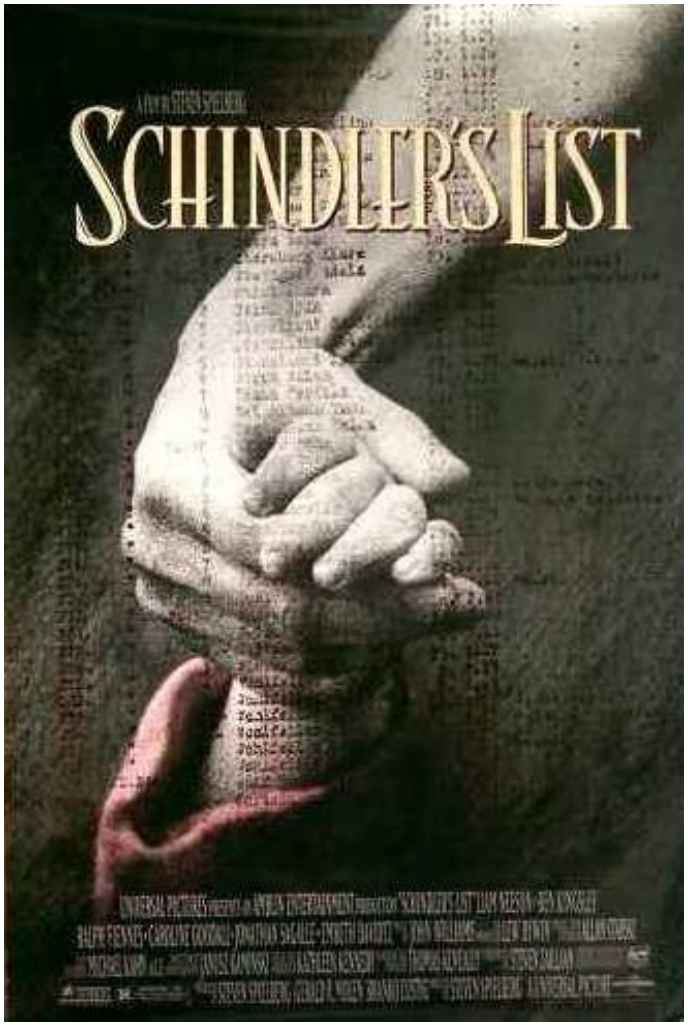
Malach: zpracování audiovizuálního archívu svědectví přeživších holocaustu

Pavel Pecina
pecina@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky, MFF UK

NMI 2015, Praha

Vše začalo v roce 1993 ...filmem ...a vizí



Vize Stevena Spielberga:

1. shromáždit a zachovat výpovědi svědků a přeživších holokaustu
2. katalogizovat tyto výpovědi a zpřístupnit je veřejnosti
3. šířit jejich obsah za účelem vzdělávání a boje proti intoleranci
4. umožnit a zjednodušit získávání podobných záznamů v budoucnu

Stručná historie archívu a jeho zpřístupnění

1993 Stephen Spielberg uvádí film Schindlerův seznam,
během natáčení se setkává s lidmi, kteří přežili holokaust, zpracovává jejich příběhy.

1994 Založena nadace Survivors of the Shoah Visual History Foundation (VHF)
s cílem zaznamenat a zpřístupnit svědectví lidí, kteří přežili holokaust.

1999 Během 5 let VHF vytvořila největší archív svého druhu na světě
obsahující 52 000 výpovědí svědků holokaustu z 57 zemí.

2000 10 % nahrávek manuálně katalogizováno za cenu 8 mil. USD,
zpracování jedné výpovědi trvá průměrně 35 hodin (indexace, sumarizace, kontrola).

2001 NSF financuje projekt Malach na automatické zpracování celého archívu,
řešitelé: University of Maryland, Johns Hopkins University, IBM; rozpočet 7,5 mil. USD.

2002 Zřízeny první přístupové body k celému archívu,
využívají rychlé počítačové sítě Internet 2 a velké mezipaměti.

2006 Z VHF se stává USC Shoah Foundation, Inst. for Visual History & Education
s obecnější misí: překonávat předsudky, netoleranci, fanatismus a utrpení, které působí.

Stručná historie archívu a jeho zpřístupnění (pokrač.)

2008 Počet přístupových míst se zvýšil na 21 po celém světě.

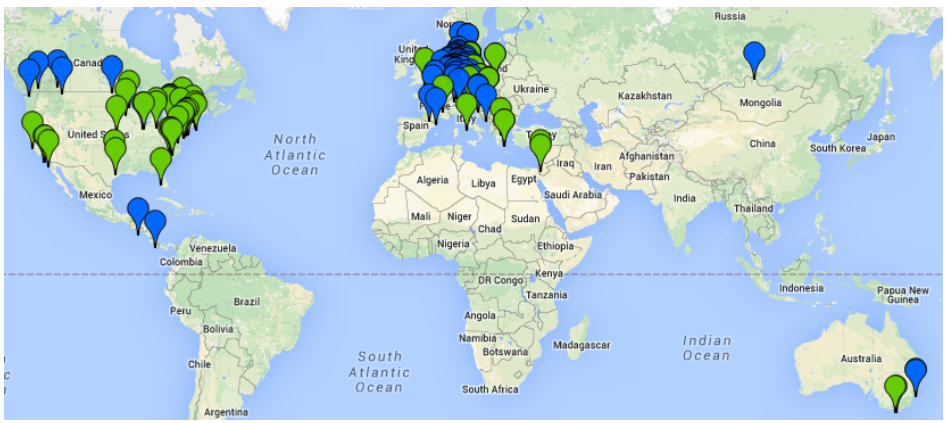
2009 Na Youtube spuštěn kanál USC Shoah Foundation.

2010 Otevřeno přístupové místo v Praze, Centrum Malach při MFF UK.

2012 Spuštěna služba VHA Online s výběrem 1,000 výpovědí přístupných on-line.

2012 MŠMT financuje projekt AMalach s cílem dál vylepšit zpřístupnění archívu.

2015 Počet přístupových míst se zvýšil na 51 (celkem ve 13 zemích).



Archív vizuální historie

Archív vizuální historie (AVH)

- spravován Shoah Foundation (nyní součást USC)
- vytvářen během let 1994–1999 za pomoci 2 300 tazatelů a 1 000 kameramanů
- obsahuje výpovědi 52 000 svědků holokaustu z 57 zemí ve 32 jazycích
- celkem 116 000 hodin VHS záznamů, 135 TB zdigitalizovaných dat
- průměrná délka výpovědi 2 h 15 min, náklady na její pořízení 2 000 USD
- výpovědi ručně indexovány pomocí tezauru o 60 000 klíčových slovech
- 3 000 výpovědí katalogizováno podrobněji (72 mil. slov)

- 573 rozhovorů proběhlo v ČR (většina v ČJ) za pomoci 38 tazatelů
- 4 500 svědectví podali lidé narozeni v ČR

Nahrávky výpovědí

- Neupravované, poskytují původní informace.
- Pokrývají život před válkou, během války i po ní, život v rodné zemi přeživších i v zemi, kam případně emigrovali.
- Zobrazují fotografie, dokumenty i jiné předměty, které se vztahují k příběhům přeživších.
- Obsahuje i pasáže z exteriéru, např. míst někdejších koncentračních táborů, ghett, masových hrobů.

- Hlavní skupiny přeživších:

  židovští přeživší (4 8848/542), *zachránci a poskytovatelé pomoci* (1 132/6), *přeživší Romové a Sintové* (407/3), *osvoboditelé a svědci osvobození* (362/1), *političtí vězňové* (261/7), *přeživší Svědkové Jehovovi* (83/0), *účastníci soudních procesů s válečnými zločinci* (62/1), *přeživší programů eugeniky* (13/0), *homosexuální přeživší* (6/0).

Podrobná (full–description) katalogizace a anotace

Na úrovni celých interview

- dotazník vyplněný před interview
- jména lidí a míst zmíněná v průběhu interview
- volný text sumarizující celé interview

Na úrovni kratších pasáží

- hranice tématických pasáží (průměrná délka 3 min
- popis obsahu: *souhrn + scratchpad*
- položky z tezauru: *jména, témata, místa, časová období*

Interview time	*Location–Time*	*Concepts*	*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*

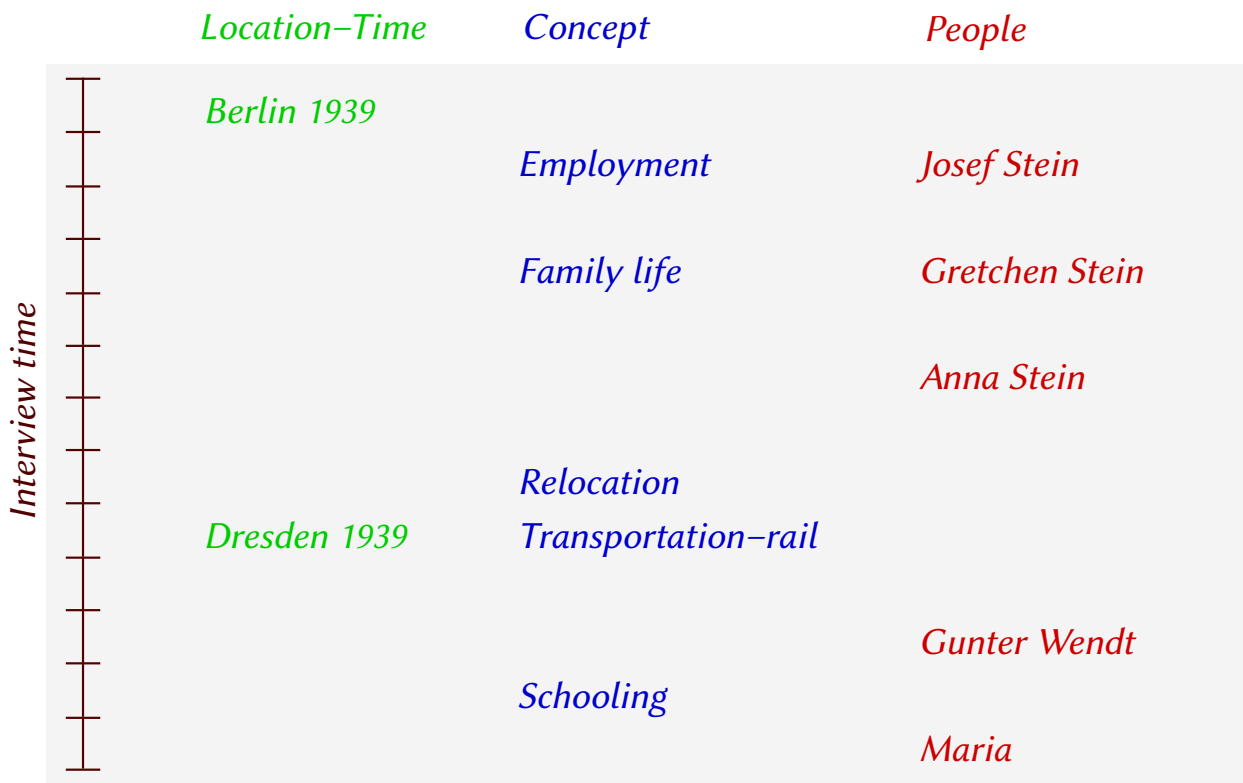
Zběžná (real–time) katalogizace a anotace

Na úrovni celých interview

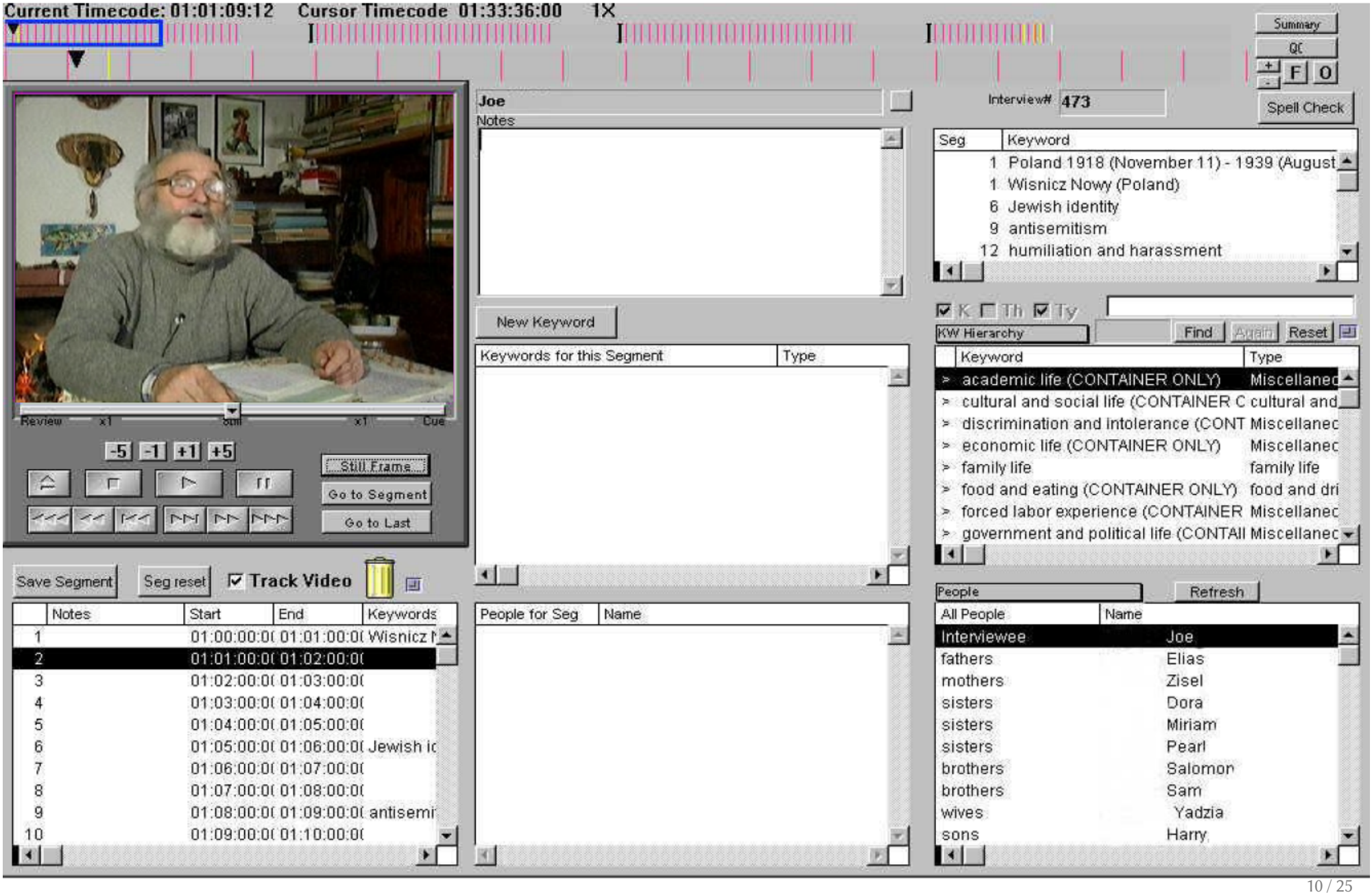
- dotazník vyplněný před interview

Průběžné anotace

- položky z tezauru přiřazené časovým okamžikům:
  jména, témata, místa, časová období

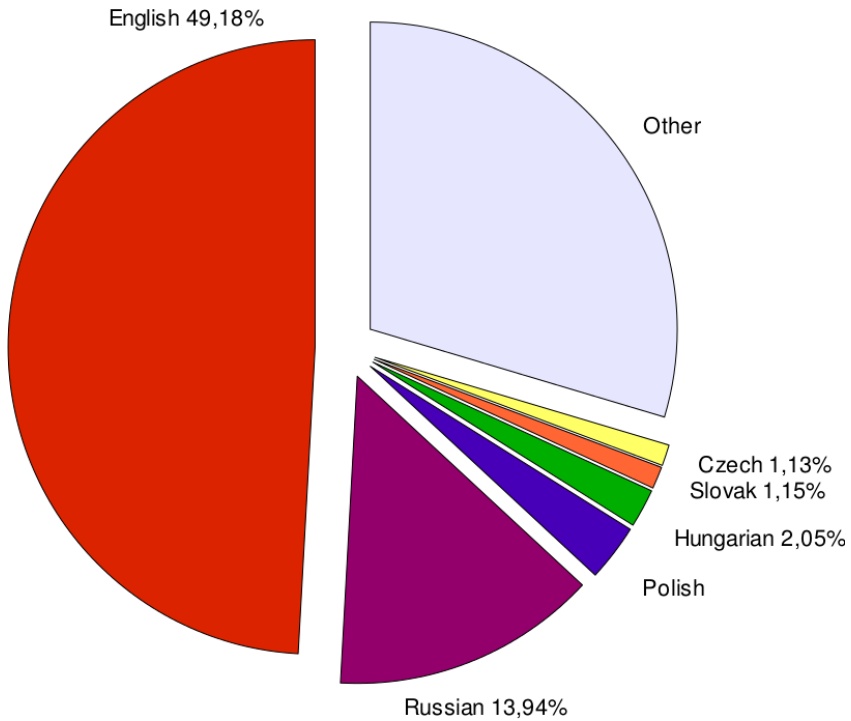


Katalogizační software



Jazyky a země výpovědí (20 nejčastějších)

počty výpovědí/jazyk	
anglicky	24 872
rusky	7 052
hebrejsky	6 126
francouzsky	1 875
polsky	1 549
španělsky	1 352
holandsky	1 077
maďarsky	1 038
německy	686
bulharsky	645
slovensky	583
česky	573
portugalsky	562
jidiš	527
italsky	433
srbsky	382
chorvatsky	353
ukrajinsky	320
řecky	301
švédsky	266



počty výpovědí/země	
Izrael	8 449
Ukrajina	3 427
Kanada	2 815
Austrálie	2 475
Francie	1 650
Polsko	1 371
Holandsko	1 044
Maďarsko	786
Argentina	726
Rusko	674
Německo	668
Slovensko	656
Bulharsko	628
Brazílie	564
Itálie	417
Chorvatsko	327
Švédsko	325
Řecko	303
Moldávie	284
Bělorusko	246

Projekt Malach

Projekt Malach

Multilingual Access to Large Spoken Archives

- projekt National Science Foundation, USA, 2001–2006
- spolufinancován Min. školství, mládeže a tělovýchovy, ČR, 2005–2006

Cíl:

- zjednodušení přístupu k archívu
- snížení nákladů na katalogizaci nahrávek

Úkoly:

1. automatické rozpoznávání spontánní řeči
   - *doslovný přepis všech rozhovorů*
2. strojový překlad doménově specifického tezauru
   - *tazaurus vytvořen přímo pro doménu výpovědí během manualní katalogizace*
3. automatická detekce témat a přiřazování metadat
   - *segmentace na tématické pasáže a přiřazování klíčových slov*
4. systém pro vícejazyčné vyhledávání informací a prohledávání archívu
   - *založené na (nedokonalých) výsledcích předchozích úloh*

Řešitelský tým projektu Malach

IBM T.J. Watson Center, New York, USA
- rozpoznávání mluvené řeči v angličtině

Johns Hopkins University (CLSP), Baltimore, USA
- rozpoznávání mluvené řeči v ostatních jazycích

University of Maryland, College Park, USA
- vyhledávání informací, prohledávání archívu, vytvoření testovací kolekce

Západočeská Univerzita (KKY, FAV), Plzeň, ČR
- rozpoznávání mluvené řeči v češtině a dalších jazycích

Univerzita Karlova v Praze (ÚFAL, MFF), ČR
- jazykové modelování, vyhledávání v mluvené řeči, testování

Rozpoznávání řeči

- doslovný přepis spontánní řeči (nezávislý na řečníkovi)
- záznamy technicky poměrně kvalitní (ale s šumy a ruchy apod.)
- řešení úlohy stěžuje jazyková kvalita (emoce, stáří, zdravotní stav, jazyk)
- specifickým problémem v češtině jsou hovorové výrazy a výslovnost

odjet	[o d j e t]	Osvětim	[o s v j e t i m]
	[v o d j e t]		[o s v e t i m]
	[o d j e c]		[v o s v j e t i m]
	[o d j e c t]		[o s v j e n č i m]
	[v o d j e c t]		[v o s v j e n č i m]
	[v o d e j e c t]		[o z v j e t i m]

- výsledky měřeny na vzorku ručně přepsaných záznamů
  – *jako poměr chybně rozpoznaných slov (WER)*

jazyk	WER (%)
angličtina	25.00
čeština	35.51
slovenština	34.49
ruština	45.75

Rozpoznávání řeči - ukázka

jméno:	???? ????Hugo Pavel
narození:	26.12. 1924
země původu:	Československo
vyznání (pre):	judaismus
vyznání (post):	N/A
klíčová slova:	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 ...

Vyhledávání v nahrávkách

Vyhledávání v mluvené řeči

- Zvláštní případ vyhledávání informací, kde informace jsou v mluvené formě.

Úlohy rozpoznávání a vyhledávání jsou odděleny

- Systém pro vyhledávání je postaven na výstupu rozpoznávače řeči.

Vyhledávání je do jisté míry odolné vůči chybám rozpoznávání

- Tolerovatelná míra chybovosti < 40% (nesprávně rozpoznaných slov)

Chyby rozpoznávání nemusí vadit systému, ale vadí uživatelům

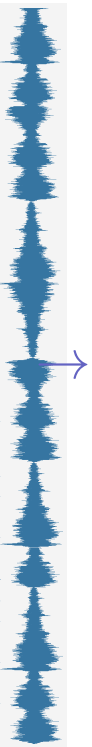
- Systém musí odkazovat na pasáže v původních nahrávkách a nikoliv na jejich autoamtické přepisy.

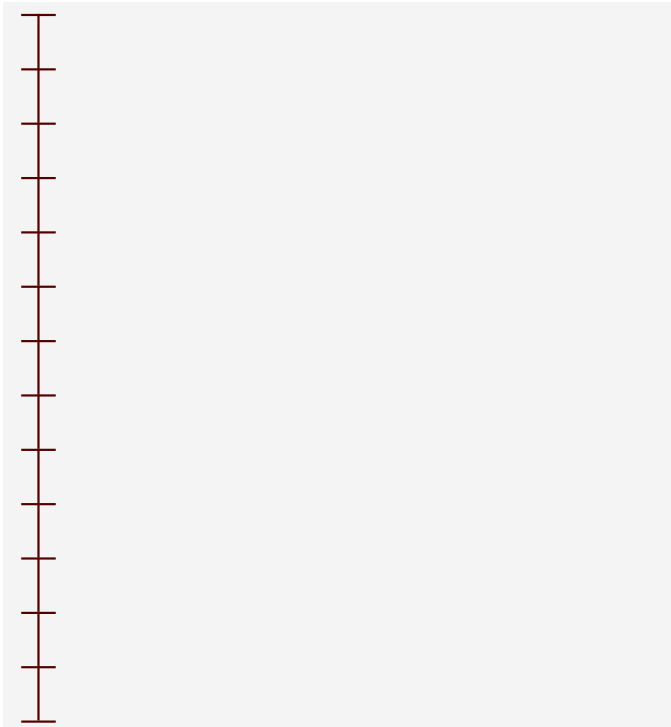
Segmentace na tématické pasáže je přínosná

- Zlepšuje kvalitu vyhledávání i spokojenost uživatele

Zpracování nahrávek

automatický přepis řeči → segmentace segmentace anotace → reprezentace segmentů





Reprezentace segmentů

Segment z anglického interview s podrobnou anotací

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 ...*

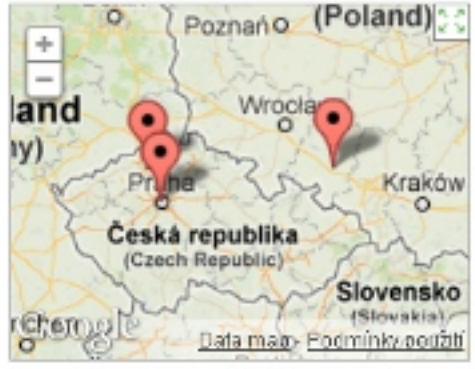
USC Shoah Foundation
The Institute for Visual History and Education

VIEWING SCREEN

Jan Stern

Video Tape 3 of 5
Slide Show

Show Map



To report a video problem click here

Segments
Biographical Profile
Indexing Terms in Testimony
People in Testimony
Search in Testimony
Donor Recognition

Segment#: 74 +

Segment#: 75 +

Segment#: 76 +
"Mischlinge"
Nuremberg Laws (September 15, 1935)

Segment#: 77 +

Maximize/Minimize Data
Next Result
Previous Result
Back to Search Results
New Search
Save to Projects
Print Testimony

Projekt Amalach

AMalach

ASR and **M**T-based **A**ccess to a **L**arge **A**rchive of **C**ultural **H**eritage

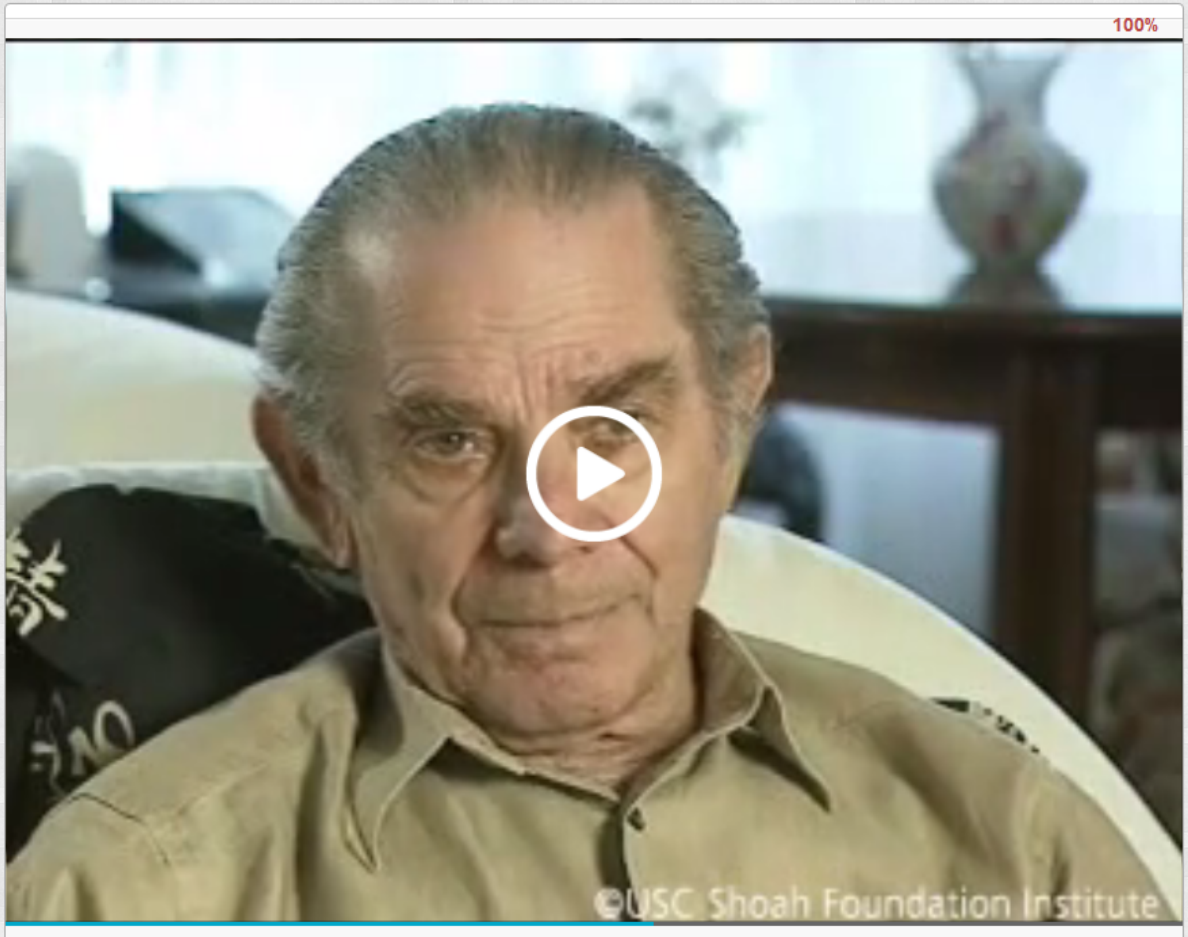
- následník projektu Malach
- projekt Univerzity Karlovy v Praze a Západočeské univerzity v Plzni
- financování Ministerstvem kultury ČR, 2012-2015

Cíle:

1. Vylepšit úspěšnosti rozpoznávání řeči v českých nahrávkách
   - *chybovost klesla z 28% na 22%*
2. Umožnit vyhledávání v „napříč“ jazyky s pomocí strojového překladu
   - *cross-lingualní vyhledávání dostupné pro CS↔EN*
   - *např. anglické dotazy lze aplikovat na česká data*
3. Vytvořit systém pro fonetické vyhledávání
   - *systém umožňuje „fultextové“ vyhledávání na úrovni fonémů, nikoliv slov*
   - *lze tedy vyhledávat slova, která nejsou ve slovníku*

AMALACH
Access to a Large Archive of Cultural Heritage

charles university | Search



00:16:31 - 00:16:42

Speakers: Bernard Firestone

Charles University	university … Charles University	Charles … university
skore 100%	skore 99%	skore 52%
od 00:16:31 od 00:16:42	od 00:04:25 od 00:04:39	od 00:14:04 od 00:14:20

DEPARTMENT OF CYBERNETICS

ÚFAL

Zpřístupnění rozsáhlého video archivu kulturního dědictví pomocí metod automatického rozpoznávání mluvené řeči a strojového překladu.
(DF12P01OVV022)
Ministerstvo kultury - Program aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI) (2011-2017)

Speakers

- [x] Bernard Firestone
- [x] William Ganz
- [x] Mitchell Gordon

Select all

Select none

Několik citací na závěr

Doug Greenberg (VHF):

- "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 (UMD):

- "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."

Odkazy

- USC Shoah Foundation
  `http://sfi.usc.edu/`
- Kanál Youtube
  `https://www.youtube.com/user/USCShoahFoundation`
- VHA Online
  `http://sfi.usc.edu/watch`
- Projekt Malach
  `http://malach.umiacs.umd.edu/`
- Projekt AMalach
  `http://ufal.mff.cuni.cz/grants/amalach/`
- Centrum vizuální historie Malach
  `http://malach-centrum.cz/`