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

LCT workshop, Prague
May 29, 2012

Cross–Language Speech Retrieval and its Evaluation

Information Retrieval

- searching a body of information for objects that match a search query
- e.g. searching for pages on the Web

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

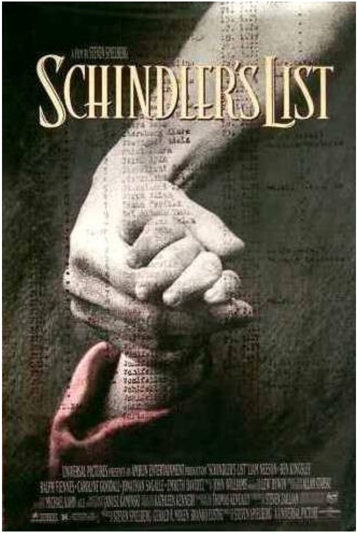
- deals with effectiveness of IR systems: how well they perform
- measures how well users are able to acquire information
- usually comparative: ranks a better system ahead of a worse system

Talk Outline

1. Introduction
2. Project Overview
3. Speech Recognition
4. Speech Retrieval
5. Evaluation
6. Conclusion

Project Overview

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.

2001 NSF project proposal with the goal to dramatically improve access to large multilingual spoken word collections

2001 The grant awarded ($7.5 million for 5 years), the project launches

2006 The project ends, results implemented in a number of archive access points

2010 CVH Malach: the first European access point opened at Charles University

2012 AMalach: 3-year project successor funded by Czech Ministry of Culture

The archive

- maintained by USC Visual History 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
- full manual cataloging of 10% data
- brief manual indexing of the rest of the interviews

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

Manual cataloging and annotation

Done for 4,000 interviews (10,000 hrs, 72 mil words); one testimony $\sim$ 35 hrs

Interview-level annotation

- pre-interview questionaire (including names of people and places)
- free text summary

Segment-level annotation

- topic boundaries (average 3 min/segment)
- descriptions: *summary, cataloguer's scratchpad*
- labels from a 30,000-keyword thesaurus: *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*

Manual indexing

Done for all the remainig interviews, real-time (i.e. $\sim 15\times$ faster)

Interview–level annotation

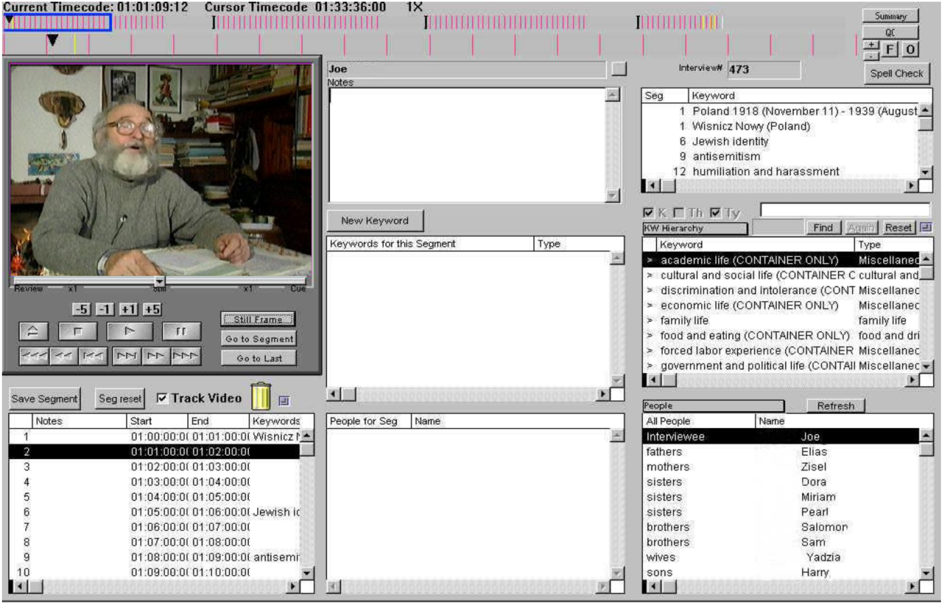
- pre–interview questionaire, no free text summary

Time-aligned annotation

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

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

Cataloging interface



Current Timecode: 01:01:09:12 Cursor Timecode 01:33:36:00 1X

Interview# 473

Joe

Notes

Seg	Keyword
1	Poland 1918 (November 11) - 1939 (August
1	Wisnicz Nowy (Poland)
6	Jewish identity
9	antisemitism
12	humiliation and harassment

☑ K □ Th ☑ Ty

KW Hierarchy

Keyword	Type
academic life (CONTAINER ONLY)	Miscellaneo
cultural and social life (CONTAINER C	cultural and
discrimination and intolerance (CONT	Miscellaneo
economic life (CONTAINER ONLY)	Miscellaneo
family life	family life
food and eating (CONTAINER ONLY)	food and dri
forced labor experience (CONTAINER	Miscellaneo
government and political life (CONTAII	Miscellaneo

Keywords for this Segment	Type

People for Seg	Name

☑ Track Video

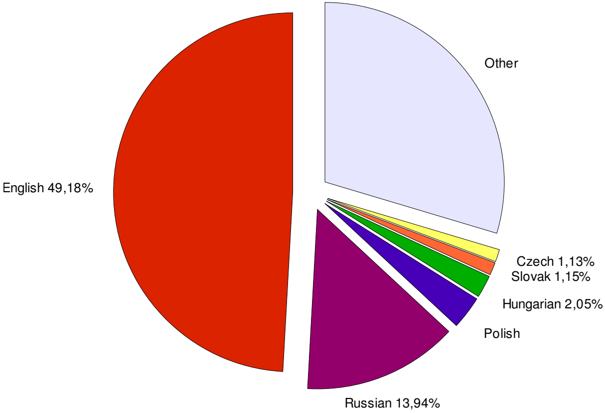
	Notes	Start	End	Keywords
1		01:00:00:0(	01:01:00:0(	Wisnicz N
2		01:01:00:0(	01:02:00:0(	
3		01:02:00:0(	01:03:00:0(	
4		01:03:00:0(	01:04:00:0(	
5		01:04:00:0(	01:05:00:0(	
6		01:05:00:0(	01:06:00:0(	Jewish ic
7		01:06:00:0(	01:07:00:0(	
8		01:07:00:0(	01:08:00:0(	
9		01:08:00:0(	01:09:00:0(	antisemi
10		01:09:00:0(	01:10:00:0(	

People

All People	Name
Interviewee	Joe
fathers	Elias
mothers	Zisel
sisters	Dora
sisters	Miriam
sisters	Pearl
brothers	Salomon
brothers	Sam
wives	Yadzia
sons	Harry

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 counts
English 24,872

Project tasks and participants

1. automatic speech recognition (multi-lingual)
2. automatic speech recognition (multi-lingual)
3. machine supported translation of domain specific thesaurus
4. automatic topic boundary tagging and time-aligned metadata assignment
5. environment for cross-language speech retrieval and browsing
6. environment for cross-language speech 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 realized by CUNI and UWB*

University of Maryland, College Park

- *archive browsing, information retrieval and its evaluation*
- *Czech test collection developed by 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
- specific issue in Czech: colloquial expressions and pronunciation

odjet	[*odjet*]	Osvětim	[*osvjetim*]
	[*odejet*]		[*vosvjetim*]
	[*odject*]		[*osvjenčim*]
	[*vodeject*]		[*ozvjetim*]

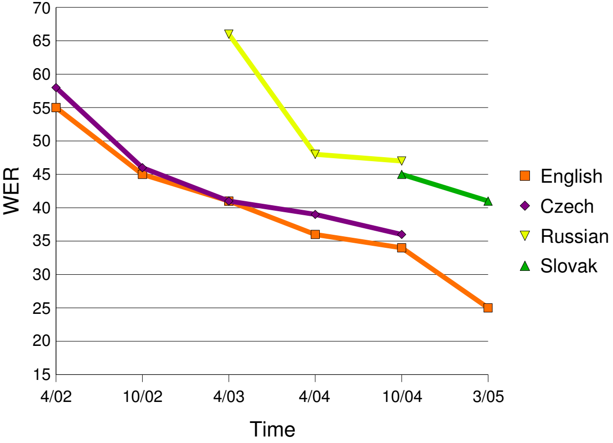
Speech recognition results

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

language	WER (%)
English	25.0
Czech	27.1
Russian	45.7
Slovak	34.5

Manual transcriptions

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



Automatic transcription example

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 ...

Speech Retrieval

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

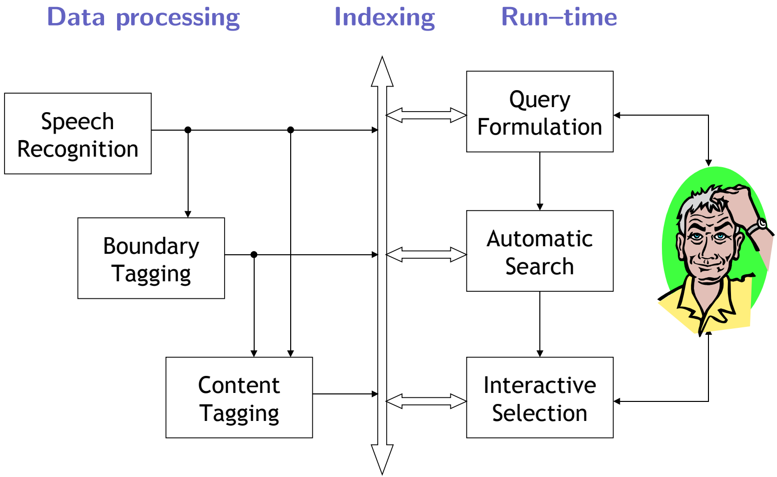
Recognition errors may not bother the system, but they do bother the user

- 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

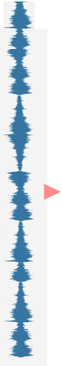
Speech Transcription

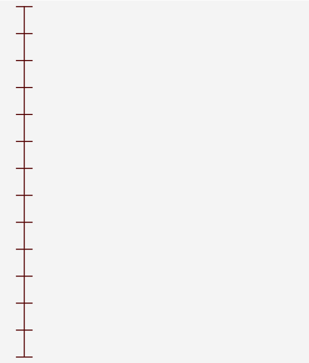
Boundary Tagging

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?

Criteria

- Effectiveness, efficiency, usability?
- 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

Malach document collections)

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 time used as document ID
- ASR: 35% mean Word Error Rate

Malach 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* → allows cross-language retrieval
- topic descriptions: title + short description + narrative description:

num 1173

title **Children's art in Terezin**

desc We are looking for the description of art-related activities of children in Terezin such as music, plays, paintings, writings and poetry.

narr *The relevant material should include discussions of such activities and how they influenced the survival and following life of the children. Any episodes where the interviewee demonstrates examples of such an art are highly relevant.*

Relevance assessment

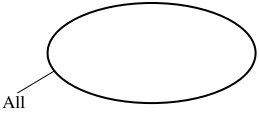
A manual process to acquire relevance judgements for document–topic pairs. Ideally for all document-topic pairs – infeasible.

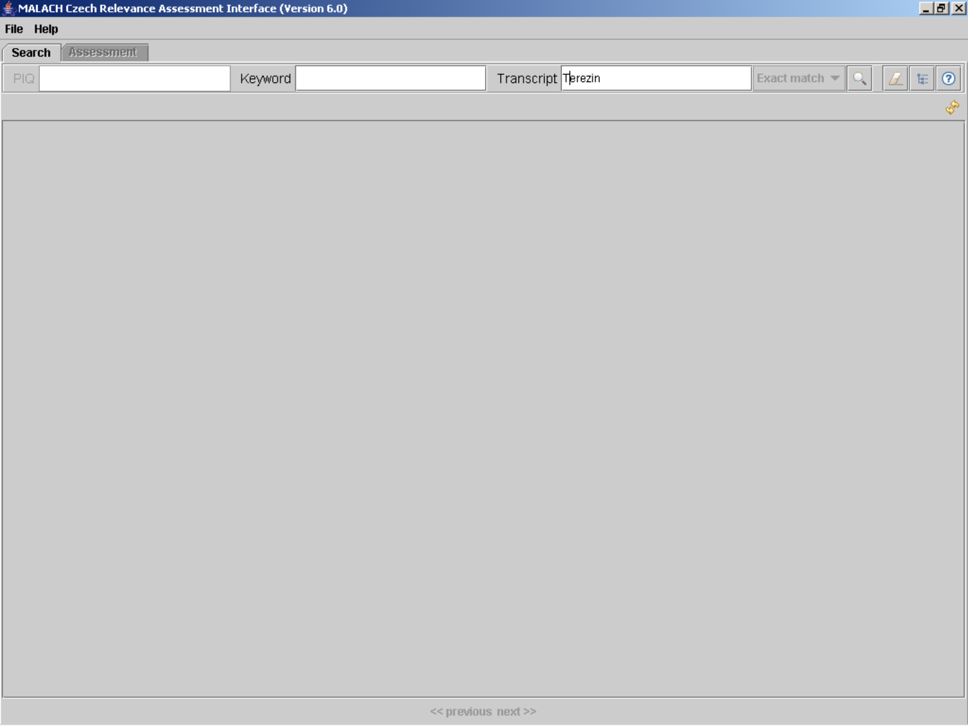
Search guided relevance assessment

- 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*

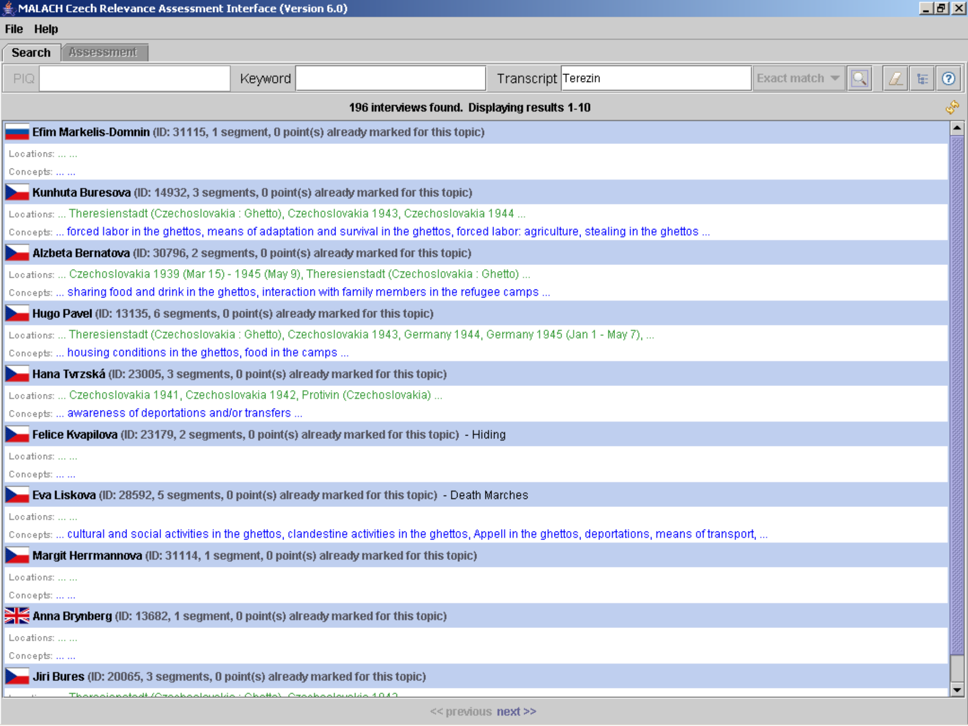
Highly ranked (pooled) relevance assessment

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





Transcript Terezin



MALACH Czech Relevance Assessment Interface (Version 6.0)

Transcript: Terezin

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: ... ...

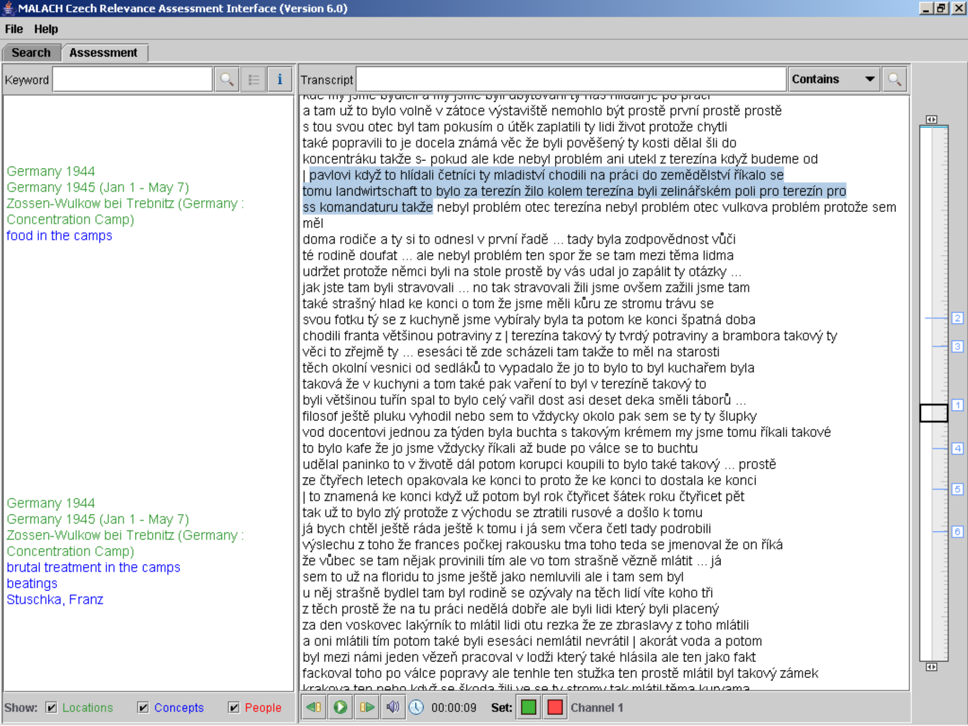
Anna Brynberg (ID: 13682, 1 segment, 0 point(s) already marked for this topic)

Locations: ... ...

Concepts: ... ...

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

... Theresienstadt (Czechoslovakia : Ghetto), Czechoslovakia 1942



MALACH Czech Relevance Assessment Interface (Version 6.0)

Germany 1944
Germany 1945 (Jan 1 - May 7)
Zossen-Wulkow bei Trebnitz (Germany : Concentration Camp)
food in the camps

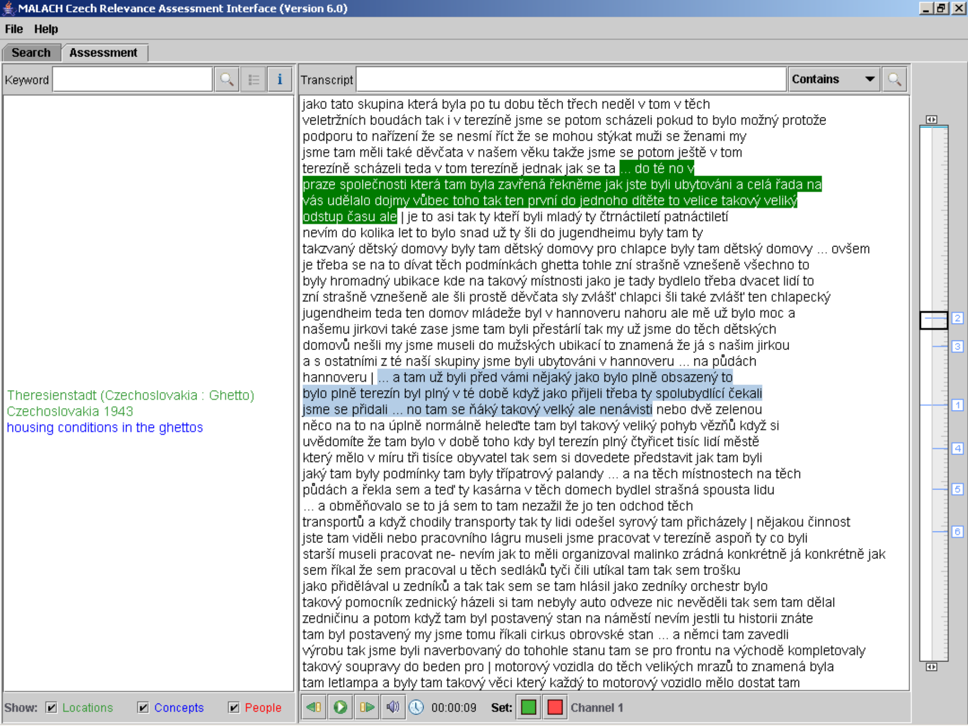
Germany 1944
Germany 1945 (Jan 1 - May 7)
Zossen-Wulkow bei Trebnitz (Germany : Concentration Camp)
brutal treatment in the camps
beatings
Stuschka, Franz

Show: ☑ Locations ☑ Concepts ☑ People

00:00:09

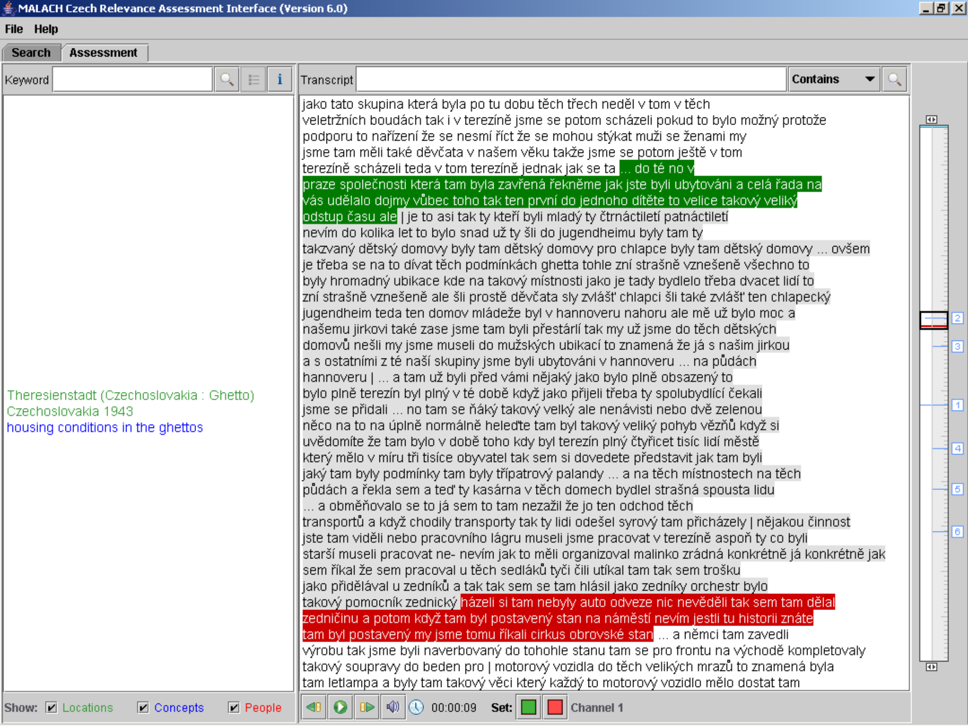
Set: 🟩 🟥 Channel 1

kde my jsme bydleli a my jsme byli ubytovaní 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 kurvama



Theresienstadt (Czechoslovakia : Ghetto)
Czechoslovakia 1943
housing conditions in the ghettos

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

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 naším 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

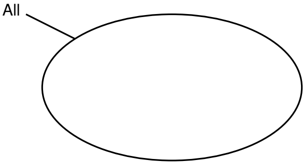
Show: ☑ Locations ☑ Concepts ☑ People

00:00:09 Set: 🟩 🟥 Channel 1

Evaluation measures: Precison and Recall

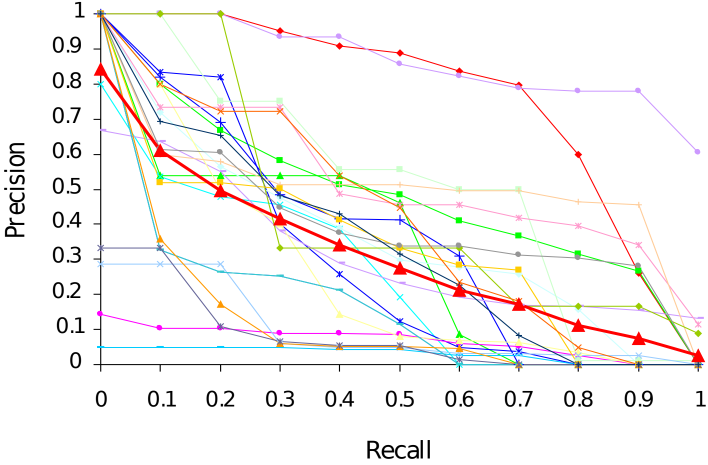
- A system ranks documents from the collection according to their relevance
- Precision/Recall calculated for a set of top n retrieved documents

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



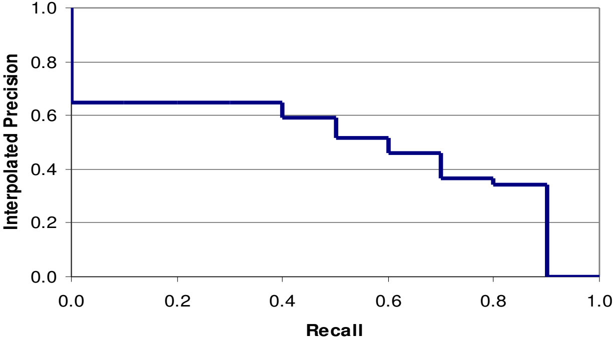
Precision-Recall curves

- Plot of Precision vs. Recall by varying the number of retrieved documents
- One curve per query

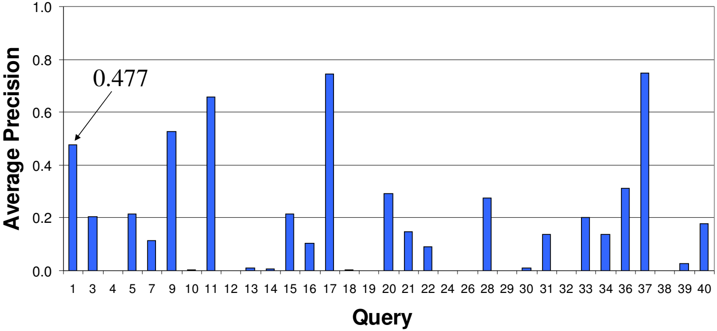


Average Precision

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



Mean Average Precision



CLEF 2007 evaluation results: English collection

Site-Run	MAP	Lang	Query	Document Fields
Dublin-1	0.2847	EN	TDN	MK,SUM
Ottawa-1	0.2761	EN	TD	MK,SUM
Brown-1	0.2577	EN	TDN	MK,SUM
Dublin-2	0.2459	EN	TD	MK,SUM,ASR06B
Brown-2	0.2366	EN	TD	MK,SUM
Brown-3	0.2348	EN	T	MK,SUM
Amsterdam-1	0.2088	EN	TD	MK,SUM,ASR06B
Dublin-3	0.1980	FR	TD	MK,SUM,ASR06B
Amsterdam-2	0.1408	NL	TD	MK,SUM,AK2,ASR06B
Ottawa-2	0.0855	EN	TD	AK1,AK2,ASR04
Ottawa-3	0.0841	EN	TD	AK1,AK2,ASR04
Brown-4	0.0831	EN	TDN	AK1,AK2,ASR06B
Dublin-4	0.0787	EN	TD	AK1,AK2,ASR06B
Brown-5	0.0785	EN	TD	AK1,AK2,ASR06B
Sinai-1	0.0737	ES	TD	ALL
Dublin-5	0.0636	FR	TD	AK1,AK2,ASR06B
Ottawa-4	0.0619	ES	TD	AK1,AK2,ASR04

CLEF 2007 evaluation results: Czech collection

Site/Run	mGAP	Query	Topic fields	Term normalization
Pilsen-1	0.0274	Auto	TDN	lemma
Pilsen-3	0.0241	Auto	TDN	lemma
Pilsen-2	0.0229	Auto	TD	stem
Chicago-2	0.0213	Auto	TD	aggressive stem
Chicago-1	0.0196	Auto	TD	light stem
Prague-4	0.0195	Auto	TD	lemma
Prague-1	0.0192	Auto	TD	lemma
Prague-2	0.0183	Manual	TD	lemma
Pilsen-4	0.0134	Auto	TD	lemma
Pilsen-5	0.0132	Auto	TD	none
Chicago-3	0.0126	Auto	TD	none
Brown-1	0.0113	Auto	TD	light stem
Brown-2	0.0106	Auto	TD	aggressive stem
Prague-3	0.0098	Manual	TD	none
Brown-3	0.0049	Auto	TD	none

That's all folks!

What 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,”