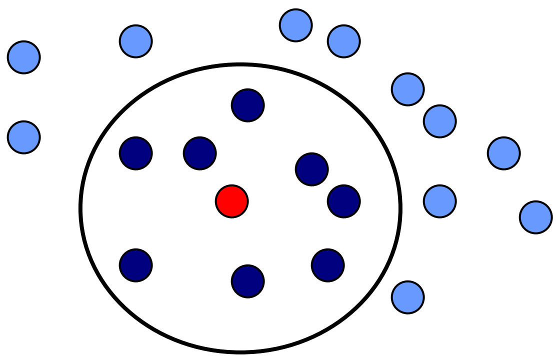
and my last year in Germany in thirty eight to thirty nine it was intense English study

- Billerbeck (Germany)
- Jewish-gentile relations
- education
- Jewish schools
- Berlin
- occupations, father's
- Germany 1933 (January 31) - 1939 (August 31)
- separation of loved ones
- flight preparations

Categorization With K-nearest Neighbors

“Segment is assigned to the same categories as the segments similar to it.”

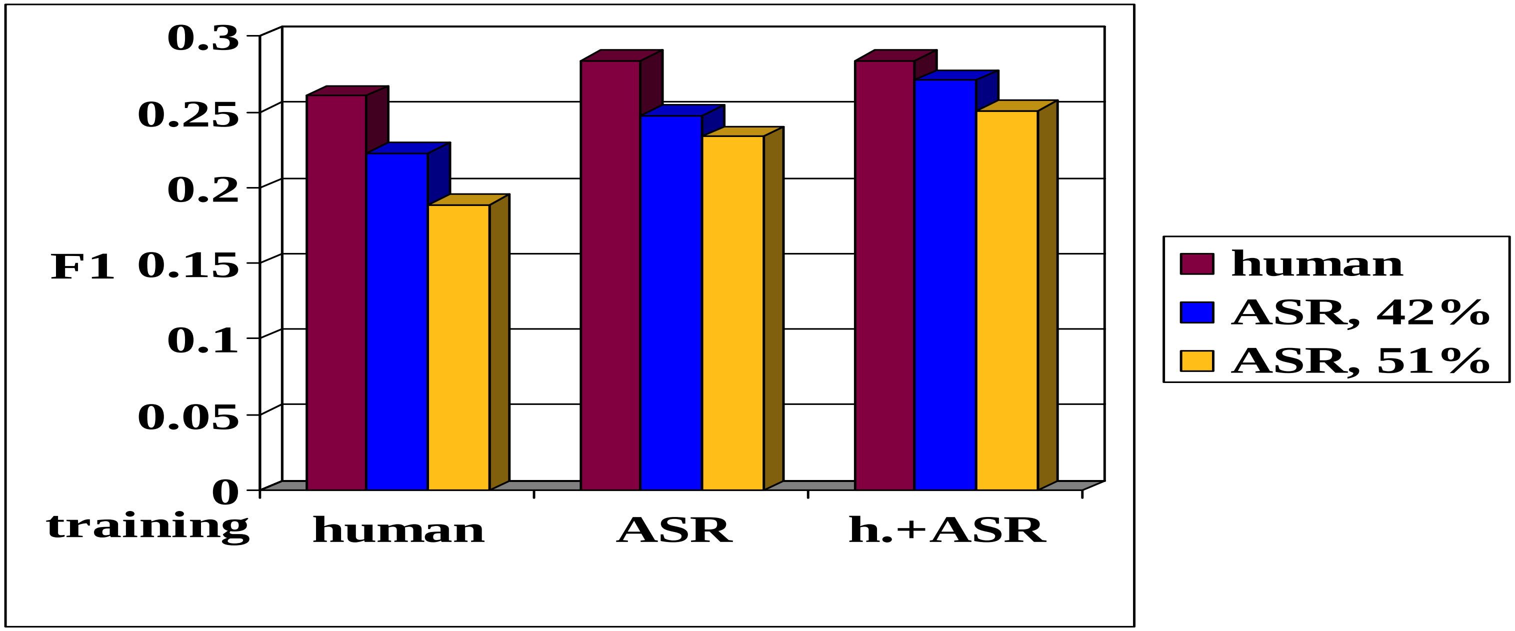
$$score(s,c) = \sum_{s_i \in kNN} sim(s, s_i) cat(s_i, c)$$



Segment-to-segment similarity, sim(s,si) is the symmetrized Okapi measure

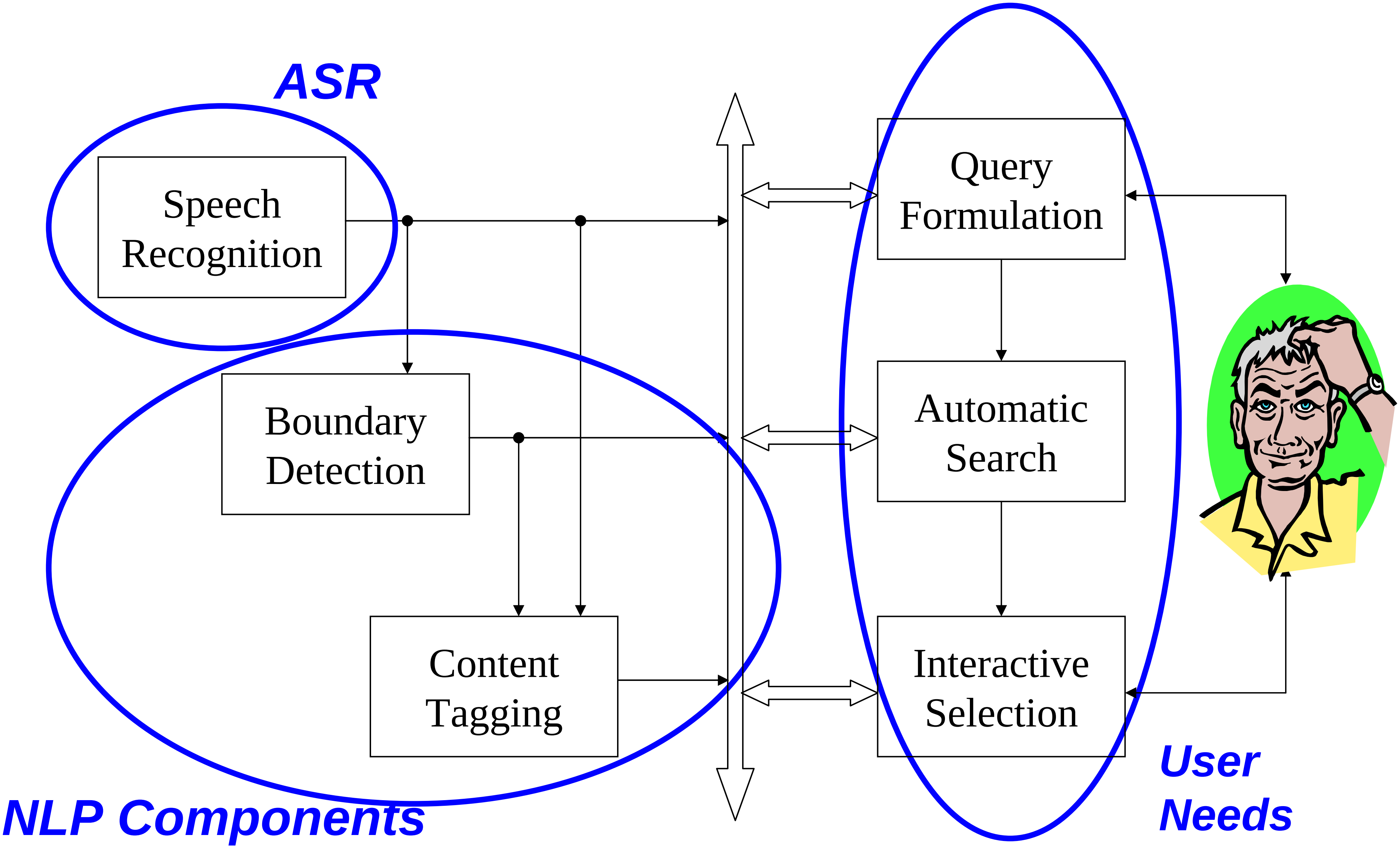
```
for each segment:
  find kNN
  for each category represented in kNN:
      compute score(s,c)
      if score(s,c) > threshold
                assign document to category
```

ASR Training in Categorization



test \ training	human	ASR	human+ASR
human	0.261	0.284	0.284
ASR, 42% WER	0.223	0.248	0.271
ASR, 51% WER	0.189	0.234	0.251

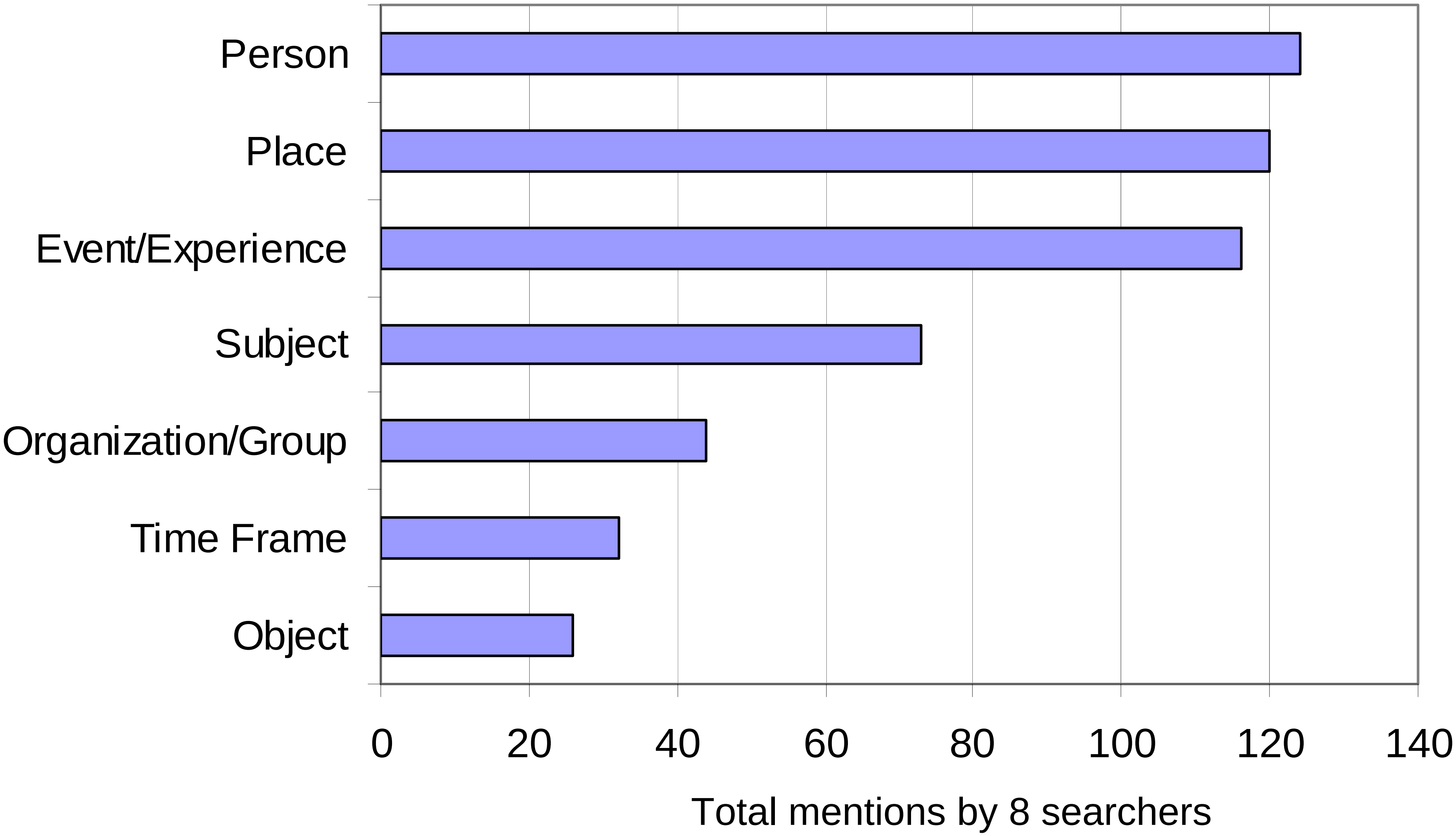
Overview


ASR
Speech Recognition
Boundary Detection
Content Tagging
NLP Components
Query Formulation
Automatic Search
Interactive Selection
User Needs

Search
Construction of topics to search for

- 600 written requests, in folders at VHF
  - From scholars, teachers, broadcasters, …
- 280 topical requests
  - Others just requested a single interview
- 50 selected for use in the collection
- 30 assessed during Summer 2004
- 28 yielded at least 5 relevant segments

What do searches look like?



Workshops 1 and 2

An Example Topic

<top>

<num> Number: 1148

<title> Jewish resistance in Europe

<desc> Description:

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

<narr> 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.

<folder> Folder Label:

Traveling exhibit on Jews in the resistance

</top>

<DOC>**<DOCNO>**
VHF00009-056149</DOCNO>

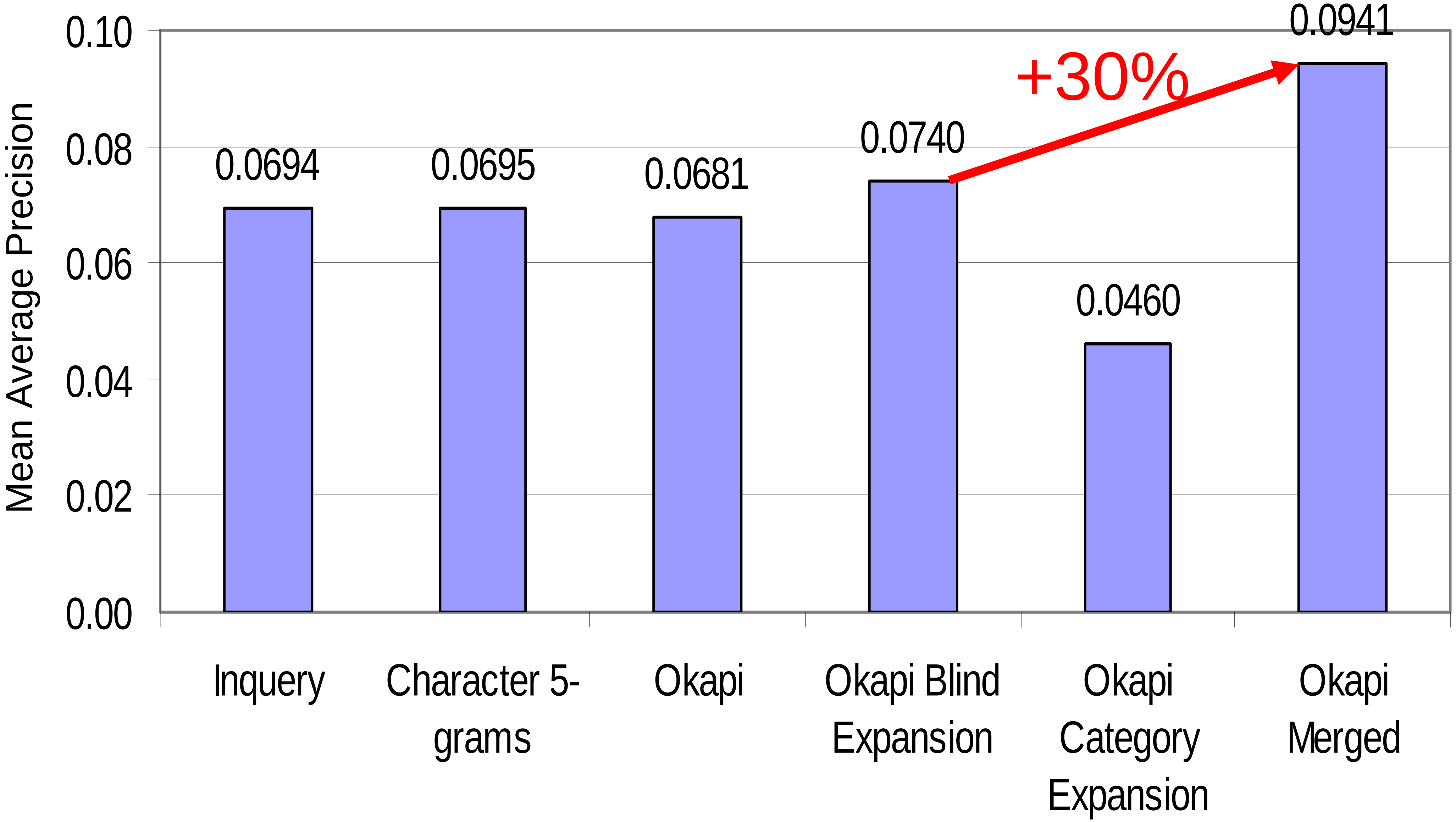
<KEYWORD> grandfathers, socioeconomic status, Przemysl (Poland), Poland 1926 (May 12) - 1935 (May 12), Poland 1935 (May 13) – 1939 (August 31), cultural and social activities </KEYWORD>

<PERSON> </PERSON>

<SUMMARY> SL remembers her grandfather. She talks about her town. SL recalls her family's socioeconomic status and her social and cultural activities. </SUMMARY>

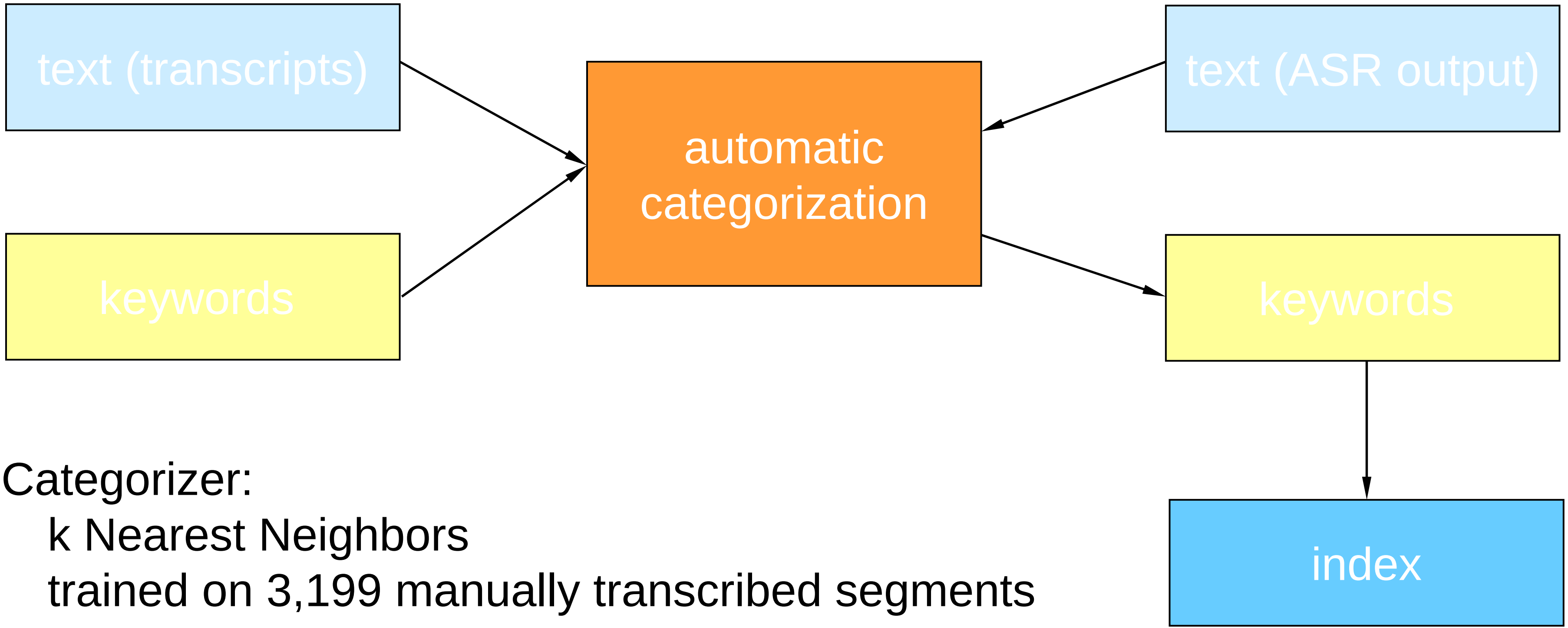
<ASRTEXT> oh i'll you know are yeah yeah yeah yeah yeah yeah yeah the very why don't we start with you saying anything in your about grandparents great grandparents well as a small child i remember only one of my grandfathers and his wife his second wife he was selling flour and the type of business it was he didn't even have a store he just a few sacks of different flour and the entrance of an apartment building and people would pass by everyday and buy a chela but two killers of flour we have to remember related times were there was no already baked bread so people had to baked her own bread all the time for some strange reason i do remember fresh rolls where everyone would buy every day but not the bread so that was the business that's how he made a living where was this was the name of the town it wasn't shammay dish he ours is we be and why i as i know in southern poland and alisa are close to her patient mountains it was rather mid sized town and uhhuh i was and the only child and the family i had a governess who was with me all their long from the time i got up until
i went to sleep she washed me practice piano she took me to ballet lessons she took me skiing and skating wherever there was else that I was doing being non reach higher out i needed other children to players and the governors were always follow me and stay with me while ours twang that i was a rotten spoiled care from work to do family the youngest and the large large family and everyone was door in the army </ASRTEXT>
</DOC>

ASR-Based Search



Title queries, topical relevance, adjudicated judgments

Automatic Categorization in Retrieval

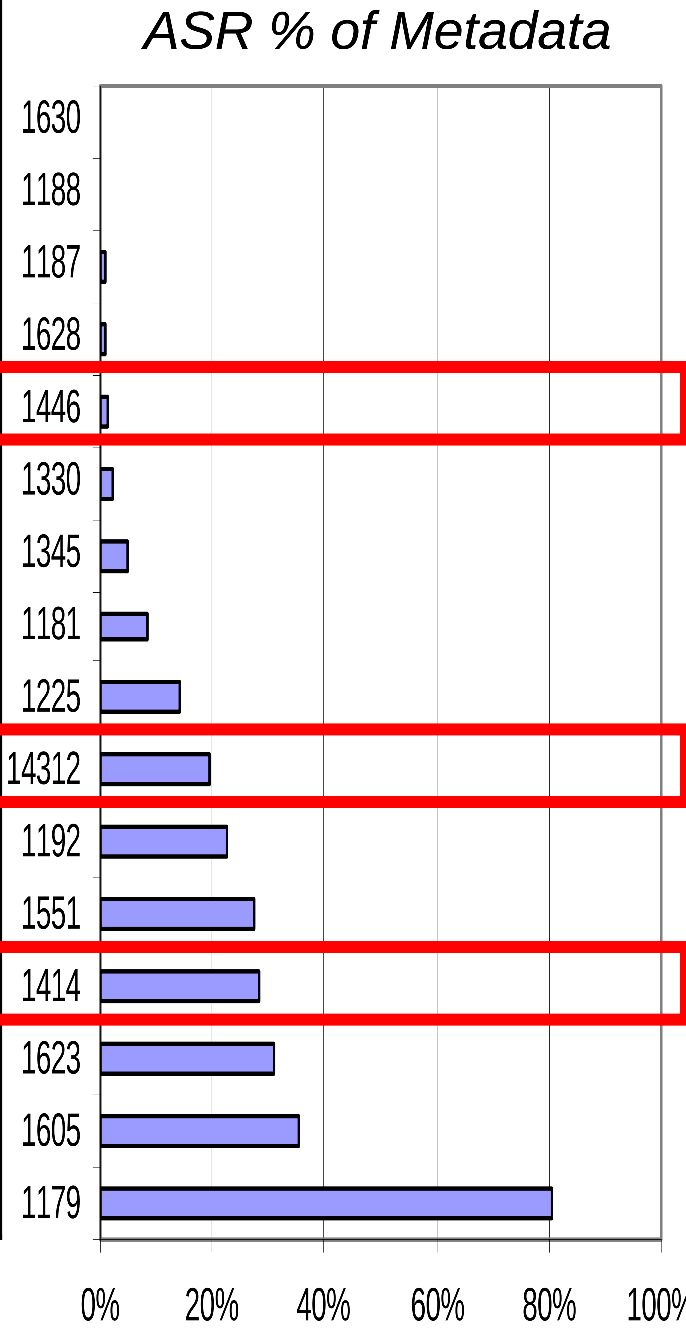


Categorizer:

k Nearest Neighbors
trained on 3,199 manually transcribed segments
micro-averaged F1 = 0.192

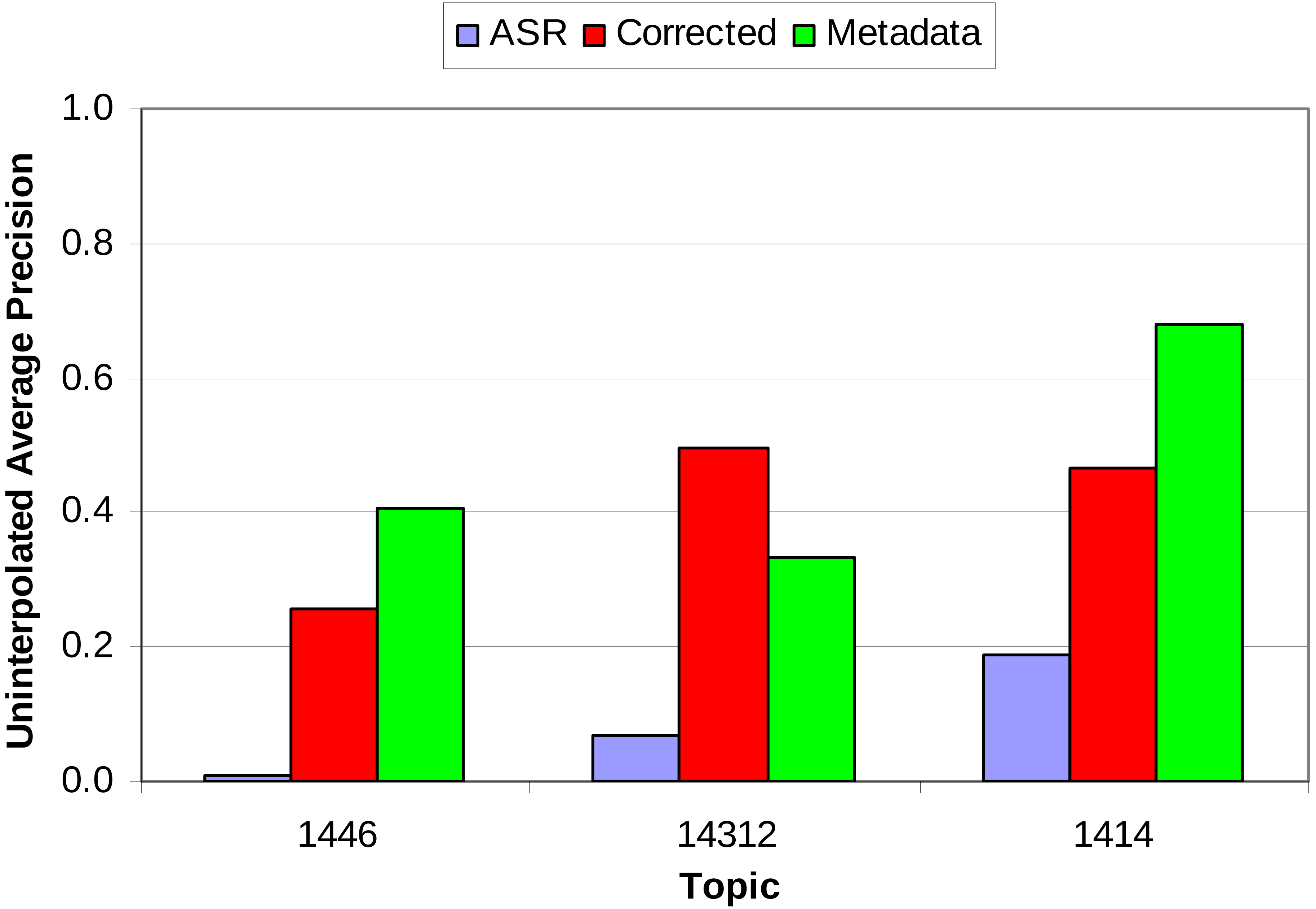
Error Analysis

*Somewhere in ASR Results (**bold** occur in <35 segments)*	*in ASR Lexicon*	*Only in Metadata*	
wit		eichmann	
jew	volkswagen		
labor camp		ig farben	
slave labor		telefunken	aeg
minsk ghetto underground			
		wallenberg eichmann	
bomb birkeneneau			
sonderkommando auschwicz			
liber buchenwald **dachau**			
jewish kapo			
kindertransport			
ghetto life			
fort ontario refugee camp			
jewish partisan poland			
jew shanghai			
bulgaria save jew			



Title queries, adjudicated judgments, Inquery

Correcting Relevant Segments



Title+Description+Narrative queries

What Have We Learned?

- IR test collection yields interesting insights
  - Real topics, real ASR, ok assessor agreement
- Named entities are important to real users
  - Word error rate can mask key ASR weaknesses
- Knowledge structures seem to add value
  - Hand-built thesaurus + text classification

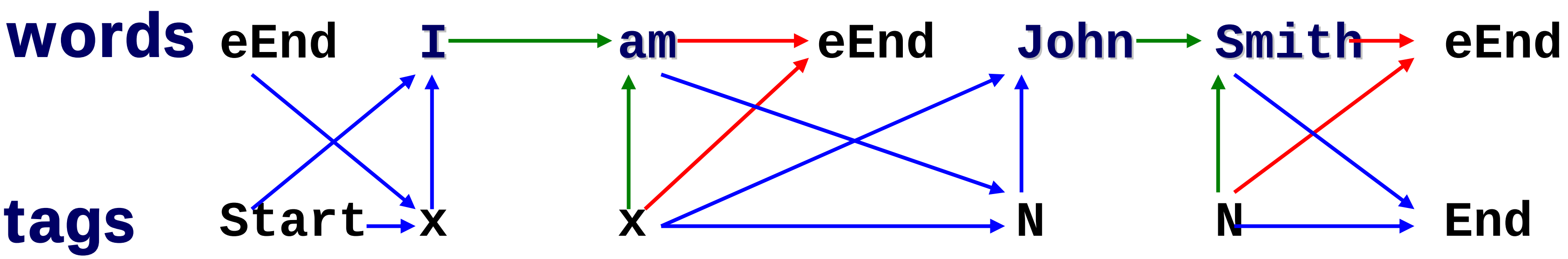
Sample Markup of “Named Entities”

my dad was a traveling salesperson man and was a good provider we I cannot complain as a child we had a pretty good life and it started in nineteen thirty three Hitler came to power and started first with the communist started trouble then started with the Jews and I felt already in school when I went to school they put me in the last row of the class because I was Jewish how how old were you when you first noticed that you were treated differently I was seven seven years old this was my first second grade going to to school it started I looked I looked fairly dark I don't look like a real German blue eyes and blond I was beaten up in in school by the youngsters and I was afraid to go to school so my father decided my mother was born in Oswiecim this became Auschwitz later on the famous infamous place to go to Oswiecim to visit her grandmother per- a lot of family live in Oswiecim so our family went to Oswiecimwe stayed there about a year and we picked up a little bit of the Polish language I started school kind of in the village and it was pretty nice we had a lot of family there cousins and and uncles and we stayed there till nineteen thirty four and my dad decided that it calmed down in Berlin we should come back we did not believe that really it will grow to something big this Hitler so we came back to Berlin and my parents put me in a a Jewish boys school was called Kaiserstrasser and we lived pretty much in the center of...

HMM-based Named Entity Detector

Maximizing probability of a sequence of **tags** given a sequence of **words**:

$P(T \mid W) = \mathbf{P(W, T)} / P(W)$



Language models to estimate probabilities of words and tags given their histories:

first word of a named entity $p(t_i \mid t_{i-1}, w_{i-1}) * p(w_i \mid t_i, t_{i-1})$

continuation $p(w_i \mid t_i, w_{i-1})$

end $p(e \mid t_i, w_{i-1})$

Named Entity Detection Results

- Data Resources
  - MALACH Corpus
    - 461K words of training data (19K entities)
    - 55K words of test data (2.5K entities)
  - Question and Answering Corpus
    - 1M words of training data from newspaper sources

NE F-measure Performance on 31 named categories
(with 3 different labeled training data sets)

	Malach (461Kw)	QA (1MW)	Both (1.5MW)
30 speakers, 15 min. each	80.9	71.8	80.5
single 2.5 hr testimony	82.1	70.6	82.1

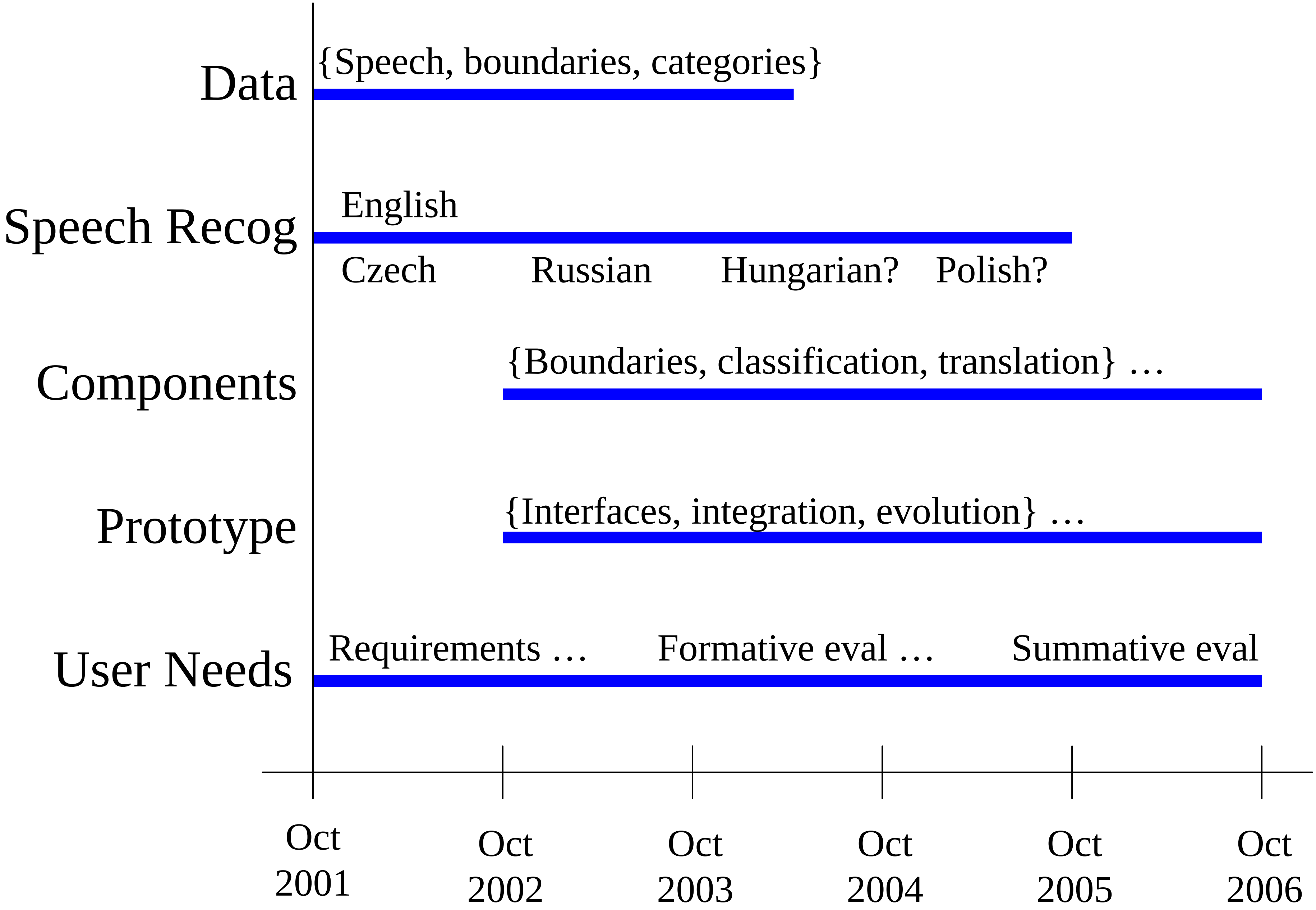
Goals

- Rich transcriptions (including lattices) of possibly the entire collection at less than 30% WER

- Information Extraction:
  - extraction and tracking of entities, events and relations from speech recognition output

- Research automatic extraction of time sequence of events

Project Timeline



Data

{Speech, boundaries, categories}

Speech Recog

English

Czech Russian Hungarian? Polish?

Components

{Boundaries, classification, translation} …

Prototype

{Interfaces, integration, evolution} …

User Needs

Requirements … Formative eval … Summative eval

Oct 2001

Oct 2002

Oct 2003

Oct 2004

Oct 2005

Oct 2006

Impact

- Being able to recognize VHF data will generate technology to enable us to handle a wide variety of tasks from different sources, accents and noisy environments.
- MALACH will also result in new approaches for use by catalogers and researchers that will substantially reduce the cost of obtaining transcripts and metadata and will significantly improve multilingual search of large audiovisual collections (digital libraries)
- With the mechanisms that MALACH will provide, scholars will be able to scan large bodies of audiovisual data and cross-index them with other audio and visual archives.
- Outreach: MALACH will lead to new international speech and language research efforts if the collection can be made public

Publications

- Journals
  - IEEE TSAP (July 2004)
- Conferences
  - ICASSP, Eurospeech, ASRU, SIGIR,TSD, JCDL
- Workshops
  - ASRU, AAAI, ISCA

http://www.clsp.jhu.edu/research/malach/malach_pubs.html