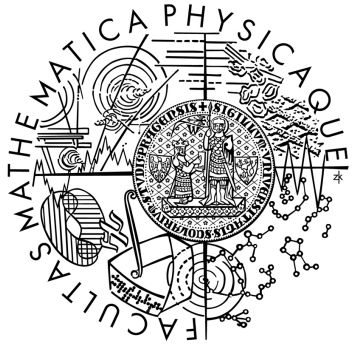
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

**DIPLOMOVÁ PRÁCA**




Petra Galuščáková

# Evaluační metody systémů pro vyhledávání v nesegmentované mluvené řeči

Ústav formální a aplikované lingvistiky

Vedúci diplomovej práce: RNDr. Pavel Pecina, Ph.D.
Študijný program: Informatika
Študijný odbor: Matematická lingvistika

Praha 2011

Rada by som poďakovala RNDr. Pavlovi Pecinovi Ph.D. za vedenie práce, cenné rady a konzultácie, svojej rodine za podporu a všetkým respondentom, ktorí sa zúčastnili prieskumu za čas, ktorý tomuto prieskumu venovali.

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Zb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova v Praze má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe dňa 15. 4. 2011 Petra Galuščáková

Názov práce: Evaluační metody systémů pro vyhledávání v nesegmentované mluvené řeči

Autor: Bc. Petra Galuščáková

Katedra: Ústav formální a aplikované lingvistiky

Vedúci diplomovej práce: RNDr. Pavel Pecina, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Práca popisuje v súčasnosti používané spôsoby evaluácie vyhľadávania v hovorenej reči. Vysvetlené sú rôzne prístupy, ktoré slúžia na vyhľadávanie v hovorenej reči, ako aj spôsoby, ktoré slúžia na evaluáciu tohoto vyhľadávania. Práca sa pritom zameriava na vyhľadávanie v nahrávkach, ktoré nie sú segmentované na kratšie úseky. Cieľom práce je overiť, či sú používané spôsoby vyhľadávania adekvátne a prípadne vylepšiť tieto spôsoby evaluácie. V práci sú použité empirické prístupy založené na tom, ako užívatelia vyhľadávanie v hovorenej reči vnímajú a ako pracujú so systémami určenými na toto vyhľadávanie. Upravené spôsoby evaluácie sú nakoniec porovnané s pôvodnými technikami.

Kľúčové slová: hovorená reč, získavanie informácií v hovorenej reči, vyhodnocovanie

Title: Evaluation methods of systems for unsegmented speech retrieval

Author: Bc. Petra Galuščáková

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Pavel Pecina, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Methods that are currently used for evaluation of speech retrieval are described in this work. Techniques that are used for speech retrieval are explained, as well as methods used for evaluation of this retrieval. Special attention is paid to processing of unsegmented records. The main aim of the work is to verify whether the methods currently used for evaluation of speech retrieval are appropriate to use and modify these methods if needed. Empirical methods based on the user‘s perception of speech retrieval is used for this verification. Modified metrics are compared with the original ones.

Keywords: speech, speech retrieval, evaluation

# Obsah

# Úvod

V dnešnej dobe máme k dispozícii obrovské množstvo dát v textovej, grafickej a audiovizuálnej podobe. Aby sme sa ale dostali k informáciám, ktoré práve potrebujeme, musíme vedieť so zdrojmi dát dobre pracovať, predovšetkým v nich musíme dobre a rýchlo vyhľadávať. Vedný odbor, ktorý sa zaoberá vyhľadávaním informácií sa nazýva *information retrieval*. Relatívne dobré výsledky sa dosahujú pri vyhľadávaní v textových dokumentoch, postupne sa zlepšuje aj kvalita vyhľadávania v obrazových a grafických prameňoch. V náväznosti na pokroky, ktoré sa dosiahli v oblasti rozpoznávania hovorenej reči, sa zlepšila aj kvalita vyhľadávania vo zvukových nahrávkach. S tým vznikla aj potreba presnejšieho merania tejto kvality.

Hlavným cieľom tejto práce je nájsť vhodný prístup, ktorý by sa mohol využiť pri vyhodnocovaní vyhľadávania v nesegmentovanej hovorenej reči. V súčasnosti sa v praxi na evaluáciu vyhľadávania v nesegmentovanej reči používa jediná metrika. Chýbajú však empirické výskumy, ktoré by potvrdili to, že je táto metrika naozaj vhodná na riešenie evaluácie. Preto chceme v tejto práci overiť vhodnosť jej použitia, prípadne túto metriku ďalej vylepšiť. Správne navrhnutá metrika, by tak mohla prispieť k zlepšeniu kvality samotného vyhľadávania.

Spôsob vyhodnocovania navrhnutý v tejto práci bude založený na empirickom posúdení správania respondentov v rôznych situáciách, ktoré počas automatického vyhľadávania nastávajú.

Úvodná časť práce je venovaná všeobecným otázkam *information retrieval* a špecifickým problémom, ktoré súvisia s využívaním *information retrieval* v hovorenej reči. Sú tu stručne charakterizované jednotlivé systémy, ktoré na *information retrieval* slúžia. V ďalšej časti sú uvedené postupy, ktoré sa najčastejšie na vyhodnocovanie *information retrieval* používajú a spôsoby vyhodnocovania špecializované na *information retrieval* v hovorenej reči. Tretia časť popisuje prípravu návrhu nového spôsobu vyhodnocovania. Priebeh a výsledky prieskumu, na základe ktorého bude nová metrika vytvorená, sú popísané v ďalšej časti práce. Konečný návrh metriky a jeho porovnanie s existujúcimi metrikami je uvedený v šiestej kapitole. V závere sa nachádza sumarizácia výsledkov, ktoré sa nám podarilo získať. Súčasťou práce sú aj prílohy. Priložené sú pokyny, ktoré boli určené respondentom pri prieskume, ukážka playlistu použitého pri prieskume, ukážka registračného formulára pre responentov, ukážka logu s prevedenými akciami respondentov a popis obsahu priloženého CD.

# 1. Vyhľadávanie v hovorenej reči

## 1.1 Information Retrieval

*Information Retrieval* (ďalej len IR) alebo automatické získavanie informácií je pomerne široký pojem, pre ktorý môžeme nájsť v literatúre viacero definícií. Podľa [18] sa IR zaoberá vyhľadávaním určitých materiálov (väčšinou dokumentov), ktoré majú neštrukturovanú podobu (teda sú to väčšinou texty) a ktoré napĺňajú nejakú potrebu získania informácií vo veľkej skupine dát (ktorá je väčšinou uložená na počítači) na základe používateľom zadanej požiadavky. Zadaná požiadavka tak musí formálne popisovať požadované informácie [7]. Výstupom IR systému by mala byť podmnožina dokumentov, prípadne úryvkov, ktoré zodpovedajú zadanej požiadavke. Výstupy bývajú najčastejšie zotriedené na základe ich relevancie [7].

Systémy slúžiace na IR nemajú podľa [16] za úlohu informovať o nejakej téme a odpovedať na zadané otázky, ale informujú o existencii dokumentov na túto tému, prípadne dávajú informáciu o tom, kde sa tieto dokumenty nachádzajú. Pracujú pritom s určitou množinou dokumentov, ktorú musia uložiť, analyzovať a následne v nich vyhľadať relevantné dokumenty na základe otázky zadanej používateľom.

Otázka môže byť položená napríklad vo forme izolovaných slov. V tom prípade je pre IR systém pomerne jednoduché zistiť, či je dokument relevantný alebo nie. IR systémy však môžu pracovať aj so zložitejšími otázkami, ako je napríklad viac kľúčových slov spojených logickými operátormi, alebo dokonca otázky položené v prirodzenom jazyku. Na základe týchto informácií môže potom IR systém zistiť, či text obsahuje požadovanú informáciu, a tak rozhodnúť, či je dokument pre zadanú otázku relevantný a prípadne na akej úrovni.

## 1.2 Existujúce IR systémy

Najvýznamnejšími IR systémami sú v súčasnosti webové fulltextové vyhľadávače ako napríklad Google[1] alebo Bing[2], medzi českými vyhľadávačmi napríklad Seznam[3]. Existuje však veľké množstvo ďalších, viac alebo menej známych IR systémov. Na rôzne systémy sú pritom kladené rôzne požiadavky. Odlišovať sa môžu napríklad typy otázok, ktoré môže používateľ klásť, rôzne môžu byť aj veľkosti a typy súborov dokumentov, s ktorými je systém schopný pracovať, relevantnosť výsledkov, ale aj rýchlosť, ktorú od systému vyžadujeme.

Ak je napríklad IR systém zameraný na vyhľadanie požadovaného dokumentu na pevnom disku alebo v e-mailovej schránke, sú naň kladené iné nároky ako na vyhľadávanie na internete. Takýto systém nemusí pracovať s takým veľkým

[1] http://www.google.com
[2] http://www.bing.com
[3] http://www.seznam.cz

množstvom súborov ako internetový vyhľadávač a nemusí pracovať tak rýchlo. Môžu byť však od neho požadované presnejšie výsledky. Pri vyhľadávaní na internete by sa malo počítať s tým, že zadaná otázka je často nepresná. V zadanej otázke sa navyše môžu nachádzať preklepy.

Pre potreby tejto práce sú však zaujímavé najmä experimentálne a akademické, prípadne open-source systémy, ktoré je možné upraviť tak, aby mohli pracovať s kolekciou dát, ktorú im zadáme. Môžu tak teda pracovať aj s automatickými prepismi hovorenej reči. Medzi takéto novšie systémy patria napríklad:

- **Lemur**[4]
  Tento systém sa v súčasnosti vyvíja v spolupráci University of Massachusetts a Carnegie Mellon University. Toolkit zahŕňa vyhľadávací systém ale aj nástroje na analýzu textu, lištu do prehliadača a dátové zdroje. Lemur je používaný pri výskume, ale aj v komerčných aplikáciách a je to open-source projekt.

- **Indri**[5]
  Indri je vyhľadávací engine, ktorý je súčasťou toolkitu Lemur, je to teda open-source software. Umožňuje vyhľadávanie v texte, pričom je možné vyhľadávať v až 500 miliónoch dokumentov. Podporuje pritom bohato štrukturované otázky.

- **Terrier**[6]
  Ďalší open source vyhľadávací engine sa nazýva Terrier. Tento software sa v súčasnosti vyvíja na University of Glasgow. Je použiteľný na veľkých kolekciách dokumentov a je ho možné použiť ako desktop alebo webovú aplikáciu. Terrier je určený najmä pre výskum a experimenty.

- **Wumpus**[7]
  Tento IR systém bol vyvinutý na University of Waterloo, posledná aktualizácia je pritom z roku 2007. Tento systém je vydávaný pod GPL licenciou. Hlavným cieľom tohto systému bolo preskúmať problémy, ktoré vznikajú v dynamických súboroch textov v prostredí s viacerými používateľmi. Bol napríklad využitý pri prehľadávaní dokumentov na počítači. Pri takomto vyhľadávaní sa kolekcia dokumentov, v ktorých sa vyhľadáva rýchlo mení, pričom počet požiadaviek je väčšinou malý. Okrem klasického použitia ako IR systému bolo teda možné Wumpus využiť aj ako systém, ktorý indexuje dokumenty na počítači a udržuje informácie o zmenách v nich.

- **Okapi**[8]
  Tento freeware program, ktorý vyvíjaný na London City University je určený najmä na výskum a experimenty v oblasti IR.

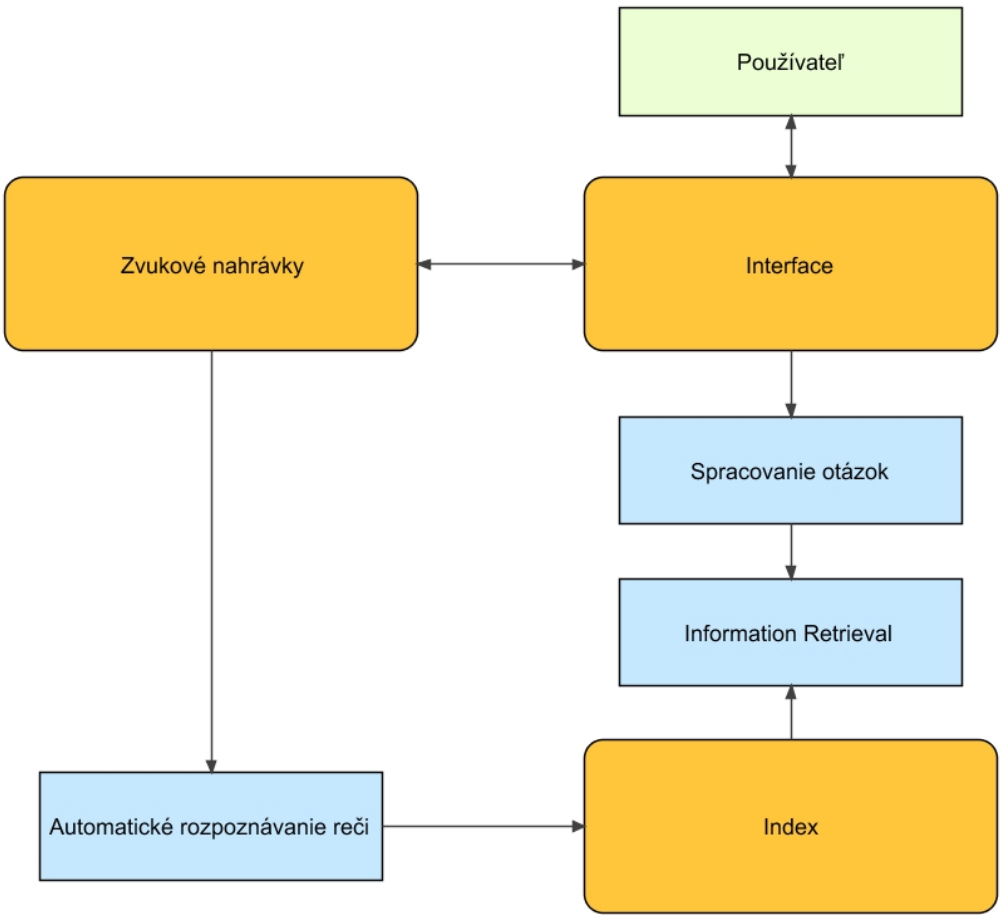
[4]http://www.lemurproject.org/
[5]http://www.lemurproject.org/indri/
[6]http://terrier.org/
[7]http://www.wumpus-search.org/
[8]http://www.soi.city.ac.uk/ andym/OKAPI-PACK/

Okrem týchto projektov samozrejme existuje veľké množstvo ďalších. Medzi inými to sú napríklad InQuery, Smart, Padre alebo Pircs.

## 1.3 Information Retrieval v hovorenej reči

Podľa [19] je cieľom IR v hovorenej reči zoradiť veľké množstvo dynamicky vygenerovaných informácií a sprístupniť tie informácie, ktoré spĺňajú zadané požiadavky. Je to teda špeciálny prípad IR, pri ktorom sú informácie v hovorenej podobe. Nezáleží pritom na spôsobe zadania požiadaviek, môžu byť zadané v podobe textu aj v podobe hovorenej reči.

V praxi sa tento problém štandardne prevádza na klasický IR problém v textoch [9]. Nahrávka sa pomocou techniky automatického rozpoznávania reči prevedie na text. Textová podoba potom umožní využiť na vyhľadávanie bežne používané IR systémy. Táto situácia je zobrazená na obrázku 1.1. Postup pri vytvorení IR systému v hovorenej reči je popísaný napríklad v [8], [9] a [11].




Obr. 1.1: *IR systém pre hovorenú reč [9].*

Pri vyhľadávaní sa navyše často postupuje tak, že sa jednotlivé nahrávky rozdelia na malé úseky, s ktorými sa potom pracuje rovnako ako s textovými dokumentami. Toto vyhľadávanie sa označuje ako vyhľadávanie v segmentovanej reči. Ak segmentácia nie je k dispozícii, ide o nesegmentované vyhľadávanie.

Presnosť systémov na automatické rozpoznávanie reči stále nie je dokonalá. Pre rozpoznávanie sú problematické najmä málo kvalitné nahrávky. Záleží tiež na doméne. Pre špecifické domény, ako napríklad lekárstvo a právo, sú dosahované výsledky rozpoznávania lepšie [5].

IR systémy môžu byť ovplyvnené kvalitou automatického rozpoznávania, ktoré používajú [27]. Podľa [12] je tento vplyv menší ak sa v nahrávke vyhľadáva pomocou izolovaných slov. Problematické sú ale kratšie nahrávky a iné spôsoby vyhľadávania ako pomocou izolovaných slov.

Ďalšou charakteristickou črtou, ktorú treba pri IR v hovorenej reči brať do úvahy je to, že hovorená reč využíva napríklad inú slovnú zásobu ako písaný text. Používajú sa napríklad rôzne výslovnostné varianty jedného slova (napr. všetko - šetko - šecko - šicko, ...) a hovorové slová (čau, zlepšovák, žúr). V hovorenej reči sa tiež nachádzajú iné objekty ako v písanom texte, napríklad pauzy alebo povzdychnutia. Zároveň môže byť niekedy ťažké rozpoznať, kde jedna veta končí a začína druhá. V reči sa naproti tomu nachádza viac informácií o hovorcovi. Vieme určiť napríklad jeho pohlavie alebo náladu [25]. Pri návrhu IR systému pre reč je treba tiež počítať s tým, že reč je spojitá a je typicky časovo náročnejšie v nej nájsť relevantný úsek ako v texte [2].

# 2. Metódy evaluácie vyhľadávania v hovorenej reči

Aby sme mohli porovnávať rôzne systémy na vyhľadávanie v hovorenej reči, potrebujeme ich najskôr evaluovať. Evaluácia je dôležitá aj pri navrhovaní nového systému, aby mal jeho tvorca prehľad, či nové pridané vlastnosti vyhľadávanie zlepšia alebo zhoršia. Vyhodnocovanie môže byť však veľmi individuálne. Tiež záleží na kontexte, v ktorom používateľ výsledky požaduje.

Podľa [18] je hlavným merítkom kvality systému spokojnosť používateľov. Najdôležitejším faktorom spokojnosti je pritom relevantnosť výsledkov. Dôležitú úlohu zohrávajú aj iné faktory, napríklad rýchlosť odozvy. Cleverdon v [6] udáva päť kritérií, efektívnosti IR systému, ktoré sú priamo spojené s prevedením systému:

1. interval medzi zadaním požiadavky a získaním odpovede, teda ako dlho musí používateľ čakať na odpoveď,
2. fyzická forma výstupu (prezentácia), teda spôsob akým používateľ zadá požiadavku a dostane odpoveď,
3. psychická alebo fyzická námaha, ktorá je vyžadovaná od používateľa,
4. schopnosť systému prezentovať všetky relevantné dokumenty (*recall*),
5. schopnosť systému odfiltrovať všetky nerelevantné dokumenty (*precision*).

Ďalšími kritériami efektívnosti systému, ktoré ale používatelia zvyčajne neberú do úvahy je politika, financovanie a ekonomická stránka systému.

Pri štandardnom spôsobe vyhodnocovania potrebujeme testovaciu sadu, ktorá podľa [18] pozostáva z troch vecí:

1. množina dokumentov,
2. popis informácií, ktoré chceme z dokumentov získať, tieto informácie sa musia dať sformulovať ako požiadavka na dokument,
3. množina správnych odpovedí, ktoré napríklad označili anotátori.

Pri vyhodnocovaní systémov môžeme použiť rôzne metriky ([18], [24]). Najčastejšie sa však pri vyhodnocovaní IR systémov používajú nasledujúce:

## 2.1 Precision a Recall

Pri použití týchto metrík sa všetky dokumenty, s ktorými sa pracuje, musia najprv označiť ako relevatné alebo nerelevatné. Systém na používateľom zadanú požiadavku vráti podmnožinu všetkých dokumentov, s ktorými pracuje. V ideálnom prípade by mal systém vrátiť všetky relevantné dokumenty a nemal by vrátiť žiadny z nerelevantných dokumentov. *Precision* potom určuje podiel relevantných dokumentov medzi tými dokumentami, ktoré systém vráti. *Recall* označuje, koľko dokumentov zo všetkých relevantných dokumentov systém vráti.

Nech teda IR systém vracia na zadanú požiadavku množinu dokumentov, ktorá je podmnožinou celej množiny dokumentov, ktorú ma systém k dispozícii, potom:

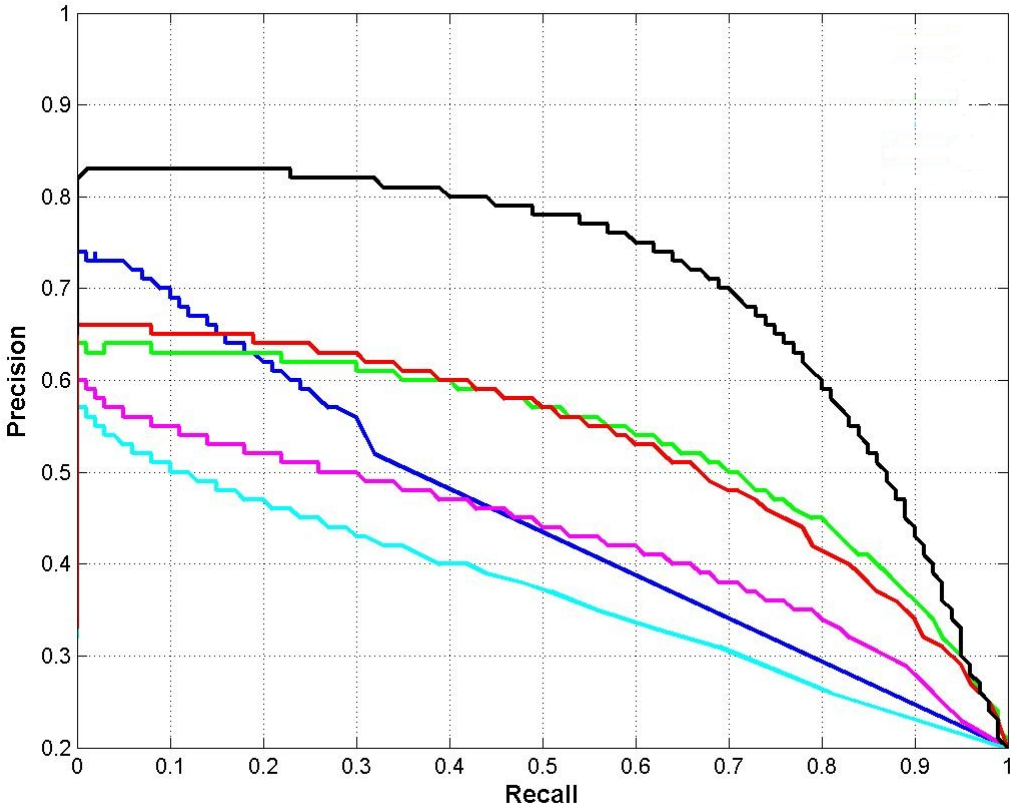
***precision*** = *počet relevantných vrátených dokumentov / počet všetkých vrátených dokumentov*
a
***recall*** = *počet relevantných vrátených dokumentov / počet všetkých relevantných dokumentov*

Vysokú hodnotu *precision* je však možné získať napríklad aj tak, že systém vráti jediný relevantný dokument, pričom existuje veľký počet relevantých dokumentov. To je ale nežiadúca situácia a hodnota *recall* bude v tomto prípade malá. Na druhej strane, vysokú hodnotu *recall* je možné získať napríklad aj tak, že IR systém predloží na výstup všetky dokumenty, ktoré má k dispozícii. Potom bude ale typicky málo z vrátených dokumentov relevantých, čo štandardne tiež nechceme. Hodnota *precision* bude v takomto prípade typicky veľmi malá. Z týchto dôvodov sa metriky *precision* a *recall* najčastejšie používajú spolu.

Tieto metriky sa často vyjadrujú v podobe závislosti hodnoty *precision* na hodnote *recall*. Táto závislosť sa vyjadruje pomocou tzv. *precsion-recall* kriviek. Príklad takejto krivky je na obrázku 2.1.

Ak sú navyše vrátené dokumenty usporiadané podľa relevantnosti, potom sa hodnoty *precision* a *recall* dajú použiť inými spôsobmi. Môžeme napríklad spočítať *precision* pre určitý počet dokumentov, ktoré IR systém označí ako najviac relevantné. Ak sú teda dokumenty pri vyhľadávaní zoradené podľa relevantnosti, stačí zobrať do úvahy prvých $n$ dokumentov a z nich spočítať *precision*. Prípadne môžeme hodnotu *precision* počítať po získaní nejakého vopred daného počtu relevantných dokumentov. Ďalšou možnosťou je počítať hodnotu *precision* pre určitú vopred zadanú hodnotu *recall* ([4], [18], [26]).

Nevýhodou *precision* a *recall* je to, že sú to dve metriky, je teda ťažké pomocou nich porovnávať viac systémov. Preto boli navrhnuté ďalšie metriky, ktoré tieto dve metriky kombinujú - *accuracy* a *F-measure*.




Obr. 2.1: *Ukážka precision recall kriviek, ktoré znázorňujú rôzne spôsoby rozpoznávania objektov na obrázkoch [1].*

## 2.2 Accuracy

Pomer všetkých správne klasifikovaných dokumentov (či už relevantných alebo nerelevantných) označujeme ako *accuracy* alebo presnosť.

Relevantné vrátené dokumenty označíme ako *true positive*, relevantné dokumenty, ktoré IR systém nevrátil ako *false negative*, nerelevantné dokumenty, ktoré napriek tomu IR systém vrátil ako *false positive* a nerelevantné dokumenty, ktoré vyhľadávač nevrátil ako *true negative*. Pojmy sú znázornené v tabuľke 2.1.

| | Predpovedaný negatívny | Predpovedný pozitívny |
|---|---|---|
| Negatívny prípad | True Negative | False Positive |
| Pozitívny prípad | False Negative | True Positive |

Tabuľka 2.1: *Rozdelenie relevantnosti dokumentov.*

*Precision* môžeme v tomto prípade vyjadriť aj ako:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

hodnotu *recall* môžeme vyjadriť ako:

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

a presnosť môžeme vyjadriť pomocou:

$$accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

Vo väčšine prípadov je počet nerelevantných dokumentov omnoho vyšší ako počet relevantných dokumentov a hodnota *true negative* je teda vysoká. Ak označíme všetky dokumenty ako nerelevantné, bude hodnota *accuracy* veľmi vysoká, čo nie je typicky vhodné. Z toho dôvodu je niekedy vhodnejšie používať metriku *F-measure*.

## 2.3 F-measure

*F-measure* je ďalšia metrika založená na hodnotách *precision* a *recall*. Je to harmonický vážený priemer hodnôt *precision* a *recall*:

$$F = \frac{1}{\alpha \cdot (\frac{1}{precision}) + (1-\alpha) \cdot (\frac{1}{recall})} = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot (precision + recall)}$$

kde

$$\beta = \frac{1-\alpha}{\alpha}$$

Výhodou tejto metriky je to, že môžeme dať väčší dôraz buď *precision* alebo *recall*, podľa toho, čo je pre nás v danom prípade dôležitejšie. Ak dáme obom metrikám rovnakú váhu, teda $\alpha = \frac{1}{2}$ (a $\beta = 1$), potom môžme *F-measure* vyjadriť tiež ako:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Ak je teda $\beta < 1$, potom je kladený väčší dôraz na *precision*, ak je $\beta > 1$, potom je kladený dôraz na *recall*. Vyšší dôraz na *precision* môžu klásť napríklad webové vyhľadávače, ktoré pracujú s veľkým počtom dokumentov. Naopak, vyšší dôraz na *recall* môže byť kladený pri IR systémoch vyhľadávajúcich napríklad dokumenty na pevnom disku, keď požadujeme, aby boli vyhľadané naozaj všetky relevantné dokumenty.

Voľný parameter $\beta$ môže byť ale aj nevýhodou tejto metriky. Ak sú pri vyhodnocovaní rôznych systémov použité rôzne parametre $\beta$, potom sa takéto systémy ťažko porovnávajú.

## 2.4 Mean Average Precision (MAP)

Táto metrika [4] je založená na hodnote *average precision*. *Average precision* pre jedinú zadanú otázku môžeme získať ako aritmetický priemer hodnôt *precision* pre množinu prvých (najviac relevantných) $m$ dokumentov. Táto hodnota sa počíta pre každý nový relevantný dokument ($d_m$), ktorý IR systém vráti. Ak je $R_k$ množina prvých $k$ výsledkov vrátených IR systémom pre zadanú otázku, potom môžeme hodnotu *average precision* vypočítať ako:

$$AP(d_m) = \frac{1}{m} \cdot \sum_{k=1}^{m} precision(R_k)$$

Hodnota MAP sa potom získa ako priemer hodnôt *average precision* pre množinu zadaných požiadaviek $Q$[18]:

$$MAP(Q) = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} AP(d_j) =$$

$$= \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} precision(R_{jk})$$

Ak sa nevyhľadá žiadny relevantný dokument, potom je hodnota rovná 0. Nevýhodou MAP je to, že vyžaduje pri testoch väčší počet rôznorodých testovacích dát. Je napríklad vhodná na simuláciu situácie, keď používateľ prezerá vyhľadané dokumenty a zastaví sa po tom, keď nájde požadovaný počet relevantných dokumentov [3].

## 2.5 mean Generalized Average Precision (mGAP)

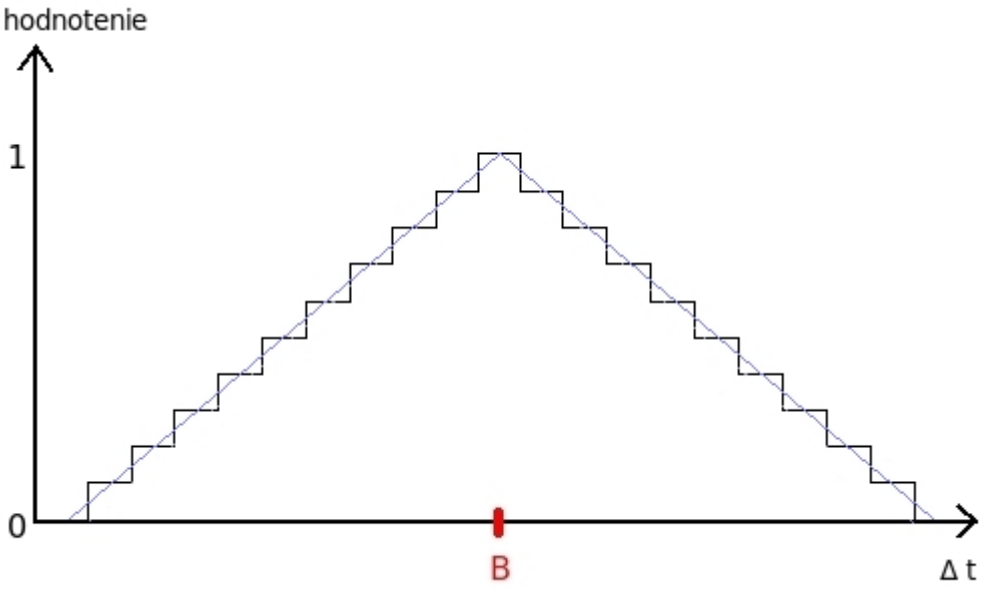
Predchádzajúce uvedené metriky boli zamerané všeobecne na IR a nebrali teda do úvahy špecifiká pri vyhľadávaní v hovorenej reči. Aby sme ich na vyhľadávanie v nahrávkach mohli použiť, musia sa najprv nahrávky nejakým spôsobom previesť na „dokumenty“. Takéto dokumenty potom sú alebo nie sú relevantné. Používateľ sa ale musí ku začiatku úseku, ktorý ho zaujímal dostať sám, čo je zvlášť náročné u dlhých nahrávok. Navyše, nahrávky sú často tématicky nekoherentné a obsahujú veľa rôznych tém. Pre používateľa je teda zaujímavejšie, ak systém vyhľadá priamo začiatok relevantného úseku. Preto v prípade dlhých nahrávok sa často nahrávky rozdeľujú na malé, prekrývajúce sa úseky. Tieto úseky je potom možné považovať za jednotlivé dokumenty, na vyhľadanie ktorých môžeme opäť aplikovať predchádzajúce metriky. Naproti tomu, mGAP je možné použiť pri vyhodnocovaní vyhľadávania začiatku relevantného úseku v nahrávkach a nie je potrebné nahrávky takýmto spôsobom segmentovať.

Podľa [17] by mal ideálny vyhľadávací systém pre hovorenú reč identifikovať začiatky tém čo najbližšie začiatkov, ktoré označili anotátori. Navyše by mali

byť správne vyhľadané body vysoko ohodnotené. Dobrá vyhodnocovacia metrika musí teda nielen reflektovať časovú blízkosť nájdeného začiatku ale ho aj správne ohodnotiť.

Výsledkom vyhľadávania je množina bodov zoradených podľa relevantnosti. mGAP berie do úvahy relevanciu vyhľadaných bodov a navyše používa penalizačnú funkciu, pomocou ktorej hodnotí presnosť odpovedí systému [17]. Penalizačná funkcia navrhnutá v [17] má tvar rovnoramenného trojuholníka. Správnejšie by sa mala táto funkcia nazývať ohodnocovacia funkcia, budeme však používať názvoslovie použité v [17], kde sa funkcia nazýva penalizačná.

Na obrázku 2.2 je znázornená implementácia tejto penalizačnej funkcie použitá v [11]. Anotovaný začiatok relevantného úseku je na tomto obrázku vyznačený bodom B. Zvuková nahrávka bola segmentovaná na úseky dlhé 15 sekúnd. Každých 15 sekúnd klesne hodnota penalizačnej funkcie o hodnotu 0,1.




Obr. 2.2: *Penalizačná funkcia použitá v [11]. $\Delta t$ vyjadruje časovú vzdialenosť vyhľadaného bodu a anotovaného začiatku.*

Penalizačnú funkciu môžeme ďalej použiť pri počítaní hodnoty mGAP:

$$mGAP = \frac{\sum_{R_k \neq 0} p_k}{N}$$

kde $N$ je počet anotovaných bodov, $R_k$ je skóre vypočítané z penalizačnej funkcie pre bod vyhľadaný na $k$-tej pozícii a $p_k$ je hodnota *precision* pre $k$-tu pozíciu:

$$p_k = \frac{\sum_{i=1}^{k} R_i}{k}$$

Hodnoty penalizačnej funkcie sa počítajú postupne pre zoznam výsledkov zo-

radený podľa relevantnosti. Anotované body sa pritom k vyhľadaným bodom priraďujú bez opakovania. Každý anotovaný bod sa teda použije iba raz, a to pre blízky bod, ktorý má najvyššiu relevantnosť. Ďalšie body v okolí tohto anotovaného bodu sa už do výsledného skóre nezapočítavajú [17].

## 2.6 Výhody a nevýhody existujúcich metrík

Takmer všetky metriky, ktoré boli spomenuté v tejto kapitole, sú určené vo všeobecnosti na vyhodnocovanie IR systémov a typicky sa používajú na vyhodnotenie systémov, ktoré v kolekcii textových dát vyhľadávajú relevantné dokumenty. Pri zvukových súboroch dokážu pracovať iba so segmentovanými nahrávkami. Na druhej strane sú používané často, a preto sú výsledky pre jednotlivé systémy porovnateľnejšie a vlastnosti týchto metrík sú dobre preskúmané.

Metrika mGAP bola špeciálne navrhnutá pre hovorenú reč. Pomocou nej je možné vyhodnocovať jednotlivé úseky nahrávky, ktoré vyhľadávač nájde. Berie sa pritom do úvahy vzdialenosť nájdeného bodu od anotovaného bodu. Je možné ju teda použiť pre vyhľadávanie v nesegmentovanej reči. mGAP má však niektoré vlastnosti, o ktorých na prvý pohľad nie je jasné, či platia. Ide napríklad o symetriu penalizačnej funkcie. Vyhľadávač, ktorý nájde začiatok minútu po anotovanom začiatku, je označený ako rovnako dobrý ako vyhľadávač, ktorý nájde začiatok minútu pred týmto anotovaným začiatkom. Empiricky sa však zdá, že to nemusí platiť. Jednou z úloh tejto práce preto je overiť, či tieto vlastnosti v praxi platia a prípadne navrhnúť lepšie riešenie.

# 3. Návrh na vytvorenie nového spôsobu vyhodnocovania

## 3.1 Popis problému

Pri návrhu novej penalizačnej funkcie by sa mali brať do úvahy niektoré charakteristické črty, ktoré sú spojené so zvukovými nahrávkami, s prácou s týmito nahrávkami a vnímaním hovorenej reči. Sú to:

- Zvukové nahrávky sú striktne lineárne. Kým v texte môžeme rýchlo prebehnúť väčší kontext a rýchlejšie tak zistiť, či je nájdený úsek relevantný, takúto možnosť pri nahrávkach nemáme. Môžme síce úseky v nahrávkach preskakovať, potom ale nevieme s určitosťou povedať, či sa tam relevantný úsek nenachádza. V niektorých prípadoch je to možné odhadnúť z kontextu. Túto nevýhodu je čiastočne možné odstrániť tým, že používateľovi zobrazíme aj prepis nahrávky. Stále však ide o prepis hovorenej reči, môžu sa v ňom teda nachádzať chyby, nejasné môžu byť napríklad hranice viet.

- Človek vníma hovorné slovo inak ako písaný text. Podľa niektorých štúdií [15] je reakčný čas na zvukové podnety typicky kratší (140–160) ako reakčný čas na zrakové podnety (180–200). Vysvetľuje sa to tým, že zvukový podnet sa dostane do mozgu rýchlejšie (8–10 ms) ako zrakový podnet (20–40 ms). Ďalšie štúdie [21] naopak dokazujú, že čas potrebný na spracovanie zvukových podnetov je dlhší, iné nenašli podstatný rozdiel medzi rýchlosťami spracovania.
  V [13] bol prevedený jednoduchý test, ktorý skúmal zvukovú a sluchovú pamäť. Respondenti si mali zapamätať niekoľko číslic. Ak bola otázka, aké číslice to boli, položená ihneď, dosahovali lepšie výsledky respondenti, ktorí počuli zvukový podnet. Ak však bola otázka položená po 10 sekundách, dosahovali lepšie výsledky respondenti, ktorí videli zrakový podnet.
  Pri počúvaní reči tiež pomáha, ak máme možnosť vidieť rečníka. Podľa [28] je rozpoznanie zrakových podnetov spoľahlivejšie ako rozpoznanie sluchových podnetov. Čítanie je rýchlejšie ako počúvanie textu. Reč zase nesie v sebe viac informácií ako písaný text (napríklad informácie o pohlaví, veku, nálade, rečníka alebo o tom, kde je reč prezentovaná).

- Kvalita systému na rozpoznávanie reči môže ovplyvniť kvalitu IR systému. Ako už bolo spomenuté, pri vyhľadávaní pomocou izolovaných slov nemá kvalita rozpoznávania príliš veľký vplyv na kvalitu IR systému. V prípade nesegmentovaných nahrávok však je dôležité úzke okolie anotovaného bodu. Preto môže mať v našom prípade kvalita rozpoznávania hovorenej reči výrazný vplyv na kvalitu celého IR systému.

## 3.2 Použité dáta

Pri riešení úlohy potrebujeme dáta, na ktorých môžeme naše predpoklady overovať a ďalej skúmať správanie ľudí. Konkrétne potrebujeme zvukové nahrávky hovorenej reči, niekoľko tém, ktoré sa v nahrávkach vyskytujú a ručne anotované hranice týchto tém v nahrávkach. Pre tieto účely môžeme použiť archív nahrávok Malach.

Malach je medzinárodný projekt, na ktorom sa zúčastňuje viacero svetových univerzít a výskumných stredísk. Jeho cieľom je sprístupniť rozsiahly audiovizuálny archív nahrávok výpovedí svedkov nacistického holokaustu vytvorených nadáciou Shoah Visual History Foundation[1]. Spolu obsahuje archív 116000 hodín digitalizovaných rozhovorov s 52000 ľudmi z celého sveta, celkom v 32 jazykoch [10]. Archív bol vytvorený počas rokov 1994 až 1999.

Na tomto projekte sa zúčastňuje aj Ústav formálnej a aplikovanej lingvistiky Matematickofyzikálnej fakulty Univerizty Karlovej. Približne 700 českých nahrávok bolo v rámci tohto projektu zdigitalizovaných. Nahrávky boli spracované systémom na rozpoznávanie hovorenej reči a následne boli nahrávky ručne anotované. Použitá bola pritom len zvuková stopa audiovizuálnych nahrávok a prepis tejto zvukovej stopy získaný pomocou automatického rozpoznávania reči. Manuálne boli označené hranice vybraných tém. Anotátori najprv naštudovali dané témy tak, aby boli schopní relevantné úseky rozlíšiť. Potom s pomocou špeciálne vytvoreného anotačného programu počúvali jednotlivé nahrávky a označovali jednotlivé úseky vo výpovediach. Označené boli začiatky a konce relevantných úsekov. Niektoré témy boli pre kontrolu spracované viacerými anotátormi [10].

My máme k dispozícii 357 českých nahrávok, pričom používame iba zvukovú stopu nahrávok. Priemerná dĺžka jednej nahrávky je pritom 95,17 minút. Nahrávky sú zaznamenané na niekoľkých (dvoch až deviatich) kazetách. Na jednej kazete je pritom nahratých približne tridsať minút záznamu. Nahrávky boli týmto spôsobom aj digitalizované, čo znamená, že jedna nahrávka sa nachádza v niekoľkých mp3 súboroch. Tieto mp3 súbory vždy spojeníme do jediného súboru. Takéto súbory majú veľkosť vo veľkej väčšine do 20 MB. Niektoré pásky však boli stratené, spolu bolo nájdených 48 takýchto nekompletných nahrávok.

V nahrávkach, je vyznačených spolu 116 tém. My by sme mali pracovať s vyznačenými začiatkami týchto tém. Konce tém by sme nemali používať, pretože pri vyhľadávaní sa typicky označujú iba začiatky. Podľa [10] bolo 32 tém anotovaných duálne (jedna téma bola anotovaná dvoma anotátormi), pričom iba pri dvoch témach sa anotátori zhodli na približne rovnakom počte začiatkov. Čo sa týka času potrebného na spracovanie segmentov a dĺžky nájdených segmentov, sú časy pomerne vyrovnané. U času, ktorý anotátori venovali počúvaniu nahrávok opäť existujú výrazné rozdiely, z ktorých je možné vysledovať štýl práce jednotlivých anotátorov. Bolo zistené [10], že štýl práce jednotlivých anotátorov súvisí s ich backgroundom. Konkrétne ide o skupiny historikov a a knihovníkov.

Podľa [10] je priemerná dĺžka označeného relevantného segmentu 167 sekúnd (najmenej 49 a najviac 502 sekúnd). Priemerný počet nájdených segmentov pre

[1] http://college.usc.edu/vhi

jednu tému je 44. Pri 19 témach nebol nájdený žiadny relevantný segment, pri piatich témach bolo nájdených viac ako 100 relevantných segmentov. Každý anotátor spracoval priemerne 25 tém. Každá téma sa skladá zo štyroch častí: číslo témy (je rovnaké pre všetky jazyky), názov témy, description (krátky popis témy, obmedzený na maximálne dve vety) a narrative (podrobnejší popis témy, obmedzený maximálne na desať viet) [10]. Témy sú napríklad „Dětské umění v Terezíně“, „Židovské děti na školách“ alebo „Kolaborace místních obyvatel“. Príklad popisu tém je možné nájsť v tabuľke 3.1

Súčasťou archívu sú aj informácie o jednotlivých nahrávkach (napr. kedy, kde a v akom jazyku bola vytvorená, kto rozhovor viedol) a o hovorcoch v danej nahrávke. Vieme napríklad rok a miesto narodenia hovorcu alebo to, či bol v pochodoch smrti, prípadne v ktorom tábore bol väznený. K dispozícii sú tiež fotografie jednotlivých hovorcov.

## 3.3 Návrh riešenia

Nahrávky použité v acrhíve Malach sú pomerne dlhé a tématicky nekoherentné. Preto je potrebné v nich vyhľadať konkrétny bod, kde sa zadaná téma začína. Nahrávky sú tiež nesegmentované. Pri návrhu spôsobu vyhodnocovania vyhľadávania v nahrávkach je našim východiskom metrika mGAP. Našim cieľom je zistiť, či nie je možné metriku mGAP upraviť tak, aby lepšie zodpovedala realite.

Penalizačná funkcia použitá v mGAP má podľa [17] tvar rovnoramenného trojuholníka. Pri návrhu tejto funkcie sa brali do úvahy aj ďalšie tvary - gaussova krivka a obdĺžnik, skúšali sa rôzne parametre týchto tvarov. Pri návrhu výslednej penalizačnej funkcie boli použité používateľmi označené začiatky tém. V okolí skutočných začiatkov tém sa pritom vygenerovali náhodné body s pravdepodobnosťami zodpovedajúcimi penalizačnej funkcii. Následne sa vyhodnocovalo ako takáto penalizačná funkcia ovplyvňuje hodnotu mGAP [17].

Podľa [17] záleží iba na vzdialenosti vyhľadaného začiatku od anotovaného začiatku. V skutočnosti sa ale zdá pravdepodobné, že záleží napríklad aj na tom, či vyhľadaný úsek leží pred alebo po anotovanom začiatku. Je napríklad pravdepodobné, že ak vyhľadávač nájde začiatok témy o niečo ďalej, po jej skutočnom začiatku, pričom sa o tejto téme ešte stále hovorí, potom bude používateľ relatívne spokojný, pretože začiatok témy už nájde jednoducho. Ak však vyhľadávač nájde začiatok témy pred skutočným začiatkom, aj keď v rovnakej vzdialenosti od neho ako v predchádzajúcom prípade, potom nebude mať používateľ dobrú predstavu, kde sa téma začína a môže mu trvať dlhšie, než ju nájde.

Jedinou možnosťou, ako overiť správnosť tejto funkcie je užívateľský prieskum. Keďže prevedenie výskumu na reálnych dátach, ktoré IR systémy vyhľadajú, by bolo príliš časovo náročné, prevedieme simuláciu takýchto IR systémov.

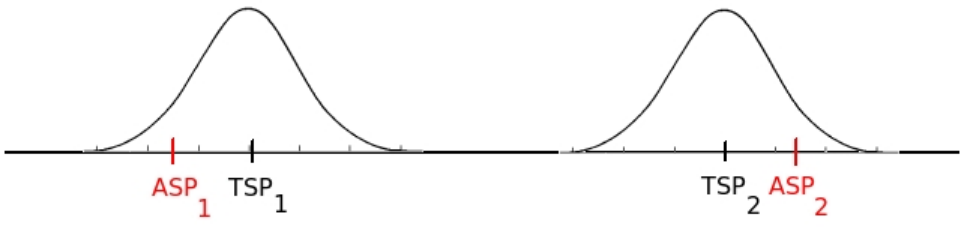
Môžeme predpokladať, že vyhľadávanie v hovorenej reči je určené pre po-

užívateľov, ktorí chcú v nahrávke čo najrýchlejšie nájsť konkrétne údaje. Preto by malo byť vyhodnotenie založené na tom, ako rýchlo sa s pri vyhľadávaní s daným IR systémom dopracujú ku požadovaným údajom. Takéto kritérium by zodpovedalo v podstate dvom kritériám hodnotenia IR systému, ktoré uvádzal Cleverdon [6] a ktoré už boli spomenuté - interval medzi zadaním požiadavky a získaním odpovede a psychická alebo fyzická námaha, ktorá je vyžadovaná od používateľa. Ak stačí používateľovi kratší čas na to, aby našiel relevantný úsek, potom sa tým skráti aj interval medzi zadaním požiadavky a získaním odpovede a od používateľa sa vyžaduje menšia psychická námaha. Je teda pravdepodobné, že sa spokojnosť používateľa odvíja od času, ktorý potrebuje na to, aby našiel relevantný úsek.

Čas, za ktorý používateľ nájde skutočný začiatok relevantného úseku by mal závisieť na vzdialenosti a smere vyhľadaného začiatku od skutočného relevantného začiatku. Zároveň nás zaujíma, ako používatelia sami hodnotia vyhľadané začiatky v závislosti od toho ako ďaleko a akým smerom sú vzdialené od anotovaných začiatkov. Aby sme získali potrebné dáta (rýchlosti hľadania a subjektívne hodnotenia) prevedieme užívateľský prieskum. Máme pritom k dispozícii ručne anotované začiatky tém. Výsledky vyhľadávania jednotlivých IR systémov nahradíme simuláciou a to tak, že vyberieme náhodne bod v okolí tohoto anotovaného bodu. Vygenerované body budú zastupovať výsledky vyhľadávania IR systémov. Zavedieme pritom tri označenia:

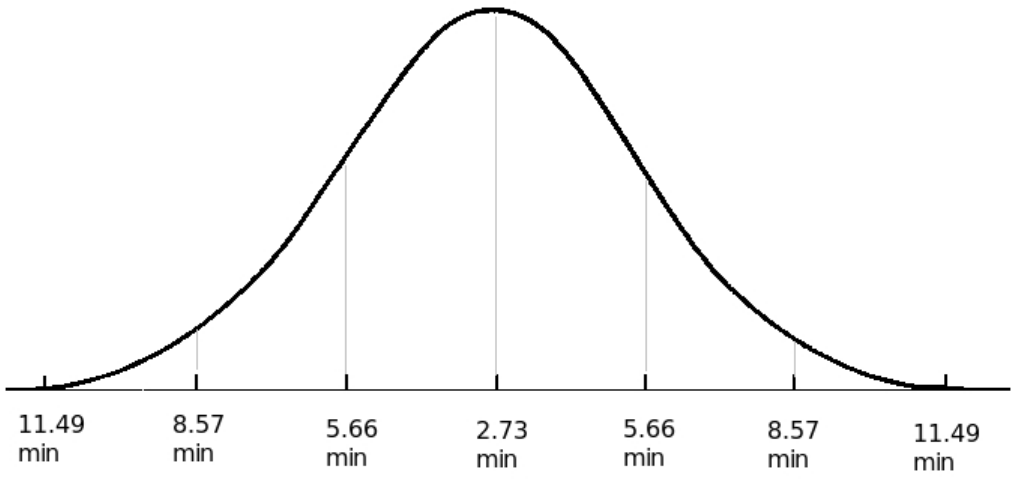
1. ručne anotované začiatky tém označíme ako True Start Points (TSP),
2. náhodne vybraný bod v okolí TSP budeme nazývať Automatic Start Point (ASP),
3. bod, ktorý vyberie respondent v užívateľskom prieskume ako začiatok relevantného úseku budeme nazývať User Start Point (USP).

Generovanie bodov ASP je znázornené na obrázku 3.1.




Obr. 3.1: *Generovanie ASP v okolí TSP.*

Náhodný výber ASP by mal čo najviac zodpovedať simulácii IR systému. Preto by malo byť pri tomto náhodnom výbere bodov ASP použité normálne rozdelenie. Stredná hodnota takéhoto rozdelenia by mala zodpovedať jednotlivým bodom TSP. Rozptyl tohto rozdelenia by mal byť určený zo skutočných dĺžok tém, ktoré máme k dispozícii. Stredná hodnota reálnych dĺžok tém je pritom 2,73 minúty a priemerná stredná odchýlka je 2,92 minúty. Normálne rozdelenie s parametrami získanými z dĺžok tém na obrázku 3.2.




Obr. 3.2: *Normálne rozdelenie s hodnotami získanými z reálnych dĺžok tém*

## 3.4 Návrh prehrávača dát

Pre navrhnutý prieskum sme sa rozhodli vytvoriť vlastný prehrávač dát, ktorý by spĺňal všetky požiadavky kladené na tento prieskum. Úlohou prehrávača by malo byť prehrávať nahrávky z projektu Malach tak, aby respondent mohol čo najrýchlejšie nájsť začiatok relevantného úseku v nahrávke pre konkrétnu tému, ktorú mu predložíme.

Respondentovi by sa najskôr mala zobraziť téma, ktorej začiatok má hľadať. V nahrávke by sa mu potom mal zobraziť bod, ktorý by znázorňoval, kde približne by sa táto téma mala začínať (bod ASP). Respondent by sa mohol následne v nahrávke pohybovať tak, aby našiel bod, kde sa téma začína podľa neho (bod USP). Nájdený bod by mal nejako označiť. Respondent by mal mať tiež možnosť subjektívne ohodnotiť kvalitu automatického vyhľadávania. Počas práce respondenta by sa mali všetky jeho akcie zaznamenávať, aby sme ich mohli neskôr analyzovať. Základom by teda mal byť audio prehrávač, ktorý by všetky tieto veci umožňoval.

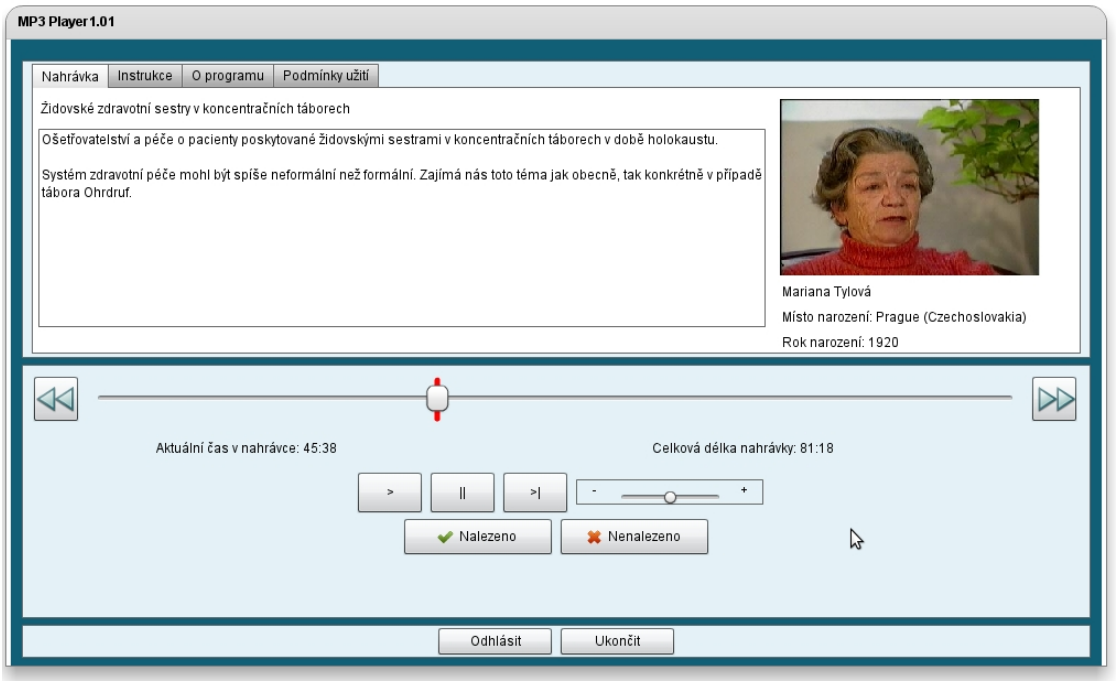
| ID | Názov témy | Description | Narrator |
|---|---|---|---|
| 1286 | Hudba v holokaustu | Svědectví o tom, zda hudba pomáhala (duševně nebo i jinak) nebo překážela vězňům internovaným v koncentračních táborech. | Popis toho, jakou roli hrála hudba v životě vězňů. |
| 1288 | Posilování víry | Posilování náboženské víry jako důsledek holokaustu. | Většina Židů, kteří byli před příchodem do tábora silně nábožensky založení, ztratila v důsledku svých zážitků svou víru v Boha. Hledáme opačné případy: lidi, jejichž víra byla v důsledku zážitků posílena. |
| 1310 | Praktiky nacistické eugeniky | Výpovědi svědků nacistických eugenických experimentů. | Očitá svědectví nacistických lékařských experimentů (včetně Mengeleho dvojčat, nucené sterilizace, eutanazie a dalších) od lidí, kteří byli nuceni se na nich podílet. |
| 1311 | Skrývající se děti a jejich zachránci | Zajímáme se o příběhy dětí, které se skrývaly bez svých rodičů, a jejich zachránců. | Přímá svědectví lidí, kteří zachránili a skrývali židovské děti v době holokaustu, nebo svědectví tehdejších židovských dětí, které byly zachráněny a skrývali se. Velmi relevantní a žádoucí je příběh ortodoxního židovského dítěte, které se skrývalo jako křesťan a eventuálně se stalo knězem nebo rabínem. |
| 1321 | Zdravotně a tělesně postižení v holokaustu | Zacházení se zdravotně a tělesně postiženými lidmi v holokaustu. | Systematické vyhlazování a zneužívání neslyšících a postižených lidí, nucená sterilizace, eugenika, rasové metody. |

Tabuľka 3.1: *Ukážka niekoľkých tém.*

# 4. Prehrávač dát

## 4.1 Užívateľské rozhranie

Pri vytváraní programu bol veľký dôraz kladený na návrh užívateľského rozhrania programu. Podľa [6] je interface jednou z veci, ktoré ovplyvňujú IR systém a každá drobnosť pri návrhu tak môže ovplyvniť celkové výsledky prieskumu. Screenshot programu sa nachádza na obrázku 4.1. Keďže sú nahrávky v češtine, rozhodli sme sa, že v celom programe použijeme češtinu.




Obr. 4.1: *Prehrávač nahrávok použitý pri užívateľskom prieskume.*

### Ovládanie

Pre základný pohyb v nahrávke bol zvolený slider. Na slideri je červenou čiarkou znázornený bod, kde sa nachádza ASP. Slider bol navrhnutý čo najdlhší, aby bola pri pohybe získaná najvyššia možná presnosť. Obmedzená však musela byť veľkosť celého prehrávača, aby ho bolo možné používať aj pri menších rozlíšeniach obrazovky. Okrem pohybu pomocou slidera môže respondent k zrýchlenému pohybu v nahrávke použiť aj tlačítka << a >>. Aj keď pri návrhu stál proti použitiu týchto tlačítok argument, že sa takéto tlačítka vo väčšine klasických prehrávačov nenachádzajú, pohyb pomocou slidera je v dlhých nahrávkach stále nepresný a tlačítka zjednodušujú navigáciu. Pri používaní týchto tlačítok môžu byť výsledky prieskumu ovplyvnené dĺžkou skoku. Tento skok bol určený na 30

sekúnd. Priemerná dĺžka témy je približne 2,7 minúty, preto by mal byť takýto skok vhodný. Existuje iba malá pravdepodobnosť, že by respondent celú tému preskočil. Použiteľnosť takéhoto skoku bola spätne odskúšaná.

Súčasťou užívateľského rozhrania boli aj tlačítka „Nalezeno“ a „Nenalezeno“, na ktoré mal respondent kliknúť v momente, keď našiel, respektíve ak nenašiel začiatok témy. Okrem toho mal respondent možnosť preskočiť nahrávku bez toho, aby jednu z týchto možností musel vybrať. Takáto situácia mohla nastať napríklad, keď respondent dobre neporozumel téme, prípadne bola nahrávka nezrozumiteľná. Informácie o téme, hovorcovi a inštrukcie, mohol respondent nájsť v štyroch záložkách.

V záložke Téma bol zobrazený popis témy, ktorej začiatok mal respondent za úlohu nájsť. Zobrazovali sa tu tri položky z popisu témy, tak ako sú uvedené v archíve Malach - názov témy, časť description a časť narrator. Vedľa témy bola zobrazená fotografia, meno, dátum a miesto narodenia hovorcu v nahrávke. Tieto údaje boli v programe zobrazené kvôli tomu, aby si respondenti mohli vytvoriť lepšiu väzbu s nahrávkou. Hodnotenia respondentov by potom mohli byť o niečo relevantnejšie. Niekedy môžu navyše tieto údaje pomôcť pri určovaní začiatku relevantnej témy.

Od respondenta sa vyžadovalo, aby si tému pred spustením nahrávky najprv prečítal. Často sa stávalo, že po sebe nasledovali dve rovnaké témy. Popis témy mal respondent pri spustení programu priamo pred sebou, pretože sa nachádzala v prvej záložke. Téma je zvyčajne popísaná niekoľkými vetami, preto by respondentovi nemalo zabrať jej naštudovanie príliš veľa času.

V druhej záložke sa nachádzali inštrukcie. Tieto inštrukcie sú uvedené v prílohe A. Ďôležité je, že si respondent mohol inštrukcie zobraziť kedykoľvek pri prehrávaní, pričom nebolo potrebné prehrávanie zastaviť. Pri vyhodnotení by sa však mala brať do úvahy aj skutočnosť, že keď mal respondent otvorenú záložku s inštrukciami, nemusel sa naplno venovať spracovaniu nahrávky.

V tretej záložke boli uvedené informácie o programe, ciele prieskumu a odkaz na autora.

V poslednej záložke boli informácie o licencii dát.

V prehrávači muselo byť umiestnené základné ovládanie nahrávky, teda jej spustenie, zastavenie, pozastavanie a ovládanie hlasitosti, ako aj tlačítka na odhlásenie a ukončenie programu. Respondent nemal možnosť sa vrátiť sa k predchádzajúcej nahrávke.

## Rozmiestnenie prvkov

Slider by mal byť najdôležitejšou súčasťou prehrávača. Preto bol umiestnený v jeho centrálnej časti. Na slideri bol červenou čiarkou vyznačený bod ASP. Vpravo a vľavo od slidera sa nachádzali tlačítka na rýchly pohyb v nahrávke. Ďalšou dôležitou časťou bolo tlačítko „Nalezeno“. Tlačítko by malo byť dostatočne veľké a na dobre umiestnené. Preto sa nachádzalo priamo pod sliderom. Vpravo od

tohto tlačítka sa nachádzalo tlačítko „Nenalezeno“. Pod týmito tlačítkom sa nachádzali ďalšie tlačítka na ovládanie nahrávky (play, pauza, skok na ďalšiu nahrávku a slider ovládajúci hlasitosť prehrávania).

V dolnej časti prehrávača sa nachádzali tlačítka na odhlásenie a ukončenie programu. Nad sliderom sa nachádzali záložky s informáciami o nahrávke a o programe. Respondent mal vždy možnosť zobraziť si jednu z nich zatiaľ čo počúval nahrávku, nemusel pritom otvárať špeciálne okno.

## Prehrávanie

Ďalej bolo treba rozhodnúť, z ktorého miesta by sa nahrávka mala začať prehrávať. Ak respondent zadá požiadavku na vyhľadávanie v IR systéme, začne sa nahrávka zvyčajne prehrávať od vyhľadaného bodu. Preto aj v našom prípade prehrávanie začínalo od červeného bodu, ktorý simuluje začiatok témy. Ak by sa nahrávka začala prehrávať od začiatku, potom by bola pravdepodobne vždy prvá akcia respondenta presun ku červenému bodu, čo by mu zbytočne zabralo nejaký čas.

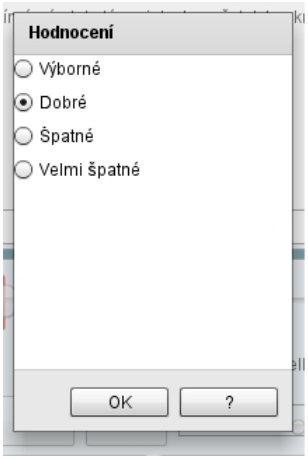
## Subjektívne hodnotenie

Súčasťou programu bolo aj subjektívne hodnotenie vyhľadávania respondentami. Respondent musel vybrať práve jedno z hodnotení *výborné*, *dobré*, *špatné* a *velmi špatné*. Zámerne neexistovala v ohodnotení neutrálna možnosť ako napríklad priemerné. V prípade, že takáto možnosť nebola na výber, bol respondent prinútený rozhodnúť, či bolo vyhľadávanie skôr dobré alebo skôr zlé. Hodnotenie bolo implementované ako *radio buttons*. Žiadne hodnotenie pritom nebolo prednastavené, aby nemal respondent tendenciu nechať ako hodnotenie prednastavenú hodnotu.

Do úvahy sme brali aj iné možnosti hodnotenia - napríklad slider, na ktorom by respondent ohodnotil vyhľadaný bod v spojitom priestore. V prípade slidera by však bolo hodnotenie menej presné - ak by chcel napríklad respondent ohodnotiť dve nahrávky rovnako, je veľmi pravdepodobné, že by sa mu to nepodarilo. Okrem toho by tu mohol opäť nastať problém s tým, že by respondenti najčastejšie vyberali neutrálne hodnotenia v strede. Screenshot hodnotenia je na obrázku 4.2.

## Nahrávky

Veľký dôraz bol kladený na výber a poradie nahrávok, ktoré boli predkladané respondentom, pretože by tak mohli byť v rôznych smeroch ovplyvnené výsledky. Spolu sme mali k dispozícii 5436 anotovaných bodov v 357 rôznych témach. Niektoré témy boli spracované viacerými anotátormi. Zo zoznamu tém boli odstránené tie, ktoré mali príliš málo výskytov (menej ako 5 výskytov). Poradie nahrávok bolo dané pevne, bolo teda rovnaké pre každého respondenta. To zaručovalo opakovateľnosť pokusu. Navyše tak jednu nahrávku spracovalo viac respondentov (napríklad prvú nahrávku spracovali všetci zúčastnení respondenti), čím sme získali referenčné nahrávky. Správanie respondentov sme tak mohli lepšie




Obr. 4.2: *Okno s vyhodnocovaním.*

porovnávať.

Potrebné bolo vyriešiť aj otázku, ako by mali byť nahrávky zoradené. Respondenti si pred prehraním nahrávky museli preštudovať tému, ktorej začiatok mali za úlohu nájsť. Keďže toto naštudovanie trvá určitý čas, nahrávky s rovnakou témou nasledovali po sebe. V rámci jednej témy boli nahrávky zoradené náhodne. Témy boli zoradené podľa počtu anotovaných začiatkov danej témy.

Pri väčšine tém však existovalo veľké množstvo jednotlivých anotovaných začiatkov. Téma „Dětské umění v Terezíně“, ktorá mala anotovaných začiatkov najviac, a ktorá sa teda respondentom zobrazovala ako prvá, bola anotovaná dokonca 162 krát. Keby sa teda respondentom zobrazili najprv všetky nahrávky s touto témou, respondenti by sa už nedostali k ďalšej téme. Preto bolo potrebné počet zobrazených nahrávok v rámci jednej témy obmedziť. Počet bol určený empiricky na sedem nahrávok. Pri veľkom počte nahrávok, napríklad viac ako desať, by sa mohol prieskum pre respondentov stať nezaujímavý. Pri malom počte, napríklad menej ako päť nahrávok, sa respondenti nestihli s danou témou dostatočne zoznámiť. Nakoniec bolo týmto spôsobom vybratých 257 anotovaných bodov, ktoré sa postupne respondentom zobrazovali.

## 4.2 Implementácia prehrávača

Dôležitou voľbou bol výber programovacieho jazyka, v ktorom mal byť prehrávač vytvorený. Podstatná bola v tomto prípade prenositeľnosť. Chceli sme dosiahnuť to, aby mohol každý respondent čo najrýchlejšie a najspoľahlivejšie prehrávač spustiť, najlepšie bez toho, aby musel niečo sťahovať alebo inštalovať. Preto sme sa nakoniec rozhodli pre Flex[1], čo je open-source framework, ktorý slúži najmä na tvorbu webových, prípadne desktopových aplikácií. Konkrétne pri bola pri použitá verzia 4. verzia Flexu. Respondent tak mohol prehrávač jednoducho spustiť v internetovom prehliadači. Všetky dáta boli tiež umiestnené na internete a

[1] http://www.adobe.com/products/flex

respondenti ich nemuseli sťahovať. Keďže boli jednotlivé nahrávky dlhé približne 20 MB, bolo spustenie prehrávania jednotlivých nahrávok dostatočne rýchle.

Program sa vyvíjal na Linuxe, konkrétne bola použitá aplikácia Eclipse[2]. Základ programu tvoril klasický audio prehrávač a ako tento základ sme prevzali freeware prehrávač[3]. Prevzaté bolo základné ovládanie prehrávania, spustenie, zastavenie nahrávky, pohyb v nahrávke, ovládanie hlasitosti a navigácia v playliste. Ďalej bol prehrávač upravený tak, aby splňoval všetky naše požiadavky.

Pri generovaní bodov ASP bol využitý skript napísaný v jazyku Python, ktorého vstupom bolo niekoľko čistých textových dokumentov, ktoré obsahovali informácie o hovorcoch, názvy tém a zoznam anotovaných bodov. Výstupom bol xml súbor, ktorý bol použitý v prehrávači. Ukážka tohto xml súboru sa nachádza v prílohe B.

Hodnotenia a všetky akcie, ktoré respondenti uskutočnili (spustenie, zastavenie nahrávky, presun v nahrávke, ale aj výber záložky) boli zapisované do logu. Do úvahy sa pritom brala aj možnosť, že sa respondent do nahrávky započúval a zabudol na svoju úlohu, prípadne od počítača odišiel a program nechal bežať. Preto sa testovala nečinnosť respondenta. V prípade, že respondent päť minút nevykonal žiadnu akciu, zaznamenala sa nečinnosť do logu. Respondenti mohli pracovať ľubovoľne dlho, bolo im ale odporučené pracovať aspoň 15 minút.

Pri prvom spustení programu sa respondentom zobrazili inštrukcie. Tieto isté inštrukcie si mohli respondenti zobraziť v príslušnej záložke.

Ku dátam z Malachu mali mať prístup iba respondenti, ktorí sa zúčastňovali prieskumu. Preto bol celý adresár s programom a všetkými dátami prístupný len s daným užívateľským menom a heslom, ktoré boli každému respondentovi poslané. Adresár bol zabezpečený pomocou .htaccess. Aby sa respondentom prihlásenie zjednodušilo, boli tieto prihlasovacie údaje odoslané v http adrese prehrávača.

Požiadavkou na prehrávač tiež bolo, aby respondent mohol svoju prácu prerušiť a potom sa opäť vrátiť ku poslednej nahrávke, ktorú nespracoval. Preto bola nutná ešte jedna autorizácia. Každý respondent najprv zaregistrovať a potom sa mohol prihlásiť pod vlastným užívateľským menom. Navyše bolo týmto spôsobom možné ku každému respondentovi, prípadne skupinám respondentov (napr. veková skupina, muži a ženy) priradiť jednotlivé spracovania nahrávok a zistiť tak napríklad rôzne taktiky pri vyhľadávani a hodnotení.

Respondent musel pri registrácii povinne zadať užívateľské meno, heslo, e-mail a musel súhlasiť s tým, že nebude nahrávky používať inak, prípadne ich ďalej šíriť. Prihlasovacie údaje boli po registrácii respondentovi odoslané na zadaný e-mail. Voliteľne mohol respondent zadať svoje meno, vek, pohlavie, oblasť, v ktorej pracuje, prípadne pridať nejakú poznámku. Registračný formulár

[2]http://www.eclipse.org

[3]http://blog.pxldesigns.com/2008/02/pxl-mp3-player-updated-with-rewindff-buttons-cairngorm-22-framework-and-progressive-download

sa nachádza v prílohe C.

Pre ukladanie súborov na server (log a posledná spracovaná nahrávka), pri prihlasovaní a registrácii respondentov boli použité php skripty. Informácie o respondentoch boli uložené v SQL databáze, informácie o poslednej spracovanej nahrávke každého respondenta a log boli uložené v čistom textovom súbore. Prehrávač bol pritom navrhnutý tak, aby zvládol súčasné prehrávanie viacerých respondentov.

# 5. Užívateľský prieskum

## 5.1 Priebeh prieskumu

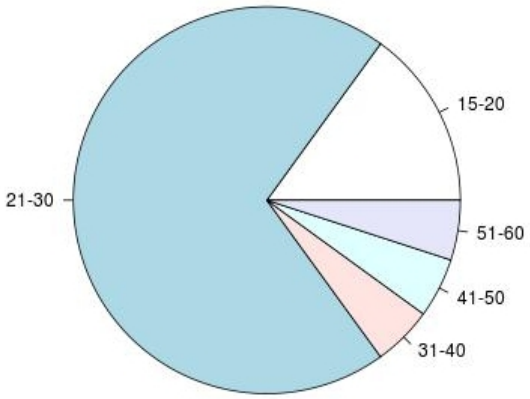
Užívateľský prieskum prebiehal od 3. 3. 2011 do 18. 3. 2011. Prieskum teda prebiehal 15 dní. Prieskum sa uskutočnil v dvoch etapách a to z toho dôvodu, aby sa po prvej etape opravili prípadné chyby. 3. 3. bol program odoslaný 10 respondentom, 7. 3. bol poslaný ďalším 28 respondentom. Výskumu sa zúčastnilo 24 respondentov, pričom počítame iba respondentov, ktorí spracovali aspoň jednu nahrávku. Niekoľko respondentov sa zaregistrovalo, prípadne prihlásilo, ale nespracovalo žiadnu nahrávku. Ďalšie štatistiky prieskumu sú uvedené v tabuľke 5.1.

| | |
|---|---|
| Počet anotovaných bodov celkom | 263 |
| Počet zúčastnených respondentov | 24 |
| Priemerný počet spracovaných nahrávok na jedného respondenta | 11 |
| Najväčší počet spracovaných nahrávok jedným respondentom | 42 |
| Čas strávený priemerne pri prieskume | 59,95 min |
| Najdlhší čas strávený pri prieskume | 4,8 hod |

Tabuľka 5.1: *Základné štatistiky prieskumu.*

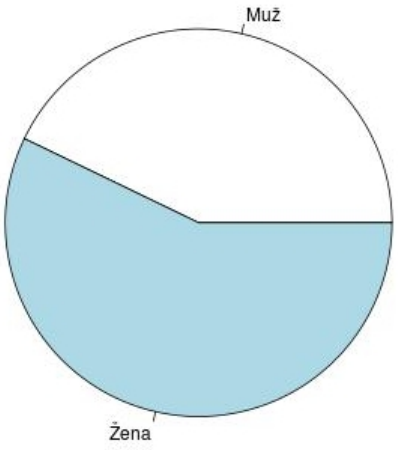
## 5.2 Respondenti

Každý respondent mohol pri registrácii zadať vek, pohlavie a odbor, v ktorom pracuje. Uvedenie týchto údajov nebolo povinné. Percentuálne zastúpenie veku respondentov, ktorí sa prieskumu zúčastnili a svoj vek zadali je uvedené v grafe 5.1. Najviac respondentov bolo vo veku 21 až 30 rokov.



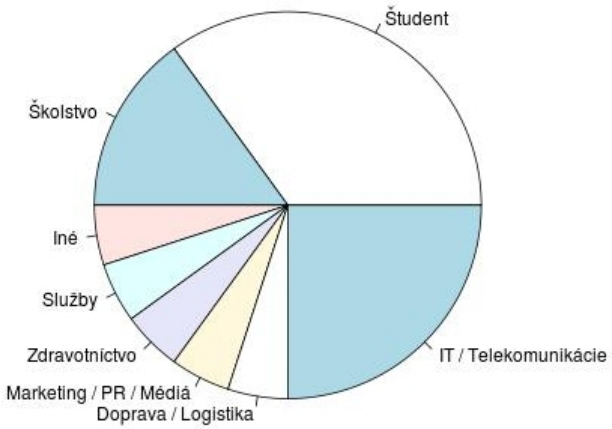

Obr. 5.1: *Vekové rozloženie respondentov.*

Percentuálne zastúpenie respondentov podľa pohlavia je uvedené na obrázku 5.2. Výskumu sa zúčastnilo 57 percent žien a 43 percent mužov.




Obr. 5.2: *Rozloženie pohlavia respondentov.*

Percentuálne zastúpenie rôznych pracovných odborov respondentov je na obrázku 5.3. Najviac respondentov boli študenti, nasledujú respondenti pracujúci v IT a v telekomunikáciách a tretiu najpočetnejšiu skupinu tvorili respondenti pracujúci v školstve.




Obr. 5.3: *Rozloženie oblasti práce respondentov.*

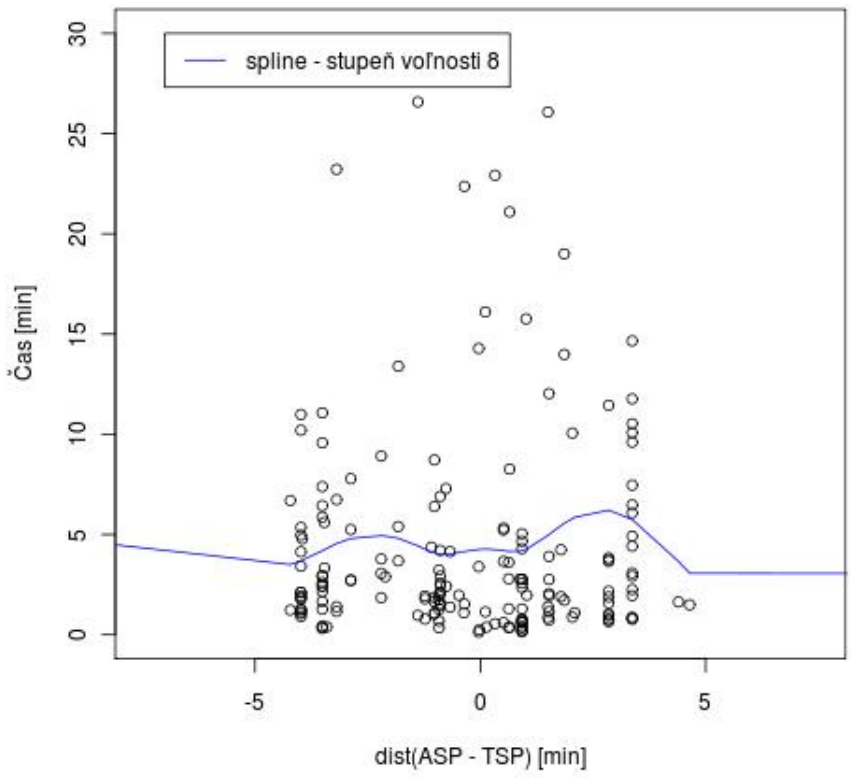
Grafy sú tvorené vždy na základe tých údajov, ktoré boli zadané. Niektorí respondenti tieto údaje nezadali vôbec, niektorí zadali iba časť.

## 5.3 Výsledky prieskumu

Pri výsledkoch nás zaujímala závislosť rýchlosti vyhľadania bodu USP a subjektívneho hodnotenia responenta od vzdialenosti anotovaného bodu (TSP) a automaticky vygenerovaného bodu (ASP) v jeho okolí. Toto generovanie by totiž malo simulovať vyhľadávanie IR systému. Pri generovaní ASP bol vygenerovaný jeden bod, ktorý je od TSP vzdialený takmer 80 minút. Toto pozorovanie považujeme ďalej za odľahlé a neberieme ho do úvahy.

### Vzdialenosť ASP a TSP

Na obrázku 5.4 je zobrazená závislosť rýchlosti ohodnotení na vzdialenosti bodov ASP a TSP. Teda na vzdialenosti skutočného začiatku témy a vygenerovaného bodu v okolí tohto začiatku. Bodmi sme ďalej preložili spline funkciu. Použili sme stupeň voľnosti 8.




Obr. 5.4: *Závislosť rýchlosti vyhľadania na vzdialenosti ASP a TSP, vybraný je úsek (-7,5; 7,5). Bodmi je preložená spline funkcia s stupňom voľnosti 8.*

Minimá funkcie ležia približne v časoch +4,5 a -4. Pri týchto časoch však máme k dispozícii podstatne menší počet pozorovaní, čo túto situáciu skresľuje.

Preto môžeme za skutočné minimum funkcie považovať lokálne minimum, ktoré leží približne v bode 0. Nízke hodnoty nadobúda funkcia v blízkom okolí tohto bodu, približne minútu pred a minútu po nulovom bode. V skutočnosti teda nehrá veľkú rolu to, či vyhľadávač nájde začiatok témy minútu pred alebo minútu po skutočnom začiatku témy. To môže súvisieť s tým, že je ťažké povedať, čo začiatok témy skutočne je. V takto blízkom okolí môže respondent nájsť začiatok relevantného úseku v priebehu približne piatich minút, čo sa môže zdať veľa, vzhľadom na to, že je začiatok vzdialený nanajvýš minútu. Často sa ale napríklad stáva, že respondent bod nájde, ale chce si overiť, že ide skutočne o správny začiatok.

Od nulového bodu funkcia rastie a to v prípade kladných aj záporných hodnôt. Maximum je dosiahnuté približne 2,5 minúty po nulovom bode. Priemerná dĺžka tém je tiež približne 2,5 minúty. Toto maximum sa teda pravdepodobne nachádza v mieste, kde sa začína hovoriť o ďalšej téme. Podobná situácia nastáva aj približne 2,5 minúty pred nulovým bodom. V tomto momente sa tiež zrejme najčastejšie hovorí ešte o predchádzajúcej téme. Lokálne maximum je v záporných hodnotách však menšie, teda respondentom spôsobuje o niečo menšie problémy ak sa prehrávanie začne pred skutočným anotovaným bodom. Z kontextu môže byť v tomto momente jasné, že sa rozprávač k danej téme blíži.

Zaujímavý je pokles funkcie za týmito maximami, približne v minútach (-4; - 2,5) a (2,5; 4). V tomto prípade ide pravdepodobne o iné témy ako bola téma v približne v čase (-2,5; -1) a (1; 2,5). Môže to byť spôsobené tým, že začiatok témy (teda body -2,5 a +2,5) sú pre respondenta kritickejšie a ťažšie sa v nich orientuje. Kontext, z ktorého je možné prípadne určiť o čom sa hovorí, je jasnejší uprostred rozprávania.

Výsledky pre vzdialenosť ASP a TSP môžu byť čiastočne ovplyvnené tým, že respondenti spracovávali rovnakú sadu nahrávok. Niektorí respondenti sa dostali ďalej, niektorí spracovali iba pár nahrávok. Preto máme pri niektorých časoch viac označených bodov, pri niektorých iba jeden alebo dva. V prípade, že máme v niektorých oblastiach málo pozorovaní, môžu byť výsledky ovplyvnené jednotlivými konkrétnymi nahrávkami. Niektoré témy môžu byť napríklad na vyhľadanie ťažšie ako iné - napríklad téma „Dětské umění v Terezíně“ bola pre respondentov zrozumiteľnejšia ako téma „Pracovní Tábory IG Farben“. To isté nastáva v prípade jednotlivých nahrávok. Niekedy môže byť veľmi ťažké určiť, či ide naozaj o danú tému, inokedy je to jasné. S týmto problémom je však potrebné počítať vždy pri vyhľadávaní v reči.

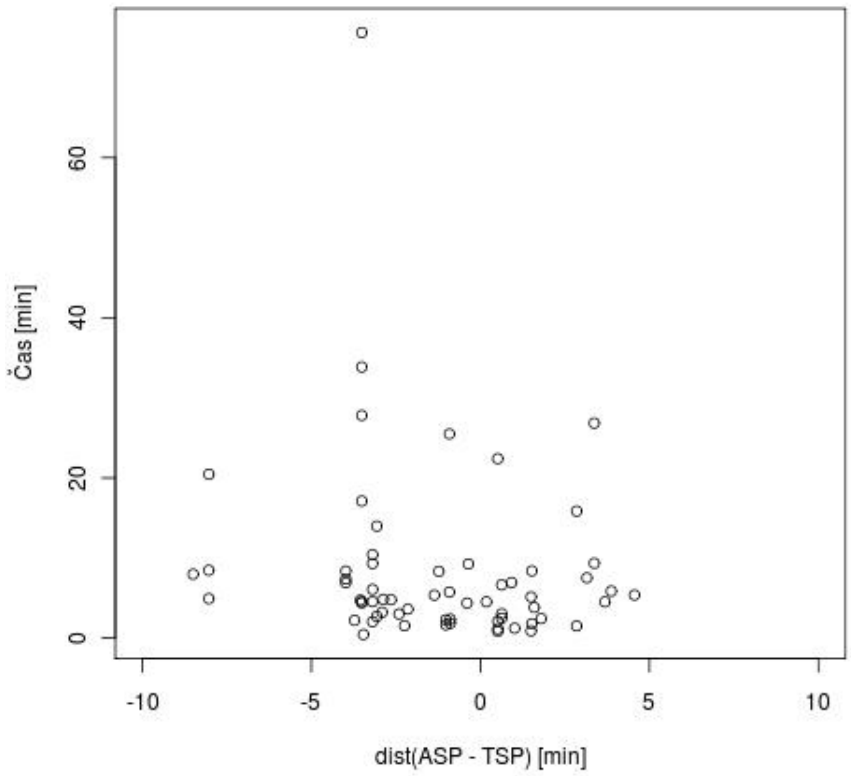
## Nenájdené začiatky

Zaujímavá je aj závislosť vzdialenosti bodov TSP a ASP a toho, že respondent začiatok označí za nenájdený. V okolí ASP sa síce musí vyskytovať ručne anotovaný začiatok témy, ale respondent môže napríklad pochopiť tému inak ako anotátor. Taktiež sa môže stať, že reálny začiatok je od ASP príliš ďaleko.

Počty nenájdených tém u jednotlivých nahrávok sú uvedené v tabuľke 5.2. Závislosť nenájdených začiatkov (a dĺžky ich hľadania) na vzdialenosť TSP a ASP je znázornená na obrázku 5.5. Časy hľadania u nenájdených nahrávok sú o niečo

| Počet nenájdených začiatkov u jednej nahrávky | Počet nahrávok |
|---|---|
| 5 | 3 |
| 4 | 1 |
| 3 | 9 |
| 2 | 2 |
| 1 | 24 |

Tabuľka 5.2: *Počty nenájdených začiatkov.*

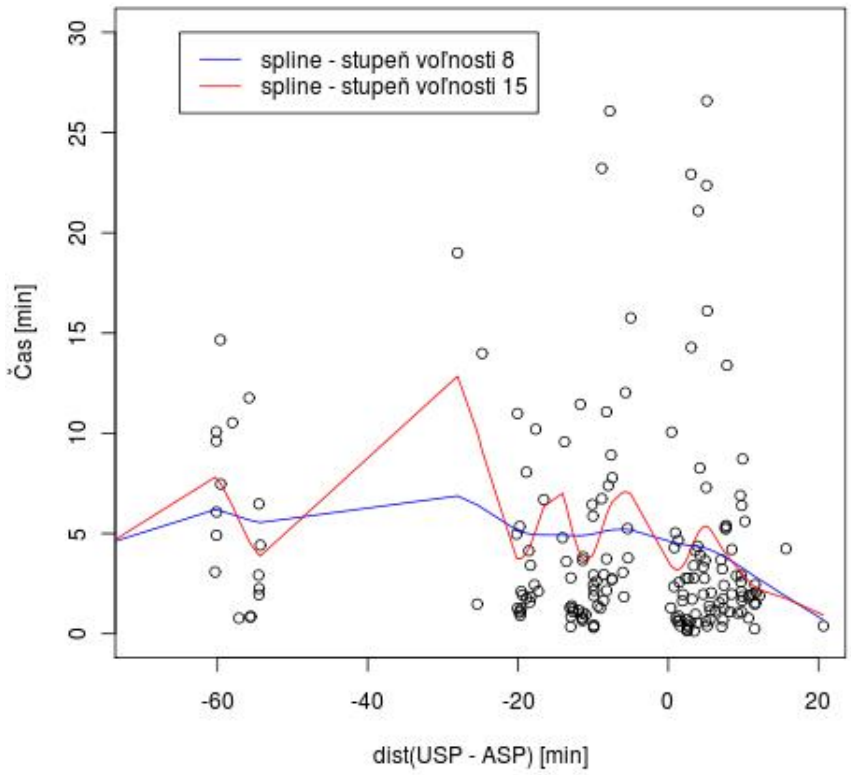



Obr. 5.5: *Závislosť rýchlosti označenia nenájdeného začiatku na vzdialenosti ASP a TSP. Vybraný je úsek, pre ktorý je vzdialenosť $|ASP - TSP|$ menšia ako 10 minút.*

vyššie ako časy u nájdených začiatkov. Nenájdené témy existujú pre všetky rôzne vzdialenosti od nulového bodu - teda aj pre prípad, že je vygenerovaný začiatok takmer zhodný s anotovaným začiatkom. Väčšie množstvo nenájdených začiatkov sa nachádza okolo 2 až 2,5 minúty od nulového bodu. To v tomto prípade súhlasí s maximami funkcie zodpovedajúcej nájdeným začiatkom v týchto bodoch.

## Vzdialenosť USP a ASP

V ďalšom prípade sme skúmali závislosť rýchlosti vyhľadania od vzdialenosti respondentom označeného bodu a automaticky vygenerovaného bodu. Tento prípad je zobrazený na obrázku 5.6.



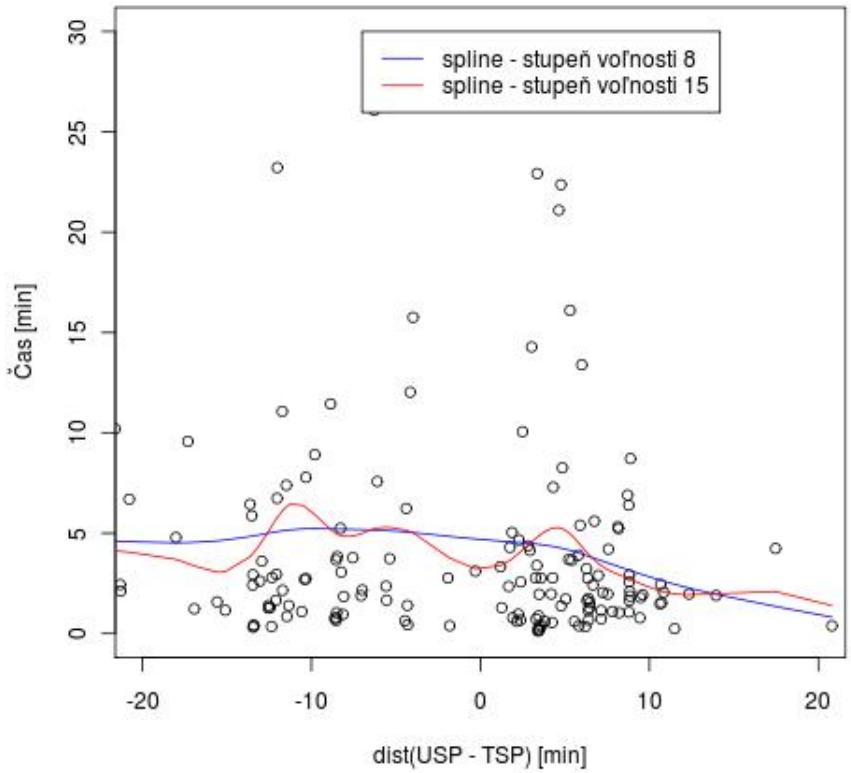

Obr. 5.6: *Závislosť rýchlosti vyhľadania na vzdialenosti ASP a USP. Zobrazený je úsek (-70; 20) min. Bodmi sú preložené spline funkcie s stupňom voľnosti 8 a 15.*

Zaujímavé na výsledkoch je, že v prípade kladných hodnôt respondenti umiestňujú začiatky maximálne 20 minút po vygenerovanom bode. V prípade záporných hodnôt však vyznačujú začiatky až do vzdialenosti približne 80 minút. Väčší rozptyl začiatkov môže znamenať menšiu istotu respondentov pri takomto označovaní bodov pred začiatkom prehrávanaia. Navyše, spline funkcia pre stupeň voľnosti 15 ukazuje veľké striedanie miním a maxím. Tie môžu byť spojené s tým, že čas, ktorý sa hovorca venuje jednotlivým témam v nahrávkach býva podobný. Aj na základe už uvedených získaných informácií potom môžeme predpokladať, že jednotlivé maximá zodpovedajú približne začiatkom tém, minimá úsekom, kde sa o nejakej téme rozpráva. Spline funkcia so stupňom voľnosti 8 naproti tomu skoro celý čas klesá. Respondenti teda v priemere rýchlejšie spracujú nahrávku, ak sa téma podľa nich začína pred vyznačeným začiatkom nahrávky (ten nemusí

zodpovedať anotovanému bodu) ako po ňom.

### Vzdialenosť USP a TSP

Pokúsili sme sa znázorniť aj závislosť rýchlosti vyhľadania od vzdialenosti respondentom označeného bodu a anotátorom označeného začiatku témy. Graf sa nachádza na obrázku 5.7.




Obr. 5.7: *Závislosť rýchlosti vyhľadania na vzdialenosti TSP a USP. Vybraný je úsek (-20; 20) min. Bodmi sú preložené spline funkcie s stupňom voľnosti 8 a 15.*

Výsledky v tomto prípade sú veľmi podobné predchádzajúcemu prípadu. Spline funkcia so stupňom voľnosti 15 tiež dosahuje striedavo maximá a minimá a spline funkcia s stupňom voľnosti 8 klesá, hoci o niečo pomalšie. Výsledky v záporných hodnotách majú tiež väčší rozptyl ako výsledky v kladných hodnotách.

## 5.4 Subjektívne hodnotenia responentov

Užívatelia mali možnosť ohodnotiť kvalitu automatického vyhľadávania, ktoré bolo v tomto prípade nahradené náhodne vybraným bodom v okolí skutočne

existujúceho začiatku témy. Počty rôznych hodnotení sú uvedené v tabuľke 5.3.

| Hodnotenie | Počet hodnotení | Boolean hodnota hodnotenia |
|---|---|---|
| Výborné | 67 | 1 |
| Dobré | 97 | 1 |
| Špatné | 25 | 0 |
| Velmi špatné | 2 | 0 |
| Nenájdený začiatok | 68 | 0 |
| Preskočená nahrávka | 34 | x |

Tabuľka 5.3: *Počty rôznych hodnotení.*

Hodnotenia sme binárne rozdelili. *Výborné* a *dobré* hodnotenia sme označili hodnotou 1. *Špatné*, *velmi špatné* a nenájdené prípady sme označili hodnotou 0. Preskočené nahrávky sme v tomto prípade nepoužili. Vzdialenosti ASP a TSP sme rovnomerne rozdelili do 30 sekundových úsekov. Pre každý z týchto úsekov sme spočítali strednú hodnotu binárnych ohodnotení. Závislosť týchto priemerných hodnôt na vzdialenosti ASP a TSP je zobrazená na obrázku 5.8.

Maximum funkcie sa nachádza v bode -4, ale v tomto bode máme iba dve hodnotenia, preto je táto situáci skreslená. Lokálne maximum sa nachádza v nulovom bode. Minimum sa nachádza približne v bode +4. Ďalšie lokálne minimum sa nachádza v bode -3. V záporných hodnotách klesá funkcia rýchlejšie ako v kladných hodnotách. Preto sú lepšie hodnotené tie prípady, keď sa vygenerovaný bod nachádza pred TSP. To zodpovedá získaným výsledkom v predchádzajúcich prípadoch.

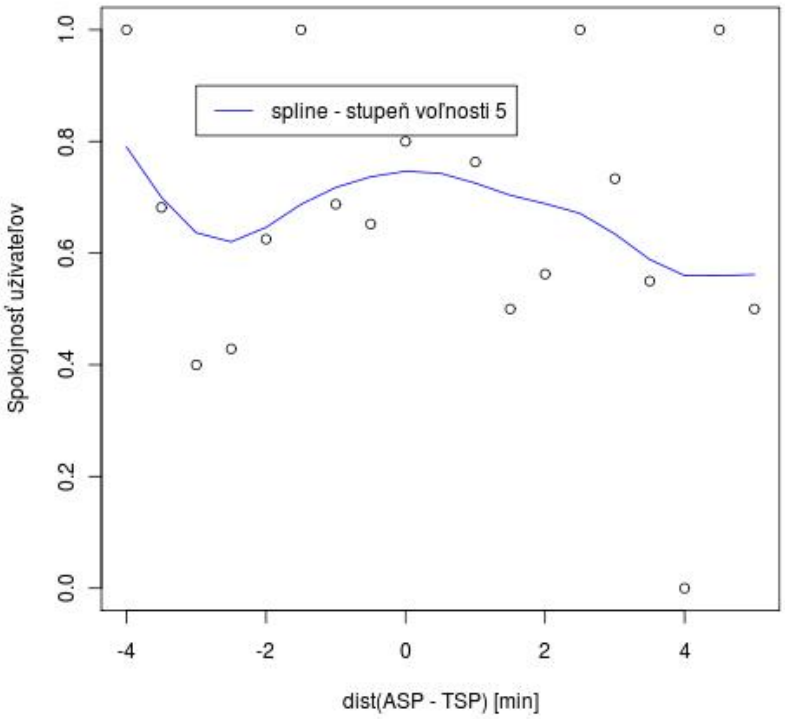
### Závislosť subjektívneho hodnotenia a rýchlosti vyhľadávania

Na obrázku 5.9 je uvedená závislosť subjektívneho hodnotenia na rýchlosti vyhľadávania.

Chceli sme overiť predpoklad, že čím rýchlejšie je respondent schopný vyhľadať začiatok témy, tým lepšie bude vyhľadávanie hodnotiť. Z obrázku to skutočne vyzerá tak, že táto hypotéza platí. *Výborné* hodnotenie je skutočne priradené najnižším časom, za ním nasleduje *dobré* hodnotenie, potom *velmi špatné* a *špatné*. *Velmi špatné* hodnotenie má síce o niečo menší medián ako špatné hodnotenie, to je však spôsobené tým, že hodnotenie veľmi špatné bolo použité len dvakrát.

## 5.5 Taktiky respondentov pri vyhľadávaní

Všetky akcie respondentov sa ukladali do logu. Formát logu je možné nájsť v prílohe D. Log bol ďalej spracovaný pomocou skriptu (v jazyku Perl) a so získanými dátami sa ďalej pracovalo v programe R. Niektoré štatistiky spracovania nahrávky sú uvedené v tabuľke 5.4.



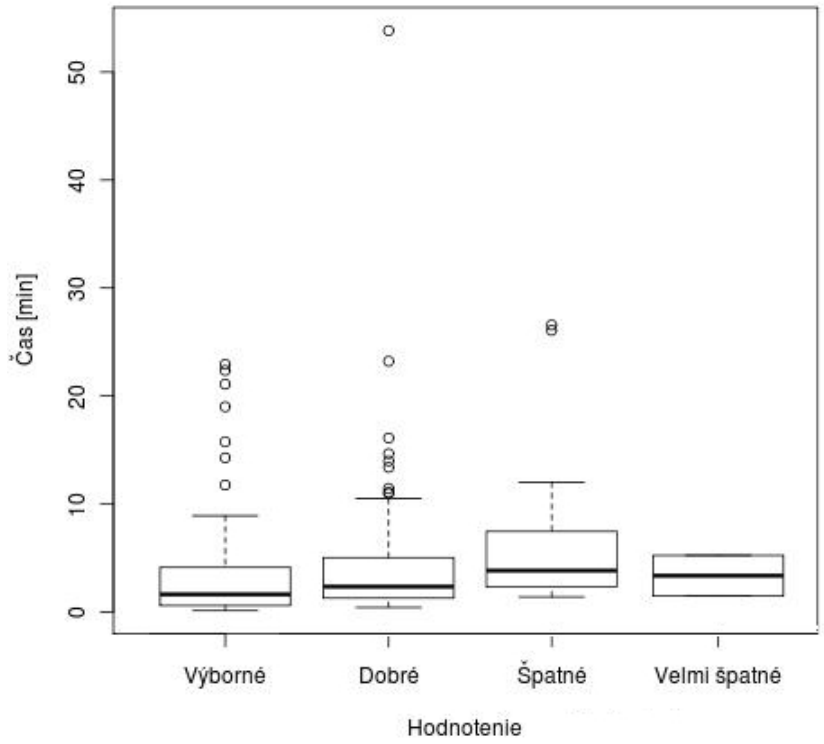

Obr. 5.8: *Závislosť subjektívneho hodnotenia respondentov na vzdialenosti bodov ASP a TSP. Vybraný je úsek (-5; 5) min. Bodmi je preložená spline funkcia s stupňom voľnosti 5. Vyššie hodnotenia v tomto prípade znamenajú lepšie priemerné hodnotenie respondnetov.*

Problémom, ktorý sme museli riešiť, bola neexistencia jednoznačných označení jednotlivých anotovaných bodov. Preto sme jednotlivé anotované body označovali kombináciou *id narávky / id témy*. V niektorých prípadoch však existovalo viac rovankých kombinácií *id narávky / id témy*. V takomto prípade sme ku *id nahrávky* pridávali písmená abecedy.

Zaznamenané akcie No Activity neboli pri vyhodnocovaní nakoniec brané do úvahy, pretože podľa odozvy respondentov dochádzalo k tejto situácii najmä vtedy, keď čakali dlhší čas na relevantný úsek. Navyše táto akcia nebola veľmi častá.

Dôležitú úlohu tiež hralo to, že anotátori, ktorí označovali relevantné úseky mali tému naštudovanú. Napríklad pri téme „Dětské umění v Terezíně“ sa objavila pasáž o umení. Anotátori vedeli na základe ďalších informácií posúdiť, že sa jednalo o detské umenie. Respondenti to však nemali často ako posúdiť, preto si neboli istí, či majú daný začiatok označiť a veľakrát ho ani neoznačili.

Zo získaných logov sme vybrali jednotlivé akcie, ktoré respondenti uskutočnili. V tabuľke 5.6 sú počty použití jednotlivý príkazov. Príkazy sú vysvetlené v tabuľke 5.5. Pri ťahaní slidera sme rozlišovali, či respondent ťahá slider smerom dopredu alebo dozadu.




Obr. 5.9: *Závislosť subjektívneho hodnotenia na rýchlosti vyhľadávania.*

Najčastejšie boli používané príkazy pre pohyb dopredu - častejší bol skok, následne potom ťahanie slidera. Popularita rýchleho skoku bola pre nás trochu prekvapujúca, pretože sme dlho zvažovali, či je vhodné ju do programu vôbec pridať. Vďaka tomu, že sú nahrávky pomerne dlhé, nie je slider pri presúvaní dostatočne presný. Preto respondenti rýchly skok tak často používajú. Pri pohybe dozadu je však, naopak o niečo populárnejší pohyb pomocou slidera.

Vyhľadávanie najčastejšie skončí tak, že respondent začiatok nájde, druhým prípadom je potom že respondent začiatok nenájde. V ďalších prípadoch respondent preskočí na ďalšiu nahrávku, program uprostred prehrávania ukončí, prípadne sa odhlási.

Z logu sme ďalej získali jednotlivé trojice po sebe idúcich príkazov. Tie sú uvedené v tabuľke 5.7. Tento prehľad by nám mal podať informácie o tom, ako respondenti pri prieskume postupovali.

Najčastejšou trojicou sú tri po sebe idúce rýchle skoky vpred. Druhým najčastejším je ťahanie slidera smerom dopredu. Respondenti teda používali buď ťahanie slidera alebo rýchly skok a menej ich už kombinovali. Tretia trojica príkazov reprezentuje spustenie nahrávky od jej stredu. Až štvrtou trojicou je viacnásobný rýchly skok vzad. Táto trojica má takmer štyrikrát menší výskyt

| | |
|---|---|
| Najrýchlejšie určený začiatok | 7,76 s |
| Najpomalšie určeny začiatok | 53,8 min |
| Priemerný čas potrebný na vyhľadanie začiatku | 5,47 min |

Tabuľka 5.4: *Štatistiky spracovania nahrávok.*

| Príkaz | Vysvetlenie |
|---|---|
| Play | Spustenie prehrávania od ASP |
| Pause | Pozastavenie nahrávky |
| Stop | Zastavenie prehrávania |
| FF | Rýchly skok dopredu |
| FB | Rýchly skok dozadu |
| DragF | Potiahnutie slidera dopredu |
| DragB | Potiahnutie slidera dozadu |
| FoundStart | Relevantný začiatok nájdený |
| NotFoundStart | Relevantný začiatok nenájdený |
| Next | Skok na ďalšiu nahrávku |
| TabChange | Vybratie záložky |
| Logout | Odhlásenie sa (uprostred prehrávania) |
| NoActivity | Žiadna zaznamenaná akcia počas piatich minút |
| NA | Žiadna predchádzajúca akcia |

Tabuľka 5.5: *Vysvetlenie jednotlivých príkazov.*

| Príkaz | Počet použití | Relatívny počet použití |
|---|---|---|
| FF | 765 | 0,27 |
| DragF | 517 | 0,18 |
| DragB | 432 | 0,15 |
| Play | 329 | 0,12 |
| FB | 236 | 0,08 |
| FoundStart | 187 | 0,07 |
| TabChange | 110 | 0,04 |
| NotFoundStart | 68 | 0,02 |
| Pause | 55 | 0,02 |
| Stop | 43 | 0,02 |
| Next | 34 | 0,01 |
| Close | 22 | 0,01 |
| NoActvity | 17 | 0,01 |
| Logout | 12 | 0,00 |

Tabuľka 5.6: *Počty použitých príkazov.*

ako skoky vpred. Respondenti teda skok dopredu preferujú, hoci pravdepodobnosť, že začiatok leží pred alebo po červenom bode je rovnaká. Skok vpred je ale

| Postupnosť príkazov | Počet použití |
|---|---|
| FF/FF/FF | 576 |
| DragF/DragF/DragF | 293 |
| NA/NA/Play | 223 |
| FB/FB/FB | 145 |
| DragB/DragB/DragB | 67 |
| NA/Play/DragB | 66 |
| DragB/DragF/DragF | 53 |
| DragF/DragF/DragB | 53 |
| NA/Play/FoundStart | 52 |
| TabChange/TabChange/TabChange | 38 |
| Play/DragB/DragB | 38 |
| DragB/DragB/DragF | 38 |
| DragF/DragB/DragB | 34 |
| DragB/DragB/FoundStart | 30 |
| NA/NA/DragB | 29 |
| DragB/FF/FF | 28 |
| DragF/DragB/DragF | 24 |
| Play/DragF/DragF | 22 |
| NA/Play/FB | 21 |
| Play/FB/FB | 20 |
| FF/FF/DragB | 20 |

Tabuľka 5.7: *Počty použití trojíc príkazov - najčastejšie použité trojice.*

prirodzenejší pretože lepšie zodpovedá prirodzenej povahe reči. Ďalšou akciou je ťahanie slidera vzad. Podobne, ako v prípade pohybu dopredu je preferovanejšia sekvencia rýchlych skokov pred viacnásobným ťahaním slidera.

Ďalšou v poradí je trojica NA/Play/DragB, ktorá reprezentuje to, že respondent spustí nahrávku a presunie sa o niečo dozadu. To znamená, že predpokladá, že sa už o téme hovorí, prípadne sa už hovorilo. Ďalšími trojicami príkazov sú kombinácie ťahania slideru dopredu a dozadu. V týchto prípadoch by sme mohli predpokladať, že ich respondent skôr používa, keď približne vie, kde leží začiatok a potrebuje ho nájsť presne. Druhým možným dôvodom môže byť to, že respondent vôbec nevie, kde môže začiatok ležať a potrebuje sa presunúť na väčšiu vzdialenosť ako mu dovoľuje rýchly skok. Trojica NA/Play/FoundStart reprezentuje situáciu, keď respondent spustí nahrávku, nechá ju bežať, a tak nájde začiatok. Častou trojicou je aj TabChange/TabChange/TabChange. Tá ukazuje na to, že keď si respondent chce prezrieť informácie o programe, pozrie si väčšinou viac záložiek naraz. Ďalším prípadom je Play/DragB/DragB, keď respondent spustí nahrávku a postupne sa presúva dozadu.

Ďalšími zaujímavým príkladom je DragB/DragB/FoundStart, z ktorého by sme mohli predpokladať, že začiatok bol najčastejšie nájdený tak, že respondent ťahal slider nejaký čas smerom dozadu. Prípad NA/NA/DragB zas nastával vtedy, keď sa respondent ešte pred spustením prehrávania chcel presunúť pred

červený bod, teda pred bod, v ktorom sa má prehrávanie spustiť. Respondent sa v tomto prípade chce vyhnúť tomu, aby sa po začatí prerávania musel vracať späť.

# 6. Návrh nového spôsobu na evaluáciu vyhľadávania v hovorenej reči

## 6.1 Návrh penalizačnej funkcie

Návrh novej penalizačnej funkcie vychádza zo zistení získaných pri užívateľskom prieskume. Do úvahy sa berie najmä čas potrebný na nájdenie počiatočného bodu témy v závislosti na vzdialenosti vygenerovaného bodu (ASP) od anotovaného bodu (TSP), teda situácia, ktorá je znázornená na obrázku 5.4. Tiež sa berú do úvahy subjektívne hodnotenia, ktorých popis je na obrázku 5.8 a funkcie, ktoré znázorňujú závislosť rýchlosti nájdeného začiatku na vzdialenosti USP a TSP (obrázok 5.7) a USP a ASP (obrázok 5.6). Navrhnutá funkcia bude mať podobu závislosti penalizačnej hodnoty na vzdialenosti vyhľadaného bodu od skutočného začiatku témy. Zároveň by mala byť funkcia dostatočne jednoduchá a dobre popísateľná, aby ju bolo možné použiť pri automatickom vyhodnocovaní. Pri návrhu sa berú do úvahy nasledujúce body:

1. Minimum funkcie na obrázku 5.4 sa nachádza okolo nulového bodu, teda približne v bodoch -1 a +1.
2. Maximá funkcie na obrázku 5.4 sa nachádzajú približne v bodoch -2,5 a +2,5.
3. Funkcie na obrázku 5.7 a obrázku 5.6 klesajú a dosahujú nižšie hodnoty v oblastiach, keď sa bod USP nachádza po bodoch ASP a TSP. Táto situácia by mala byť v novo vytvorenej penalizačnej funkcie preferovanejšia.
4. Pri približovaní do nekonečna by mala funkcia postupne rásť, hoci to z grafov nie je zrejmé. Je to ale spôsobené malým počtom hodnotení v týchto oblastiach.

Praktickejšie je však penalizačnú funkciu navrhnúť tak, aby jej maximum bolo tam, kde je začiatok vyhľadaný najpresnejšie. A naopak minimum tam, kde je najmenej vhodné začiatok nájsť. Od určitých bodov, ktoré sú príliš vzdialené od anotovaného bodu je potom penalizačná funkcia nulová. V dôsledku toho sa vymenia minimá a maximá popísanej funckie.

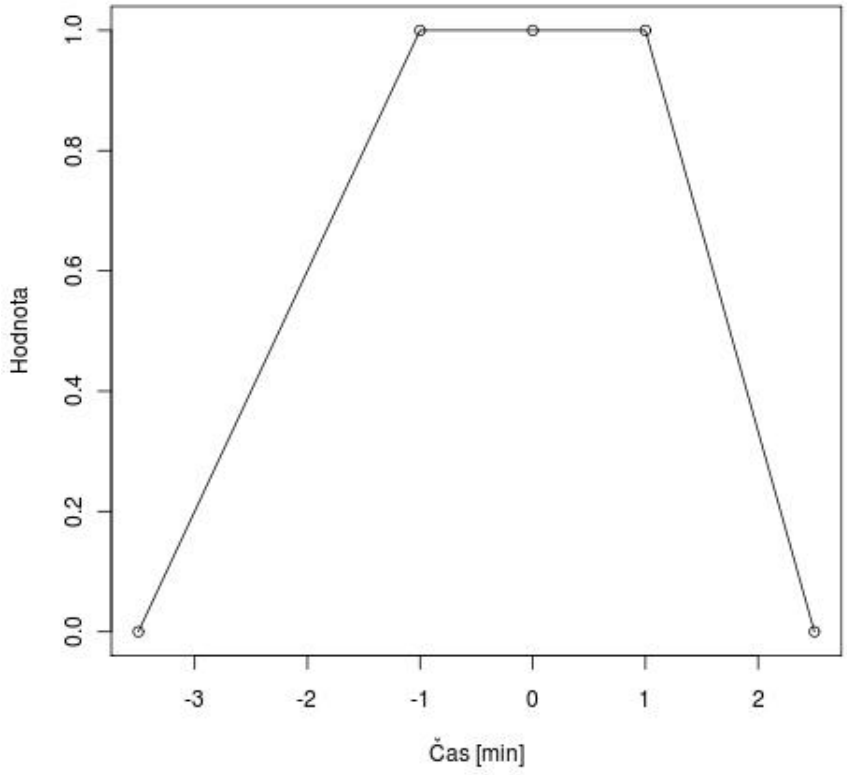
Na základe týchto požiadaviek sme navrhli body, ktorými musí výsledná penalizačná funkcia prechádzať.

1. Maximum funkcie stanovíme rovné 1. Maximum by malo ležať v bode, kde je vyznačený anotovaný začiatok (teda v bode 0). Podľa prieskumu však rovnako dobre vychádza blízke okolie bodu 0

Takto vytvoríme tri body, ktorými by mala prechádzať penalizačná funkcia: (-1; 1), (0; 1) a (+1; 1).

2. Minimá funkcii môžeme stanoviť v bodoch -2,5 a +2,5. Keďže preferujeme záporné hodnoty, posunieme minimum v záporných hodnotách do bodu -3,5. Hodnota funkcie v minimách je nulová. Dostaneme tak body (-3,5; 0) a (2,5; 0).

Nakoniec týmito bodmi preložíme spojnicu. Táto krivka vyjadruje výslednú penalizačnú funkciu. Je znázornená na obrázku 6.1.




Obr. 6.1: *Návrh výslednej penalizačnej funkcie.*

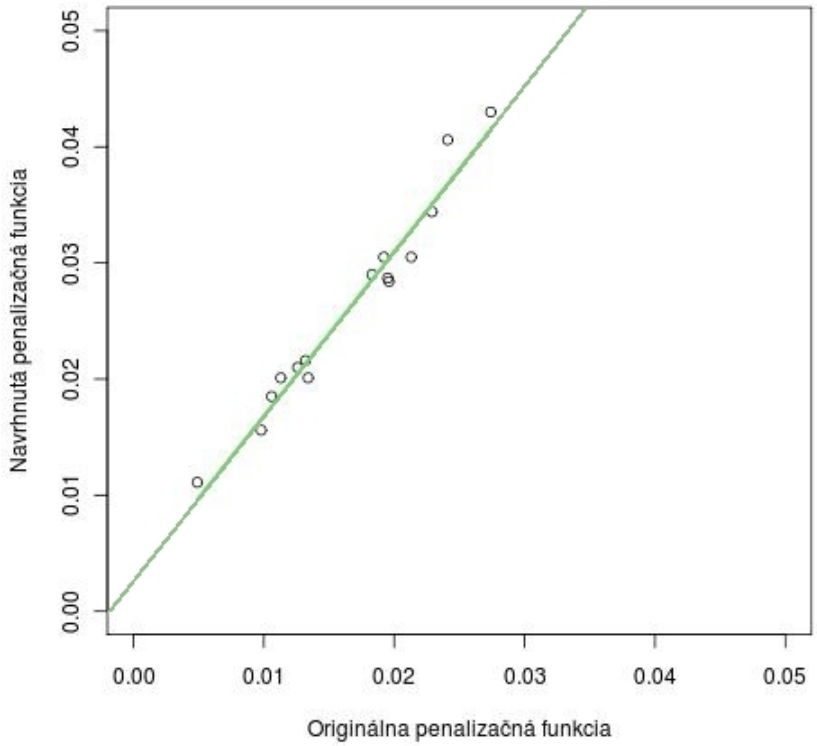
## 6.2 Porovnanie s mGAP

Vytvorenú penalizačnú funkciu nakoniec porovnáme s penalizačnou funkciou navrhnutou v [17]. Na toto porovnanie použijeme dáta získané v rámci projektu CLEF [11]. K dispozícii máme ručne anotované body a body vyhľadané pomocou rôznych IR systémov vytvorených na rôznych univerzitách. Spolu máme k dispozícii výstupy od 15 rôznych IR systémov. Z týchto výsledkov spočítame hodnoty

mGAP pre obe penalizačné funkcie. Pri oboch funkciách bola časová os diskretizovaná na 15 sekúnd dlhé úseky.

Pôvodná penalizačná funkcia má tvar trojuholníka. Táto penalizačná funkcia je implementovaná tak, že každých 15 sekúnd sa zhorší hodnotenie systému o hodnotu 0,1. Ak sa vyhľadávací nástroj úplne zhoduje s anotovaným bodom, potom má funkcia hodnotu 1, ak vyhľadávací nástroj nájde začiatok viac ako 150 sekúnd od anotovaného bodu, potom je hodnota penalizačnej funkcie nulová. Funkcia je znázornená na obrázku 2.2.

Pôvodná penalizačná funkcia je prísnejšia ako nami navrhnutá penalizačná funkcia. Napríklad, ak je vyhľadaný bod vzdialený od anotovaného bodu 60 sekúnd, potom má podľa pôvodnej penalizačnej funkcie hodnotu 0,6, podľa navrhnutej penalizačnej funkcie má hodnotu 1. Preto sú hodnoty získané pomocou našej penalizačnej funkcie v priemere vyššie. Korelácia výsledkov pre rôzne penalizačne funkcie je znázornená na obrázku 6.2. V tabuľke 6.1 sú uvedené jednotlivé hodnoty mGAP pre obe penalizačné funkcie pre rôzne vyhľadávacie systémy. Brown Univeristy a Karlova univerzita použili IR systém Indri, Západočeská univerzita použila systém Lemur a Univerity of Chicago použila systém InQuery [20].




Obr. 6.2: *Korelácia systémov používajúcich rôznu penalizačnú funkciu. Zelená čiara vyjadruje najvyššiu koreláciu [23].*

| IR systém | Univerzita | Pôvodná penalizačná funkcia | Nová penalizačná funkcia |
|---|---|---|---|
| brown.f | Brown University | 0,0049 | 0,0111 |
| prague03 | Karlova univerzita | 0,0098 | 0,0156 |
| brown.sA.f | Brown University | 0,0106 | 0,0185 |
| brown.s.f | Brown University | 0,0113 | 0,0201 |
| UCunstTD3 | University of Chicago | 0,0126 | 0,0210 |
| UWB_2-1_td_w | Západočeská univerzita | 0,0132 | 0,0216 |
| UWB_3-1_td_l | Západočeská univerzita | 0,0134 | 0,0201 |
| prague02 | Karlova univerzita | 0,0183 | 0,0290 |
| prague01 | Karlova univerzita | 0,0192 | 0,0305 |
| prague04 | Karlova univerzita | 0,0195 | 0,0287 |
| UCcslTD1 | Univeristy of Chicago | 0,0196 | 0,0284 |
| UCcsaTD2 | University of Chicago | 0,0213 | 0,0305 |
| UWB_2-1_td_s | Západočeská univerzita | 0,0229 | 0,0344 |
| UWB_3-1_tdn_l | Západočeská univerzita | 0,0241 | 0,0406 |
| UWB_2-1_tdn_l | Západočeská univerzita | 0,0274 | 0,0430 |

Tabuľka 6.1: *Hodnoty mGAP pre rôzne penalizačné funkcie pre niekoľko IR systémov.*

Nakoniec spočítame hodnotu Kendallovho korelačného koeficient $\tau$, ktorý vyjadruje vzťah medzi dvoma množinami ohodnotených dát [22]. Táto hodnota je 0,856, teda korelácia oboch penalizačných funkcií je pomerne vysoká. Na základe tejto korelácie môžme vidieť, že pôvodný návrh penalizačnej funkcie v [14] je pomerne dobrý a dobre zodpovedá tomu, ako reálne ľudia posudzujú IR systémy na vyhľadávanie v hovorenej reči. Tiež z obrázku 2.2 je zrejmé, že zmena poradia dvoch systémov pri ohodnotení pomocou rôznych penalizačných funkcií nastáva len v niekoľkých prípadoch a rozdiely v hodnoteniach nie sú signifikantné.

Rozdiel je v tom, že ľudia sú ochotní hľadať skutočný začiatok v širšom okolí vyhľadaného bodu, ako sme sa pôvodne domnievali. Navyše sme zistili, že ak je vyhľadaný bod vzdialený nanajvýš minútu od anotovaného bodu, nemá to vplyv na kvalitu vyhľadávania.

# Záver

V práci sme najskôr popísali techniky a jednotlivé systémy, ktoré sa v súčasnosti používajú na vyhľadávanie v hovorenej reči. Následne sme rozobrali jednotlivé metriky, ktoré umožňujú vyhodnocovanie tohto vyhľadávania. Posúdili sme ich výhody a nevýhody. Snažili sme sa pritom zamerať na oblasť nesegmentovantovaných nahrávok, ktorá je pomerne málo preskúmaná. V rámci tejto oblasti sme analyzovali najmä metriku mGAP. Chceli sme overiť, či je použitie tejto metriky v reálnych podmienkach opodstatnené a v prípade, že by boli nájdené nedostatky tejto metriky, navrhnúť lepšie riešenie.

V rámci overenia adekvátnosti metriky mGAP sme uskutočnili užívateľský prieskum. Pri prieskume sme zisťovali, ako môžu respondenti vnímať rôzne IR systémy. Namiesto reálnych IR systémov sme však použili ich simuláciu. Skúmali sme najmä závislosť času a spokojnosti respondentov vzhľadom na vzdialenosť a smer bodu vyhľadaného IR systémom a skutočného začiatku témy. Do prieskumu sa zapojilo 24 respondentov, pričom každý z týchto respondentov strávil v priemere pri prieskume takmer hodinu. Vzhľadom na to, že respondenti boli dobrovolníci, ktorí neboli finančne ohodnotení, je to pomerne dosť. Získaný počet anotovaných bodov je dostatočný, aj keď pri väčšom počte anotácií by boli výsledky presnejšie.

Na základe prieskumu sme penalizačnú funkciu použitú v metrike mGAP mierne upravili. Zistili sme, že pokiaľ je vyhľadaný bod vzdialený najviac jednu minútu od skutočného začiatku relevantného úseku, neovlpyvňuje to hodnotenie respondentov. Tiež sme zistili, že respondenti uprednostňujú situáciu, keď sa vyhľadaný bod nachádza pred relevantným úsekom pred situáciou, keď sa vyhľadaný bod nachádza až po relevantnom začiatku úseku. Ďalej sme zistili, že užívatelia sú schopní začiatok relevantného úseku označiť aj vtedy, ak je vyhľadaný pomerne ďaleko od skutočného začiatku relevantného úseku.

Nakoniec sme porovnali pôvodnú a nami vytvorenú penalizačnú funkciu. Korelácia výsledkov rôznych IR systémov je pre tieto dve penalizačné funkcie vysoká. To znamená, že pôvodná penalizačná funkcia navrhnutá v mGAP je praxi dobre použitelná a výsledky získané pomocou metriky mGAP by mali korelovať s vnímaním užívateľov IR systémov.

V práci sme tiež popísali postupy, ktoré užívatelia používajú pri vyhľadávaní relevantného úseku v nahrávke. Získané poznatky by tak mohli byť užitočné nielen pri úprave penalizačnej funkcie v metrike mGAP, ale mohli by byť aj ďalej použité pri návrhu systémov pre vyhľadávanie v hovorenej reči.

# Literatúra


[1] ACHANTA, R. et al. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, s. 1597–1604, Miami Beach, Florida, USA, 2009.

[2] ARONS, B. SpeechSkimmer: a system for interactively skimming recorded speech. 4, s. 3–38, New York, NY, USA, 1997. ACM.

[3] BLANKEN, H. M. et al. *Multimedia Retrieval*. Data Centric Systems and Applications. Springer Berlin Heidelberg, 2007. ISBN 3642091997.

[4] BUCKLEY, C. – VOORHEES, E. M. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, s. 33–40, New York, NY, USA, 2000. ACM.

[5] CHOTIMONGKOL, A. et al. Toward Benchmarking a General-domain Thai LVCSR System. In *Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI-CON)*, s. 1080–1084, Chiang Mai, Thailand, 2010.

[6] CLEVERDON, C. W. – MILLS, J. – KEEN, M. Factors determining the performance of indexing systems. 1966, 2, Test results.

[7] FRAKES, W. B. – BAEZA-YATES, R. A. (Ed.). *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall, 1992. ISBN 9780134638379.

[8] GLAVITSCH, U. – SCHÄUBLE, P. A system for retrieving speech documents. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, s. 168–176, New York, NY, USA, 1992. ACM.

[9] HEEREN, W. F. L. et al. Easy Listening: Spoken Document Retrieval in CHoral. *Interdisciplinary Science Reviews*. 2009, 34, 2-3, s. 236–252.

[10] HOFFMANNOVÁ, P. Malach(Multilingual Access to Large Spoken Archives). *Rukopis*. 2007.

[11] IRCING, P. et al. Information retrieval test collection for searching spontaneous Czech speech. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, s. 439–446, Berlin, Heidelberg, 2007. Springer-Verlag.

[12] JAMES, A. Perspectives on Information Retrieval and Speech. In *Information Retrieval Techniques for Speech Applications*, s. 1–10, London, UK, 2002. Springer-Verlag.

[13] JENSEN, A. R. Individual differences in visual and auditory memory. *Journal of Educational Psychology*. 1971, 62, 2, s. 123–131.

[14] KEKÄLÄINEN, J. – JÄRVELIN, K. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*. 2002, 53, s. 1120–1129.

[15] KOSINSKI, R. J. A Literature Review on Reaction Time, 2008. [Online] `http://biae.clemson.edu/bpc/bp/Lab/110/reaction.htm`.

[16] LANCASTER, F. W. *Information retrieval systems: characteristics, testing and evaluation*. John Wiley & Sons, 1978. ISBN 0471512400.

[17] LIU, B. – OARD, D. W. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, s. 673–674, New York, NY, USA, 2006. ACM.

[18] MANNING, C. D. – RAGHAVAN, P. – SCHŰTZE, H. *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press, 2008. ISBN 0521865719.

[19] OARD, D. Speech Retrieval Defined, 1997. [Online] `http://terpconnect.umd.edu/~dlrg/speech/speechretrieval_definition.html`.

[20] PECINA, P. et al. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. s. 674–686, Berlin, Heidelberg, 2008. Springer-Verlag.

[21] PENNEY, T. B. – GIBBON, J. – MECK, W. H. Differential Effects of Auditory and Visual Signals on Clock Speed and Temporal Memory. *Journal of Experimental Psychology: Human Perception and Performance*. 2000, 26, 6, s. 1770–1787.

[22] SHESKIN, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 4 edition, 2007. ISBN 1584888148.

[23] THOM, J. A. – SCHOLER, F. A Comparison of Evaluation Measures Given How Users Perform on Search Tasks. In *Proceedings of the Australasian Document Computing Symposium*, s. 100–103, Melbourne, Australia, 2007. RMIT University.

[24] RIJSBERGEN, C. J. *Information Retrieval*. London : Butterworths, 2 edition, 1979. ISBN 0408709294.

[25] VOGT, T. – ANDRÉ, E. Improving automatic emotion recognition from speech via gender differentiation. In *Language Resources and Evaluation Conference (LREC) 2006*, s. 1123–1126, Genoa, Italy, 2006. ELRA.

[26] VOORHEES, E. M. – HARMAN, D. Overview of the Seventh Text REtrieval Conference TREC-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC)-7*, s. 1–24. NIST Special Publication, 1999.

[27] WITBROCK, M. J. – HAUPTMANN, A. G. Speech recognition and information retrieval: experiments in retrieving spoken documents. In *Proceedings of the DARPA Speech REcognition Workshop*, s. 160–164, Chantilly, Virginia, USA, 1997. NIST Special Publication.

[28] WITTEN, I. B. – KNUDSEN, E. I. Why Seeing Is Believing: Merging Auditory and Visual Worlds. *Neuron*. 2005, 48, 3, s. 489–496.

# Zoznam tabuliek

# Zoznam použitých skratiek

| Skratka | Vysvetlenie | Zavedenie |
|---|---|---|
| AP | Average Precision, metrika používaná pri vyhodnocovaní IR systémov | 2.4 |
| ASP | Automatic Start Point, bod ktorý bol náhodne vygenerovaný v okolí bodu TSP | 3.3 |
| IR | Information Retrieval | 1.1 |
| MAP | Mean Average Precision, metrika používaná pri vyhodnocovaní IR systémov | 2.4 |
| mGAP | mean Generalized Average Precision, metrika používaná pri vyhodnocovaní IR systémov pre nesegmentovanú hovorenú reč | 2.5 |
| TSP | True Start Point, anotovaný začiatok témy | 3.3 |
| USP | User Start Point, respondentom označený začiatok témy | 3.3 |

# A. Pokyny pre užívateľov

Děkujeme, že jste si našli čas a účastníte se tohoto průzkumu. Jeho cílem je zlepšit kvalitu metod pro automatické vyhledávání v mluvené řeči.

V rámci průzkumu pracujete se zvukovými nahrávkami českých výpovědí svědků holocaustu. Nahrávky pocházejí z mezinárodního projektu Malach, jehož cílem je zpřístupnit rozsáhlý archiv těchto nahrávek široké veřejnosti včetně možnosti vyhledávání tématicky relevantních pasáží.

Praktickým příkladem může být situace, kdy učitel dějepisu chce v archivu najít informace o umění, které vzniklo během holocaustu, a konkrétní pasáže potom prezentovat žákům v hodinách dějepisu. Učitel popíše téma, které chce vyhledat, zadá jej do vyhledávacího systému a ten automaticky identifikuje pasáže, které jsou pro toto téma relevantní. Ne všechny takové odpovědi budou ovšem správné a přesné. Vaším úkolem bude ohodnotit přesnost a správnost takových výsledků pro předem určená témata. Naším cílem je zjistit, jak by učitel v naší příkladové situaci byl s výsledky svého vyhledávání spokojen.

Na začátku práce budete seznámeni s náhodně vybraným tématem, které souvisí s tématikou holocaustu, např. "Dětské umění v Terezíně", "Kolaborace místních obyvatel"nebo "Židovské děti ve školách". Každé téma bude popsáno natolik detailně, aby bylo možno jednoznačně identifikovat pasáže nahrávek, které jsou relevantní. Než začnete pracovat, tak si dané téma pozorně prostudujte.

Dále vám bude předložena nahrávka nějaké výpovědi (včetně jména a fotografie mluvčího) a konkrétní okamžik, který byl automaticky identifikován jako začátek pasáže, kde se o daném tématu hovoří. Tento okamžik je označen na časové ose červeným bodem. Kliknutím na tlačítko ">"spustíte přehrávání od tohoto místa. Vaším úkolem je pozorně poslouchat nahrávku a v okamžiku, kdy poznáte, že se o daném tématu začalo mluvit, stisknout tlačítko "Nalezeno".

Toto místo nemusí být shodné s označeným červeným bodem. Relevantní pasáž může začít později nebo i dříve, než je indikováno červeným bodem na časové ose. V okně, které se Vám po stisknutí tlačítka "Nalezeno"zobrazí, ohodnoťte jak jste byli spokojeni s automaticky vyhledaným začátkem tématu. Vyberte právě jedno z uvedených hodnocení tak, aby podle Vás nejlépe vystihovalo kvalitu automatického vyhledávání. Klikněte na jedno ze čtyř hodnocení: *Výborné*, *Dobré*, *Špatné* nebo *Velmi špatné* a stiskněte tlačítko Ok.

Přehrávání můžete pozastavit tlačítkem "||". Pokud se nahrávka přehrává, pak tlačítkem »"spustíte znovu přehrávání od červeného bodu. V nahrávce se můžete libovolně pohybovat posunováním ukazatele přehrávání na časové ose, tak, abyste co nejrychleji našli relevantní úsek. Tlačítky «¡ä »¿"se můžete v nahrávce pohybovat zrychleně. V případě, že jste relevantní úsek nenalezli, můžete přejít k další nahrávce, tlačítkem "Nenalezeno". Pokud chcete nahrávku přeskočit z jakéhokoliv jiného důvodu, můžete to udělat stisknutím tlačítka "> |". K již zpracované nahrávce není možné se vrátit.

Pracovat můžete libovolně dlouho, je však vhodné, abyste pracovali nejméně 15 minut. Po odhlášení a opětovném přihlášení se Vám jako první zobrazí nahrávka, kterou jste jako první nezpracovali.

Při práci se pokuste vžít do role učitele z našeho příkladu výše. I on musí projít nalezené úseky jeden po druhém, spustit přehrávání a ověřit, jestli je opravdu relevantní nebo ne. V některých případech bude s vyhledáváním spokojen a relevantní pasáž začne téměř okamžitě, jindy bude muset chvíli čekat, případně se v přehrávání o několik okamžiků vrátit.

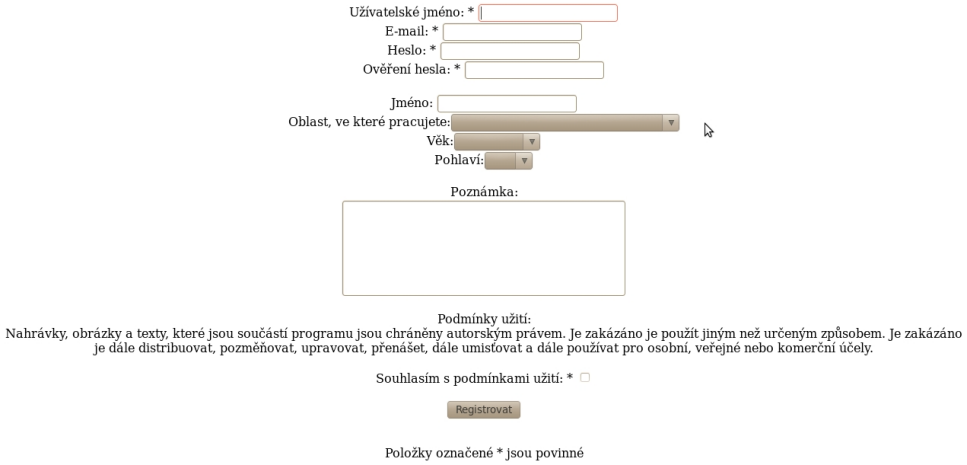
# B. Ukážka playlistu


```
−<track>
  <id>13876</id>
  <speaker>Marie Sandová</speaker>
  <birthyear>1916</birthyear>
  <birthplace>Austria-Hungary</birthplace>
 −<data>
    http://guest:fN2m5dhw@ufallab.ms.mff.cuni.cz/~galuscakova/player/assets/sounds/13876_MP3WRAP.mp3
  </data>
 −<image>
    http://guest:fN2m5dhw@ufallab.ms.mff.cuni.cz/~galuscakova/player/assets/images/13876.jpg
  </image>
  <length>6870540.0</length>
  <starts>2684673.86536</starts>
  <title>Dětské umění v Terezíně</title>
 −<about>
    Hledáme popis uměleckých aktivit dětí v Terezíně, jako např. hudby, divadla, malování, poezie a jiných psaných děl. 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.
  </about>
 </track>
−<track>
  <id>27065</id>
  <speaker>Gerda Pavlíková</speaker>
  <birthyear>1918</birthyear>
  <birthplace>Moravská Ostrava (Czechoslovakia)</birthplace>
 −<data>
    http://guest:fN2m5dhw@ufallab.ms.mff.cuni.cz/~galuscakova/player/assets/sounds/27065_MP3WRAP.mp3
  </data>
 −<image>
    http://guest:fN2m5dhw@ufallab.ms.mff.cuni.cz/~galuscakova/player/assets/images/27065.jpg
  </image>
  <length>4656300.0</length>
  <starts>2621064.82686</starts>
  <title>Dětské umění v Terezíně</title>
```


**Kde:**

- *track* - nahrávky zodpovedajúce jednotlivým ASP
- *id* - id nahrávky
- *speaker* - hovorca nahrávky
- *birthyear* - rok narodenia hovorcu
- *birthplace* - miesto narodenia hovorcu
- *data* - odkaz na audio nahrávku
- *image* - odkaz na fotografiu hovorcu
- *length* - dĺžka nahrávky (v ms)
- *starts* - začiatok relevantného úseku (v ms)
- *about* - popis témy

# C. Registračný formulár



# D. Ukážka logu

*22495a*
*Dětské umění v Terezíně*
*majakukova*
*1.01*
*Mon Mar 7 13:38:55 GMT+0100 2011*
*Login*

———-
*22495a*
*Dětské umění v Terezíně*
*majakukova*
*1.01*
*Mon Mar 7 13:39:31 GMT+0100 2011*
*** DragF 46.657830510413206 *** DragB 46.174698795180724*
*** Play 0 46.41566265060241 *** FoundStart 15625 46.870279443995216*
*** Stop 15625 46.870279443995216 *** Next 22547 46.870279443995216*
*22547*
*46.870279443995216*
*Výborné*
———-
*16024*
*Dětské umění v Terezíně*
*majakukova*
*1.01*
*Mon Mar 7 13:45:48 GMT+0100 2011*
*** Play 0 48.18453882030029 *** TabChange 160437 1*
*** TabChange 177734 0 *** Drag 50.388856652669595 223953 40.704819277108435*
*224531*
*** Drag 40.80962604592195 234547 42.825301204819276 235250*
*** Drag 43.10061212552036 263078 44.873493975903614 263859*
*** Drag 45.42594532065508 320140 46.945783132530120 320656*
*** Drag 47.00988767647119 326609 46.626506024096386 327187*
*** Drag 46.65448179558439 329359 45.6566265060241 329843*
*** FoundStart 359734 45.95716922977213 *** Stop 359734 45.95716922977213*
*** Next 364297 45.95716922977213*
*364297*
*45.95716922977213*
*Dobré*

**Kde:**

- *22495a* - Id nahrávky
- *Dětské umění v Terezíně* - Téma
- *majakukova* - Užívateľský login
- *1.01* - Použitá verzia programu
- *Mon Mar 7 13:39:31 GMT+0100 2011* - Čas akcie
- **** DragF 46.657830510413206 *** DragB 46.174698795180724 *** Play 0 46.41566265060241 *** FoundStart 15625 46.870279443995216 *** Stop 15625 46.870279443995216 *** Next 22547 46.870279443995216* - Prevedené akcie
- *22547* - Čas potrebný na spracovanie nahrávky (v ms)
- *46.870279443995216* - Označený začiatok nahrávky
- *Výborné* - Subjektívne hodnotenie

# E. Obsah priloženého CD

Priložené CD je rozdelené na niekoľko adresárov:

### audio-player

- *audio-player/bin* - výsledný audio prehrávač
- *audio-player/src* - zdrojové súbory audio prehrávača
- *audio-player/scripts* - online skripty používane audio prehrávačom
- *audio-player/generating* - skripty použité pri generovaní playlistu

### data

- *data/log.txt* - log akcií prevedených respondentami

### doc

- *doc/text.pdf* - text tejto práce

### eval

- *eval/comparison* - skripty použité pri porovnaní penalizačných funkcií
- *eval/evaluation* - skripty použité pri vyhodnocovaní