Počítač a Holocaust

Počítač a Holocaust

S neúprosným tokem času ubývá očitých svědků událostí, jež jsou pro historii důležité nejen jako fakta, ale především jako memento. To si také v průběhu přípravných prací na natáčení filmu Schindlerův seznam, mapujícího osudy Židů za druhé světové války, uvědomil světoznámý americký režisér Steven Spielberg a reagoval nedocenitelným počinem: založil nadaci SHOA VISUAL HISTORY FOUNDATION.

V současné době je nadace přetvořena v Shoa Foundation Institute for Visual History a působí při University of South California. Jejím cílem bylo shromáždit svědectví přeživších, osvoboditelů a svědků nacistického holocaustu.

Výsledkem je velký mnohojazyčný soubor audiovizuálních nahrávek, obsahující 116 000 hodin digitalizovaných záznamů rozhovorů ve dvaatřiceti jazycích s 52 000 respondenty. Pro zajímavost uveďme, že nejvíce nahrávek je z USA – 19 841 – a z Izraele – 8504; Česko přispělo 566 a Slovensko 656 dokumenty. Tato čísla ovšem nic neříkají o jazycích dokumentů.

Soubor je nedocenitelným zdrojem informací nejen pro historiky, ale i pro sociology, psychology, lingvisty, učitele všech stupňů škol, národopisce, samozřejmě filmaře (z archívu již byla použita řada svědectví pro další dokumentární filmy, například i českého režiséra Vojtěcha Jasného) atd. Dají se tu i po letech najít informace a souvislosti, které třeba unikly i při soudních procesech v minulosti.

INFORMAČNÍ LABYRINT

Orientovat se v rozsáhlém souboru vzpomínek svědků nacistického vyvražďování Židů je velmi obtížné a důležitou roli tu dnes hraje komputační lingvistika s počítačovým modelem „porozumění“ (understanding) psaným i mluveným projevům. Problém porozumění přirozenému jazyku má mnoho vrstev, a to jak co do složitosti, tak co do rozsahu: od vyhledávání netriviální informace v textech až po úplné pochopení daného sdělení se schopností vyvozování důsledků. Bez možnosti počítačového vyhledávání by nebylo v lidských silách tento archív zpracovat a následně v souboru vyhledávat relevantní informace. Původní odhad nadace Shoa byl, že by to stálo víc než 150 miliónů dolarů.

Jako příklad systému integrujícího schopnost „porozumění“ mluvené řeči a schopnost vyhledat požadovanou informaci v rozsáhlých vícejazyčných dokumentech můžeme uvést projekt MALACH (akronym pro úplný název Multilingual Access to Large Spoken ArCHives), jehož cílem je umožnit přístup právě k dokumentům shromážděným Spielbergovou nadací. Nejde o úlohu snadnou: různí mluvčí odkazují často k téže situaci, ale z jiného úhlu pohledu, v různých jazycích i souvislostech atd. Jde o data výjimečně bohatá co do jejich charakteru: ne vždy mluvčí užívali svůj mateřský jazyk (po osvobození z koncentračních táborů se mnozí nevrátili do rodné země, někteří z ní emigrovali v poválečném období, jejich znalost nově získaného jazyka ovšem stále nesla stopy po jazyku mateřském). Jindy naopak mluvili jazykem mateřským, ale již s většími nebo menšími vlivy jazyka země, v níž žijí. Jejich promluvy byly pochopitelně velmi emotivní, prokládali je i německými slovy, která byla v táborech běžná. Videonahrávky umožňují sledovat i vztah mezi gesty a mimikou obličeje a promluvou atd.

CO SE LZE DOZVĚDĚT

Výsledky projektu mají vést k získání odpovědí na dotazy jako Co se dělo v táboře Treblinka v září 1943? nebo Kterými tábory prošla paní X. Y.? Odpovědi budou někdy širší, než by dotaz vyžadoval (např. odpověď na první otázku může zahrnovat delší období než jen měsíc září), nebo neúplné (např. výčet táborů při druhé otázce). Podkladem pro odpověď samozřejmě budou dokumenty získané od různých mluvčích, s různým kontextem, ale vždy s nějakou informací relevantní pro daný dotaz. Předpokládá se, že dotazy, stejně jako sdělení v souboru obsažená, mohou být formulovány ve kterémkoli z dvaatřiceti jazyků.

Projekt zahrnuje automatické rozpoznávání mluvené řeči i počítačem podporovaný oborově specifický překlad na základě mnohojazyčného zásobníku (multilingválního tezauru). Vytváří dosud nevídané možnosti pro výzkum daného období pro historiky, ale přináší i bohatý a jinak nedostupný materiál pro výzkum účinné katalogizace promluv i obecně pro vyhledávání a využití informace atd. Projekt je velmi dobrým příkladem spojení vědeckého výzkumu s aplikační oblastí.

PODÍLNÍCI PROJEKTU

Projektu se účastní několik výzkumných pracovišť: vedle zmíněného institutu je to IBM Thomas J. Watson Research Center v Yorktown Heights, N. Y., dvě americké univerzity (Johns Hopkins University v Baltimoru a University of Maryland) i Ústav formální a aplikované lingvistky na **MFF Univerzity Karlovy** v Praze a katedra kybernetiky fakulty aplikovaných věd Západočeské univerzity v Plzni. Je třeba zdůraznit, že obě česká pracoviště se zúčastnila

výběrového procesu vypsaného americkou grantovou agenturou NSF. Český řešitelský tým má přitom na starost češtinu, slovenštinu, ruštinu, polštinu a maďarštinu.