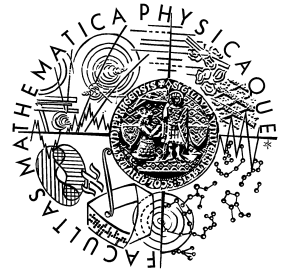
Cross–Language Speech Retrieval and its Evaluation

in the Malach Project

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles Univesity, Prague



Seminář formální lingvistiky
ÚFAL, 20. 11. 2006

Cross–Language Speech Retrieval and its Evaluation

Cross–Language Speech Retrieval and its Evaluation

Information Retrieval

- searching for information in documents or for documents themselves
- searching a body of information for objects that match a search query
- the science and practice of identification and efficient use of recorded data

Cross–Language Speech Retrieval and its Evaluation

Information Retrieval

- searching for information in documents or for documents themselves
- searching a body of information for objects that match a search query
- the science and practice of identification and efficient use of recorded data

Speech Retrieval

- a special case of IR in which the information is in spoken form

Cross–Language Speech Retrieval and its Evaluation

Information Retrieval

- searching for information in documents or for documents themselves
- searching a body of information for objects that match a search query
- the science and practice of identification and efficient use of recorded data

Speech Retrieval

- a special case of IR in which the information is in spoken form

Cross–Language

- retrieving information in a language different from the language of the user's query

Cross–Language Speech Retrieval and its Evaluation

Information Retrieval

- searching for information in documents or for documents themselves
- searching a body of information for objects that match a search query
- the science and practice of identification and efficient use of recorded data

Speech Retrieval

- a special case of IR in which the information is in spoken form

Cross–Language

- retrieving information in a language different from the language of the user's query

Evaluation

- deals with effectiveness of IR systems: how well they perform
- measures how well users are able to acquire information
- usually comparative: ranks a better system ahead of a worse system

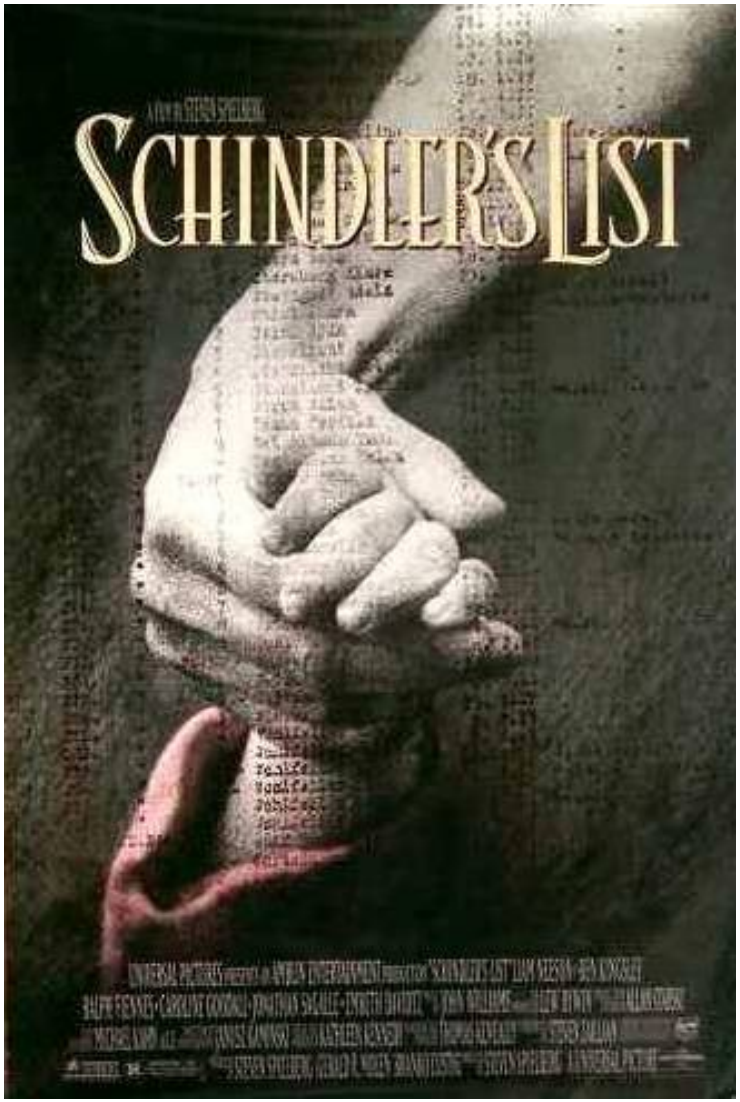
Roadmap

1. Introduction
2. Project Overview
   - *History*
   - *Tasks and Goals*
3. Speech Recognition
   - *Chalanges and Results*
   - *Demo*
4. Speech Retrieval
   - Overview
5. Evaluation
   - *Test Collection*
   - *Evaluation Measures*
   - *CLEF 2006*
6. Conclusions

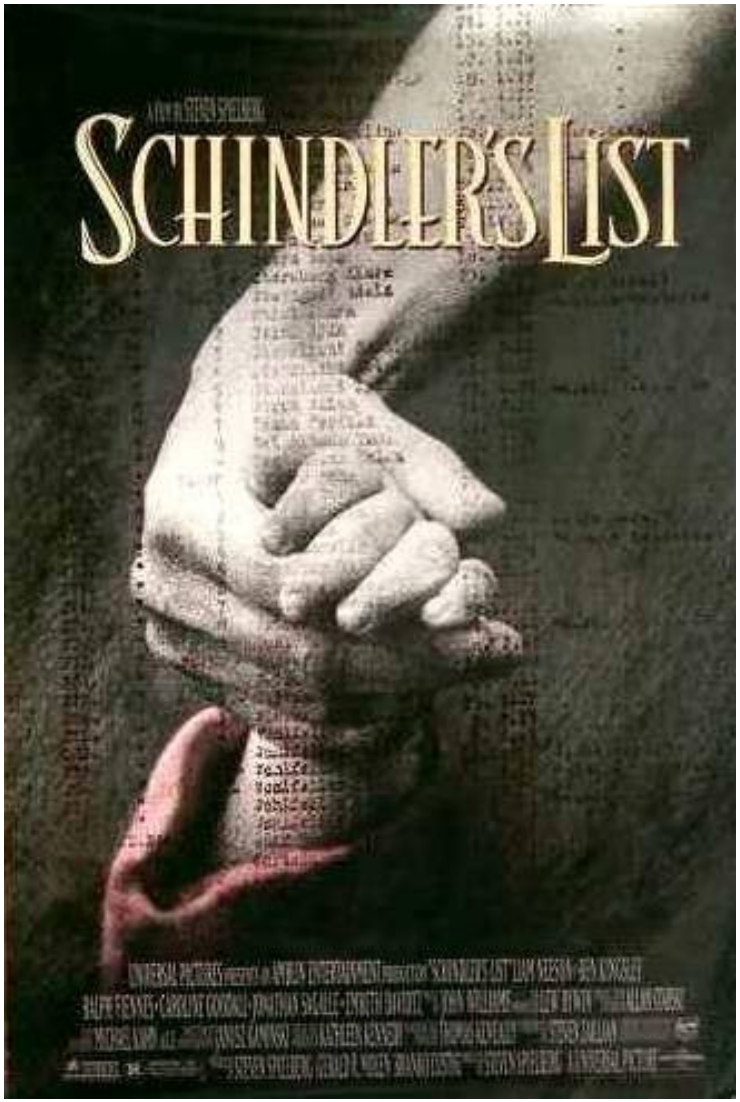
Project Overview

The story begins in 1993

The story begins in 1993 with a movie



The story begins in 1993 with a movie and a vision



Steven Spielberg's vision of:

1. collecting and preserving survivor and witness testimony of the Holocaust
2. cataloging those testimonies to make them available
3. disseminating the testimonies for educational purposes to fight intolerance
4. enabling others to collect testimonies of other atrocities and historical events or perhaps do so itself

A brief history of the project

1993 Stephen Spielberg releases Schindler's List.
He is approached by survivors who want him to listen their stories of Holocaust.

1994 Spielberg starts Survivors of the Shoah Visual History Foundation
to videotape and preserve testimonies of Holocaust survivors and witnesses.

1999 VHF assembled the world's largest archive of videotaped oral histories
with interviews from 52,000 survivors, liberators, and rescuers from 57 countries.

2000 10 % interviews manually catalogized by VHF at a cost of $8 million.
One single testimony consumes an average of 35 hrs (index, summary, review).

2001 NSF project proposal with the goal to dramatically improve access to large multilingual spoken word collections. *Principal investigators: UMD, JHU, IBM*

2001 Grant awarded; project **M**ultilingual **A**ccess to **L**arge Spoken **Arch**ives launches
$7.5 million budget distributed over five years, CU and UWB part of the team.

The archive

- maintained by Visual History Foundation/Institute
- assembled in 1994–1999 by 2,300 interviewers and 1,000 photographers
- contains testimonies of 52,000 survivors from 57 countries in 32 languages
- total of 116,000 hours of VHS tapes, 180 TB of MPEG-1 digitalized video
- average duration of a testimony 2:15 hours, total cost per interview $2,000
- extensive human cataloging completed for over 3,000 interviews (72 mil words)
- manually indexed (time–aligned descriptors from a 30,000-keyword thesaurus)

The archive

- maintained by Visual History Foundation/Institute

SURVIVORS OF THE SHOAH VISUAL HISTORY FOUNDATION

- assembled in 1994–1999 by 2,300 interviewers and 1,000 photographers
- contains testimonies of 52,000 survivors from 57 countries in 32 languages
- total of 116,000 hours of VHS tapes, 180 TB of MPEG-1 digitalized video
- average duration of a testimony 2:15 hours, total cost per interview $2,000
- extensive human cataloging completed for over 3,000 interviews (72 mil words)
- manually indexed (time–aligned descriptors from a 30,000-keyword thesaurus)

- 573 interviews recorded in the Czech Republic by 38 interviewers
- 4 500 testimonies provided by people born in the Czech Republic

Full–description cataloging and anotations

Interview–level annotation

- pre–interview questionaire
- names of people and places mentioned in the course of
- free text summary an interview

Segment–level annotation

- topic boundaries (average 3 min/segment)
- descriptions: *summary, cataloguer's scratchpad*
- thesaurus labels: *names, topic, locations, time periods*

	Location–Time	**Concept**	**People**
	Berlin 1939	*Employment*	*Josef Stein*
	Berlin 1939	*Family life*	*Gretchen Stein* *Anna Stein*
	Dresden 1939	*Relocation* *Transportation–rail*	
	Dresden 1939	*Schooling*	*Gunter Wendt* *Maria*

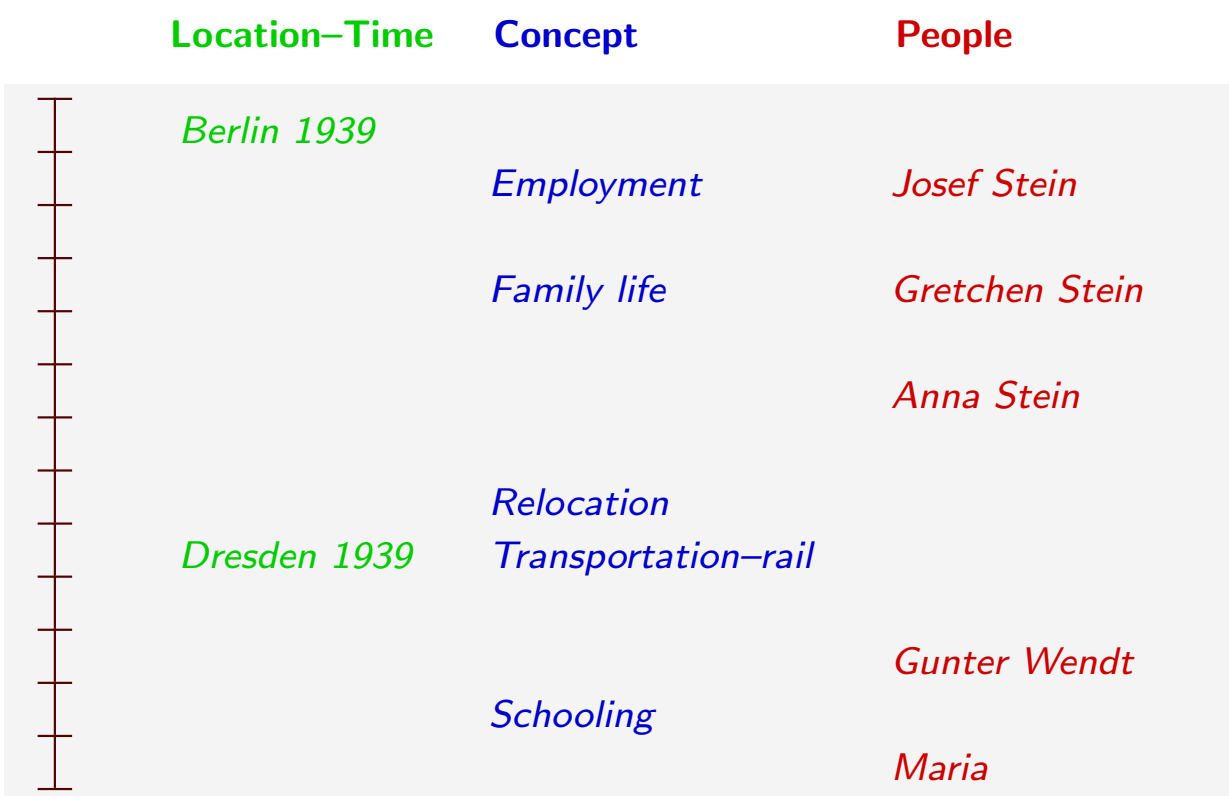
"Real–time" cataloging and anotations

Interview–level annotation

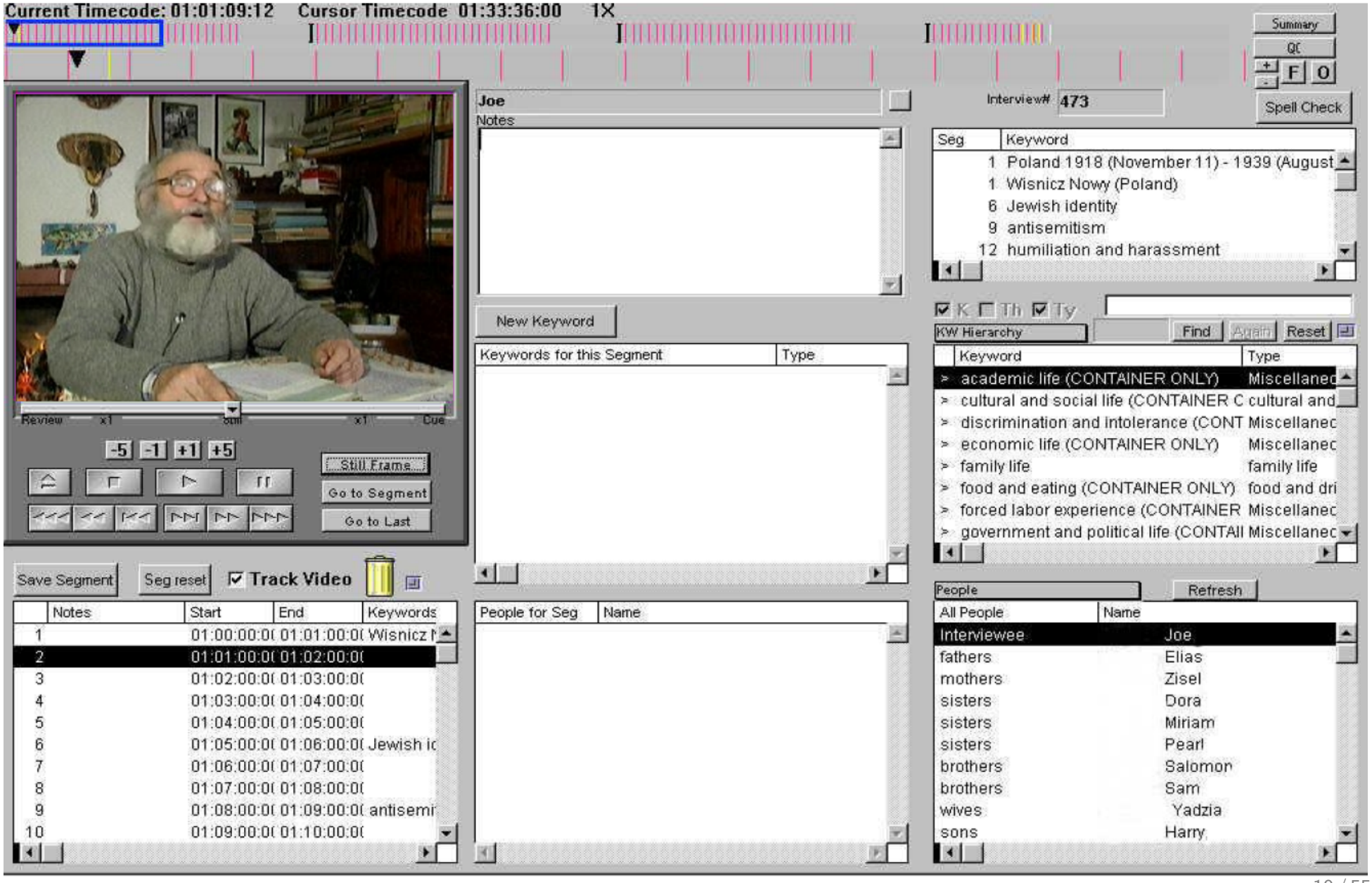
- pre–interview questionaire

Time–aligned annotation

- thesaurus labels: *names*, *concepts*, *locations*, *time periods*

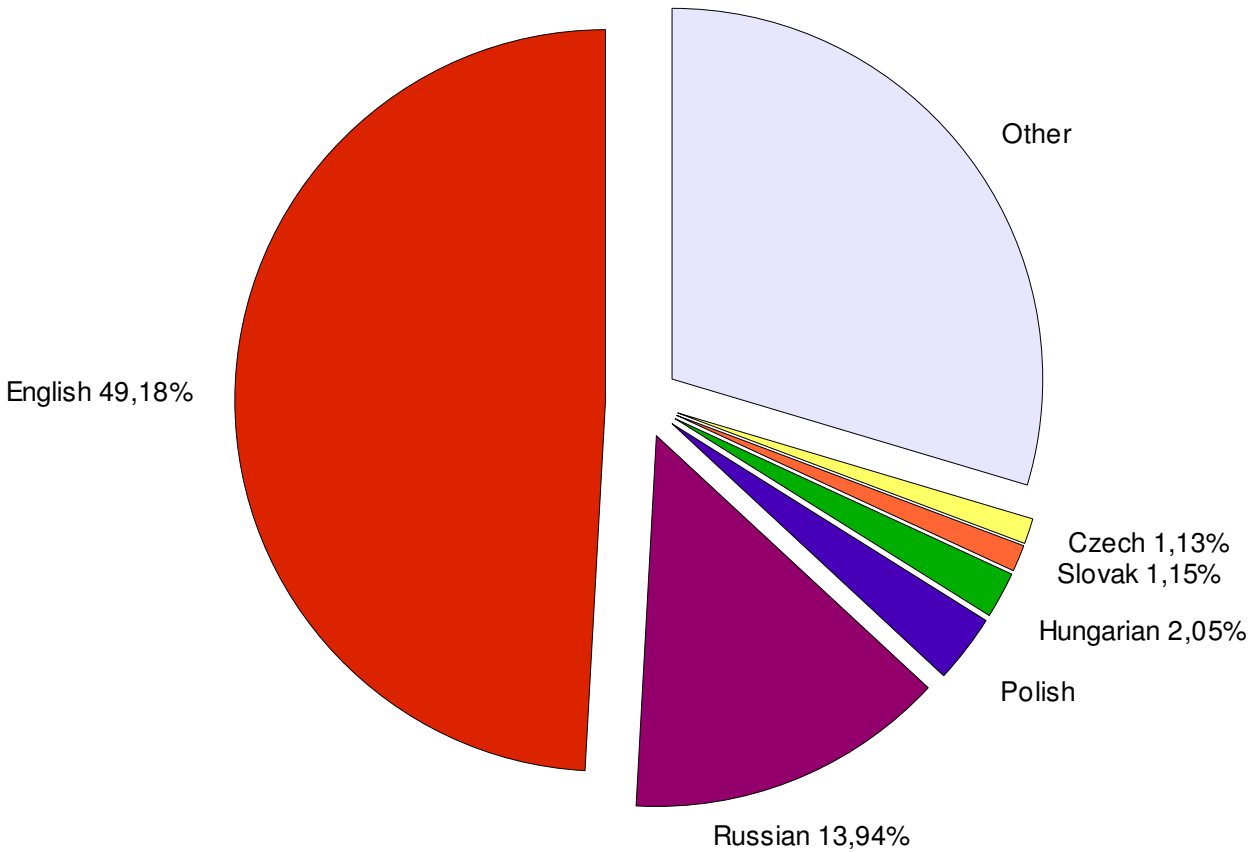


Cataloging interface



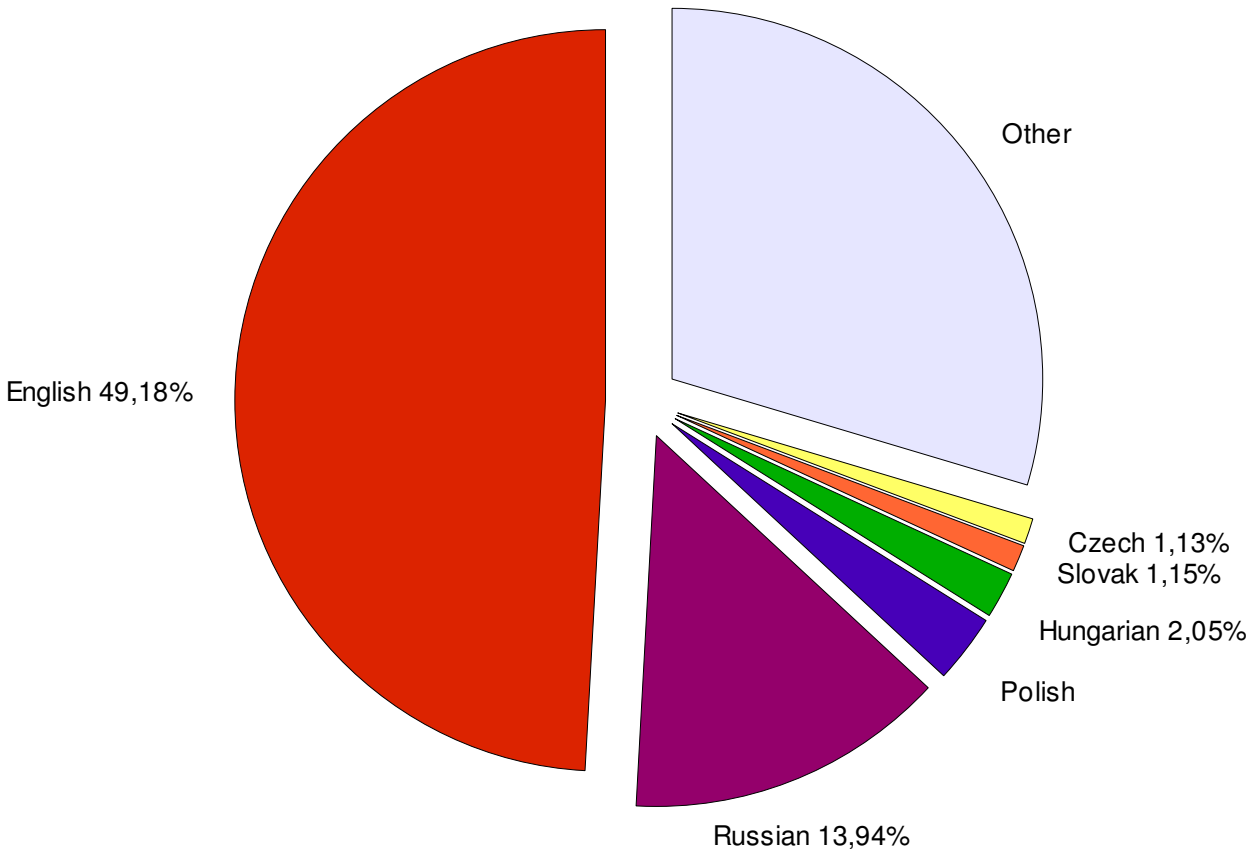
Interview languages (top 20)

interview counts	
English	24,872
Russian	7,052
Hebrew	6,126
French	1,875
Polish	1,549
Spanish	1,352
Dutch	1,077
Hungarian	1,038
German	686
Bulgarian	645
Slovak	583
Czech	573
Portuguese	562
Yiddish	527
Italian	433
Serbian	382
Croatian	353
Ukrainian	320
Greek	301
Swedish	266



Interview languages (top 20)

interview counts	
English	24,872
Russian	7,052
Hebrew	6,126
French	1,875
Polish	1,549
Spanish	1,352
Dutch	1,077
Hungarian	1,038
German	686
Bulgarian	645
Slovak	583
Czech	573
Portuguese	562
Yiddish	527
Italian	433
Serbian	382
Croatian	353
Ukrainian	320
Greek	301
Swedish	266



Project tasks and participants

1. automatic recognition of spontaneous speech (multi–lingual)
2. machine supported translation of domain specific thesaurus
3. automatic topic boundary tagging and time–aligned metadata asignment
4. environment for cross–language information retrieval and browsing

IBM T.J. Watson Center, New York
- *speech recognition in English*

Center for Speech and Language Processing, JHU, Baltimore
- *speech recognition in other languages*
- *Czech and other Slavic languages subcontracted to CU and UWB*

University of Maryland, College Park
- *archive browsing, information retrieval and its evaluation*
- *Czech test collection subcontracted to Charles University*

Speech Recognition

Project challenges

- complete speaker independent recognition of spontaneous speech
- relatively high technical quality of recordings
- low "language quality": difficult even for a human listener:
  - spontaneous, emotional, disfluent, and whispered speech from elders
  - speech with background noise and frequent interruptions
  - heavily accented speech that switches between languages
  - speech with words such as names, obscure locations, unknown events

Project challenges

- complete speaker independent recognition of spontaneous speech
- relatively high technical quality of recordings
- low "language quality": difficult even for a human listener:
  - spontaneous, emotional, disfluent, and whispered speech from elders
  - speech with background noise and frequent interruptions
  - heavily accented speech that switches between languages
  - speech with words such as names, obscure locations, unknown events
- specific issue in Czech: colloquial expressions and pronunciation

oběd	[o b j e d]	Osvětim	[o s v j e t i m]
	[o b j e t]		[v o s v j e t i m]
	[v o b j e d]		[o s v j e n č i m]
	[v o b j e t]		[o z v j e t i m]

Speech recognition results

Word error rate estimated on a sample of manually trascribed data as a ratio of misrecognized words.

language	WER (%)
English	25.0
Czech	35.0
Russian	45.7
Slovak	34.5

Speech recognition results

Word error rate estimated on a sample of manually trascribed data as a ratio of misrecognized words.

language	WER (%)
English	25.0
Czech	35.0
Russian	45.7
Slovak	34.5

Manual transcriptions

language	TrData (h)
English	200
Czech	84
Russian	100
Slovak	

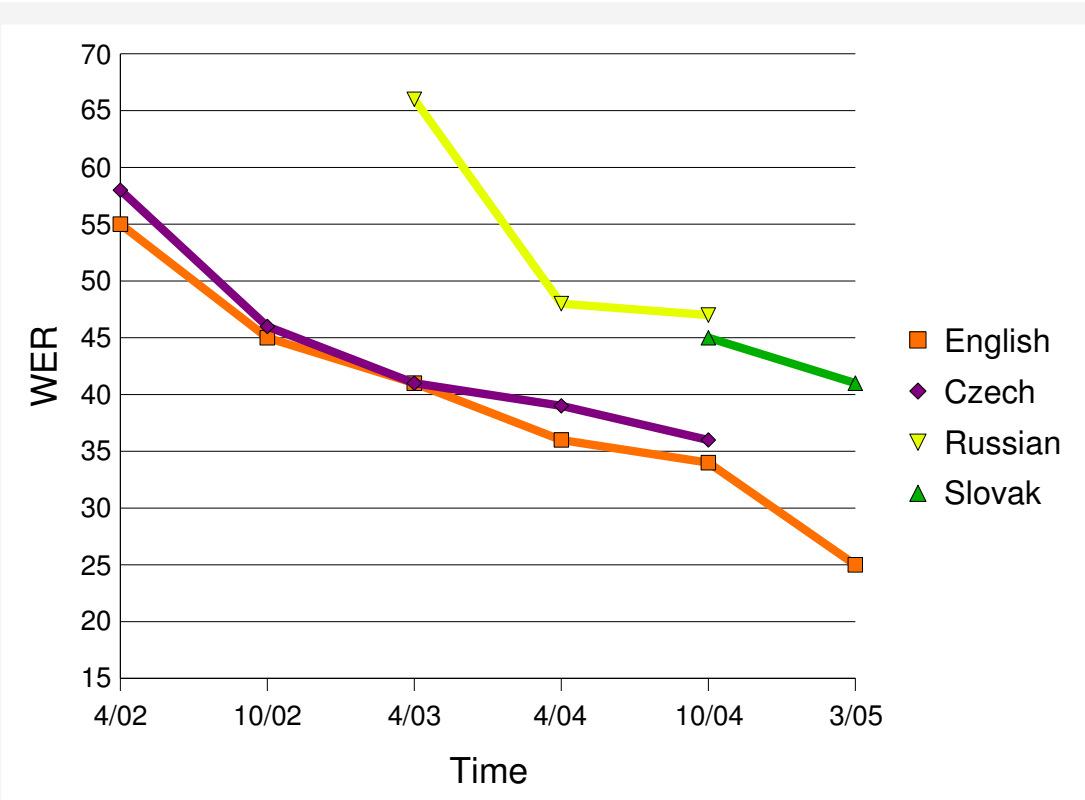
Speech recognition results

Word error rate estimated on a sample of manually trascribed data as a ratio of misrecognized words.

language	WER (%)
English	25.0
Czech	35.0
Russian	45.7
Slovak	34.5

Manual transcriptions

language	TrData (h)
English	200
Czech	84
Russian	100
Slovak	



Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	* * *
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	* * *
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name: * * *
day of birth: Dec 26, 1924
country: Czechoslovakia
religion: judaism
keywords: hiding/death marches
underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	* * *
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	* * *
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

Pane Pavle, začal jste historku o srncích a tatínkovi bez hvězdy. Jak to pokračovalo? Bylo, pokračovalo to tím způsobem, že tatínek si sundal hvězdu, pan doktor Jeřáb mu napsali skupinku na Kladně. To bylo báječný doktor, ten a fandila. Náš tatínek se vydal na cestu na Křivoklátsko, aby upekla že sem se Pochopitelně, že strejda Prošek s tím nechtěl nic mít. Za to byly krutý tresty, za to se tenkrát popravovalo. Takže strejda Prošek nepytlačil a bál se. Tady všude v lesích byli Němci. Střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na. To byl pes – vlčák, s kterém dříve Prošek nepytlačil, a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	* * *
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

pane pavle začal jste historku o srncích a tatínkovi bez hvězdy jak to pokračovalo bylo pokračovalo to tím způsobem že tatínek si sundal hvězdu pan doktor jeřáb mu napsali skupinku na kladně to bylo báječný doktor ten a fandila náš tatínek se vydal na cestu na křivoklátsko aby upekla že sem se ... pochopitelně že strejda prošek s tím nechtěl nic mít za to byly krutý tresty za to se tenkrát popravovalo takže strejda prošek nepytlačil a bál se tady všude v lesích byli němci střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na to byl pes vlčák s kterém dříve prošek nepytlačil a ten prostě každého sem se nepytlačil ...

Speech recognition demo

name:	Hugo Pavel
day of birth:	Dec 26, 1924
country:	Czechoslovakia
religion:	judaism
keywords:	hiding/death marches underground/resistance

pane pavle začal jste historku o srncích a tatínkovi bez hvězdy jak to pokračovalo bylo pokračovalo to tím způsobem že tatínek si sundal hvězdu pan doktor jeřáb mu napsali skupinku na kladně to bylo báječný doktor ten a fandila náš tatínek se vydal na cestu na křivoklátsko aby upekla že sem se ... pochopitelně že strejda prošek s tím nechtěl nic mít za to byly krutý tresty za to se tenkrát popravovalo takže strejda prošek nepytlačil a bál se tady všude v lesích byli němci střílelo se ... a náš táta se vydal na tuhle cestu a ubytoval se mnou slůvko toho v roce sem opravdu podařilo u pytlačit – za pomoci legendární a volal na to byl pes vlčák s kterém dříve prošek nepytlačil a ten prostě každého sem se nepytlačil ...

Information Retrieval

Bag–of–words representation

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

Bag–of–words representation

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

Bag–of–words representation

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

word	d_1
a	1
Afghanistan	0
ahead	1
arrival	1
Blair	0
British	0
Bush	1
flew	0
huge	1
Indonesia	1
Minister	0
Monday	0
of	1
on	0
operation	1
President	1
Prime	0
security	1
started	1
the	1
to	0
Tony	0

Bag–of–words representation

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

word	d_1	d_2
a	1	0
Afghanistan	0	1
ahead	1	0
arrival	1	0
Blair	0	1
British	0	1
Bush	1	0
flew	0	1
huge	1	0
Indonesia	1	0
Minister	0	1
Monday	0	1
of	1	0
on	0	1
operation	1	0
President	1	0
Prime	0	1
security	1	0
started	1	0
the	1	0
to	0	1
Tony	0	1

Bag–of–words representation / Vector space model

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

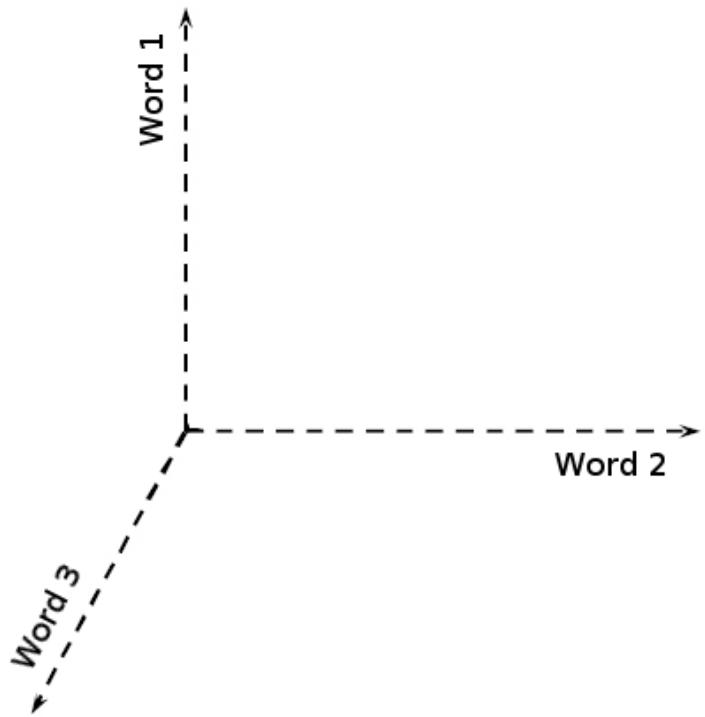
document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

word	d_1	d_2
a	1	0
Afghanistan	0	1
ahead	1	0
arrival	1	0
Blair	0	1
British	0	1
Bush	1	0
flew	0	1
huge	1	0
Indonesia	1	0
Minister	0	1
Monday	0	1
of	1	0
on	0	1
operation	1	0
President	1	0
Prime	0	1
security	1	0
started	1	0
the	1	0
to	0	1
Tony	0	1



Bag–of–words representation / Vector space model

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

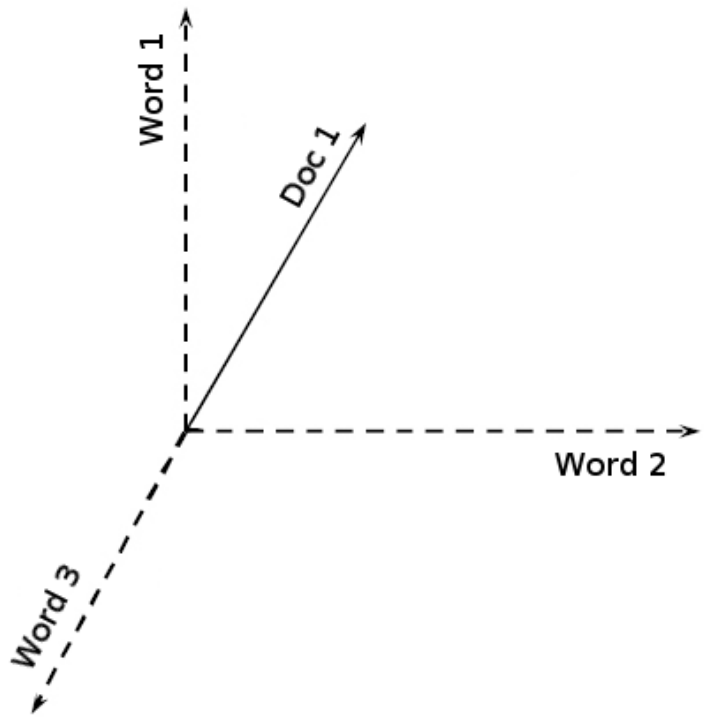
document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

word	d_1	d_2
a	1	0
Afghanistan	0	1
ahead	1	0
arrival	1	0
Blair	0	1
British	0	1
Bush	1	0
flew	0	1
huge	1	0
Indonesia	1	0
Minister	0	1
Monday	0	1
of	1	0
on	0	1
operation	1	0
President	1	0
Prime	0	1
security	1	0
started	1	0
the	1	0
to	0	1
Tony	0	1



Bag–of–words representation / Vector space model

- Simple strategy for representing documents
- Count how many times each word occurs (regardless of word order)
- Distribution over fixed vocabulary

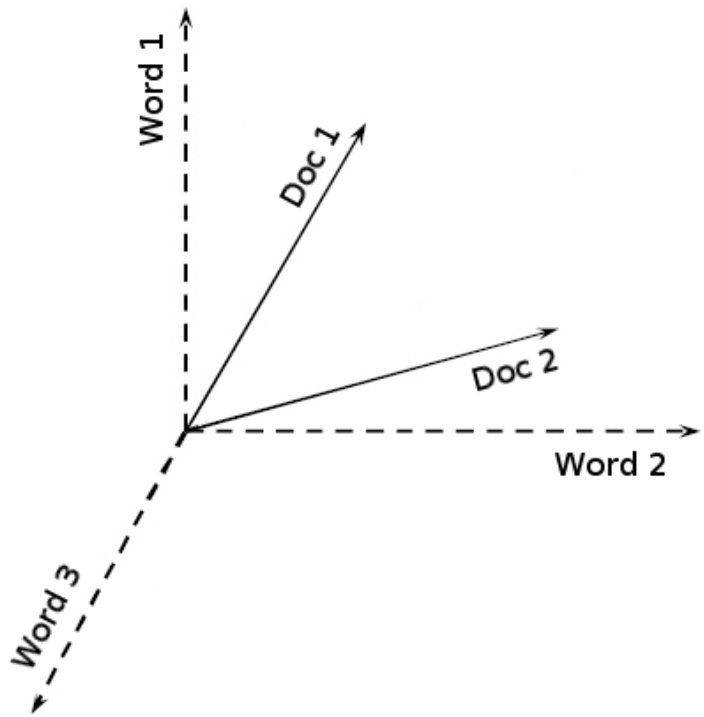
document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

word	d_1	d_2
a	1	0
Afghanistan	0	1
ahead	1	0
arrival	1	0
Blair	0	1
British	0	1
Bush	1	0
flew	0	1
huge	1	0
Indonesia	1	0
Minister	0	1
Monday	0	1
of	1	0
on	0	1
operation	1	0
President	1	0
Prime	0	1
security	1	0
started	1	0
the	1	0
to	0	1
Tony	0	1



Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

query

Prime Minister

Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

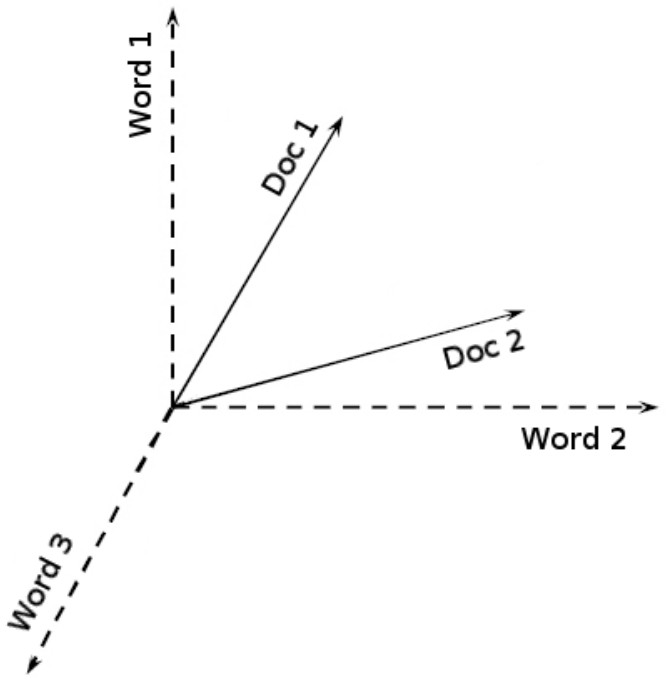
document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

query

Prime Minister

word	d_1	d_2
a	1	0
Afghanistan	0	1
ahead	1	0
arrival	1	0
Blair	0	1
British	0	1
Bush	1	0
flew	0	1
huge	1	0
Indonesia	1	0
Minister	0	1
Monday	0	1
of	1	0
on	0	1
operation	1	0
President	1	0
Prime	0	1
security	1	0
started	1	0
the	1	0
to	0	1
Tony	0	1



Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

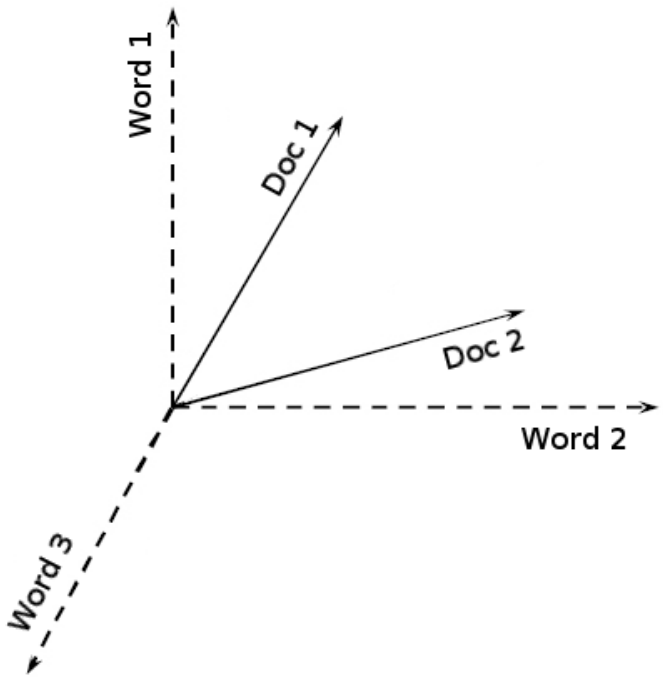
document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

query

Prime Minister

word	d_1	d_2	q
a	1	0	0
Afghanistan	0	1	0
ahead	1	0	0
arrival	1	0	0
Blair	0	1	0
British	0	1	0
Bush	1	0	0
flew	0	1	0
huge	1	0	0
Indonesia	1	0	0
Minister	0	1	**1**
Monday	0	1	0
of	1	0	0
on	0	1	0
operation	1	0	0
President	1	0	0
Prime	0	1	**1**
security	1	0	0
started	1	0	0
the	1	0	0
to	0	1	0
Tony	0	1	0



Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

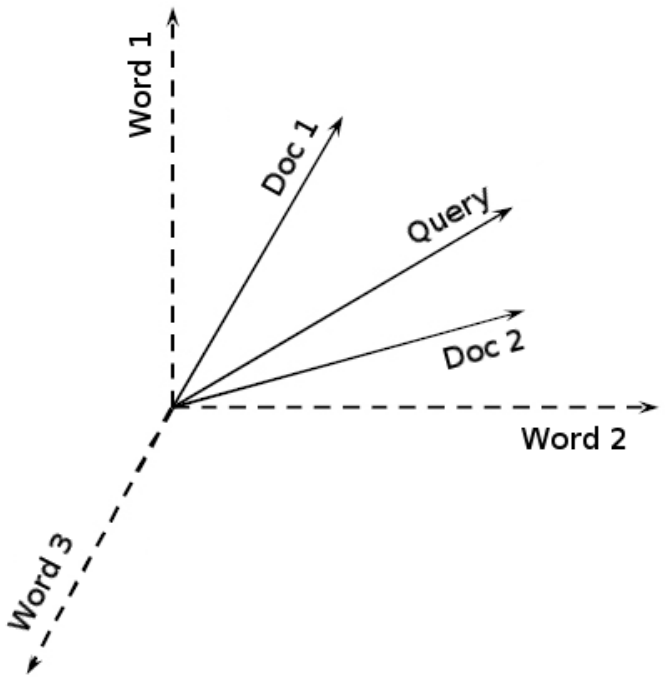
document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

query

Prime Minister

word	d_1	d_2	q
a	1	0	0
Afghanistan	0	1	0
ahead	1	0	0
arrival	1	0	0
Blair	0	1	0
British	0	1	0
Bush	1	0	0
flew	0	1	0
huge	1	0	0
Indonesia	1	0	0
Minister	0	1	**1**
Monday	0	1	0
of	1	0	0
on	0	1	0
operation	1	0	0
President	1	0	0
Prime	0	1	**1**
security	1	0	0
started	1	0	0
the	1	0	0
to	0	1	0
Tony	0	1	0



Similarity–based ranking

1. Treat the query as if it were a document: *create a query bag–of–terms*
2. Find the similarity of each document: *compute inner product*
3. Rank order the documents by similarity: *most similar to the query first*

document d_1

Indonesia started a huge security operation ahead of the arrival of President Bush

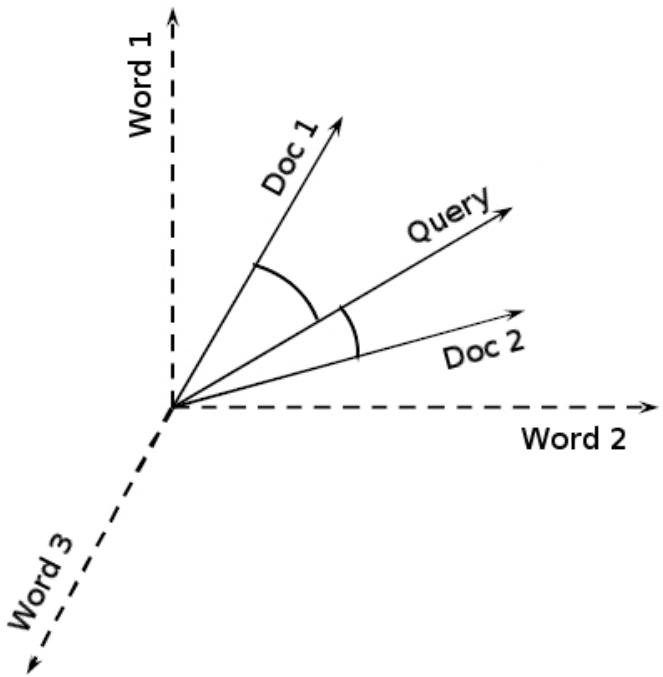
document d_2

British Prime Minister Tony Blair flew to Afghanistan on Monday

query

Prime Minister

word	d_1	d_2	q
a	1	0	0
Afghanistan	0	1	0
ahead	1	0	0
arrival	1	0	0
Blair	0	1	0
British	0	1	0
Bush	1	0	0
flew	0	1	0
huge	1	0	0
Indonesia	1	0	0
Minister	0	1	**1**
Monday	0	1	0
of	1	0	0
on	0	1	0
operation	1	0	0
President	1	0	0
Prime	0	1	**1**
security	1	0	0
started	1	0	0
the	1	0	0
to	0	1	0
Tony	0	1	0



What to use (and not use) as "words"?

Substrings
- overlapping character n–grams

Tokens
- white–space–delimited word forms

Normalized forms
- lemmas, stems

Shingles
- overlapping word n–grams

Multiwords
- collocations, multiword lexemes

Stopwords
- too frequent or too rare words
- closed–class words (*prepositions*, *conjunctions*, *etc.*)

Query expansion

Finds terms that could have been in the query

- synonyms
- other related terms

Blind relevance feedback is widely used

1. search once using the original query
2. find discriminating terms in top–ranked documents
3. add new query terms / reweight existing terms

Several alternative approaches

- thesaurus–based expansion
- collocations
- LSI

Back to Speech Retrieval

Some key insights

Speech Retrieval (recap)

- A special case of IR in which the information is in spoken form

Some key insights

Speech Retrieval (recap)

- A special case of IR in which the information is in spoken form

Recognition and retrieval can be decomposed

- Build IR system on ASR output.

Some key insights

Speech Retrieval (recap)

- A special case of IR in which the information is in spoken form

Recognition and retrieval can be decomposed

- Build IR system on ASR output.

Retrieval is robust with recognition results

- Up to 40% word error rate is tolerable (*TREC result*)

Some key insights

Speech Retrieval (recap)

- A special case of IR in which the information is in spoken form

Recognition and retrieval can be decomposed

- Build IR system on ASR output.

Retrieval is robust with recognition results

- Up to 40% word error rate is tolerable (*TREC result*)

Recognition errors may not bother the system, but they do bother the user

- Retrieval based on ASR output should return playback points.

Some key insights

Speech Retrieval (recap)

- A special case of IR in which the information is in spoken form

Recognition and retrieval can be decomposed

- Build IR system on ASR output.

Retrieval is robust with recognition results

- Up to 40% word error rate is tolerable (*TREC result*)

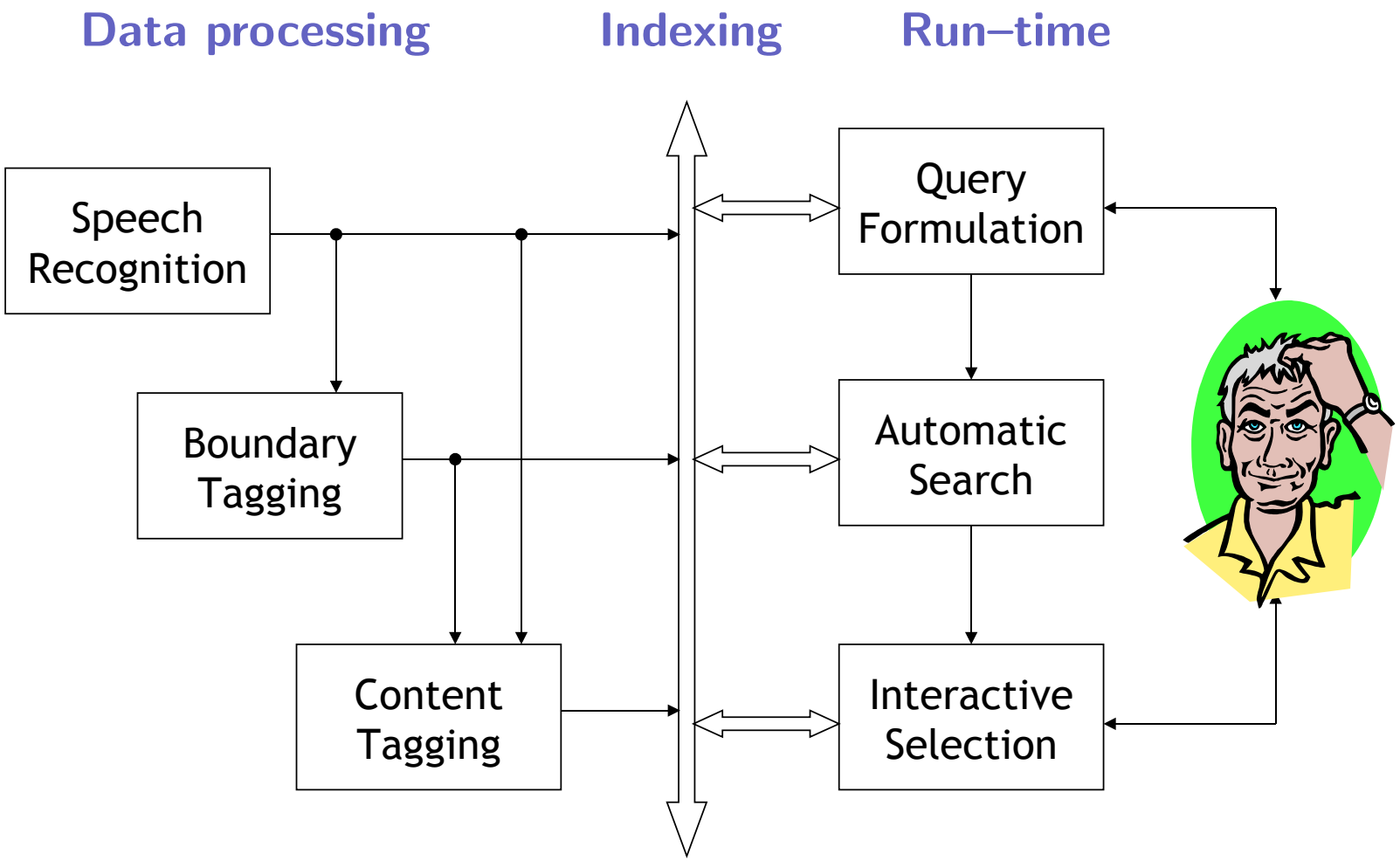
Recognition errors may not bother the system, but they do bother the user

- Retrieval based on ASR output should return playback points.

Segment–level indexing/summary is usefull

- Vocabulary shift/pauses provide strong cues for boundary tagging

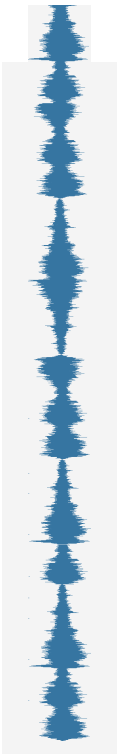
System overview



credit to Doug Oard

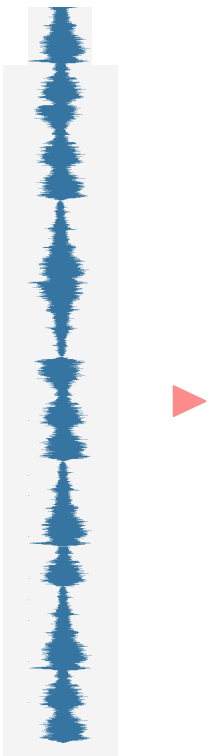
Document processing

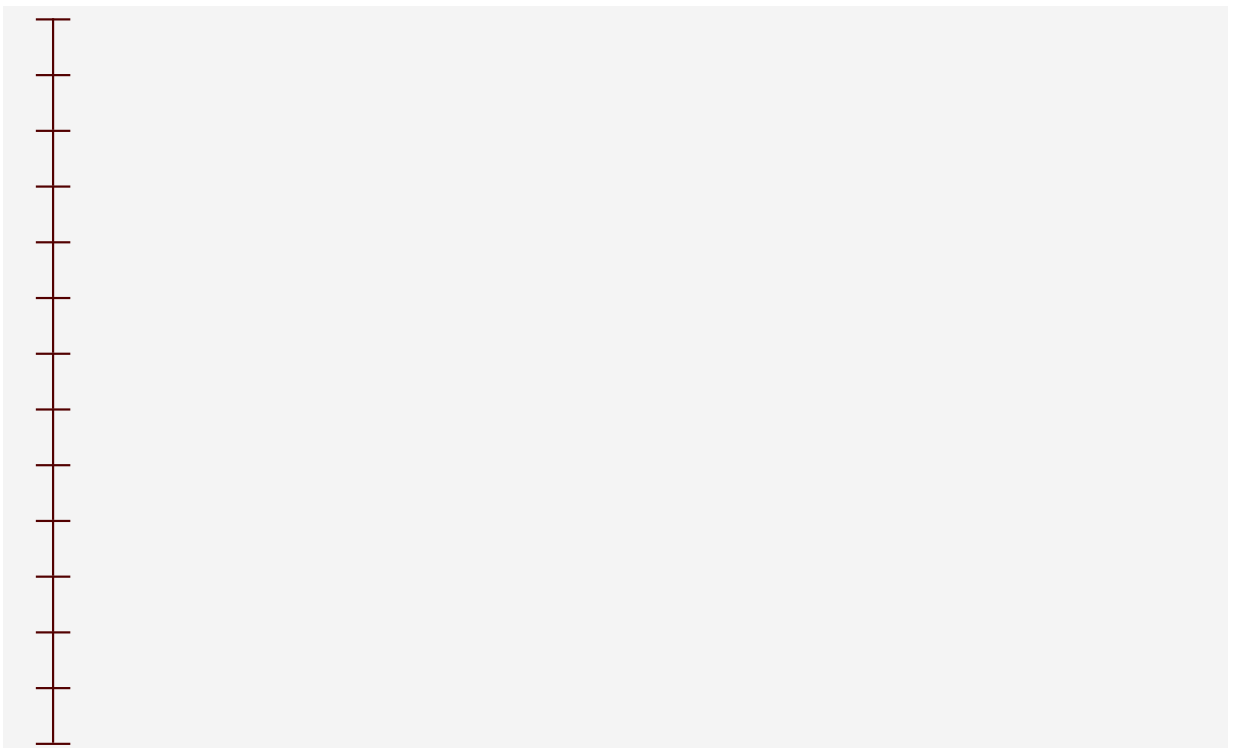
Speech



Document processing

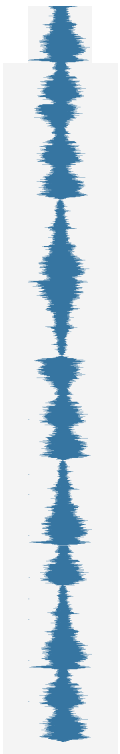
Speech
Transcription





Document processing

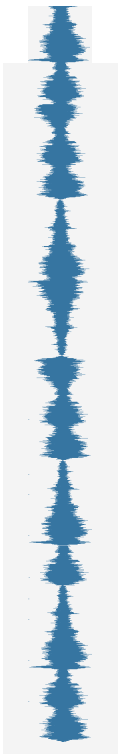
Speech Transcription ▶ Boundary Tagging



▶

Document processing

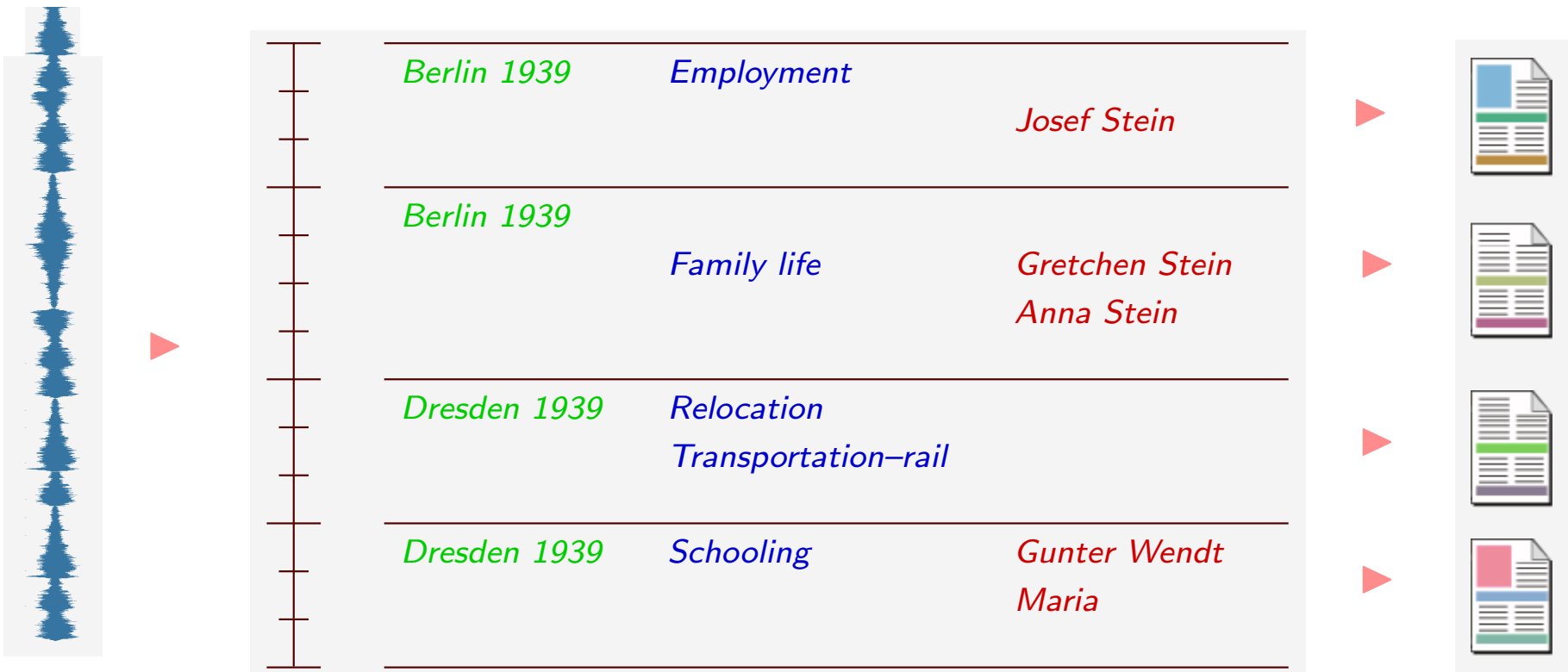
Speech Transcription ▶ Boundary Tagging Content Tagging



Berlin 1939	*Employment*	*Josef Stein*
Berlin 1939	*Family life*	*Gretchen Stein* *Anna Stein*
Dresden 1939	*Relocation* *Transportation–rail*	
Dresden 1939	*Schooling*	*Gunter Wendt* *Maria*

Document processing

Speech Transcription ▶ Boundary Tagging Content Tagging ▶ Document Representation



Spoken document example

Passage from English interview with full annotation

doc no	00009-056150.002
interview data	Sidonia L., 1930
name	Issac L., Cyla L.
manual keyword	family businesses, family life, food, Przemysl (Poland)
summary	SL describes her parents and their roles in the family business. She remembers her home and she recalls her responsibilities. . . .
asr text	*were to tell us about that my mother's name was sell us c y l a new and her maiden name was leap shark l i e b b a c h a r d my mother was a dress . . .*
auto keyword	*family businesses, family homes, means of adaptation and survival, extended family members . . .*

Spoken document example

Passage from Czech interview with brief annotation

doc no	01539-025217.003
interview data	Alois P., 1927
name	-na-
manual keyword	-na-
summary	-na-
asr text	*a když nějaká ta dívenka na na a pláži svolávali Stalin že jo a nu náš fotograf přišel úplně sem byl strašně hrdý že se mě rozuměla a pak sem věděl takže se říká dva zešílení to je že na jedné kde už pak už bylo vše říkalo že venku koho jste si vzal jak jste se seznámili dobře zůstala v Polsku . . .*
auto keyword	-na-

Evaluation

Evaluation

Question:

- How well do we perform?

Evaluation

Question:

- How well do we perform?

Criteria

- Effectiveness, efficiency, usability?

Evaluation

Question:

- How well do we perform?

Criteria

- Effectiveness, efficiency, usability

Evaluation

Question:

- How well do we perform?

Criteria

- Effectiveness, efficiency, usability

User–centered strategy

- Given several users and at least two retrieval systems
- Have each user try the same task on both systems
- Measure which system works the "best"

Evaluation

Question:

- How well do we perform?

Criteria

- Effectiveness, efficiency, usability

User–centered strategy

- Given several users and at least two retrieval systems
- Have each user try the same task on both systems
- Measure which system works the "best"

System–centered strategy

- Given documents, queries, and relevance judgements
- Try several variations on the retrieval systems
- Measure which ranks more good docs near the top

Evaluation

Question:

- How well do we perform?

Criteria

- Effectiveness, efficiency, usability

User–centered strategy

- Given several users and at least two retrieval systems
- Have each user try the same task on both systems
- Measure which system works the "best"

System–centered strategy

- Given documents, queries, and relevance judgements = test collection
- Try several variations on the retrieval systems
- Measure which ranks more good docs near the top

Test collection design

Documents

- Representative sources (*interviewees*)
- Representative topics (*stories*)

Topics

- Detailed and structured description of actual information needs
- Used as a basis for query formulation

Relevance judgements

- Document–topic relations (binary relevance)
- Created by humans, perpetually valid
- Sampled, focus on documents likely be retrieved

Document collections (Malach)

English

- 297 interviews
- known–boundary condition (segments)
- 8,104 topically–coherent segments
- average 503 words/segment
- ASR: 25% mean Word Error Rate

Czech

- 350 interviews
- unknown segment boundaries
- 3-minute automatically generated passages, with 67% overlap
- start times used as DOCNO for easy results
- ASR: 35% mean Word Error Rate

Distributed to track participants by ELDA

Topic construction

- 115 representative topics developed from actual user requests:
- Scholars, educators, documentary film makers, and others produced 250 topic–oriented written requests for materials from the collection.
- from English translated to *Czech, French, German, Spanish*, and *Dutch*
- TREC-like topic descriptions consists of *title*, a *short description* and a *narrative description*:

Topic construction

- 115 representative topics developed from actual user requests:
- Scholars, educators, documentary film makers, and others produced 250 topic–oriented written requests for materials from the collection.
- from English translated to *Czech, French, German, Spanish*, and *Dutch*
- TREC-like topic descriptions consists of *title*, a *short description* and a *narrative description*:

num 1173

title **Dětské umění v Terezíně**

desc Hledáme popis uměleckých aktivit dětí v Terezíně, jako např. hudby, divadla, malování, poezie a jiných psaných děl.

narr *Relevantní materiál by měl obsahovat diskuse o těchto aktivitách a to, jak ovlivnily přečkání holokaustu a následný život dětí. Zejména jsou žádané příběhy, ve kterých účastník rozhovoru uvádí příklady takových aktivit.*

Topic construction

- 115 representative topics developed from actual user requests:
- Scholars, educators, documentary film makers, and others produced 250 topic–oriented written requests for materials from the collection.
- from English translated to *Czech, French, German, Spanish*, and *Dutch*
- TREC-like topic descriptions consists of *title*, a *short description* and a *narrative description*:

num 1431

title **Denní život v Terezíně**

desc Popište denní život v táboře Terezín.

narr *Anekdoty, příběhy nebo detaily o denním životě v táboře. Relevantní jsou příběhy, které se zmiňují o modlitbách, svátcích, volném čase vězňů a strukturách vnitřní vlády v táborech.*
Příběhy, které popisují neobvyklé události v životě vězňů, relevantní nejsou.

Topic construction

- 115 representative topics developed from actual user requests:
- Scholars, educators, documentary film makers, and others produced 250 topic–oriented written requests for materials from the collection.
- from English translated to *Czech, French, German, Spanish*, and *Dutch*
- TREC-like topic descriptions consists of *title*, a *short description* and a *narrative description*:

num 1431

title **Denní život v Terezíně**

desc Popište denní život v táboře Terezín.

narr *Anekdoty, příběhy nebo detaily o denním životě v táboře. Relevantní jsou příběhy, které se zmiňují o modlitbách, svátcích, volném čase vězňů a strukturách vnitřní vlády v táborech.*
Příběhy, které popisují neobvyklé události v životě vězňů, relevantní nejsou.

Relevance assessment

A manual process to acquire relevance judgements for document–topic pairs.
Ideally for all document–topic pairs – infeasible.

Search guided relevance assessment

- for each topic a set of documents to be judged is restricted to those potentially relevant by full–text search
- *topic research* $\rightarrow$ *query formulation* $\rightarrow$ *search* $\rightarrow$ *judging*

Highly ranked (pooled) relevance assessment

- Restriction based on results of actual evaluation runs
- *n*-deep pools from *m*-systems

Relevance assessment

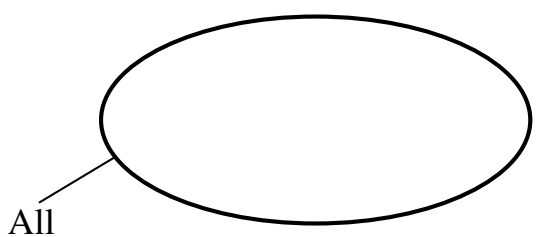
A manual process to acquire relevance judgements for document–topic pairs.
Ideally for all document–topic pairs – infeasible.

Search guided relevance assessment

- for each topic a set of documents to be judged is restricted to those potentially relevant by full–text search
- *topic research* → *query formulation* → *search* → *judging*

Highly ranked (pooled) relevance assessment

- Restriction based on results of actual evaluation runs
- *n*-deep pools from *m*-systems



Relevance assessment

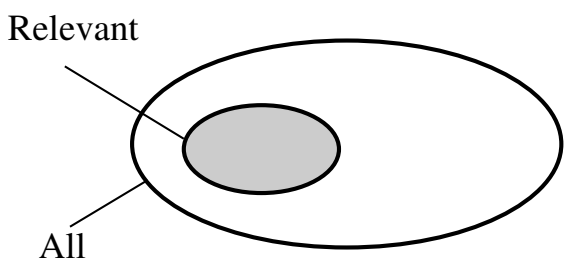
A manual process to acquire relevance judgements for document–topic pairs.
Ideally for all document–topic pairs – infeasible.

Search guided relevance assessment

- for each topic a set of documents to be judged is restricted to those potentially relevant by full–text search
- *topic research* → *query formulation* → *search* → *judging*

Highly ranked (pooled) relevance assessment

- Restriction based on results of actual evaluation runs
- *n*-deep pools from *m*-systems



Relevance assessment

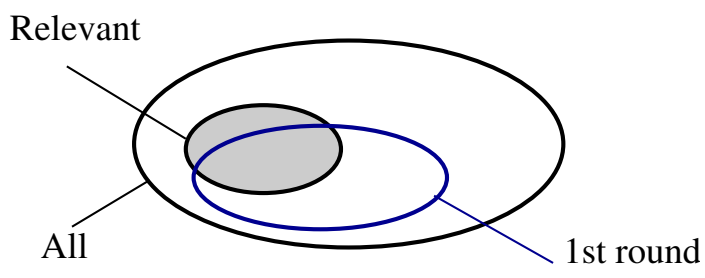
A manual process to acquire relevance judgements for document–topic pairs.
Ideally for all document–topic pairs – infeasible.

Search guided relevance assessment

- for each topic a set of documents to be judged is restricted to those potentially relevant by full–text search
- *topic research* → *query formulation* → *search* → *judging*

Highly ranked (pooled) relevance assessment

- Restriction based on results of actual evaluation runs
- *n*-deep pools from *m*-systems



Relevance assessment

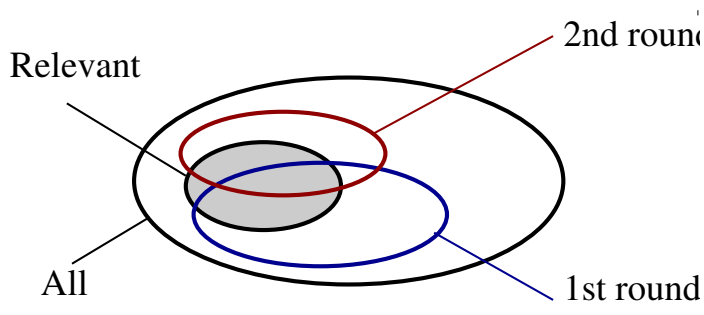
A manual process to acquire relevance judgements for document–topic pairs. Ideally for all document–topic pairs – infeasible.

Search guided relevance assessment

- for each topic a set of documents to be judged is restricted to those potentially relevant by full–text search
- *topic research* → *query formulation* → *search* → *judging*

Highly ranked (pooled) relevance assessment

- Restriction based on results of actual evaluation runs
- *n*-deep pools from *m*-systems



Relevance categories

Direct Direct evidence for what the user asks for
Talks about food given to Auschwitz inmates

Indirect Indirect evidence on the topic; data from which one could infer something about the topic
Talks about seeing emaciated people in Auschwitz

Context Background/context for the topic; sets the stage for what the user asks for; sheds additional light on the topic
Talks about physical labor of Auschwitz inmates

Comparison Information on similar/parallel/contrasting situation
Talks about food in Warsaw ghetto

Relevance categories

Direct Direct evidence for what the user asks for

Talks about food given to Auschwitz inmates

Indirect Indirect evidence on the topic; data from which one could infer something about the topic

Talks about seeing emaciated people in Auschwitz

Context Background/context for the topic; sets the stage for what the user asks for; sheds additional light on the topic

Talks about physical labor of Auschwitz inmates

Comparison Information on similar/parallel/contrasting situation

Talks about food in Warsaw ghetto

Overall Strictly from the point of view of finding out about the topic, how useful is this segment for the requester

Talks about food in Auschwitz

Relevance categories

Direct Direct evidence for what the user asks for
Talks about food given to Auschwitz inmates

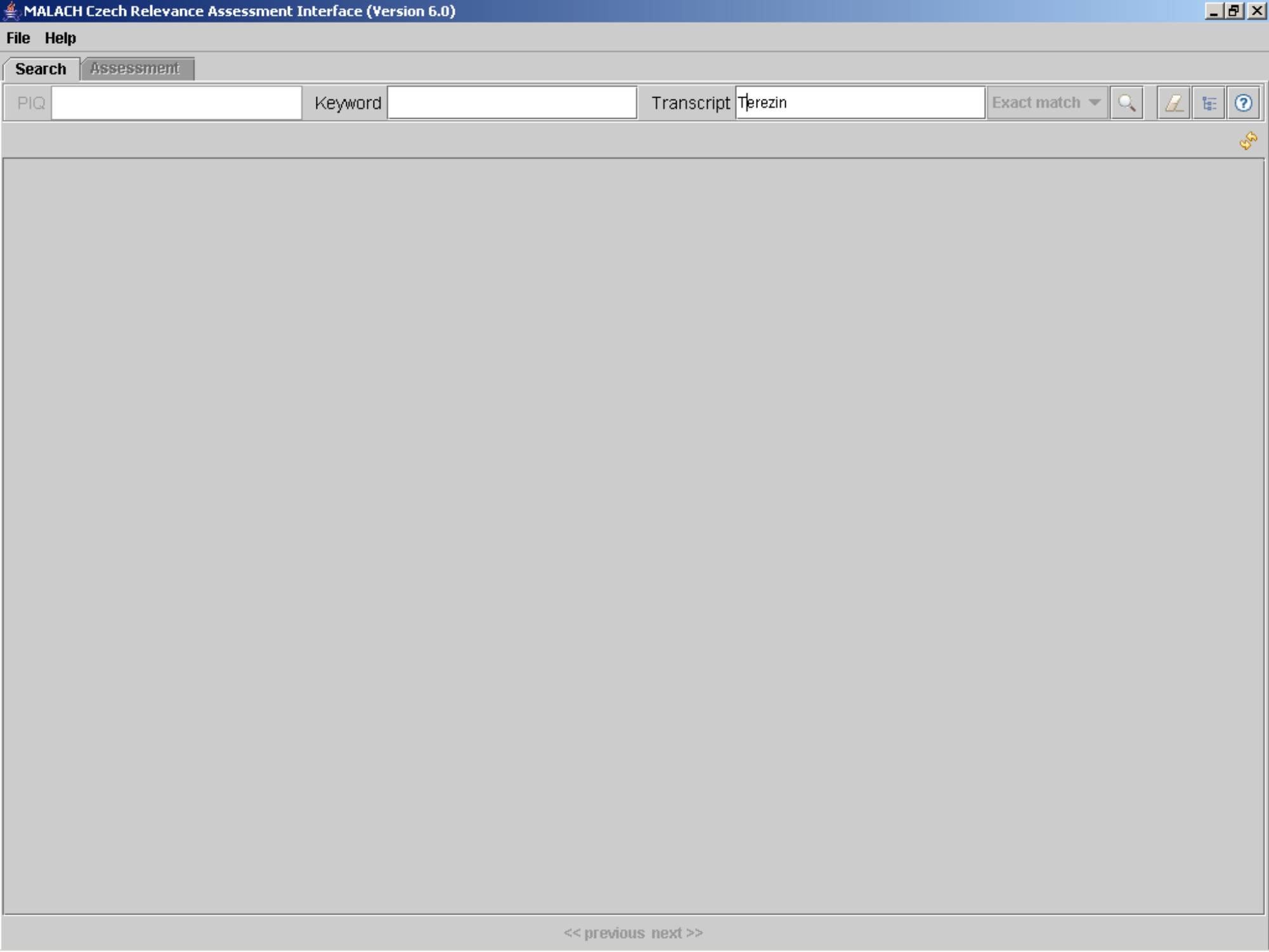
Indirect Indirect evidence on the topic; data from which one could infer something about the topic
Talks about seeing emaciated people in Auschwitz

Context Background/context for the topic; sets the stage for what the user asks for; sheds additional light on the topic
Talks about physical labor of Auschwitz inmates

Comparison Information on similar/parallel/contrasting situation
Talks about food in Warsaw ghetto

Overall Strictly from the point of view of finding out about the topic, how useful is this segment for the requester
Talks about food in Auschwitz

- Five levels of relevance for each (0=none, 4=highly)
- Collapsed to binary relevance using $\{\text{Direct} \geq 2\} \cup \{\text{Indirect} \geq 2\}$


MALACH Czech Relevance Assessment Interface (Version 6.0)
File
Help
Search
Assessment
PIQ
Keyword
Transcript
Terezin
Exact match
<< previous
next >>

MALACH Czech Relevance Assessment Interface (Version 6.0)

File Help

Search | Assessment

PIQ | Keyword | Transcript: Terezin | Exact match

196 interviews found. Displaying results 1-10

Efim Markelis-Domnin (ID: 31115, 1 segment, 0 point(s) already marked for this topic)

Locations: ... ...

Concepts: ... ...

Kunhuta Buresova (ID: 14932, 3 segments, 0 point(s) already marked for this topic)

Locations: ... Theresienstadt (Czechoslovakia : Ghetto), Czechoslovakia 1943, Czechoslovakia 1944 ...

Concepts: ... forced labor in the ghettos, means of adaptation and survival in the ghettos, forced labor: agriculture, stealing in the ghettos ...

Alzbeta Bernatova (ID: 30796, 2 segments, 0 point(s) already marked for this topic)

Locations: ... Czechoslovakia 1939 (Mar 15) - 1945 (May 9), Theresienstadt (Czechoslovakia : Ghetto) ...

Concepts: ... sharing food and drink in the ghettos, interaction with family members in the refugee camps ...

Hugo Pavel (ID: 13135, 6 segments, 0 point(s) already marked for this topic)

Locations: ... Theresienstadt (Czechoslovakia : Ghetto), Czechoslovakia 1943, Germany 1944, Germany 1945 (Jan 1 - May 7), ...

Concepts: ... housing conditions in the ghettos, food in the camps ...

Hana Tvrzská (ID: 23005, 3 segments, 0 point(s) already marked for this topic)

Locations: ... Czechoslovakia 1941, Czechoslovakia 1942, Protivin (Czechoslovakia) ...

Concepts: ... awareness of deportations and/or transfers ...

Felice Kvapilova (ID: 23179, 2 segments, 0 point(s) already marked for this topic) - Hiding

Locations: ... ...

Concepts: ... ...

Eva Liskova (ID: 28592, 5 segments, 0 point(s) already marked for this topic) - Death Marches

Locations: ... ...

Concepts: ... cultural and social activities in the ghettos, clandestine activities in the ghettos, Appell in the ghettos, deportations, means of transport, ...

Margit Herrmannova (ID: 31114, 1 segment, 0 point(s) already marked for this topic)

Locations: ... ...

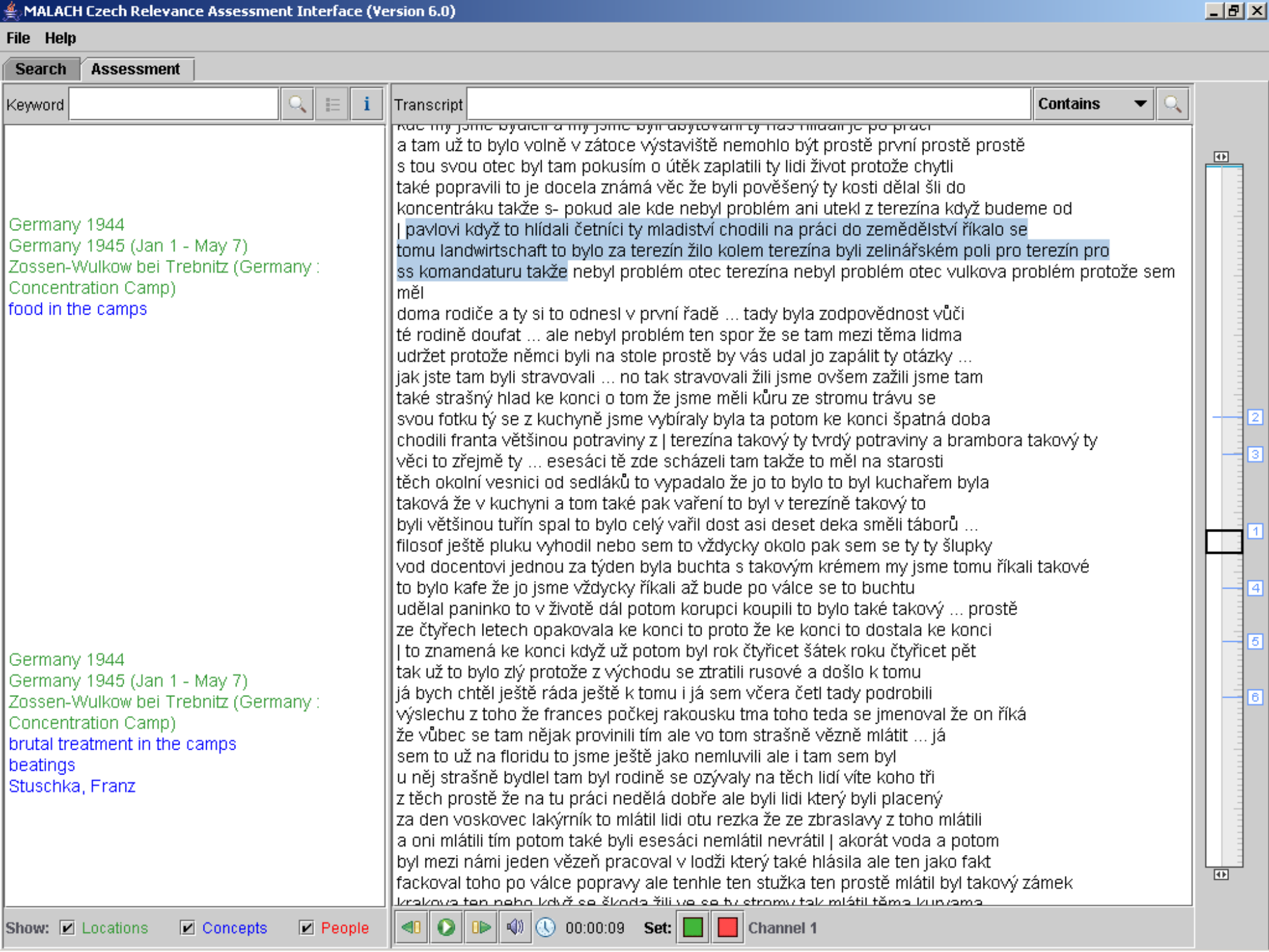
Concepts: ... ...

Anna Brynberg (ID: 13682, 1 segment, 0 point(s) already marked for this topic)

Locations: ... ...

Concepts: ... ...

Jiri Bures (ID: 20065, 3 segments, 0 point(s) already marked for this topic)


MALACH Czech Relevance Assessment Interface (Version 6.0)
File
Help
Search
Assessment
Keyword
Transcript
Contains
Germany 1944
Germany 1945 (Jan 1 - May 7)
Zossen-Wulkow bei Trebnitz (Germany : Concentration Camp)
food in the camps
Germany 1944
Germany 1945 (Jan 1 - May 7)
Zossen-Wulkow bei Trebnitz (Germany : Concentration Camp)
brutal treatment in the camps
beatings
Stuschka, Franz
a tam už to bylo volně v zátoce výstaviště nemohlo být prostě první prostě prostě
s tou svou otec byl tam pokusím o útěk zaplatili ty lidi život protože chytli
také popravili to je docela známá věc že byli pověšený ty kosti dělal šli do
koncentráku takže s- pokud ale kde nebyl problém ani utekl z terezína když budeme od
| pavlovi když to hlídali četníci ty mladiství chodili na práci do zemědělství říkalo se
tomu landwirtschaft to bylo za terezín žilo kolem terezína byli zelinářském poli pro terezín pro
ss komandaturu takže nebyl problém otec terezína nebyl problém otec vulkova problém protože sem
měl
doma rodiče a ty si to odnesl v první řadě ... tady byla zodpovědnost vůči
té rodině doufat ... ale nebyl problém ten spor že se tam mezi těma lidma
udržet protože němci byli na stole prostě by vás udal jo zapálit ty otázky ...
jak jste tam byli stravovali ... no tak stravovali žili jsme ovšem zažili jsme tam
také strašný hlad ke konci o tom že jsme měli kůru ze stromu trávu se
svou fotku tý se z kuchyně jsme vybíraly byla ta potom ke konci špatná doba
chodili franta většinou potraviny z | terezína takový ty tvrdý potraviny a brambora takový ty
věci to zřejmě ty ... esesáci tě zde scházeli tam takže to měl na starosti
těch okolní vesnici od sedláků to vypadalo že jo to bylo to byl kuchařem byla
taková že v kuchyni a tom také pak vaření to byl v terezíně takový to
byli většinou tuřín spal to bylo celý vařil dost asi deset deka směli táborů ...
filosof ještě pluku vyhodil nebo sem to vždycky okolo pak sem se ty ty šlupky
vod docentovi jednou za týden byla buchta s takovým krémem my jsme tomu říkali takové
to bylo kafe že jo jsme vždycky říkali až bude po válce se to buchtu
udělal paninko to v životě dál potom korupci koupili to bylo také takový ... prostě
ze čtyřech letech opakovala ke konci to proto že ke konci to dostala ke konci
| to znamená ke konci když už potom byl rok čtyřicet šátek roku čtyřicet pět
tak už to bylo zlý protože z východu se ztratili rusové a došlo k tomu
já bych chtěl ještě ráda ještě k tomu i já sem včera četl tady podrobili
výslechu z toho že frances počkej rakousku tma toho teda se jmenoval že on říká
že vůbec se tam nějak provinili tím ale vo tom strašně vězně mlátit ... já
sem to už na floridu to jsme ještě jako nemluvili ale i tam sem byl
u něj strašně bydlel tam byl rodině se ozývaly na těch lidí víte koho tři
z těch prostě že na tu práci nedělá dobře ale byli lidi který byli placený
za den voskovec lakýrník to mlátil lidi otu rezka že ze zbraslavy z toho mlátili
a oni mlátili tím potom také byli esesáci nemlátil nevrátil | akorát voda a potom
byl mezi námi jeden vězeň pracoval v lodži který také hlásila ale ten jako fakt
fackoval toho po válce popravy ale tenhle ten stužka ten prostě mlátil byl takový zámek
Show:
Locations
Concepts
People
00:00:09
Set:
Channel 1

MALACH Czech Relevance Assessment Interface (Version 6.0)

File Help

Search Assessment

Speaker Information

Hugo Pavel

Personal details

Date of birth Dec 26, 1924
Country of birth Czechoslovakia
Other names Not available
Other attributes/roles Not available
Religious identity (PIQ) none
☒ **Hiding**
☒ **Underground/Resistance/Partisan**
☒ **Death Marches**

Family details

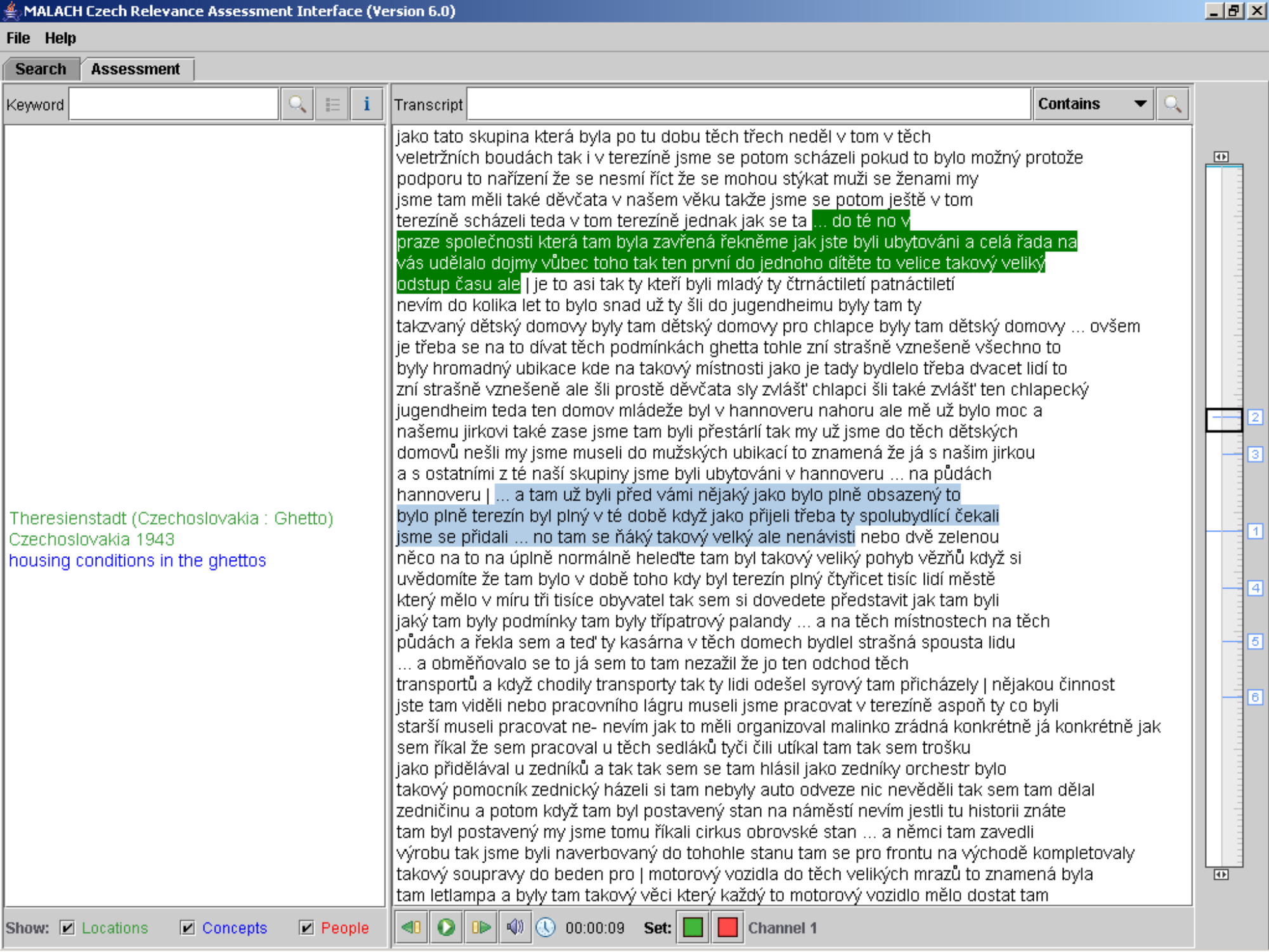
None Available

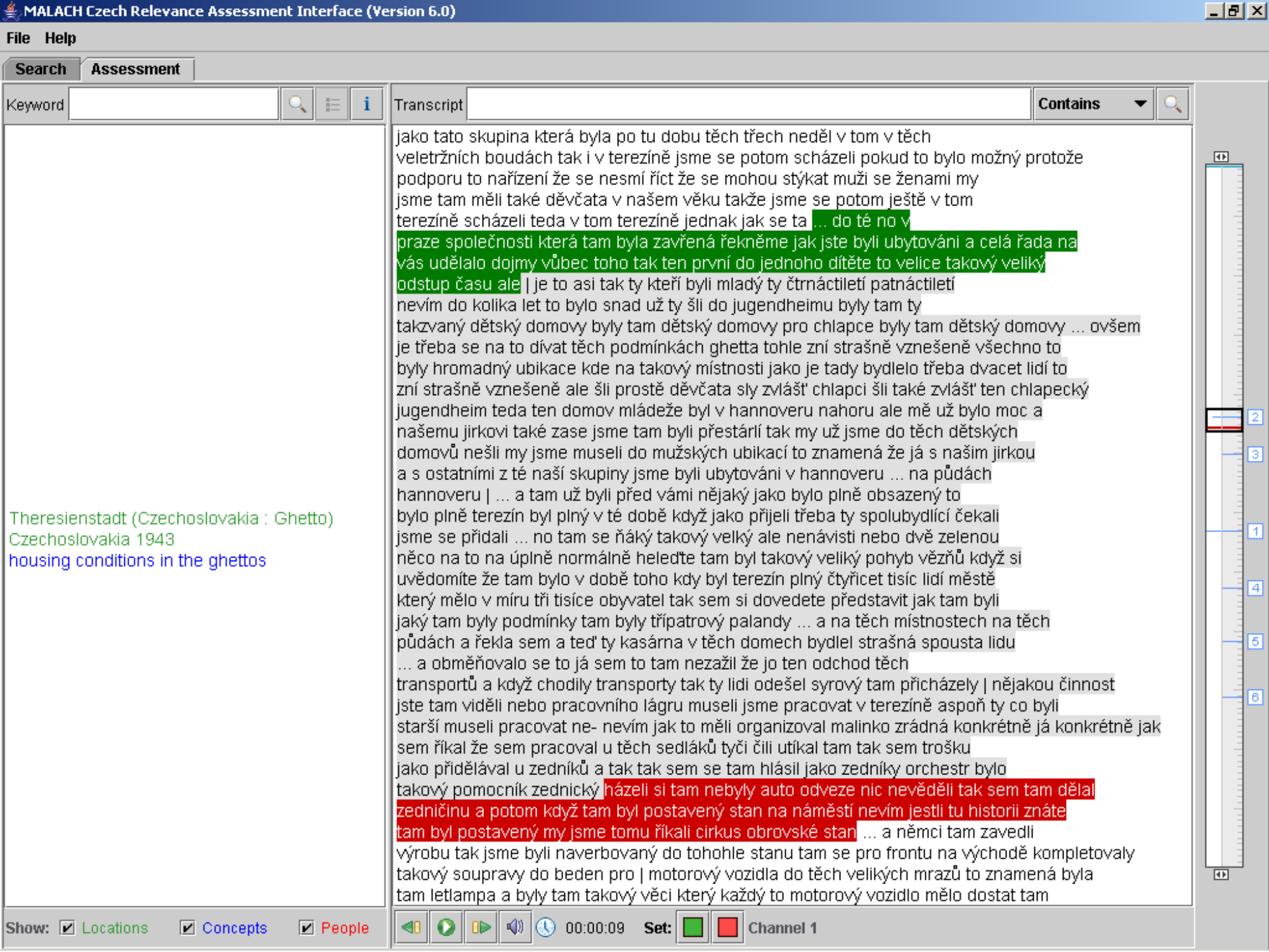
Transcript

Contains

[illegible] byli jsme byli ubytováni ty nás hlídali je po práci
a tam už to bylo volně v zátoce výstaviště nemohlo být prostě první prostě prostě
s tou svou otec byl tam pokusím o útěk zaplatili ty lidi život protože chytli
také popravili to je docela známá věc že byli pověšený ty kosti dělal šli do
koncentráku takže s- pokud ale kde nebyl problém ani utekl z terezína když budeme od
| pavlovi když to hlídali četníci ty mladiství chodili na práci do zemědělství říkalo se
tomu landwirtschaft to bylo za terezín žilo kolem terezína byli zelinářském poli pro terezín pro
ss komandaturu takže nebyl problém otec terezína nebyl problém otec vulkova problém protože sem
měl
doma rodiče a ty si to odnesl v první řadě ... tady byla zodpovědnost vůči
té rodině doufat ... ale nebyl problém ten spor že se tam mezi těma lidma
udržet protože němci byli na stole prostě by vás udal jo zapálit ty otázky ...
jak jste tam byli stravovali ... no tak stravovali žili jsme ovšem zažili jsme tam
také strašný hlad ke konci o tom že jsme měli kůru ze stromu trávu se
svou fotku tý se z kuchyně jsme vybíraly byla ta potom ke konci špatná doba
chodili franta většinou potraviny z | terezína takový ty tvrdý potraviny a brambora takový ty
věci to zřejmě ty ... esesáci tě zde scházeli tam takže to měl na starosti
těch okolní vesnici od sedláků to vypadalo že jo to bylo to byl kuchařem byla
taková že v kuchyni a tom také pak vaření to byl v terezíně takový to
byli většinou tuřín spal to bylo celý vařil dost asi deset deka směli táborů ...
filosof ještě pluku vyhodil nebo sem to vždycky okolo pak sem se ty ty šlupky
vod docentovi jednou za týden byla buchta s takovým krémem my jsme tomu říkali takové
to bylo kafe že jo jsme vždycky říkali až bude po válce se to buchtu
udělal paninko to v životě dál potom korupci koupili to bylo také takový ... prostě
ze čtyřech letech opakovala ke konci to proto že ke konci to dostala ke konci
| to znamená ke konci když už potom byl rok čtyřicet šátek roku čtyřicet pět
tak už to bylo zlý protože z východu se ztratili rusové a došlo k tomu
já bych chtěl ještě ráda ještě k tomu i já sem včera četl tady podrobili
výslechu z toho že frances počkej rakousku tma toho teda se jmenoval že on říká
že vůbec se tam nějak provinili tím ale vo tom strašně vězně mlátit ... já
sem to už na floridu to jsme ještě jako nemluvili ale i tam sem byl
u něj strašně bydlel tam byl rodině se ozývaly na těch lidí víte koho tři
z těch prostě že na tu práci nedělá dobře ale byli lidi který byli placený
za den voskovec lakýrník to mlátil lidi otu rezka že ze zbraslavy z toho mlátili
a oni mlátili tím potom také byli esesáci nemlátil nevrátil | akorát voda a potom
byl mezi námi jeden vězeň pracoval v lodži který také hlásila ale ten jako fakt
fackoval toho po válce popravy ale tenhle ten stužka ten prostě mlátil byl takový zámek
krakova ten nebo když se škoda žili ve se ty stromy tak mlátil těma kupama

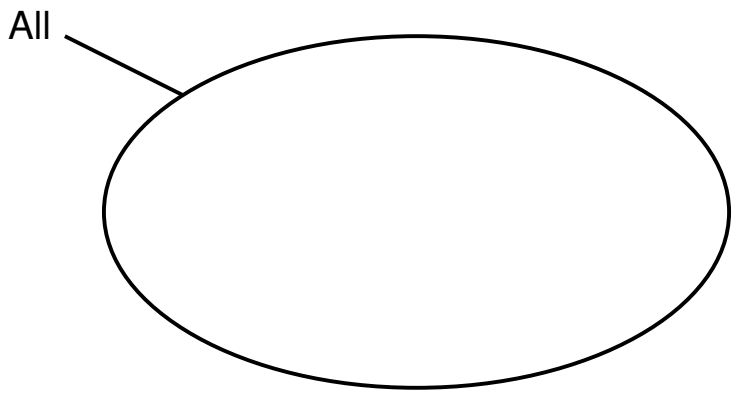
00:00:09 Set: Channel 1


MALACH Czech Relevance Assessment Interface (Version 6.0)
File
Help
Search
Assessment
Keyword
Transcript
Contains
jako tato skupina která byla po tu dobu těch třech neděl v tom v těch
veletržních boudách tak i v terezíně jsme se potom scházeli pokud to bylo možný protože
podporu to nařízení že se nesmí říct že se mohou stýkat muži se ženami my
jsme tam měli také děvčata v našem věku takže jsme se potom ještě v tom
terezíně scházeli teda v tom terezíně jednak jak se ta ... do té no v
praze společnosti která tam byla zavřená řekněme jak jste byli ubytováni a celá řada na
vás udělalo dojmy vůbec toho tak ten první do jednoho dítěte to velice takový veliký
odstup času ale | je to asi tak ty kteří byli mladý ty čtrnáctiletí patnáctiletí
nevím do kolika let to bylo snad už ty šli do jugendheimu byly tam ty
takzvaný dětský domovy byly tam dětský domovy pro chlapce byly tam dětský domovy ... ovšem
je třeba se na to dívat těch podmínkách ghetta tohle zní strašně vznešeně všechno to
byly hromadný ubikace kde na takový místnosti jako je tady bydlelo třeba dvacet lidí to
zní strašně vznešeně ale šli prostě děvčata sly zvlášť chlapci šli také zvlášť ten chlapecký
jugendheim teda ten domov mládeže byl v hannoveru nahoru ale mě už bylo moc a
našemu jirkovi také zase jsme tam byli přestárlí tak my už jsme do těch dětských
domovů nešli my jsme museli do mužských ubikací to znamená že já s našim jirkou
a s ostatními z té naší skupiny jsme byli ubytováni v hannoveru ... na půdách
hannoveru | ... a tam už byli před vámi nějaký jako bylo plně obsazený to
bylo plně terezín byl plný v té době když jako přijeli třeba ty spolubydlící čekali
jsme se přidali ... no tam se ňáký takový velký ale nenávisti nebo dvě zelenou
něco na to na úplně normálně heleďte tam byl takový veliký pohyb vězňů když si
uvědomíte že tam bylo v době toho kdy byl terezín plný čtyřicet tisíc lidí městě
který mělo v míru tři tisíce obyvatel tak sem si dovedete představit jak tam byli
jaký tam byly podmínky tam byly třípatrový palandy ... a na těch místnostech na těch
půdách a řekla sem a teď ty kasárna v těch domech bydlel strašná spousta lidu
... a obměňovalo se to já sem to tam nezažil že jo ten odchod těch
transportů a když chodily transporty tak ty lidi odešel syrový tam přicházely | nějakou činnost
jste tam viděli nebo pracovního lágru museli jsme pracovat v terezíně aspoň ty co byli
starší museli pracovat ne- nevím jak to měli organizoval malinko zrádná konkrétně já konkrétně jak
sem říkal že sem pracoval u těch sedláků tyči čili utíkal tam tak sem trošku
jako přidělával u zedníků a tak tak sem se tam hlásil jako zedníky orchestr bylo
takový pomocník zednický házeli si tam nebyly auto odveze nic nevěděli tak sem tam dělal
zedničinu a potom když tam byl postavený stan na náměstí nevím jestli tu historii znáte
tam byl postavený my jsme tomu říkali cirkus obrovské stan ... a němci tam zavedli
výrobu tak jsme byli naverbovaný do tohohle stanu tam se pro frontu na východě kompletovaly
takový soupravy do beden pro | motorový vozidla do těch velikých mrazů to znamená byla
tam letlampa a byly tam takový věci který každý to motorový vozidlo mělo dostat tam
Theresienstadt (Czechoslovakia : Ghetto)
Czechoslovakia 1943
housing conditions in the ghettos
Show:
Locations
Concepts
People
00:00:09
Set:
Channel 1
2
3
1
4
5
6

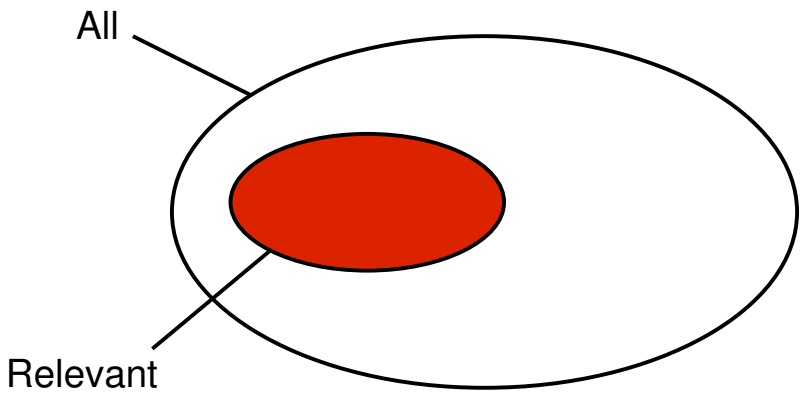

MALACH Czech Relevance Assessment Interface (Version 6.0)
File
Help
Search
Assessment
Keyword
Transcript
Contains
jako tato skupina která byla po tu dobu těch třech neděl v tom v těch
veletržních boudách tak i v terezíně jsme se potom scházeli pokud to bylo možný protože
podporu to nařízení že se nesmí říct že se mohou stýkat muži se ženami my
jsme tam měli také děvčata v našem věku takže jsme se potom ještě v tom
terezíně scházeli teda v tom terezíně jednak jak se ta ... do té no v
praze společnosti která tam byla zavřená řekněme jak jste byli ubytováni a celá řada na
vás udělalo dojmy vůbec toho tak ten první do jednoho dítěte to velice takový veliký
odstup času ale | je to asi tak ty kteří byli mladý ty čtrnáctiletí patnáctiletí
nevím do kolika let to bylo snad už ty šli do jugendheimu byly tam ty
takzvaný dětský domovy byly tam dětský domovy pro chlapce byly tam dětský domovy ... ovšem
je třeba se na to dívat těch podmínkách ghetta tohle zní strašně vznešeně všechno to
byly hromadný ubikace kde na takový místnosti jako je tady bydlelo třeba dvacet lidí to
zní strašně vznešeně ale šli prostě děvčata sly zvlášť chlapci šli také zvlášť ten chlapecký
jugendheim teda ten domov mládeže byl v hannoveru nahoru ale mě už bylo moc a
našemu jirkovi také zase jsme tam byli přestárlí tak my už jsme do těch dětských
domovů nešli my jsme museli do mužských ubikací to znamená že já s našim jirkou
a s ostatními z té naší skupiny jsme byli ubytováni v hannoveru ... na půdách
hannoveru | ... a tam už byli před vámi nějaký jako bylo plně obsazený to
bylo plně terezín byl plný v té době když jako přijeli třeba ty spolubydlící čekali
jsme se přidali ... no tam se ňáký takový velký ale nenávisti nebo dvě zelenou
něco na to na úplně normálně heleďte tam byl takový veliký pohyb vězňů když si
uvědomíte že tam bylo v době toho kdy byl terezín plný čtyřicet tisíc lidí městě
který mělo v míru tři tisíce obyvatel tak sem si dovedete představit jak tam byli
jaký tam byly podmínky tam byly třípatrový palandy ... a na těch místnostech na těch
půdách a řekla sem a teď ty kasárna v těch domech bydlel strašná spousta lidu
... a obměňovalo se to já sem to tam nezažil že jo ten odchod těch
transportů a když chodily transporty tak ty lidi odešel syrový tam přicházely | nějakou činnost
jste tam viděli nebo pracovního lágru museli jsme pracovat v terezíně aspoň ty co byli
starší museli pracovat ne- nevím jak to měli organizoval malinko zrádná konkrétně já konkrétně jak
sem říkal že sem pracoval u těch sedláků tyči čili utíkal tam tak sem trošku
jako přidělával u zedníků a tak tak sem se tam hlásil jako zedníky orchestr bylo
takový pomocník zednický házeli si tam nebyly auto odveze nic nevěděli tak sem tam dělal
zedničinu a potom když tam byl postavený stan na náměstí nevím jestli tu historii znáte
tam byl postavený my jsme tomu říkali cirkus obrovské stan ... a němci tam zavedli
výrobu tak jsme byli naverbovaný do tohohle stanu tam se pro frontu na východě kompletovaly
takový soupravy do beden pro | motorový vozidla do těch velikých mrazů to znamená byla
tam letlampa a byly tam takový věci který každý to motorový vozidlo mělo dostat tam
Theresienstadt (Czechoslovakia : Ghetto)
Czechoslovakia 1943
housing conditions in the ghettos
Show:
Locations
Concepts
People
00:00:09
Set:
Channel 1
2
3
1
4
5
6

Effectiveness measures

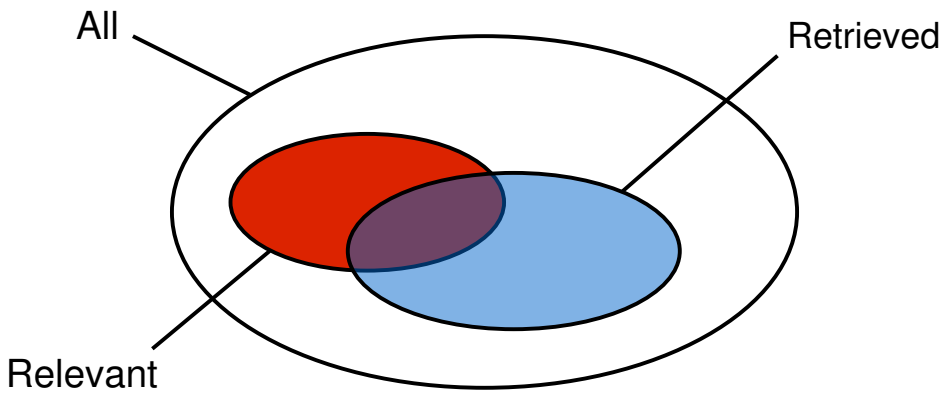
Effectiveness measures



Effectiveness measures

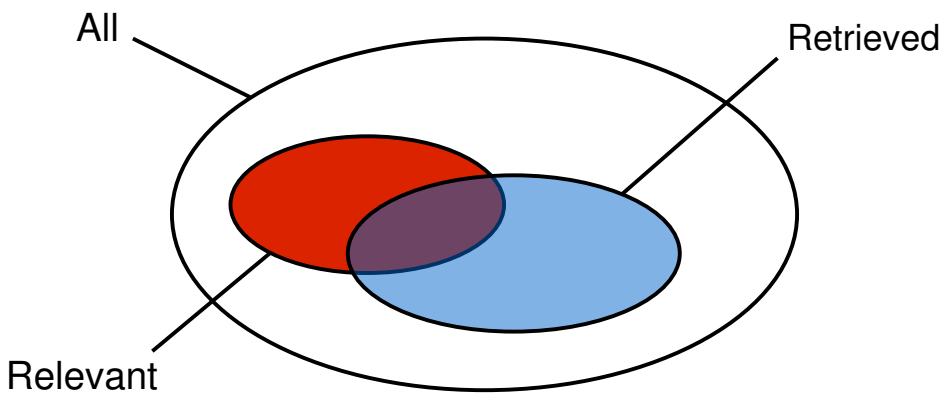


Effectiveness measures



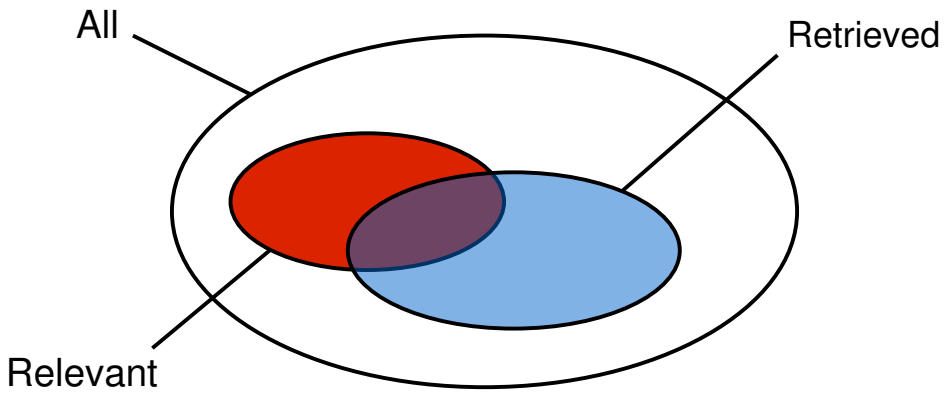
Effectiveness measures

document	retrieved	not retrieved
relevant	**relevant retrieved**	**relevant missed**
not relevant	**false alarm**	**irrelevant rejected**



Effectiveness measures

document	retrieved	not retrieved
relevant	**relevant retrieved**	**relevant missed**
not relevant	**false alarm**	**irrelevant rejected**



$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	*0.96*
int.7264-seg.068	*0.93*
int.3012-seg.142	*0.88*
int.9595-seg.119	*0.85*
int.2306-seg.120	*0.79*
int.9262-seg.131	*0.75*
int.4188-seg.033	*0.73*
int.6805-seg.007	*0.65*
int.2591-seg.123	*0.62*
int.5325-seg.015	*0.61*
int.6042-seg.038	*0.55*
int.1066-seg.149	*0.54*
int.3215-seg.071	*0.52*
int.4404-seg.023	*0.45*
int.2012-seg.121	*0.34*
int.6707-seg.116	*0.21*

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	*0.96*
int.7264-seg.068	*0.93*
int.3012-seg.142	*0.88*
int.9595-seg.119	*0.85*
int.2306-seg.120	*0.79*
int.9262-seg.131	*0.75*
int.4188-seg.033	*0.73*
int.6805-seg.007	*0.65*
int.2591-seg.123	*0.62*
int.5325-seg.015	*0.61*
int.6042-seg.038	*0.55*
int.1066-seg.149	*0.54*
int.3215-seg.071	*0.52*
int.4404-seg.023	*0.45*
int.2012-seg.121	*0.34*
int.6707-seg.116	*0.21*

treshold

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

treshold

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	0
int.9262-seg.131	0
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

treshold

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	0
int.9262-seg.131	0
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

treshold ↙

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	0
int.9262-seg.131	0
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	0
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	0
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	0
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	0
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	1
int.2591-seg.123	0
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	1
int.2591-seg.123	1
int.5325-seg.015	0
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	1
int.2591-seg.123	1
int.5325-seg.015	1
int.6042-seg.038	0
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %

Evaluation metrics: example

$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.96
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	1
int.2591-seg.123	1
int.5325-seg.015	1
int.6042-seg.038	1
int.1066-seg.149	0
int.3215-seg.071	0
int.4404-seg.023	0
int.2012-seg.121	0
int.6707-seg.116	0

precision	recall
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %
72 %	100 %

Evaluation metrics: example

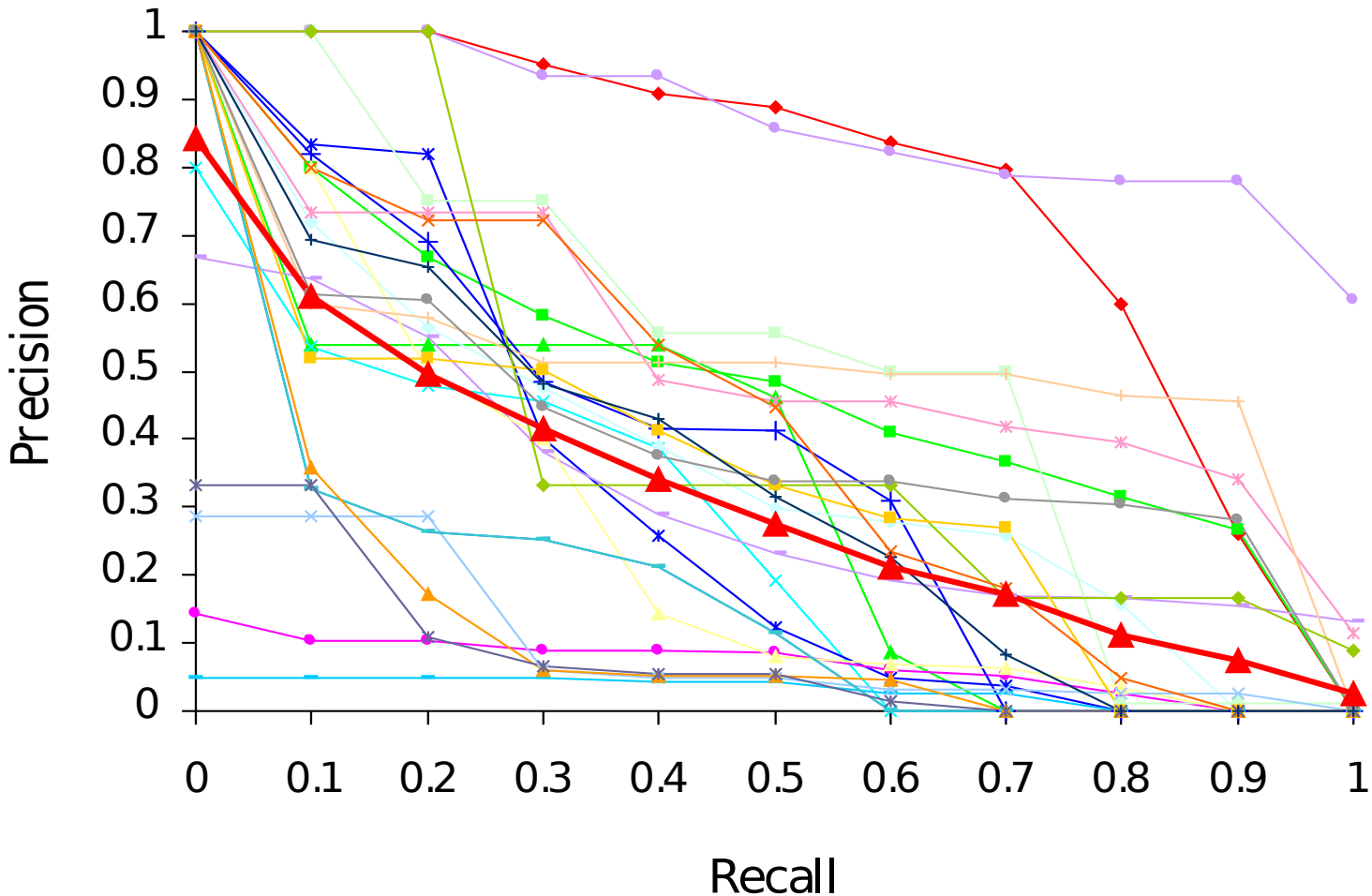
$$Precision = \frac{|relevant\ retrieved|}{|retrieved|} \qquad Recall = \frac{|relevant\ retrieved|}{|relevant|}$$

system results	
int.3431-seg.125	0.91
int.7264-seg.068	0.93
int.3012-seg.142	0.88
int.9595-seg.119	0.85
int.2306-seg.120	0.79
int.9262-seg.131	0.75
int.4188-seg.033	0.73
int.6805-seg.007	0.65
int.2591-seg.123	0.62
int.5325-seg.015	0.61
int.6042-seg.038	0.55
int.1066-seg.149	0.54
int.3215-seg.071	0.52
int.4404-seg.023	0.45
int.2012-seg.121	0.34
int.6707-seg.116	0.21

classification confusion	
int.3431-seg.125	1
int.7264-seg.068	1
int.3012-seg.142	1
int.9595-seg.119	1
int.2306-seg.120	1
int.9262-seg.131	1
int.4188-seg.033	1
int.6805-seg.007	1
int.2591-seg.123	1
int.5325-seg.015	1
int.6042-seg.038	1
int.1066-seg.149	1
int.3215-seg.071	1
int.4404-seg.023	1
int.2012-seg.121	1
int.6707-seg.116	1

precision	recall
100 %	12 %
100 %	25 %
100 %	37 %
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %
72 %	100 %
66 %	100 %
61 %	100 %
57 %	100 %
53 %	100 %
50 %	100 %

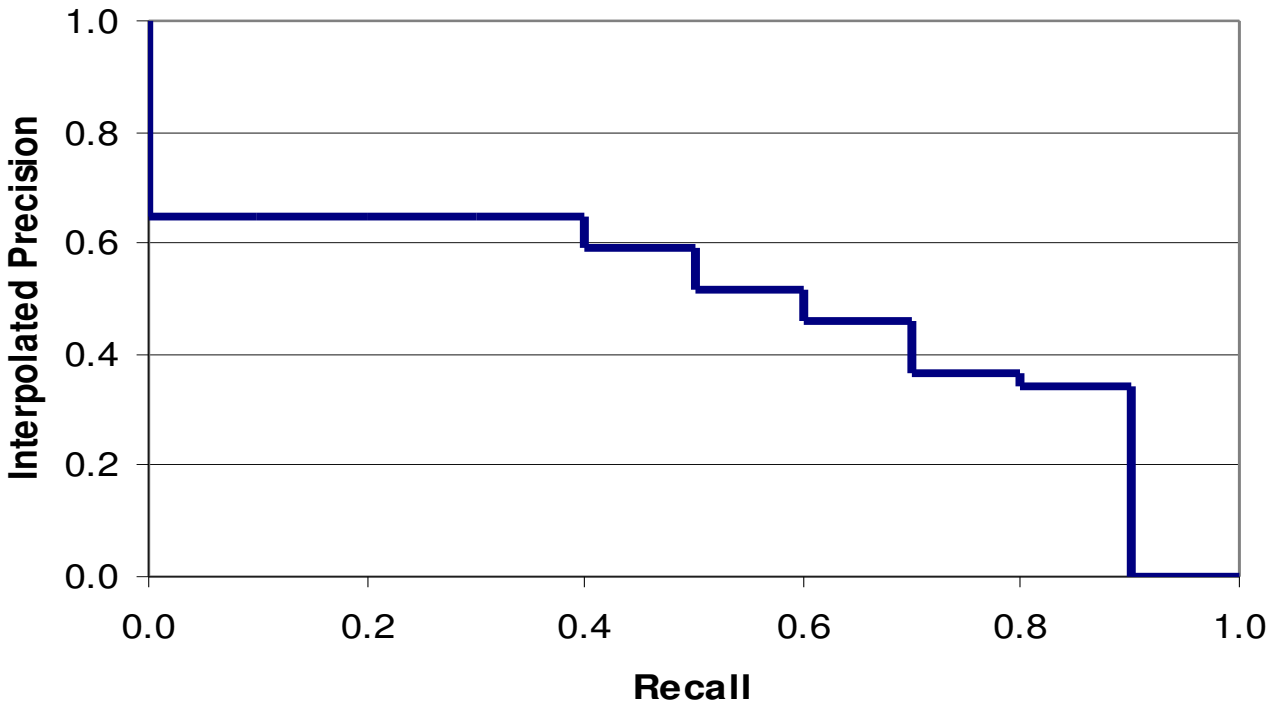
Precision–recall curves



credit to Doug Oard

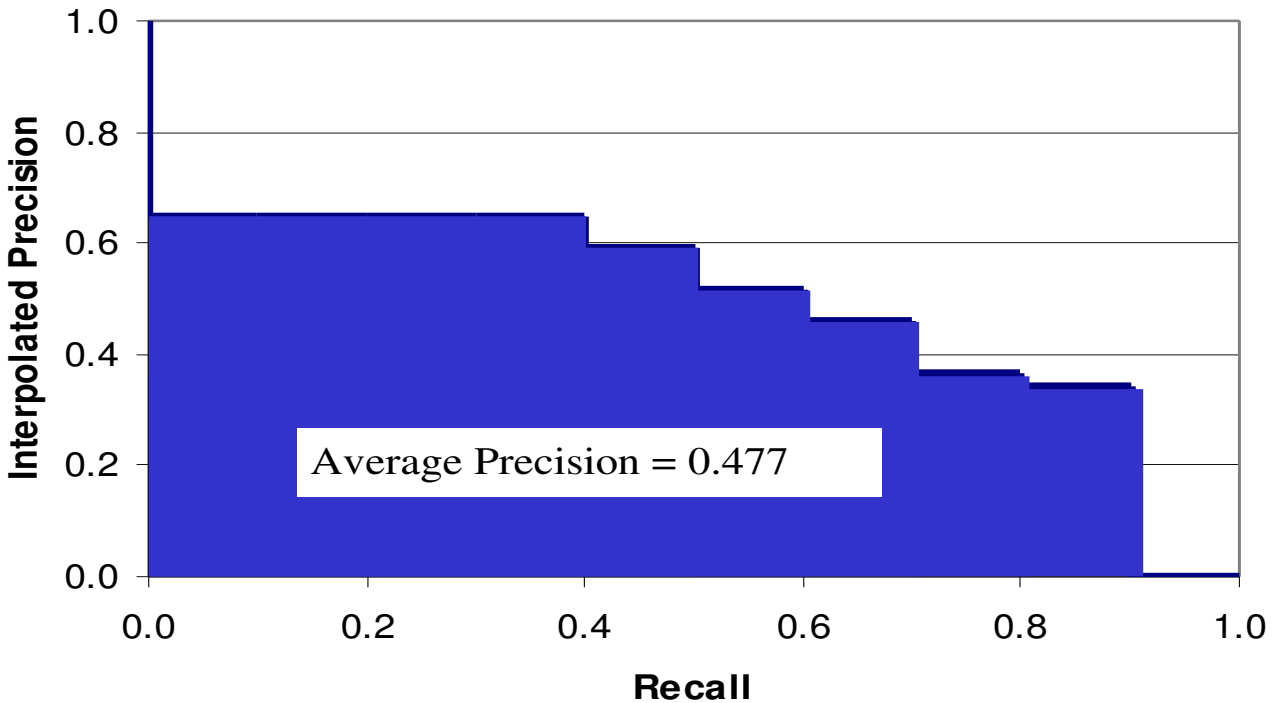
Average precision

- the expected value of precision for all possible values of recall
- equal to the area under the precision–recall curve (AUC)

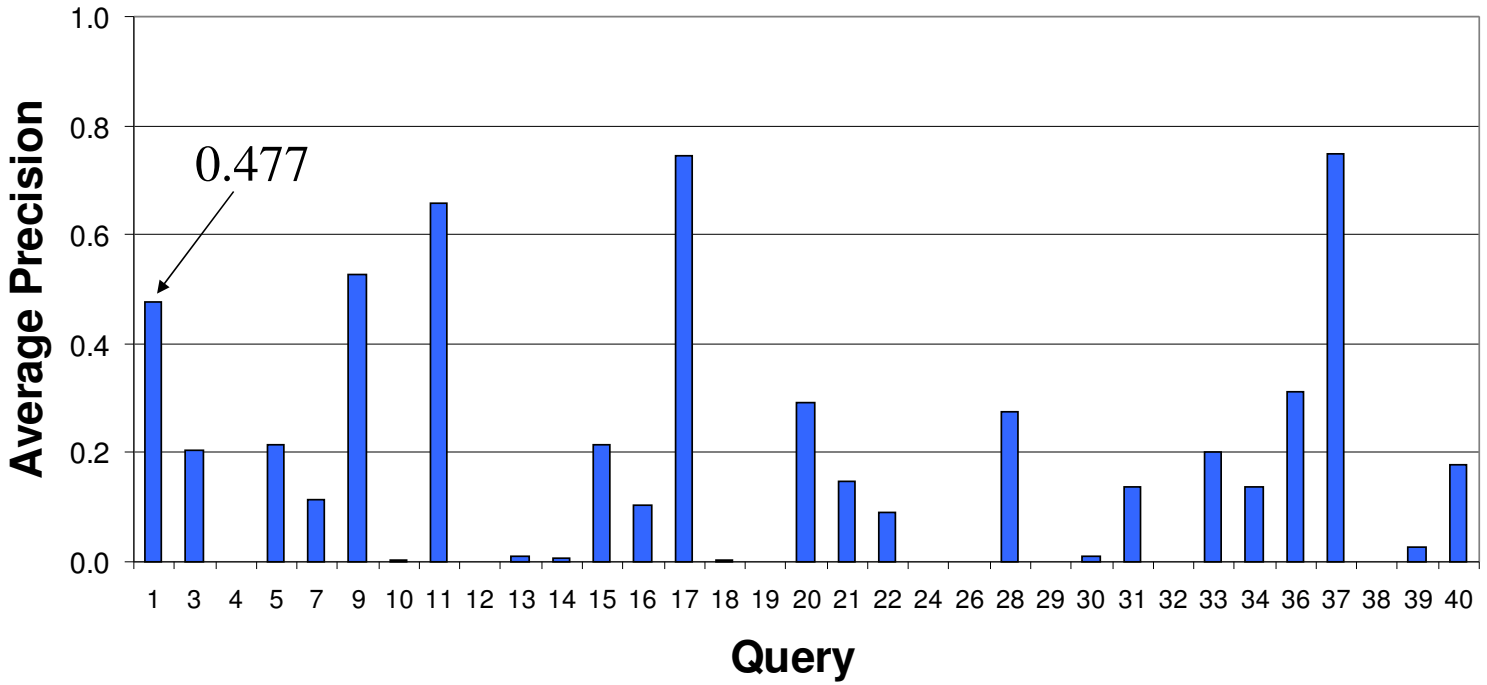


Average precision

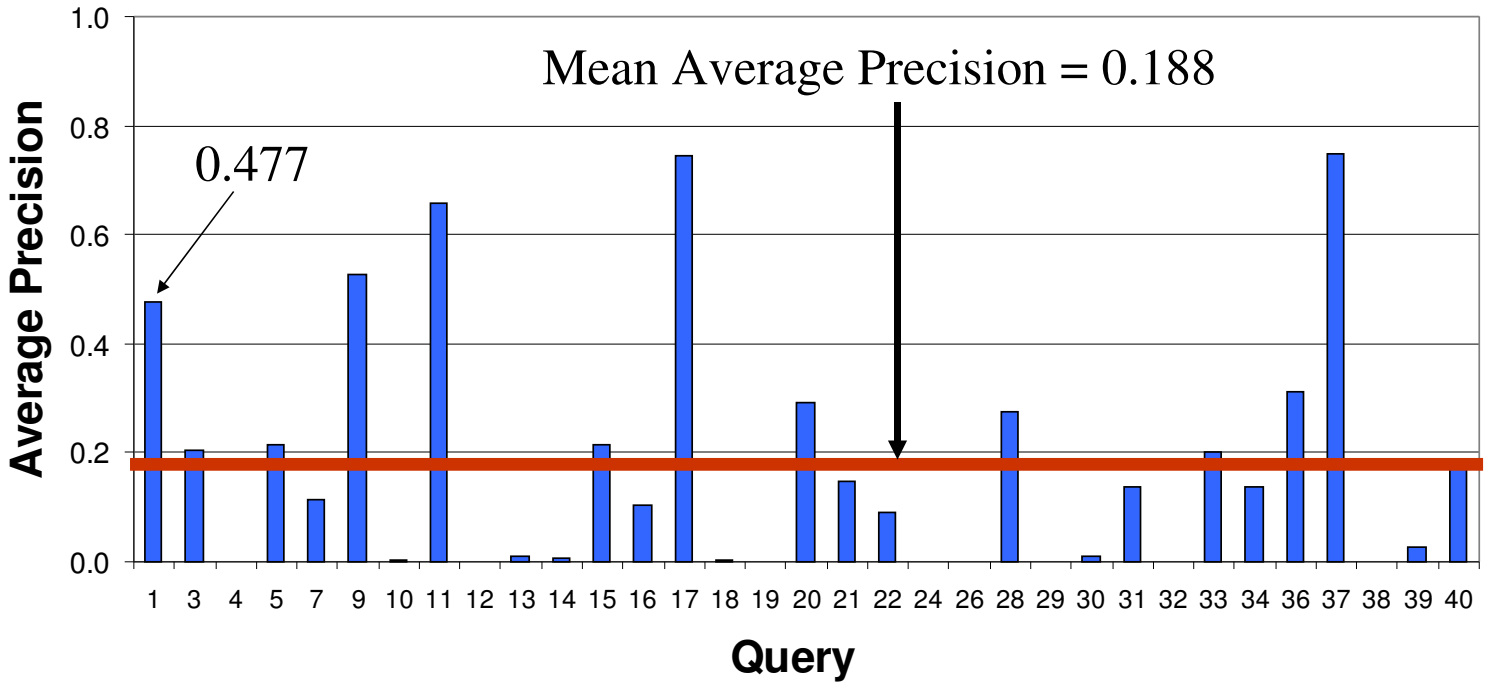
- the expected value of precision for all possible values of recall
- equal to the area under the precision–recall curve (AUC)



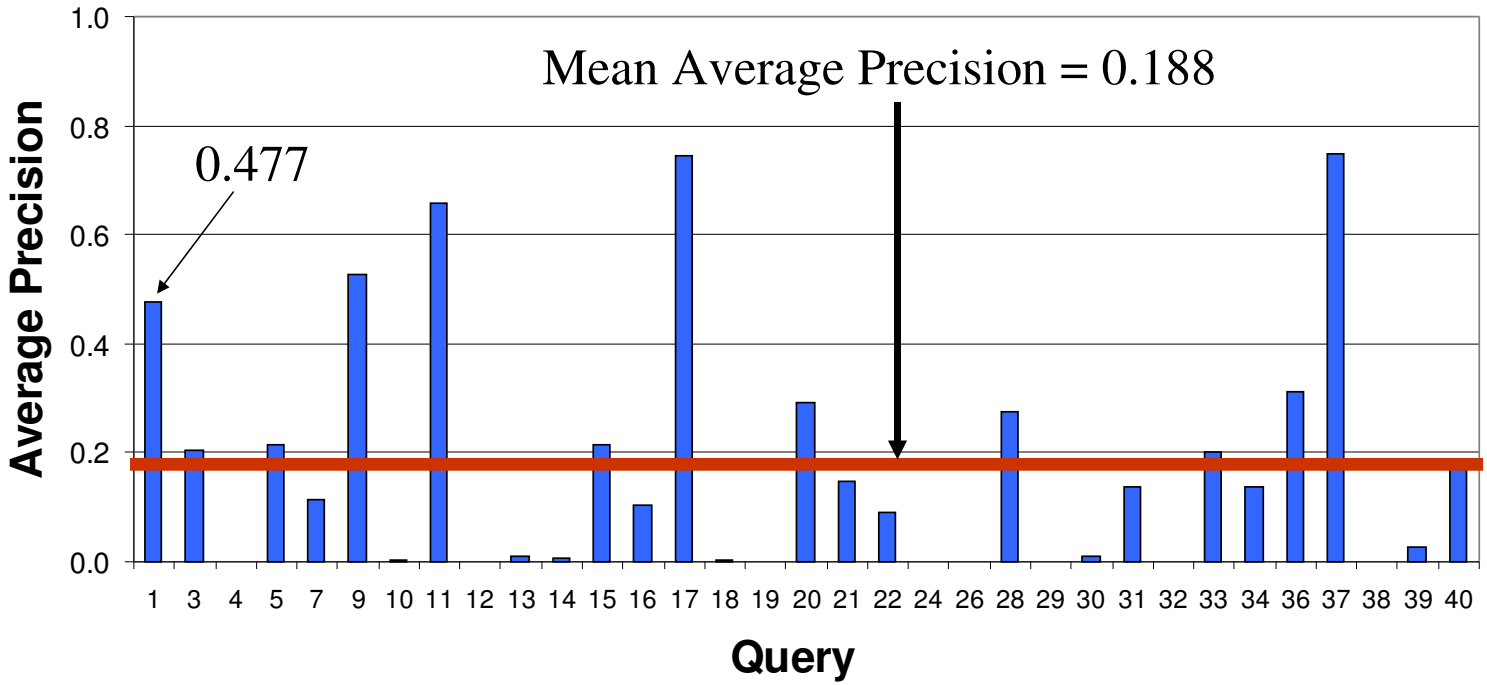
Mean average precision



Mean average precision



Mean average precision



Measuring improvement

- Meaningful improvement: *0.05 is noticable, 0.1 makes a difference*
- Reliable improvement: *Wilcoxon signed rank test for paired samples*
- Maximum precision limit: *set by inter–assessors agreement*

credit to Doug Oard

Cross-Language Evaluation Forum (CLEF)

- an activity of the *DELOS Network of Excellence for Digital Libraries* under the *Sixth Framework Programme of the European Commission*
- develops an infrastructure for testing, tuning, and evaluation of IR systems on European languages in both monolingual and cross–language contexts
- creates test–suites of reusable data which can be employed by system developers for benchmarking purposes.
- offers a series of evaluation tracks to test different aspects of cross–language information retrieval system development:

1. Mono–, Bi– and Multilingual Document Retrieval on News Collections
2. Mono and Cross–Language IR on Structured Scientific Data
3. Interactive Cross–Language IR
4. Multiple Language Question Answering
5. Cross–Language Retrieval in Image Collections
6. Multilingual Web Track (WebCLEF)
7. Cross–Language Speech Retrieval
8. Cross–Language Geographical Retrieval

CLEF 2006 Cross–language speech retrieval track overview

Two tasks: English segments, Czech start times

- Max of 5 official runs per team per task
- Baseline English run: ASR / English TD topics

7 teams / 6 countries

- Canada: *Ottawa* (EN, CZ)
- Czech Rep: *West Bohemia* (CZ)
- Ireland: DCU (EN)
- Netherlands: *Twente* (EN)
- Spain: *Alicante* (EN), *UNED* (EN)
- USA: *Maryland* (EN, CZ)

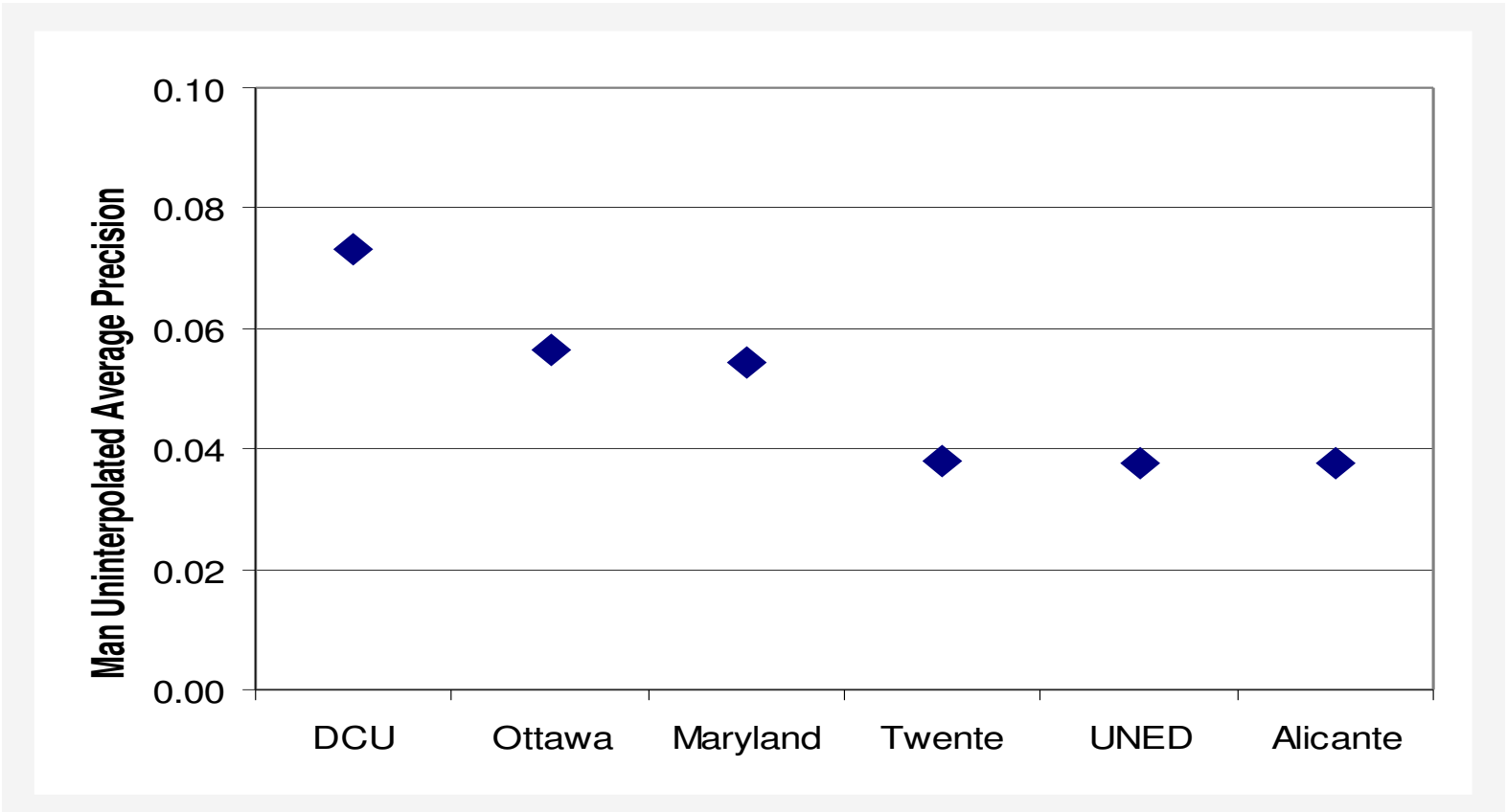
Schedule

Jan 15	Registration opened
Apr 14	Data release
May 1	Topic release
Jun 6	Submission of runs by participants
Aug 1	Release of relevance assessments and individual results
Aug 8	Submission of paper for Working Notes
Sep 20	Workshop (Alicante, Spain)

English results (MAP)

Queries: title + description

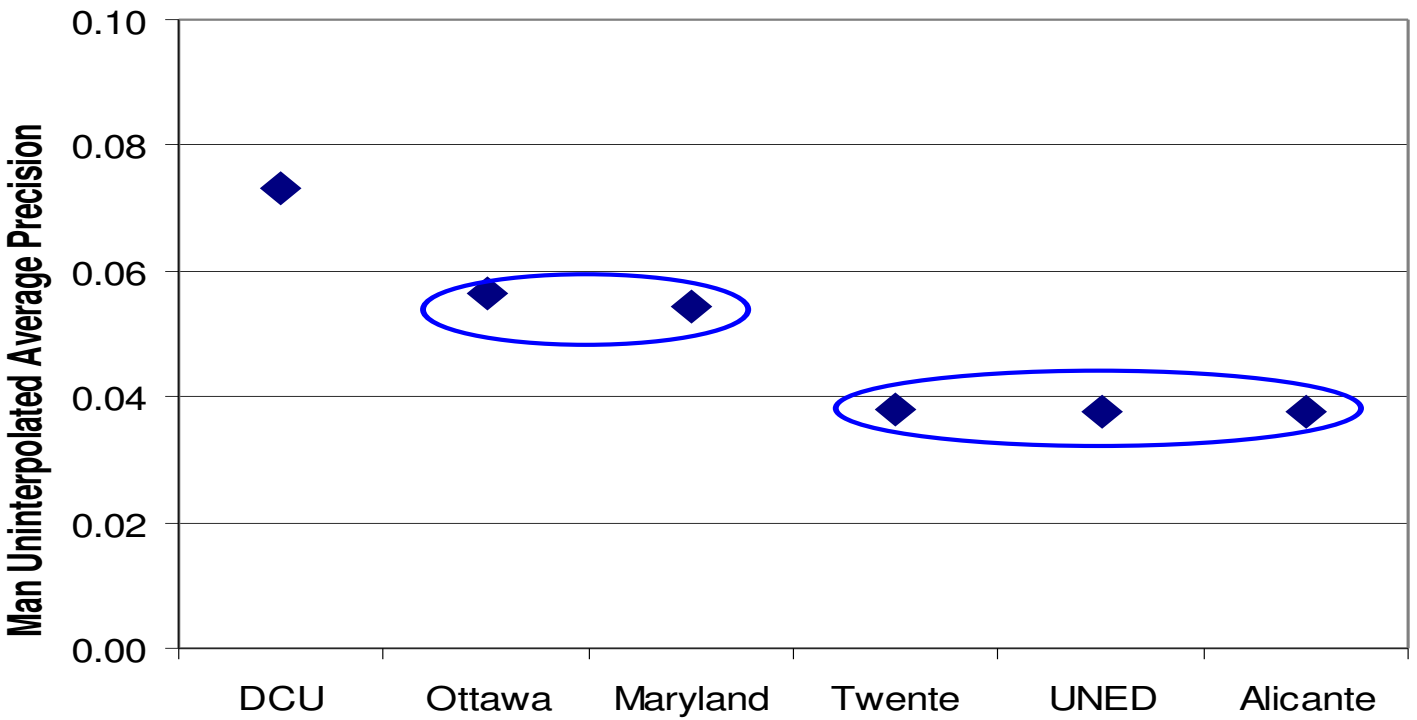
Documents: ASR output



English results (MAP)

Queries: title + description

Documents: ASR output



Monolingual vs. cross–language comparisons (MAP)

Queries: title/description/narrative

Documents: ASR output, metadata, both

Team	Q	D	English	French	Dutch	Spanish
Ottawa	TDN	Meta	0.2902			
Twente	T	Meta	0.2058		80%	
DCU	TD	Both	0.2015	79%		
Maryland	TD	Meta	0.2350	44%		
UNED	TD	Meta	0.1766			51%
Ottawa	TDN	ASR	0.0768	83%		81%
DCU	TD	ASR	0.0733	63%		
Twente	T	ASR	0.0495		77%	
UNED	TD	ASR	0.0376			68%
Maryland	TD	ASR	0.0543	38%		

Conclusion

Call for participation: CLEF 2007

- Registration Opens - *15 January 2007*
- Data Release - *15 February 2007*
- Topic Release - *1 April 2007*
- Submission of Runs by Participants - *15 June 2007*
- Release of Relevance Assessments and Individual Results - *15 July 2007*
- Submission of Paper for Working Notes - *15 August 2007*
- Workshop - *19–21 September 2007*

1. Multilingual Document Retrieval on News Collections
2. Scientific Data Retrieval
3. Interactive Cross–Language Information Retrieval
4. Multiple Language Question Answering
5. Cross–Language Image Retrieval
6. Cross–Language Spoken Retrieval
7. CLEF Web Track
8. Cross–Language Geographical Information Retrieval

Thank You!

People said ...

Doug Greenberg:

- "We don't edit any of these interviews. It's completely raw footage taken directly from interviews with survivors. It will be broadly accessible, but it won't be edited."
- "Our mission now is to use the archive in educational settings to overcome prejudice and bigotry."

Doug Oard:

- "There's a lot more oral history than anybody even knows about".
- "It isn't as good as a human cataloging, but it's $100 million cheaper."
- "When you develop this type of technology, you open a lot of doors,"

Links

- Shoah Foundation / Visual History Institute
  `http://www.usc.edu/schools/college/vhi/`
- Malach Project
  `http://malach.umiacs.umd.edu/`
- Cross–Language Evaluation Forum
  `http://www.clef-campaign.org/`
- Cross–Language Speech Retrieval track of CLEF
  `http://clef-clsr.umiacs.umd.edu/`