Možnosti automatickej transkripcie v platforme Transkribus na príklade správ o vybavovaní sťažností občanov v období komunistickej diktatúry[1]

MATEJ SMIDA

Katedra histórie, Filozofická fakulta, Univerzita Mateja Bela v Banskej Bystrici

Possibilities of automatic transcription in the Transkribus platform using the example of reports on the handling of citizen complaints during the communist dictatorship

Abstract: This thesis discusses the basic options of work in an application Transkribus, about its benefits, pitfalls and innovations that we presented in the selected sample of archival documents with the topic of citizen complaints during the governance of the Communist party in Czechoslovakia in the 1948-1989. Gradually we have documented the process of working with historical sources in Transkribus, preparation, transcription and the results. Because we focused on typewritten documents only, we compared the results with other ways and options of transcription, we evaluated their usability, which leads to interesting conclusions and results. The thesis is also supported by statistical indicators and pictorial schedules on the basis of which we have did specific conclusions. In the thesis we also present the issue of citizen complaints as a topic of historical research.
Keywords: Transkribus, automatic transcription, digital humanities, citizen complaints, typewritten documents.

DOI: https://doi.org/10.24040/ahn.2023.26.01.125-148

Digitalizácia prameňov, ich presun do virtuálneho prostredia a s tým spojená ochrana a zálohovanie originálnych dokumentov je jednou z najväčších priorít a výziev ohľadom starostlivosti a rozvoji pramennej základne historického výskumu na Slovensku. V mnohých ohľadoch a rámcoch je prístup k prameňom veľmi komplikovaný, spájaný s veľkou dávkou byrokracie a v konečnom dôsledku napokon dochádza k značnému opotrebovaniu či ničeniu originálnych prameňov. Komplikácie nastávajú aj v samotnej digitalizácii prameňov, ktorá vo väčšine prípadov spočíva „len" vo fotografovaní a skenovaní materiálov. Častokrát im chýba edičná hodnota a možnosť pracovať s informáciami z prameňa priamo

[1] Tento text je výstupom z riešenia projektu APVV-19-0456 SKRIPTOR – Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov.

vo virtuálnom priestore. Okrem rukopisných dokumentov sa táto problematika týka aj obrovského množstva strojopisných dokumentov, ktorými disponujú najmä archívne fondy s prameňmi z novších dejín (primárne 20. storočie). Z toho dôvodu vidíme potenciál (okrem samotnej digitalizácie) aj v automatickej transkripcii strojopisných dokumentov, z ktorých mnohé stále len čakajú na spracovanie v archívoch. V tomto dokážu byť nápomocné rôzne softvéry venujúce sa transkripcii či preformátovaniu dokumentu do prostredia Wordu, Excelu prípadne PDF (OCR softvéry). Jedným z takýchto nástrojov je aj softvér *Transkribus*, ktorý je komplexnou platformou na digitalizáciu a rozpoznávanie textu prameňov. Taktiež slúži na prepis a vyhľadávanie historických prameňov z akéhokoľvek miesta. Samotný softvér využíva nástroj umelej inteligencie HTR+ (Handwritten Text Recognition) a PyLaia.[2] Proces transkripcie dokumentu v *Transkribe* je však stále pomerne komplikovaný, spájaný v mnohých prípadoch s komplexnou a veľmi podrobnou prípravou dokumentu na transkripciu (segmentácia) a s vyhľadaním alebo vytvorením vlastného modelu (najmä v prípade rukopisov). Na jeho základe potom môže prebehnúť samotná transkripcia. Ďalej je potrebné vyhodnotenie transkripcie, po ktorom môže nasledovať práca s metadátami, a nakoniec publikovanie transkribovaného dokumentu.

V našom prípade sme sa rozhodli pre dokumenty siahajúce do 20. storočia a problematiky sťažností počas vlády komunistického režimu v Československu v rokoch 1948 – 1989. Vybrali sme si pramene, ktoré reflektujú stav problematiky sťažností občanov v tomto období. Podrobnejšie sa jej venujeme v ďalšej časti štúdie. Z formálneho hľadiska ide práve o strojopisné dokumenty s početnými štatistikami a tabuľkami, čím chceme zvýrazniť možnosť využitia metód digitálneho spracovania dokumentov na príklade strojopisu. Považujeme za prínosné načrtnúť rôzne možnosti práce so strojopisnými dokumentmi (či už prostredníctvom softvéru *Transkribus,* alebo iných nástrojov), ktoré by mohli byť využiteľné v rámci ich digitalizácie a spracovávania v archívoch a iných pamäťových inštitúciách.

Téme digitálnej transkripcie sa venujeme v rámci pomocnej vedeckej činnosti na Katedre histórie Univerzity Mateja Bela v Banskej Bystrici a výskumného projektu SKRIPTOR (APVV-19-0456 Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov).[3]

[2] MALINIAK, Pavol – NAGY, Imrich (eds.): *Digital humanities : nástroje sprístupňovania historického dedičstva : zborník abstraktov*. Banská Bystrica : ŠVK v Banskej Bystrici, 2022, 73 s. [cit. dňa 2023-04-26]. Dostupné na: https://repo.umb.sk/handle/123456789/39

[3] Touto cestou sa chcem poďakovať mojej priateľke za nepretržitú podporu a pomoc a doc. Mgr. Imrichovi Nagyovi, PhD. za odborné tútorstvo v rámci našej výskumnej činnosti.

Výskumné ciele

Ciele štúdie majú všestranný rámec. Na ich základe chceme čitateľovi ponúknuť čo najkomplexnejší obraz o pertraktovanej problematike na základe vlastných skúseností z realizovaného výskumu, konkrétne:

1. predstaviť softvér *Transkribus* na príklade správ o vybavovaní sťažností občanov,
2. zhodnotiť prácu v *Transkribe* a poukázať na výhody a nástrahy pracovania so strojopisnými prameňmi,
3. porovnať *Transkribus* s inými OCR softvérmi a poukázať na jeho špecifiká a využiteľnosť,
4. vzhľadom na nízku mieru spracovania témy sťažností občanov poukázať na špecifickú historickú hodnotu tohto typu prameného dokumentu.

V nasledujúcich častiach štúdie sme postupne rozpracovali nami zadané cieľové roviny.

Sťažnosti občanov v období totality

Sťažnosti občanov a listy pracujúcich sú pojmy, ktoré sa človeku len veľmi ťažko spájajú s komunistickou totalitou v povojnovom Československu rokoch 1948 – 1989. Napriek tomu existovala aj v tomto období pre občanov možnosť, aby sa mohli do istej miery domáhať práva a spravodlivosti. Svedčia o tom množstvá sťažností, žiadostí, podnetov na odvolanie a listov pracujúcich, ktoré môžeme nájsť v archívnych fondoch. Ide tiež o početnú korešpondenciu súkromných osôb, snažiacich sa odvolávať na zodpovedné inštitúcie, a žiadajúcich o nápravu, pomoc a spravodlivosť. Obracali sa väčšinou na národné výbory, pod ktorých správu konkrétny občan spadal. Tu patrili mestské národné výbory (MsNV), miestne národné výbory (MNV), okresné národné výbory (ONV) či krajské národné výbory (KNV), ktoré sa prostredníctvom svojich odborov s konkrétnou špecializáciou ďalej jednotlivými sťažnosťami zaoberali. Taktiež sa odvolávali na Kanceláriu prezidenta republiky, prípadne na iné zodpovedné orgány, ktoré v danom období mali v rámci Československa pôsobnosť (Slovenská národná rada, vláda a ministerstvá). V neposlednom rade je potrebné spomenúť aj často využívanú možnosť podávať sťažnosti prostredníctvom tlače a jej rubrík.

V súvislosti s predmetmi jednotlivých sťažností a podnetov občanov sa nám naskytuje veľmi široké spektrum rôznych tém a problematík zasahujúcich do bežného života obyvateľov krajiny. Tu môžeme spomenúť najmä oblasť financií, obchodu, dopravy, zdravotníctva, sociálnych záležitostí, školstva, výstavby, bytového hospodárstva, kádrových vecí, kultúry a pod.

Na základe našej bádateľskej činnosti v archívoch (Banská Bystrica, Zvolen, Žilina) môžeme rozčleniť dokumenty venujúce sa sťažnostiam a podnetom občanov do nasledujúcich základných celkov:

- samotné sťažnosti, podnety, žiadosti, listy pracujúcich a pod.,
- reakcie na podnety občanov zo strany zodpovedných orgánov (ONV, KNV a pod.),
- previerky a správy o vybavovaní podnetov občanov zo strany zodpovedných orgánov (väčšinou za konkrétny časový interval),
- štatistické výkazy o sťažnostiach (väčšinou za konkrétny časový interval).

Charakteristika dokumentov

Na prácu v platforme *Transkribus* sme si zvolili previerky a správy o vybavovaní podnetov občanov a štatistické výkazy o sťažnostiach z rôznych časových období (50., 60. a 70. roky 20. storočia). Výber týchto časových horizontov bol zvolený s úmyslom poukázať na konkrétne trendy v rámci problematiky sťažností v jednotlivých obdobiach komunistickej totality. Celková vzorka dokumentov pre prácu v *Transkribe* má 97 strán (A4). Všetky dokumenty pochádzajú zo Štátneho archívu v Banskej Bystrici, s ktorým sme v tejto veci spolupracovali. Bolo nám umožnené dané dokumenty zdigitalizovať, a následne samotné digitalizáty využiť na účely spracovania v prostredí *Transkribus*. Je potrebné podotknúť, že po formálnej stránke ide o pramene písané strojopisom. Každý dokument má isté vonkajšie odchýlky v type strojopisu, úprave strán, vonkajšej kompozícii a pod., čo do veľkej miery ovplyvnilo našu ďalšiu prácu. Tomuto problému sa venujeme aj v ďalších častiach príspevku.

Prvý dokument so svojimi 52 stranami je rozsahovo zároveň ten najväčší. Ide o *Rozbor poznatkov z vybavovania sťažností, oznámení a podnetov občanov podaných v roku 1975 ministerstvám, ústredným orgánom, národným výborom v Slovenskej socialistickej republike*. Tento dokument sa nachádza v archívnom fonde *Stredoslovenský národný výbor v BB – sekretariát predsedu 1971 – 1990* a člení sa na niekoľko častí. Hneď v úvode je vyjadrený záujem riešiť problematiku maximálneho uspokojovania potrieb spoločnosti a nastolenie citlivého riešenia oprávnených požiadaviek, problémov a pripomienok občanov. Úvodná časť taktiež naráža na ústavné zakotvenie práva občanov a organizácií obracať sa na zastupiteľské a štátne orgány s návrhmi, podnetmi, sťažnosťami, o čom pojednáva článok 29 Ústavy Československej socialistickej republiky (z roku 1960).[4] Samotná hlavná časť dokumentu sa venuje ministerstvám,

[4] Ústava Československej socialistickej republiky (ústavný zákon 100/1960 Zb.). In: *Slov-lex: právny a informačný portál* [online], 1960, [cit. 2023-04-13]. Dostupné na internete: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/1960/100/19600711.html

ústredným orgánom a národným výborom v Slovenskej socialistickej republike, ktorým sú sťažnosti a podnety občanov adresované. Konkrétne charakterizuje trend sťažností, ich opodstatnenosť, adresnosť, príčiny a pod. Ďalej sa venuje aj problematike anonymných sťažností a porovnáva štatistické údaje venujúce sa podávaniu podnetov týmto organizáciám s predošlými obdobiami. Nasleduje časť, ktorá sa zaoberá jednotlivým predmetmi sťažností, ich frekvenciou, poznatkami nadobudnutými z vybavovania týchto podnetov a i. Ide najmä o sťažnosti venujúce sa zásobovaniu, hospodáreniu s bytmi a národným majetkom, problémom investičnej výstavby a stavebným konaniam a problematike nedostatkov v súdnom konaní. Posledná časť hodnotí naplnenie plánov a opatrení v súvislosti s vybavovaním sťažností a ich nedostatky, a zároveň charakterizuje konkrétne závery vychádzajúce z nadobudnutých poznatkov. K samotnému rozboru sú pripojené aj štatistické prehľady, ktoré sa venujú počtu podaných sťažností v komparácii s predošlými rokmi, ich príčinám, anonymným sťažnostiam, opatreniam a pod.[5]

Druhým dokumentom, ktorý sme zahrnuli do našej vzorky je *Správa o vybavovaní sťažností pracujúcich v zmysle uznesenia vlády republiky Československej z 27. apríla 1954 č. 731/54 a uznesenia Zboru poverenícov z 25. mája 1954 č. 233/54*. Ide o správu Predsedníctva Zboru poverenícov s rozsahom 4 strany. Poukazuje na množstvo podávaných sťažností a kritických oznámení na národné výbory a ich funkcionárov, ktorým je venovaná značná časť dokumentu. Nasleduje prehľad o predmetoch podávaných sťažností, zahrňujúcich najmä bytové a majetkové záležitosti, a samotné podnety na verejných funkcionárov. Tie sú rozvinuté v konkrétnych prípadoch porušovania zákona zo strany funkcionárov národných výborov v jednotlivých krajoch na Slovensku. Správa je ukončená analýzou problematiky stavu vybavovania sťažností a štatistikou nevybavených sťažností za krajské národné výbory. Z týchto poznatkov sa následne vyviedli náležité závery.[6]

Ďalším dokumentom je *Rozbor stavu a úrovne vybavovania sťažností a oznámení občanov, podaných v roku 1967 štátnym orgánom a organizáciám na Slovensku*. Táto správa, ktorá má rozsahovo 16 strán, sa člení na 3 základné okruhy charakterizujúce problematiku sťažností a podnetov

[5] Štátny archív (ŠA) v Banskej Bystrici, fond (f.) Stredoslovenský krajský národný výbor (StKNV) v B. Bystrici – sekretariát predsedu (sekr. preds.) 1971 – 1990, škatuľa (šk.) 10, číslo (č.) P241/76. Rozbor poznatkov z vybavovania sťažností, oznámení a podnetov občanov podaných v roku 1975 ministerstvám, ústredným orgánom, národným výborom v Slovenskej socialistickej republike.

[6] ŠA v B. Bystrici, f. Krajský národný výbor (KNV) v B. Bystrici – sekr. preds. 1951 – 1960, šk. 5, č. Zn: 86/1955. Správa o vybavovaní sťažností pracujúcich v zmysle uznesenia vlády republiky Československej z 27. apríla 1954 č. 731/54 a uznesenia Zboru poverenícov z 25. mája 1954 č. 233/54.

občanov v roku 1967. Rozbor bol vypracovaný na základe podkladov národných výborov, povereníctiev a komisií Slovenskej národnej rady a iných orgánov. Taktiež do neho boli zahrnuté aj poznatky z previerok úrovne vybavovania sťažností v jednotlivých organizáciách. Prvá časť sa venuje vývoju prijatých sťažností a ich opodstatnenosti, k čomu sú priložené aj štatistické údaje. Zhodnocuje, že celkovo má podávanie sťažností a podnetov občanov klesajúcu tendenciu. Pokles bol zaznamenaný aj vo veci opodstatnenosti konkrétnych podnetov. Na základe analýzy dát usudzujeme, že v roku 1967 došlo k zlepšeniu úrovne vybavovania sťažností, čo bolo prisudzované najmä zlepšeniu hospodárskej a sociálnej situácie v krajine, a vylúčeniu rôznych druhov podaní, ktoré nebolo možné považovať za sťažnosti (reklamácie). Táto časť dokumentu sa venuje aj podielu anonymných oznámení a ich vývoju. Druhá sekcia správy charakterizuje problematiku sťažností prostredníctvom jednotlivých inštitúcií. Ako prvé je v nej analyzované vybavovanie sťažností komisiami a povereníctvami Slovenskej národnej rady (SNR) za rok 1967 v komparácii s rokmi 1965 a 1966. K tejto analýze je pripojená aj štatistika počtu sťažností, prijatými konkrétnymi povereníctvami SNR. Druhá časť sa ďalej venuje prevencii proti sťažnostiam a odstraňovaniu príčin sťažností. Tretia sekcia rozboru hovorí o oznámeniach, ktoré boli obdržané národnými výbormi. Poskytuje štatistické údaje o podávaní sťažností z hľadiska ich obsahu, orgánov, ktorým boli adresované, príčin sťažností a pod. Na konci správy sú koncipované výsledky a závery analýzy.[7]

Štvrtý dokument s názvom *Rozbor vybavovania sťažností a listov pracujúcich za III. štvrťrok 1958* sa skladá z dvoch častí, ktorých rozsah je spolu 5 strán. Jeho autorom je Rada KNV v Banskej Bystrici – kontrolný odbor. Prvé dve strany hovoria o návrhu uznesenia rady KNV zahrňujúcom zistenia vyplývajúce z rozboru, konkrétne závery, opatrenia, nápravy, odporučenia radám ONV a pod. Nasleduje samotná správa vyplývajúca z rozboru vybavovania sťažností, ktorá obsahuje štatistiky jednotlivých odborov rady KNV, konkrétne opatrenia na odstránenie nedostatkov pri vybavovaní sťažností, postupy ONV v problematike, komparácia 3. štvrťroku 1958 s predchádzajúcimi obdobiami a pod. Na záver správy bola vytýčená snaha o zlepšenie daného stavu a náležitá prevencia zo strany KNV. Súčasťou dokumentu je aj *Výkaz o stave sťažností v kraji Banská Bystrica za 3. štvrťrok 1958*, ktorý zahrňuje štatistické údaje vo veci podávania sťažností za odbory rady KNV, jednotlivé ONV atď.[8]

[7] ŠA v B. Bystrici, f. StKNV v B. Bystrici – sekr. preds. 1960 – 1969, šk. 35, č. P-207/68. Rozbor stavu a úrovne vybavovania sťažností a oznámení občanov, podaných v roku 1967 štátnym orgánom a organizáciám na Slovensku.

[8] ŠA v B. Bystrici, f. KNV v B. Bystrici – odbor kontrolný (odb. kontr.) 1958-1960, šk. 1, č. KO-70/1958. Rozbor vybavovania sťažností a listov pracujúcich za III. štvrťrok 1958.

Nasleduje *Správa o vybavovaní sťažností za I. polrok 1959* zo strany Rady ONV v Hnúšti – odbor kontrolný a *Vybavovanie sťažností, pripomienok a návrhov za I. polrok 1959* za Radu ONV v Modrom Kameni – odbor kontrolný. Oba dokumenty s rozsahom 7 strán sú štatistickým rozborom o vybavovaní sťažností. Priamo sa venujú počtu prijatých sťažností za dané časové obdobie, ich opodstatnenosti, sociálnemu a triednemu začleneniu sťažovateľov, samotnému stavu úrovne vybavovania sťažností, predmetom jednotlivých podnetov, a taktiež kritike jednotlivých odborov ONV pri narábaní so sťažnosťami.[9]

Posledná séria dokumentov zahrnutá do nášho súboru, s ktorým sme pracovali v prostredí *Transkribus*, sú *Mesačné výkazy o sťažnostiach za rok 1958* zo strany Rady KNV v Banskej Bystrici – odbor kontrolný. Ide o 12 strán výkazov (každá strana pre daný mesiac v roku) v tabuľkovom formáte s vopred predpísanými ukazovateľmi, ku ktorým sa zapisovali len samotné číselné údaje. Ide o konkrétne ukazovatele ako: počet sťažností a ich vybavenosť, triedne rozloženie sťažovateľov, povaha sťažností (na funkcionárov, trestné, bytové, pracovné, na zásobovanie atď.), anonymné sťažnosti a pod.[10]

Vyhľadanie dokumentov a ich digitalizácia

Pre prácu v platforme *Transkribus*, smerujúcej k samotnej transkripcii, je najprv potrebné vyhľadať a zhromaždiť pramene, s ktorými chceme pracovať a pripraviť digitalizáty, ktoré budú disponovať dostatočnou kvalitou a ostrosťou snímky. Ide o prvotnú predispozíciu pre úspešnú transkripciu dokumentu, bez ktorej by sme uspokojivé výsledky dosiahli len veľmi ťažko.

Po ich jednotlivom sprístupnení v rámci bádateľne archívu sme pristúpili k samotnej digitalizácii dokumentov prostredníctvom zariadenia *ScanTent*.[11] Ide o prenosné, ľahko manipulovateľné zariadenie pripomínajúce stan. Jeho funkcia spočíva v samotnom „stane", brániacom vonkajším elementom, ktoré by mohli digitalizáciu znehodnotiť, a LED osvietení, smerujúcom priamo na digitalizovaný dokument a zabezpečujúcom kvalitu a ostrosť pripravovanej snímky. V hornej časti zariadenia sa nachádza otvor slúžiaci na snímanie dokumentu, ktorý sa nachádza vo vnútri *ScanTentu*. V našom prípade nám snímanie dokumentov s rozsahom 97

[9] ŠA v B. Bystrici, f. KNV v B. Bystrici – odb. kontr. 1958-1960, šk. 13, č. 326/KO-1959. Správa o vybavovaní sťažností za I. polrok 1959, Vybavovanie sťažností, pripomienok a návrhov za I. polrok 1959.

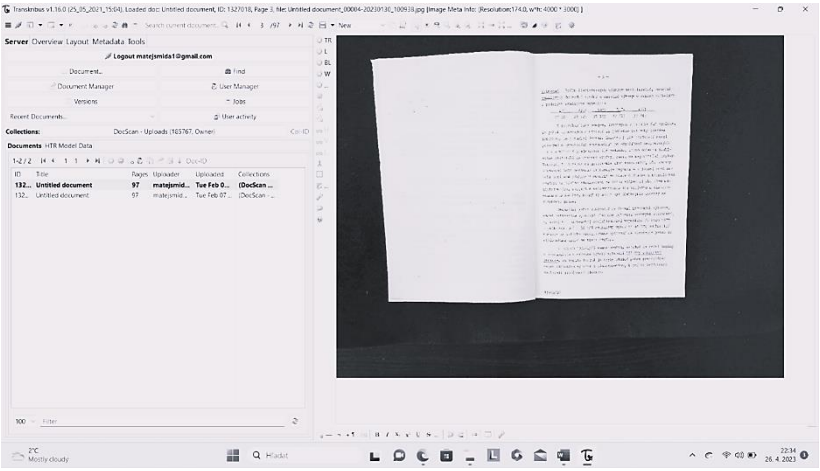
[10] ŠA v B. Bystrici, f. KNV v B. Bystrici – odb. kontr. 1958-1960, šk. 13, sign. kontr. 60/1. Mesačné výkazy o sťažnostiach za rok 1958.

[11] The ScanTent : Professional scanning with your smartphone. [2023-06-01]. Dostupné na internete: https://readcoop.eu/scantent/

strán trvalo asi 2 hodiny s tým, že so samotným *ScanTentom* sme manipulovali prvýkrát. Pramene sme fotografovali prostredníctvom mobilného telefónu s operačným systémom Android a aplikácie *DocScan*,[12] ktorá slúži priamo na snímanie digitalizátov a ich nahrávanie do prostredia *Transkribus*. V prípade práce s *Transkribom* odporúčame jej využitie, keďže kvalita vyhotovených snímok bola lepšia a výrazným spôsobom uľahčuje aj presun materiálov priamo do konta užívateľa *Transkribu*. Zdigitalizované dokumenty sme po kontrole ich kvality, prostredníctvom vzájomnej spolupráce, sprístupnili aj archívu a poslúžia mu v rámci postupnej digitalizácie ich archívnych fondov.

Ďalším dôležitým bodom v našom výskume bolo nahranie nafotografovaných dokumentov na platformu *Transkribus*. V tom nám výrazným spôsobom pomohla už spomínaná aplikácia *DocScan*, prostredníctvom ktorej je možné prihlásiť sa na používateľské *Transkribus* konto a nafotené dokumenty priamo doň stiahnuť. Samotné nahrávanie sa nám podarilo až po niekoľkých pokusoch a zabralo nám zhruba hodinu.

Obr. č. 1.: *Ukážka dokumentu v prostredí Transkribus priamo po nahratí*



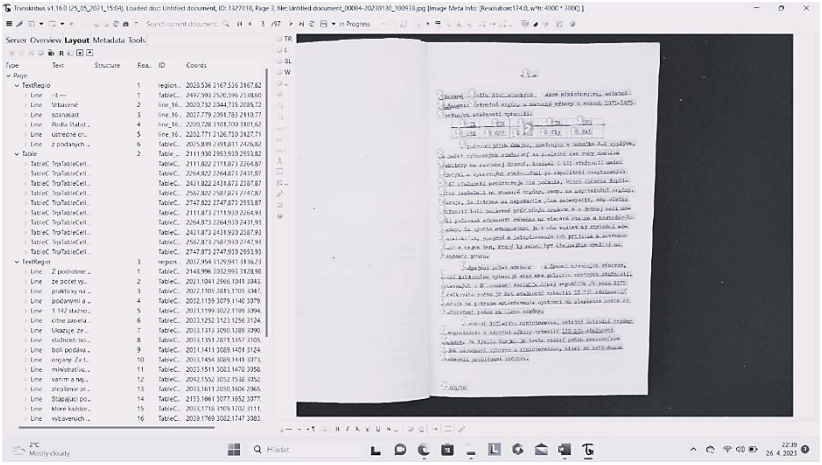
Príprava dokumentov na proces transkripcie

Následne sme mohli pristúpiť k samotnej úprave dokumentov a jej príprave na automatickú transkripciu – k segmentácii. Ide o rozčlenenie textových polí jednotlivých strán dokumentu, určenie ich poradia a

[12] Google Play: DocScan. [cit. 2023-06-01]. Dostupné na internete: https://play.google.com/store/apps/details?id=at.ac.tuwien.caa.docscan&hl=sk&gl=US&pli=1

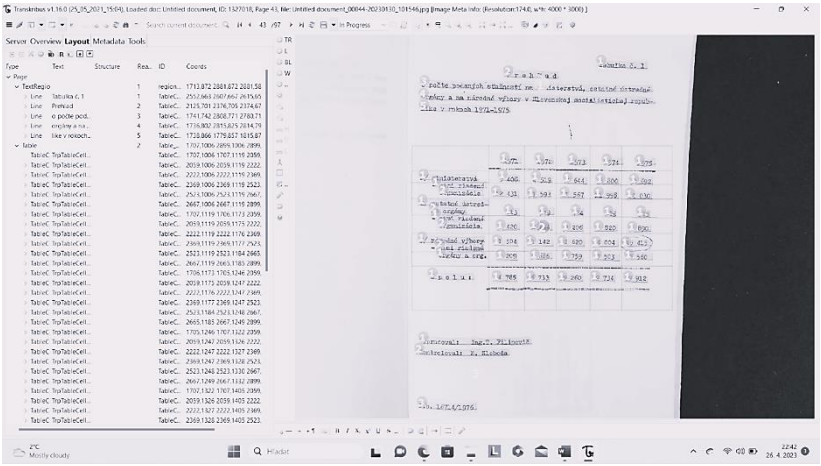
smeru čítania, ktorým sa má *Transkribus* riadiť pri transkripcii. Ten ponúka široké množstvo nástrojov potrebných na segmentáciu a následnú transkripciu. Ide o nástroje na označovanie jednotlivých textových regiónov a riadkov v dokumente. Špecifickým prípadom, s ktorým sme sa v rámci našej výskumnej vzorky stretli, sú tabuľky, ktoré je potrebné označovať spôsobom určeným priamo pre prácu s tabuľkami. Po kontrole správnosti nahrania dokumentov a dodržania ich poradia, sme sa po konzultácii rozhodli pre strany dokumentov, na ktorých sa nenachádzali štatistické tabuľky, zvoliť možnosť tzv. automatickej segmentácie, ktorá sa nám žiaľ neosvedčila. Po aplikovaní automatickej segmentácie na konkrétnu stranu sme narazili na problém rozčlenenia jedného textového poľa na viacero regiónov, čo bolo v našom prípade nežiadúce. Rozhodli sme sa preto ísť cestou manuálneho označovania textových regiónov. Vzhľadom na to, že výsledné textové rámce strany, na ktorej sme aplikovali automatickú segmentáciu, bolo potrebné následne opravovať, z časového hľadiska a z hľadiska množstva vynaloženého úsilia bolo výhodnejšie textové rámce označovať manuálnou formou. Následne sme pokračovali už automatickým označovaním riadkov textových polí, s ktorým neboli takmer žiadne problémy. Niekedy bolo potrebné daný riadok predĺžiť alebo rozčleniť, čo však bolo len v zanedbateľnej miere a v konečnom dôsledku hodnotíme túto časť procesu segmentácie v pozitívne. Na záver sme vždy skontrolovali aj poradie čítania textových rámcov a riadkov, keďže sa po úpravách zvyklo meniť.

Obr. č. 2.: *Ukážka dokumentu v Transkribe po segmentácii*



Iný postup sme museli zvoliť pri segmentácii tabuliek, ktoré sa v rámci našej série dokumentov vyskytovali v hojnom počte. Zmena nastala už pri samotnom označovaní textových rámcov. Tu *Transkribus* ponúka konkrétny nástroj na označovanie častí tabuliek. V tejto situácii do veľkej miery záleží od komplikovanosti a veľkosti tabuľky, ktorú segmentujeme. S problémami sme sa stretávali v prípadoch, ak išlo o veľmi podrobnú tabuľku s mnohými údajmi a jej celkové rozloženie, tak ako aj rozloženie textu v rámci strany, bolo zdeformované (nedôsledná úprava strojopisu a jeho odchýlenie od zvislého a kolmého smerovania), s čím *Transkribus* pôvodne neráta. Segmentácia tabuľky ako takej spočíva v ohraničení jej konkrétnych častí prostredníctvom nástroja *Transkribus* na označovanie tabuliek. Následne tieto rámce je potrebné zoradiť v takom poradí, v akom chceme, aby ich *Transkribus* čítal (čo v našom prípade bolo väčšinou automaticky). Ďalej je potrebné označiť samotný text tabuľky. To je možné urobiť automaticky alebo manuálne. Nám sa automatické označovanie pri práci s tabuľkami neosvedčilo z dôvodu príliš veľkej chybovosti. Častokrát dochádzalo k nesprávnemu označeniu riadkov, čo si vyžiadalo početné úpravy. Konkrétne išlo najmä o spájanie riadkov naprieč niekoľkými textovými rámcami. Ďalej sa vyskytovali problémy s nesprávnym označovaním riadkov, prípadne s jeho úplnou absenciou. Vzhľadom na pomerne veľkú chybovosť automatického označovania textu v rámci tabuliek, sa nám omnoho viac osvedčil manuálny spôsob. Čas, ktorý sme strávili opravovaním nesprávnych označení, bol neúmerne vyšší oproti dĺžke trvania manuálneho označovania danej tabuľky. Z toho dôvodu sme napokon väčšinu textu v tabuľkách označili manuálne.

Obr. č. 3.: *Segmentácia tabuľky v prostredí Transkribus*



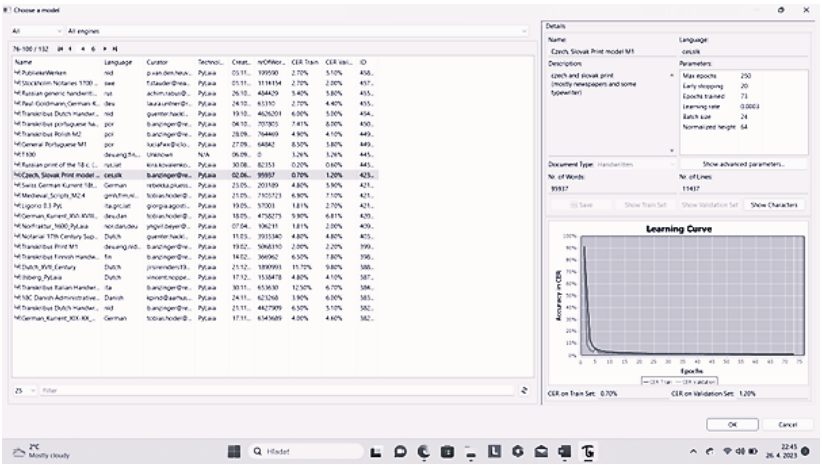
Po tomto procese a po odkontrolovaní poradia čítania jednotlivých textových rámcov a riadkov textu sme ukončili segmentáciu dokumentov a mohli sme pristúpiť k ich samotnej transkripcii.

Transkripcia dokumentov

1. Transkripcia prostredníctvom už vytvoreného modelu

Transkripciu dokumentov sme vykonávali v rámci metódy PyLaia, ktorú *Transkribus* ponúka. Vzhľadom na to, že v našej vzorke sme pracovali výhradne so strojopisom, rozhodli sme sa, že spočiatku nepôjdeme cestou vytvárania vlastného modelu pre transkripciu, ale využijeme už prístupný model, ktorý je v rámci verejných modelov dostupný pre používateľa v *Transkribe* a bol dobre využiteľný aj pre náš typ prameňov. Zvolili sme si model s názvom *Czech, Slovak, Print model M1*, vytvorený v júni 2022. Primárne je určený pre transkripciu slovenskej a českej tlače s udávanou chybovosťou v tréningovom sete 0,70 % a vo validačnom (kontrolnom) sete 1,20 %. V tomto prípade je ale nutné podotknúť, že tento model nie je určený priamo na transkripciu strojopisov. Po zvolení modelu sme spustili samotný proces transkripcie, ktorý prebehol bez problémov a zabral nám asi 30 minút. Následne bol pod každým digitalizátom prístupný výsledok jeho transkripcie (prepísaný text z konkrétnej strany dokumentu).

Obr. č. 4.: *Parametre nami vybraného modelu transkripcie Czech, Slovak, Print model M1*



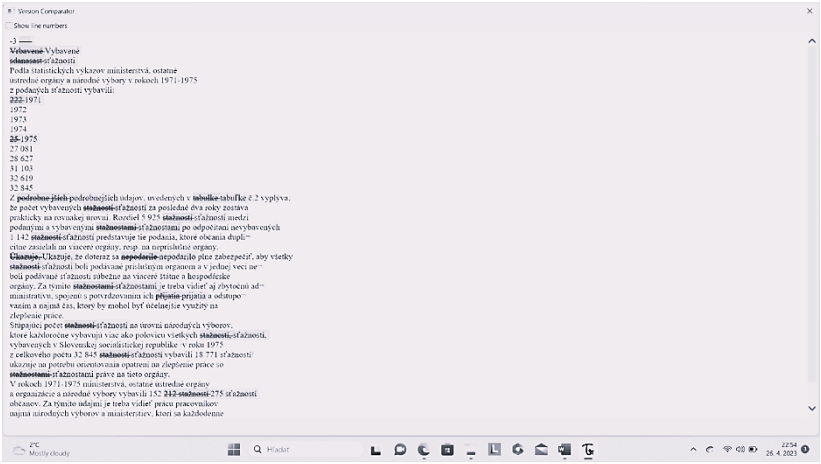
Ďalšia fáza spočívala v oprave chýb a nepresností z transkribovaného textu tak, aby bol v súlade s originálnym dokumentom. Tým nám vznikla

porovnávacia vzorka, na základe ktorej sme následne mohli zistiť percentuálnu mieru chybovosti transkripcie a jej výsledky. Počas opráv transkripcie jednotlivých dokumentov sme tieto chyby priebežne sondovali a snažili sme sa vytýčiť hlavné prvky, s ktorými mal *Transkribus* a nami zvolený model problémy. Časté chyby v transkripcii dokumentov, s ktorými sme pracovali, sú nasledovné:

A) *Rozbor poznatkov z vybavovania sťažností, oznámení a podnetov občanov podaných v roku 1975 ministerstvám, ostatným ústredným orgánom, národným výborom v Slovenskej socialistickej republike*

- deformácia slov a číslic
- nadbytočné medzery uprostred slov
- zle umiestnená alebo chýbajúca diakritika
- problém s transkripciou textu, ktorý je podčiarknutý (najmä v prípade nadpisov)
- časté zamieňanie nasledovných písmen:
  „o“ → „e“; „o“ → „0“; „r“ → „ř“; „e“→ „ě“; „X“→ „K“
- opakujúci sa problém s číslicou „1“, ktorú platforma zamieňala za písmeno „i“
- zamieňanie symbolu „%“ za číslicu „7“ a pod.
- v rámci tabuliek problém so správnou transkripciou číselných údajov

Obr. č. 5.: *Ukážka chybovosti v texte dokumentu po manuálnej oprave transkripcie*



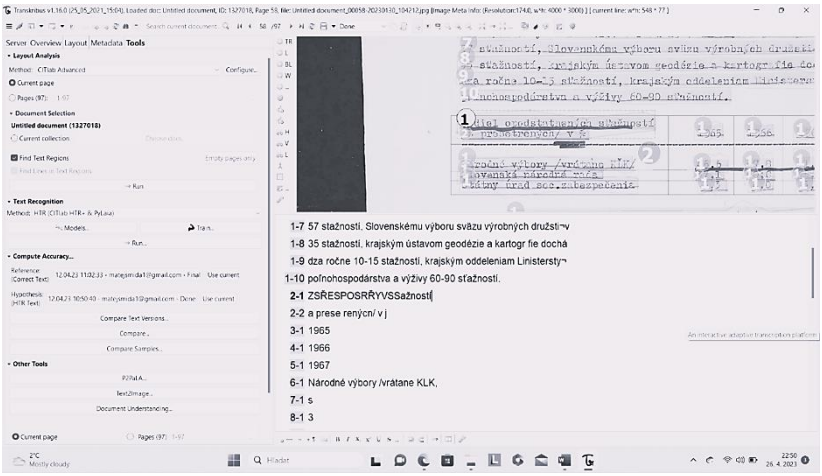
B) *Správa o vybavovaní sťažností pracujúcich v zmysle uznesenia vlády republiky Československej z 27. apríla 1954 č. 731/54 a uznesenia Zboru povereníkov z 25. mája 1954 č. 233/54*

- zle umiestnená alebo chýbajúca diakritika
- v niektorých pasážach dokumentu nedostatočná kvalita strojopisu, ktorá deformovala samotný výsledok transkripcie
- časté zamieňanie nasledovných písmen:
  „s“ → „e“; „s“ → „a“; „B“ → „E“
  „o“ → „e“; „u“ → „ů“; „N“→ „M“
- zamieňanie písmena „i“ za číslice „1“ a „4“

C) *Rozbor stavu a úrovne vybavovania sťažností a oznámení občanov, podaných v roku 1967 štátnym orgánom a organizáciám na Slovensku*

- zle umiestnená alebo chýbajúca diakritika
- v niektorých pasážach dokumentu nedostatočná kvalita strojopisu, ktorá deformovala samotný výsledok transkripcie
- zásahy do textu dokumentu jeho používateľmi (podčiarkovania, poznámky), ktorá negatívne ovplyvňovala kvalitu transkripcie
- časté zamieňanie písmen: „r“ → „ř“; „o“ → „e“; „o“ → „c“;
  „e“ → „ě“; „d“ → „c“; „ä“ → „á“
- zamieňanie symbolu „%“ za číslicu „7“
- časté zamieňanie písmen a číslic za pomlčku (-)
- nadbytočné medzery uprostred slov (spôsobené nedostatočnou kvalitou strojopisu – chýbajúce alebo nevýrazné grafémy)
- nesprávne transkribované číselné údaje v rámci tabuliek

Obr. č. 6.: *Ukážka deformácie textu po zásahu do dokumentu zo strany používateľa*



D) *Rozbor vybavovania sťažností a listov pracujúcich za III. štvrťrok 1958*

- nízka kvalita strojopisu
- deformácia transkripcie textu, ktorý je podčiarknutý (najmä nadpisy)
- problém so správnou transkripciou číselných údajov
- chýbajúca alebo zle umiestnená diakritika
- časté zamieňanie písmen: „o“ → „e“; „e“ → „a“; „ú“ → „ů“; „H“ → „E“; „z“ → „s“; „N“ → „W“
- veľký problém nám spôsobovala štatistická tabuľka vo formáte A4, ktorá sa nachádzala na konci dokumentu.[13]

E) *Správa o vybavovaní sťažností za I. polrok 1959*

- nadbytočné medzery uprostred slov (spôsobené nedostatočnou kvalitou strojopisu – chýbajúce alebo nevýrazné grafémy)
- deformácia transkripcie podčiarknutého textu (najmä nadpisy)
- zle umiestnená alebo chýbajúca diakritika
- nesprávna transkripcia údajov z tabuliek (číselné údaje)
- časté zamieňanie písmen a číslic: „o“ → „e“; „a“ → „e“; „ú“ → „ů“; „N“ → „M“; „2“ → „4“
- problém s transkripciou skratiek KNV, MNV, ONV a pod.
- v rámci transkripcie časté vynechávanie číslice „5“

F) *Mesačné výkazy o sťažnostiach za rok 1958*

V tomto prípade ide o špecifický prípad dokumentu, ktorý je čisto v tabuľkovom formáte. Išlo o veľmi podrobné a svojou štruktúrou komplikované tabuľky, ktoré *Transkribu* pri transkripcii robili veľké problémy, čo sa odzrkadlilo aj na veľmi veľkej chybovosti. Tá bola už len v prípade prvej strany vyše 30 %, čo nie je ani z ďaleka uspokojivý výsledok. Najväčšie problémy týkajúce sa transkripcie boli:

- nadbytočné medzery uprostred slov (spôsobené nedostatočnou kvalitou strojopisu – chýbajúce alebo nevýrazné grafémy)
- častý prechod medzi dvoma typmi strojopisu, keďže ide o vopred predpísanú šablónu tabuľky, do ktorej sa následne vpisovali údaje za konkrétny mesiac.[14]
- problém s čítaním číselných údajov
- nízka kvalita strojopisu

[13] Vzhľadom na jej podrobnosť, nízku kvalitu strojopisu, nešťastné rozloženie na strane (zakrivenie) atď., automatická transkripcia tejto strany dokumentu vykázala obrovskú chybovosť a v tomto prípade je veľmi zle aplikovateľná.

[14] To podľa nášho názoru bolo pre *Transkribus* mätúce a spôsobovalo to časté deformácie transkribovaného textu.

Od začiatku práce s dokumentami sme sa stretávali aj s rôznymi typmi používaného strojopisu, či s jeho problematickou kvalitou, ktorá v niektorých častiach prameňov nebola z nášho pohľadu dostatočná. Nastala preto obava, či nebude mať *Transkribus* problém s transkripciou týchto pasáží, čo sa napokon aj potvrdilo. Na niektorých stranách sme narazili aj na poznámky používateľov daného dokumentu, ktoré boli vpisované rukou, prípadne podčiarknutia celých riadkov, ktoré tiež mohli mať negatívny dopad na výsledky transkripcie. To sa pri opravovaní transkripcie javilo ako opodstatnené, keďže dochádzalo často k deformácii textu, do ktorého používateľ priamo zasiahol buď podčiarkovaním jednotlivých riadkov, alebo vpisovaním vedľajších poznámok. Ďalším závažným problémom pri transkripcii bola zle umiestnená alebo absentujúca diakritika (najviac to bolo viditeľné pri slove „sťažnosti“, ktoré *Transkribus* opakovane čítal ako „stažnosti“). S týmto problémom sme sa stretli pri oprave každého dokumentu. Na častej chybovosti dokumentov sa odzrkadlili aj neopodstatnené medzery uprostred slov, zlá kvalita transkripcie tabuliek a zamieňanie grafém. Chybovosť jednotlivých strán sa zväčšovala priamoúmerne so zníženou kvalitou strojopisu a väčším výskytom tabuliek, ktorých transkripcia bola najviac neuspokojivá. V prípadoch, kde dokument spočíval výhradne na tabuľkách, bola miera chybovosti mimoriadne vysoká (nad 30 %). Na zvýšený výskyt chýb sme narazili aj pri dokumentoch, kde absentovalo vodorovné smerovanie textového poľa a dostávalo sa do odchýlky s rozložením strany. Tu percentuálna chybovosť znakov (CER) siahala od 3 % až do 12 %.

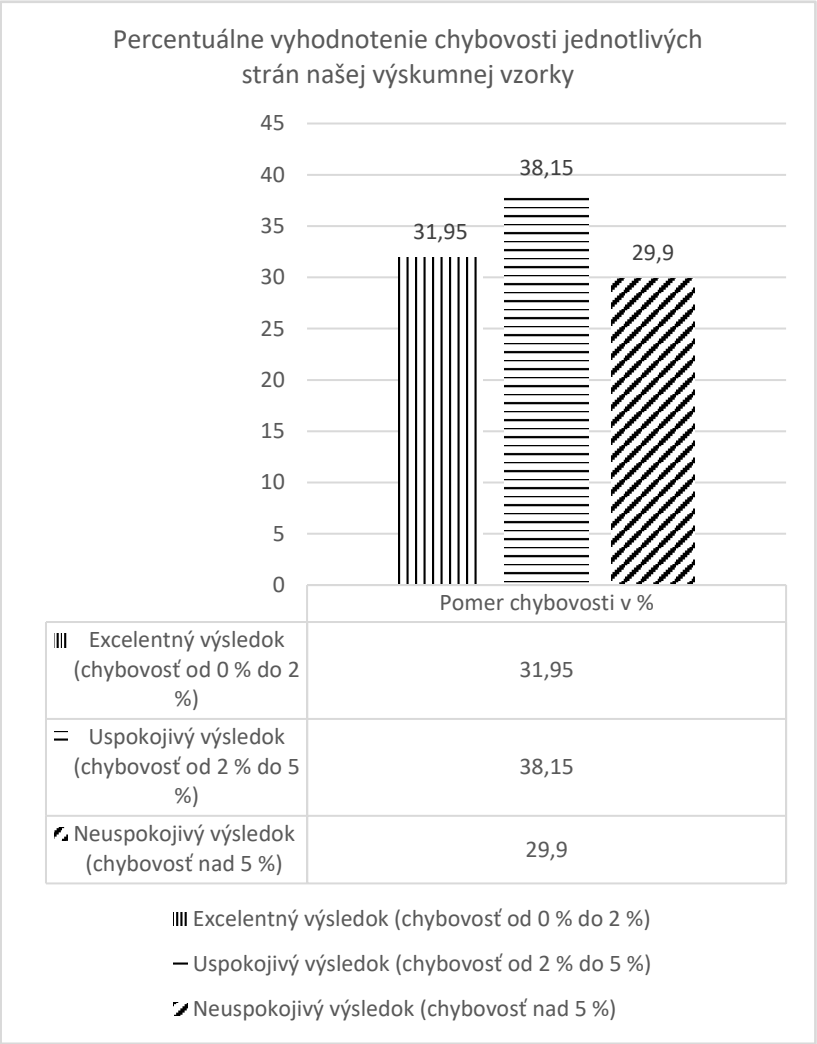
Samotná oprava transkribovaných dokumentov trvala v priemere cca 10 minút na jednu stranu A4 (v prípade tabuliek a dokumentov, ktorých chybovosť bola vysoká, to mohlo trvať aj niekoľkonásobne dlhšie).

Štatistické vyhodnotenie transkripcie

Po ukončení opráv transkripcie sme pristúpili k hodnoteniu a analýze jej výsledkov. V prvom rade bolo potrebné určiť si základné parametre, s ktorými sme pri vyhodnocovaní pracovali. Zostavili sme si jednoduchú percentuálnu stupnicu chybovosti znakov (CER) na jednu transkribovanú stranu:

1. 0 % - 2 % → excelentný výsledok
2. 2 % - 5 % → uspokojivý výsledok
3. 5 % a viac → neuspokojivý výsledok

Na základe tejto stupnice nám vyšla nasledovná štatistika chybovosti, ktorú sme kvôli názornosti vyjadrili nasledujúcim grafom:



Na základe tejto štatistiky môžeme usúdiť, že najväčší počet transkribovaných strán dosiahol uspokojivý výsledok (38,15 %). Nasledujú strany, ktorých transkripcia dosiahla excelentný výsledok (31,95 %) a strany s neuspokojivým výsledkom (29,9 %).

Výsledky transkripcie, žiaľ, nenaplnili naše očakávania. Zaznamenali sme veľmi veľkú mieru chybovosti, ktorá sa v niektorých ojedinelých prípadoch šplhala až k 50 % (tabuľky), čo podľa nás naštrbuje opodstatnenosť využitia *Transkribu* pri takomto type dokumentov.

2. Transkripcia prostredníctvom nami vytvoreného modelu

Po týchto výsledkoch automatickej transkripcie našej výskumnej vzorky, prostredníctvom voľne dostupného modelu *Czech, Slovak, Print model M1*, ktoré nenaplnili naše očakávania, sme sa rozhodli pre vytvorenie nášho vlastného modelu, ktorý by bol vhodný pre našu konkrétnu vzorku dokumentov.

Vytvorenie nového modelu v prostredí *Transkribus* je pomerne jednoduché a samotný model je ľahko aplikovateľný. V prvom rade je potrebné v *Transkribe* otvoriť možnosť pre trénovanie modelov, zadať názov nového modelu, ktorý chceme vytvoriť, a jeho stručnú charakteristiku. Ďalej sme do trénovania zakomponovali aj nami využívaný model *Czech, Slovak, Print model M1*, ktorý mal z hľadiska procesu tvorby nového modelu pozíciu základného „base" modelu. Najdôležitejším krokom bol výber vzoriek do tréningového a validačného setu. Celkovo sme vytvorili tri modely:

1. *slovak_typewriting*
2. *slovak_typewriting002*
3. *slovak_typewriting003*

V rámci prvého modelu sme do tréningového setu zaradili 45 strán z našej vzorky dokumentov, zatiaľ čo do validačného setu sme zaradili 5 strán. Po utvorení nového modelu sme však narazili na pomerne vysokú chybovosť. Z hľadiska chybovosti tréningového setu sme dosiahli vynikajúci výsledok 0,20 %. Problém nastal pri overovacom (validačnom) sete, kde sme dosiahli chybovosť 3,71 %, čo je na model určený pre strojopis neuspokojivý výsledok. To prisudzujeme tomu, že do tréningového setu sme dali strany s minimálnym počtom tabuliek, zatiaľ čo vo validačnom sete boli dokumenty čisto tabuľkového formátu, a to mohlo mať dopad na celkové neuspokojivé výsledky.

Z toho dôvodu sme s naším tútorom usúdili, že bude potrebné vytvoriť nový model, kde zmeníme pomer tabuliek v tréningovom a validačnom sete. V modeli *slovak_typewriting002* bol pomer strán v tréningovom sete 40 a vo validačnom sete 5 s tým, že do validačného setu sme zakomponovali skôr dokumenty, ktoré neobsahovali tabuľky. Chybovosť v tréningovom sete ostala na 0,20 % a vo validačnom sa rapídne znížila na 1,10%. Z toho dôvodu sme pristúpili k samotnej transkripcii. Vybrali sme porovnávaciu vzorku s rozsahom 7 strán, s ktorými sme v rámci transkripcie s voľne prístupným modelom nemali uspokojivé výsledky. Problémom tohto modelu bol zas charakter dokumentov, ktorý sme zaradili do procesu jeho tréningu. Zakomponovali sme do neho len strany prvého dokumentu v rámci našej pramennej základne, ktorý disponuje odlišným typom strojopisu ako ostatné dokumenty. Tým pádom, keď sme týmto modelom transkribovali strany v rámci

tohto dokumentu, dosiahli sme skvelú chybovosť, ale v prípade iných prameňov, ktoré boli zakomponované v našej vzorke a vykazovali isté odlišnosti v type strojopisu, jeho kvalite alebo v iných faktoroch, ktoré sme v predchádzajúcich častiach štúdie charakterizovali, sa chybovosť znížila len neuspokojivo alebo sa naopak zvýšila. Konkrétne percentuálne výsledky ponúkame v porovnávacej tabuľke nižšie.

Pokúsili sme sa preto o tretiu variantu nášho modelu *slovak_typewriting003*. Pri jeho tréningovom procese sme použili strany zo všetkých dokumentov v rámci našej pramennej základne, a to v pomere 31 strán v tréningovom sete a 8 strán vo validačnom. Výsledky ale opäť neboli uspokojivé. V tréningovom sete sme dosiahli chybovosť 0,70 % a vo validačnom až 5,10 %. Pre porovnanie sme sa rozhodli vybrané strany transkribovať aj cez tento model. Paradoxne strany, ktoré v druhom modeli mali lepšie výsledky, tu dosiahli horšie, a v dokumentoch dlhodobo s veľmi zlou chybovosťou sme zas zaznamenali výrazný pokles.

Tab. č. 1: *Percentuálne porovnanie chybovosti nami vybranej vzorky dokumentov, ktoré sme transkribovali jednak prostredníctvom verejného modelu (Czech, Slovak, Print model M1) a nami vytvorených modelov (Slovak_typewriting002 a Slovak_typewriting003):*

	Czech, Slovak, Print model M1 (chybovosť CER/WER)	Slovak_typewriting002 (chybovosť CER/WER)	Slovak_typewriting003 (chybovosť CER/WER)
Strana číslo 1	16,86 % / 28,09 %	0,94 % / 3,70 %	1,55 % / 6,48 %
Strana číslo 2	12,04 % / 24,16 %	1,33 % / 6,04 %	2,08 % / 9,06 %
Strana číslo 3	0,77 % / 4,51 %	0,35 % / 1,75 %	0,88 % / 4,86 %
Strana číslo 4	12,28 % / 31,16 %	10,45 % / 44,81 %	10,51 % / 44,81 %
Strana číslo 5	7,45 % /33,43 %	11, 52 % / 47,51 %	11,52 % / 47,51 %
Strana číslo 6	45,38 % / 89,45 %	29,00 % / 61,72 %	22,50 % / 47,27 %
Strana číslo 7	30,14 % / 62,80 %	46,52 % / 89,76 %	37,16 % / 64,42 %

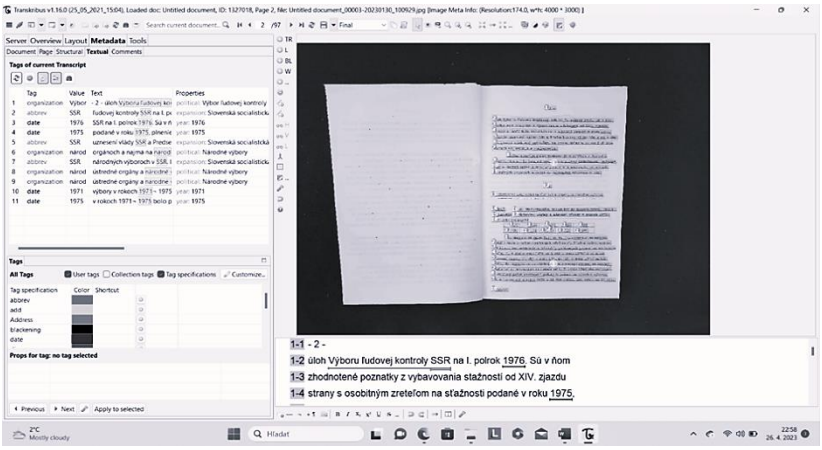
Prácu s vlastnými modelmi v konečnom dôsledku hodnotíme pozitívne, a to najmä z toho dôvodu, že vidíme veľký potenciál transkripcie v prípade veľmi rozsiahleho dokumentu, ktorý je z hľadiska vonkajšej

formy monotónny a stály. V prípade, ak v ponuke voľne dostupných modelov nebude k dispozícii taký, ktorý by bol „šitý na mieru“ práve nášmu prameňu, je výhodné vytvoriť si vlastný model, a tým eliminovať chybovosť na minimum tak, ako sme to štatisticky dokázali v prípade nášho modelu, ktorý sme trénovali len na stranách jedného strojopisného dokumentu. Tým pádom sme pri transkripcii ďalších strán tohto dokumentu s využitím daného modelu dosiahli veľmi solídnu chybovosť založenú na desatinách percenta (viď. tabuľka vyššie).

Ďalšie možnosti Transkribu

Možnosti práce s dokumentmi v *Transkribe* nekončia ukončením procesu transkripcie. Tento softvér totiž ponúka aj ďalšie možnosti pre prácu s dokumentmi vedúce k ich editovaniu a prezentovaniu. V tomto smere má *Transkribus* k dispozícii možnosť práce s tzv. metadátami, ktoré umožňujú charakterizovať prameň, s ktorým pracujeme, označiť jeho vnútorné a vonkajšie členenie, charakterizovať určité špecifické výrazy vyplývajúce z textu, skratky, dátumy, názvy organizácií, mená osobností, názvy miest a pod. V prípade našej pramennej vzorky môže ísť o rôzne názvy organizácií štátnych orgánov (Komunistická strana Československa, miestne, okresné, či krajské národné výbory a pod.), dátumov, mien alebo skratiek (ČSSR, ONV, KNV a pod.). Tie je možné následne kategorizovať a vytvoriť z nich akúsi databázu, ktorá bádateľovi pomáha v lepšej orientácii v prameni. Všetky tieto možnosti vytvárajú pridanú hodnotu *Transkribu* v oblasti digitalizácie prameňov, prístupu k nim a edičnej činnosti, čím je vlastne naplnená jedna zo základných úloh archivárov a historikov, ktorou je ochrana prameňov a šetrná manipulácia s nimi.

Obr. č. 7.: *Ukážka práce s metadátami v Transkribe (kategorizácia výrazov)*



Transkripcia prostredníctvom softvéru Adobe Acrobat a ABBYY FineReader

Na komparáciu s prostredím *Transkribus* sme sa rozhodli využiť dva softvéry, ktoré disponujú funkciou OCR, slúžiacou na optické rozpoznávanie znakov snímok a digitalizovaných skenov a ich prenos do prostredia Wordu, Excelu alebo do formátu PDF.[15] Pre tento experiment sme si zvolili vzorku dokumentov v rámci *Rozboru poznatkov z vybavovania sťažností, oznámení a podnetov občanov podaných v roku 1975 ministerstvám, ústredným orgánom, národným výborom v Slovenskej socialistickej republike* v rozsahu 52 strán, s ktorou sme pracovali aj priamo v *Transkribe*.

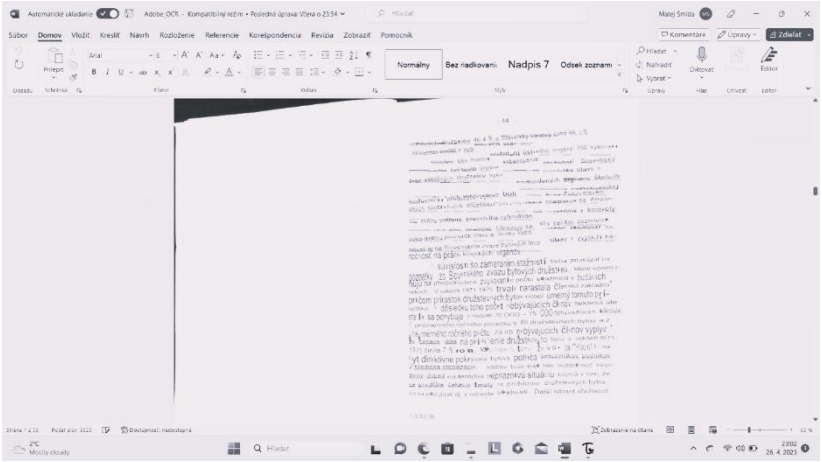
Ako s prvým sme pracovali so softvérom *Adobe Acrobat*, ktorý nám bol poskytnutý prostredníctvom nášho konzultanta. Samotná manipulácia s ním spočívala len v nahratí skenov do prostredia *Adobe Acrobat* a spustení procesu OCR. Celkovo nám to zabralo zhruba pol hodiny. Výsledky transkripcie sme uložili vo formáte dokumentov Word a Excel.

Výsledky práce s programom *Adobe Acrobat* neboli z nášho pohľadu uspokojivé. Prvé strany dokumentu s pomerne dobrou kvalitou strojopisu vykazovali ešte pomerne nízku chybovosť (rovnako ako pri *Transkribe*). Následne sa ale začali vyskytovať časté preklepy, nesprávne čítanie niektorých grafém (ktoré začali aj absentovať) a cca od 6. strany dokumentu začalo dochádzať k totálnej deformácii textu, ktorý sa stal nečitateľný a nepoužiteľný pri prípadnej ďalšej manipulácii s ním. Tento problém prisudzujeme zakriveniu textového poľa v rámci strany, s ktorým mal veľké problémy aj samotný *Transkribus*. Ďalší problém nastal pri čítaní tabuliek, s ktorými sa softvér nedokázal vysporiadať. Za veľký nedostatok považujeme aj fakt, že sme narážali na nestálosť formátovania textových polí a zmeny veľkostí a typov písma v rámci jedného riadku, čo pôsobilo veľmi zmätočne. Taktiež v mnohých ohľadoch výsledky nadobúdajú skôr podobu skenu, ako dokumentu v programe Word. V tomto prípade sme využili možnosť formátovať dokument aj do programu Excel, ktorý bol, žiaľ, pre tento prípad absolútne nepoužiteľný.

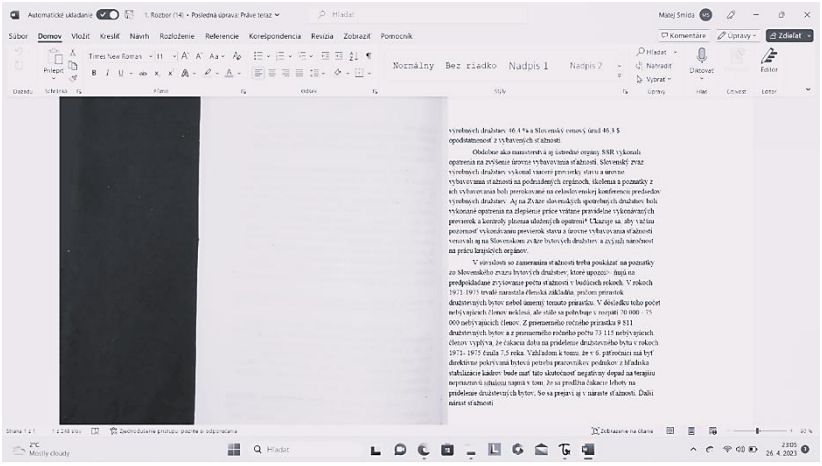
Z týchto dôvodov musíme usúdiť, že transkripcia našej vzorky v prostredí *Adobe Acrobat* sa neosvedčila, a aj napriek nízkej časovej náročnosti (na rozdiel od *Transkribu* nie je potrebná segmentácia dokumentov), výsledky tohto experimentu sú absolútne neuspokojivé. Tak ako v prostredí *Transkribus* to pripisujeme jednak úsekom s nízkou kvalitou strojopisu, problémom s transkripciou tabuliek, zakrivením textového poľa, čo je pre OCR softvér veľmi zmätočné (potvrdené aj v *Transkribe*).

[15] SLOVÁK, Lukáš: Používate OCR programy? Na týchto veciach si dajte záležať. In: *Lexika – preklady a tlmočenie* [online], 2020, [cit. 2023-04-25]. Dostupné na internete: https://www.lexika.sk/blog/pouzivate-ocr-programy-na-tychto-veciach-si-dajte-zalezat/

Obr. č. 8.: *Ukážka prepisu v Adobe Acrobat (str. 14)*

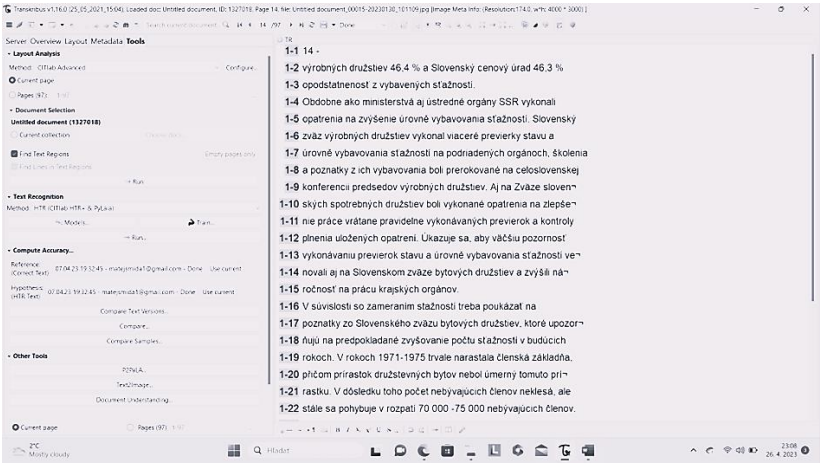


Obr. č. 9.: *Ukážka prepisu v ABBYY FineReader (str. 14)*



Úplne inú skúsenosť máme so softvérom *ABBYY FineReader*, ktorý nám bol poskytnutý prostredníctvom Univerzitnej knižnice UMB v Banskej Bystrici. Tak ako pri Adobe Acrobat, tak aj v tomto prípade sme zvolili rovnakú 52 stranovú vzorku, ktorú sme nahrali priamo do prostredia *ABBYY FineReader* a spustili jej transformáciu do programu Word a Excel. Celkovo nám tento proces trval len niekoľko minút a taktiež, ako v predchádzajúcom prípade, nebola nutná segmentácia dokumentu.

Obr. č. 10.: *Ukážka neopravenej transkripcie v prostredí Transkribus (str. 14)*



Výsledky však boli diametrálne odlišné, ako v prípade Adobe Acrobat. Text už na prvý pohľad pôsobil uhladene (ustálená veľkosť písma 10 a typ Times New Roman), preto s ním bolo možné prakticky okamžite ďalej pracovať. Narazili sme taktiež na veľmi nízku, miestami vôbec žiadnu chybovosť (občas sme však narážali na zamieňanie grafém, chýbajúcu diakritiku alebo nesprávne umiestnenie medzier). Ďalej musíme vyzdvihnúť excelentné čítanie tabuliek, ktorých vonkajšia štruktúra sa v porovnaní s originálom takmer nezmenila. Najväčší problém, na ktorý sme pri práci s *ABBYY FineReaderom* narazili, bola potreba správnej orientácie snímok. V našom prípade pri snímkach, ktorých text bol prispôsobený pre vodorovnú orientáciu snímky a jej orientácia nebola tomu uspôsobená, dochádzalo k jej chybnému čítaniu a tieto strany našej vzorky boli nepoužiteľné.

V konečnom dôsledku ale prácu s *ABBYY FineReaderom* môžeme zhodnotiť ako uspokojivú a po skúsenosti so softvérom *Adobe Acrobat* aj nad očakávania. Samozrejme, tak ako aj v predchádzajúcom prípade, sme sa stretli s problémom nekvalitného strojopisu prameňa, čo mohlo samotnú transkripciu ovplyvniť, ale pri tomto experimente to bolo v omnoho menšej miere. Vyzdvihujeme tiež excelentnú úpravu textu a tabuliek vo Worde, zanedbateľnú chybovosť, nízku časovú náročnosť a okamžitú možnosť tento preformátovaný dokument používať, bez nutnosti väčšej kontroly. Je potrebné podotknúť, že OCR softvéry sú vhodné na prácu so strojopisnými dokumentmi len v prípade, ak nám ide čisto len o prepis textu a jeho digitálnu transformáciu, keďže ide o softvéry založené na inej báze ako Transkribus a vznikli s iným cieľom.

Záver

Na základe analýzy a komparácie výstupov z nášho výskumu, ktorého atribúty sme zhrnuli v úvode, môžeme vyvodiť nasledujúce závery, ktoré podopierame naším výskumným procesom:

- *Transkribus* ponúka obrovskú pridanú hodnotu historikovi, ktorý sa prostredníctvom práce v ňom zoznamuje s možnosťami analýzy a prezentácie prameňov vo virtuálnom prostredí, s digitalizovaním archívnych dokumentov prostredníctvom zariadenia ScanTent, či automatickou transkripciou týchto prameňov. V dnešnej zdigitalizovanej dobe, ktorá sa podpisuje už aj na postupoch a metódach práce historika, je to veľmi cenná skúsenosť.
- Z hľadiska pracovného nasadenia a doby trvania celého procesu však nemôžeme potvrdiť jeho efektivitu (z tohto hľadiska sa nám viac osvedčil OCR softvér *ABBYY FineReader*). Je potrebné zdôrazniť, že toto tvrdenie je uplatniteľné len v prípade strojopisného dokumentu a aj to do veľkej miery závisí od jeho vonkajších parametrov. Z časového rámca bol veľmi vyčerpávajúci pomerne zložitý proces segmentácie dokumentov a opráv, pri ktorých sme strávili celé hodiny. V tomto prípade bol omnoho efektívnejší *ABBYY FineReader* a *Adobe Acrobat*, kde nebola potrebná žiadna segmentácia, čo sa podpísalo na značne kratšom procese transkripcie (cca pol hodiny).
- Analýza a porovnanie výsledkov transkripcie hovorí jasne v prospech *ABBYY FineReaderu*, ktorý sa nám javil z hľadiska času, námahy a napokon aj konečných výsledkov transkripcie ako bezkonkurenčný. Ako najhoršiu možnosť hodnotíme sorftvér Adobe Acrobat, ktorého prepis bol do značnej časti nepoužiteľný. Po oprave transkripcie v prostredí *Transkribus* sme narazili na veľmi vysokú chybovosť dokumentov (viď. štatistický rozbor), ktorá bola oproti *ABBYY FineReaderu* neporovnateľne väčšia. Tento stav prisudzujeme rôznym faktorom, z ktorých najviac zvýrazňujeme nízku kvalitu strojopisu, zásahy do dokumentov, nevodorovné rozloženie textového poľa na strane, značné problémy s transkripciou tabuliek. Ďalší objektívny faktor je aj meniaca sa podoba strojopisu v jednotlivých dokumentoch, čo sa odzrkadľovalo aj na rozdielnej chybovosti jednotlivých prameňov. Ani po vytvorení vlastných modelov na transkripciu sme, žiaľ, nenarazili na nejaký prevratný pokles v chybovosti. Tú sme dosiahli len čiastočne pri jednom konkrétnom dokumente v rámci našej pramennej vzorky, na ktorej bol daný model postavený. Pri iných dokumentoch, naopak, tento pokles nebol zaznamenaný v takej miere alebo došlo skôr k nárastu chybovosti. Preto pri tvorbe vlastného modelu odporúčame, aby

používateľ pracoval s dokumentom, ktorý má rovnaký typ písma, rovnakú skladbu, vonkajšiu formu a pod. V tom prípade predpokladáme uspokojivejšie výsledky.
- Na rozdiel od *Adobe Acrobatu* alebo *ABBYY FineReaderu* obrovská pridaná hodnota *Transkribu* je práca s metadátami, ktoré ponúkajú ďalšie možnosti zužitkovania transkripcie a informácií vyplývajúcich z prameňa.

Na základe predostretých záverov môžeme tvrdiť, že v prípade, ak používateľovi ide iba o samotnú transkripciu strojopisu a jeho preformátovanie do programu Word alebo PDF, je jednoznačne výhodnejšie využiť na tento účel iný OCR softvér, z ktorých sa nám osvedčil *ABBY FineReader*. Ušetrený čas, námaha a aj samotný výsledok transkripcie hovorí jednoznačne v jeho prospech.

Avšak v prípade snahy o hlbšiu analýzu prameňa a možnosti jeho editovania má *Transkribus* nezanedbateľnú pridanú hodnotu, preto z tohto hľadiska je prínos oftvéru fenomenálny. Dôležité je tiež poznamenať, že v *Transkribe* je práca so strojopisným typom prameňa len v začiatkoch a je tu potenciál na to, aby sa po vytvorení kvalitného modelu mohol rovnať alebo dokonca predbehnúť OCR softvéry.

V každom prípade nám ale nedá nepodotknúť, že základom každej úspešnej transkripcie je vysoká kvalita textu a kvalita rozlíšenia samotného digitalizátu, bez ktorého je každá obdobná snaha zbytočná.