Obrazová data a strojové učení

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně

Obsah

- Úvod
- Projekt PERO
- Identifikace obrázků na stránce
- Systém VISE

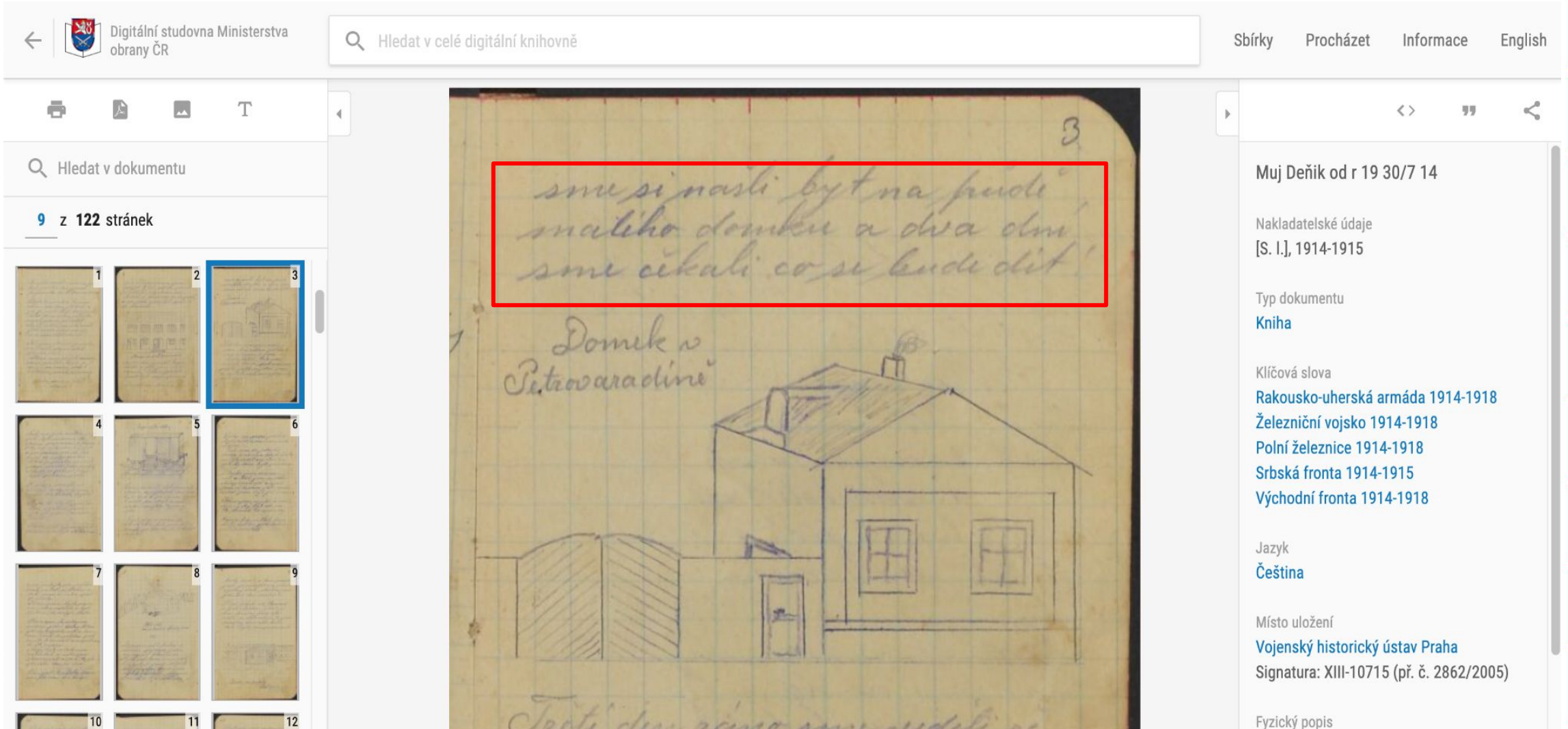
Strojové učení a proces digitalizace

- Ořez, vyrovnání stránky, barevné podání
- Scelování dokumentů
- Struktura dokumentu
- OCR
- ...

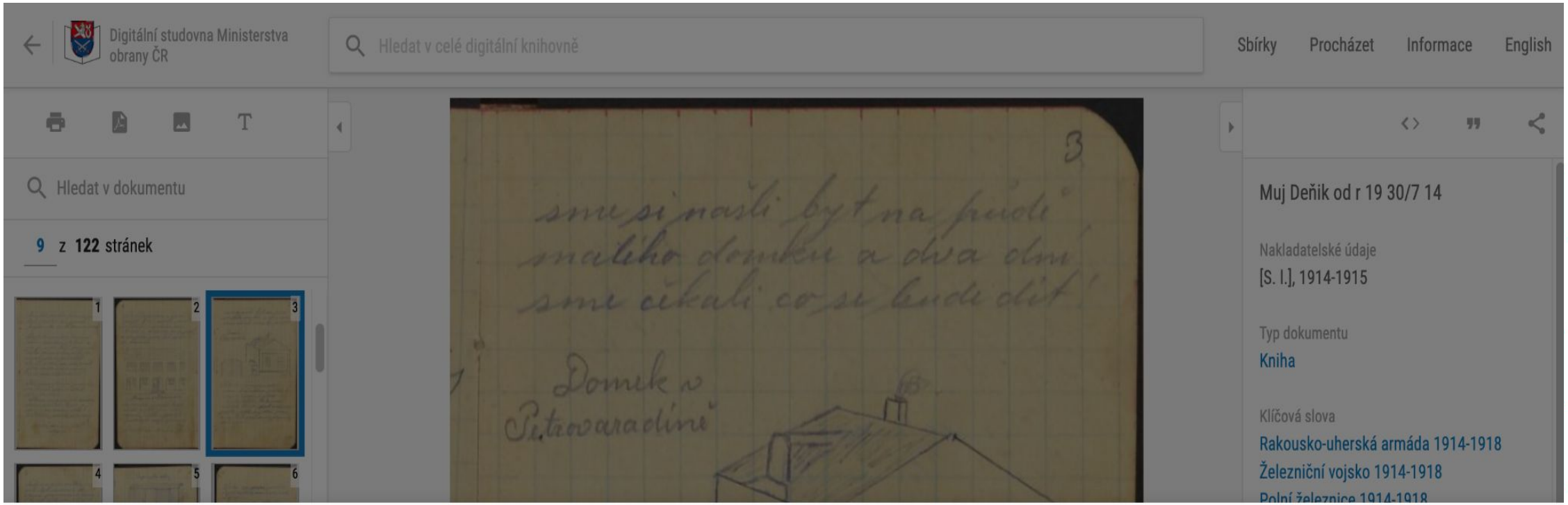
Strojové učení a zpřístupňování

- Vyhledávání obrázků
- Vyhledávání a podobnost částí stránek
- Identifikace obsahu obrázku
- ...

OCR - projekt PERO



OCR - projekt PERO

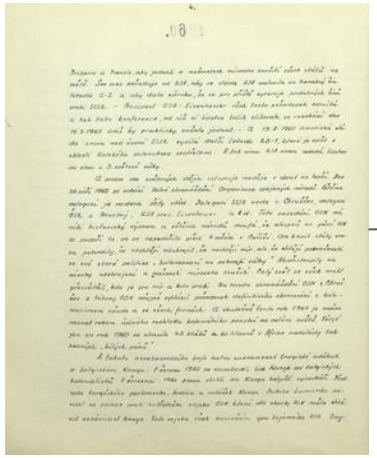


sme si našli byt na půdě malého domku a dva dní sme čekali co se bude dít!

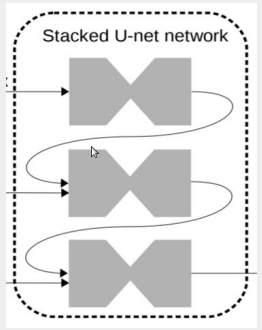
Muj Deňik od r 19 30/7 14. [S. l.], 1914-1915, s. 3. Dostupné také z: http://www.digitalniknihovna.cz/dsmo/uuid/uuid:32f7ba5e-a323-11ea-95c0-001b63bd97ba

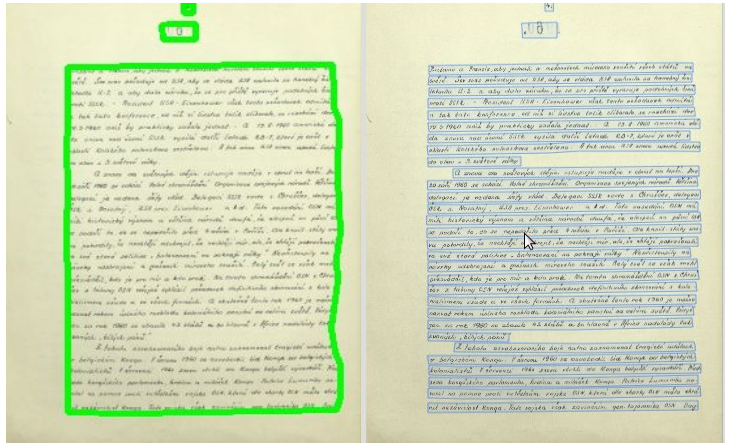
OCR - projekt PERO

Řetězec procesu rozpoznání textu

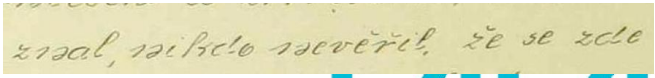


Layout





Crop line



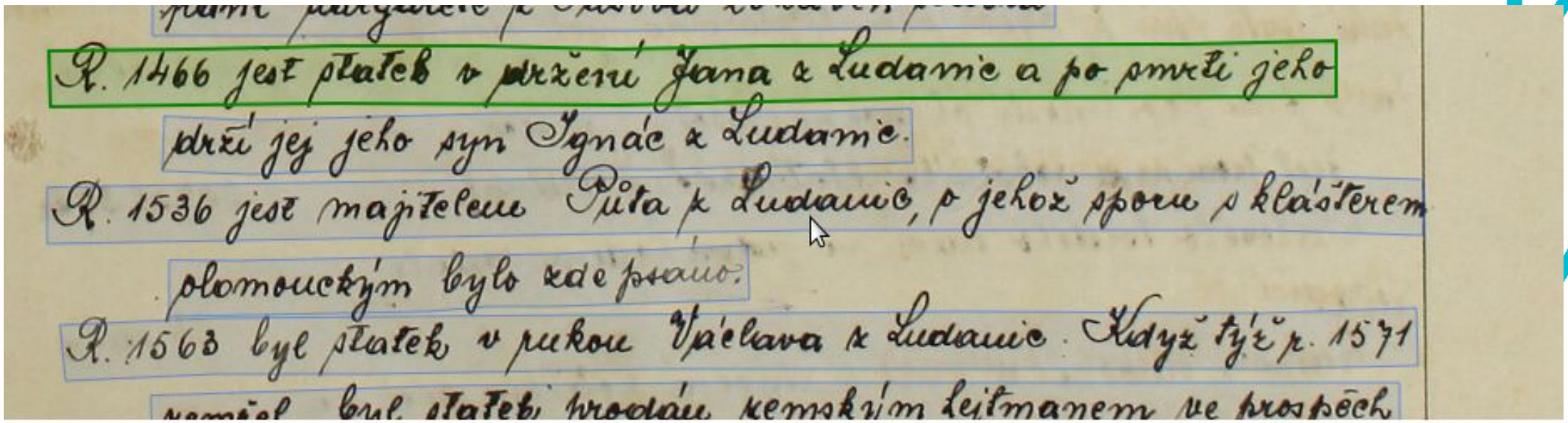
OCR network

Language model

Roku 1906 se stal předsedou spolku přátel školy a podílel se na financování a výstavbě nové školní budovy v Ratibořech nad Pekárkou. Ze své funkce byl odvo-

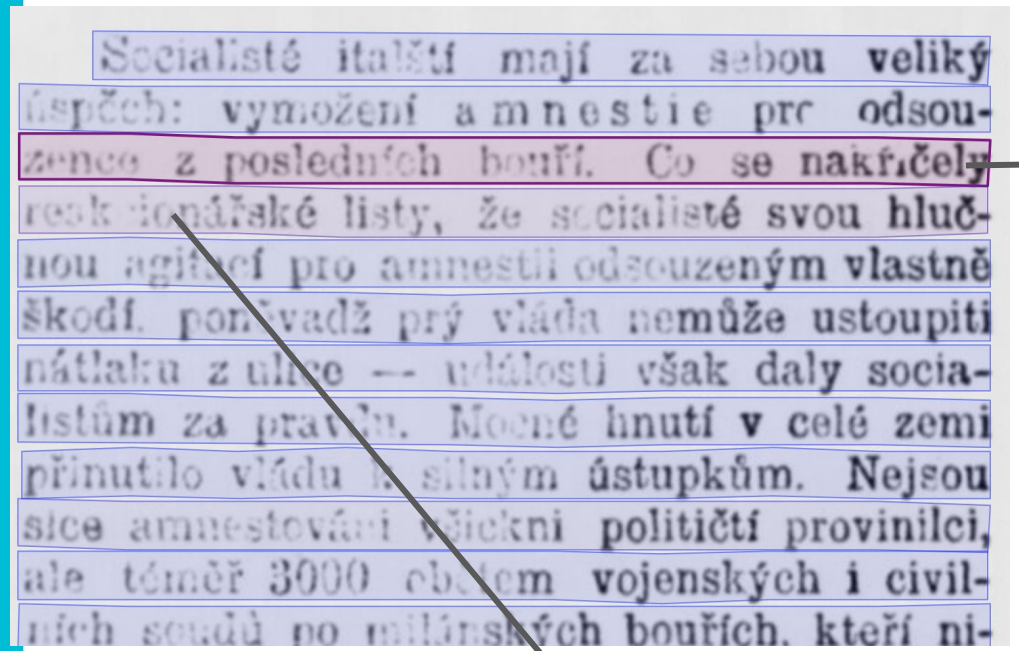
Web application

České kroniky 20. století



R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho
drží jej jeho syn Ignác z Ludanic.
R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášterem
olomouckým bylo zde psáno.
R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571

Periodika na mikrofilmech

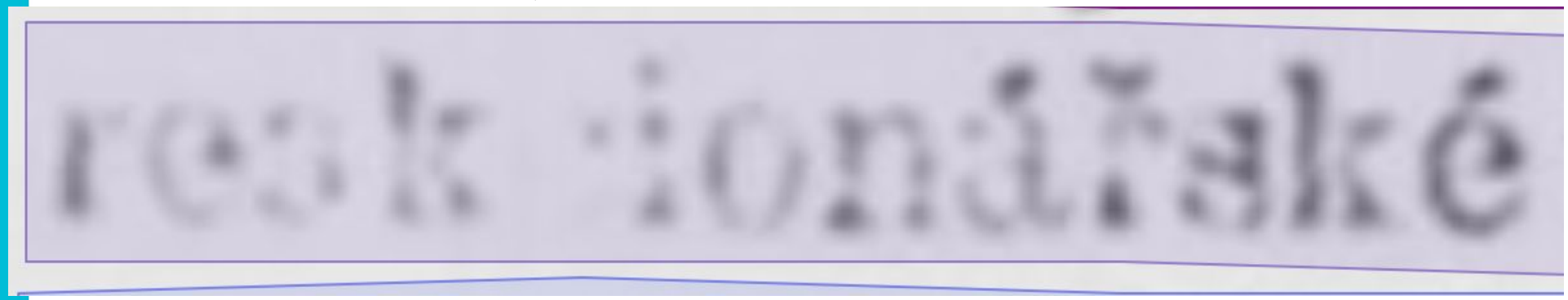


Socialisté italští mají za sebou veliký úspěch: vymožení amnestie pro odsouzence z posledních bouří. Co se nakřičely reskcionářské listy, že socialisté svou hlučnou agitací pro amnestii odsouzeným vlastně škodí, poněvadž prý vláda nemůže ustoupiti nátlaku z ulice — události však daly socialistům za pravdu. Mocné hnutí v celé zemi přinutilo vládu k silným ústupkům. Nejsou sice amnestováni všichni političtí provinilci, ale téměř 3000 obětem vojenských i civilních soudů po milánských bouřích, kteří ni-

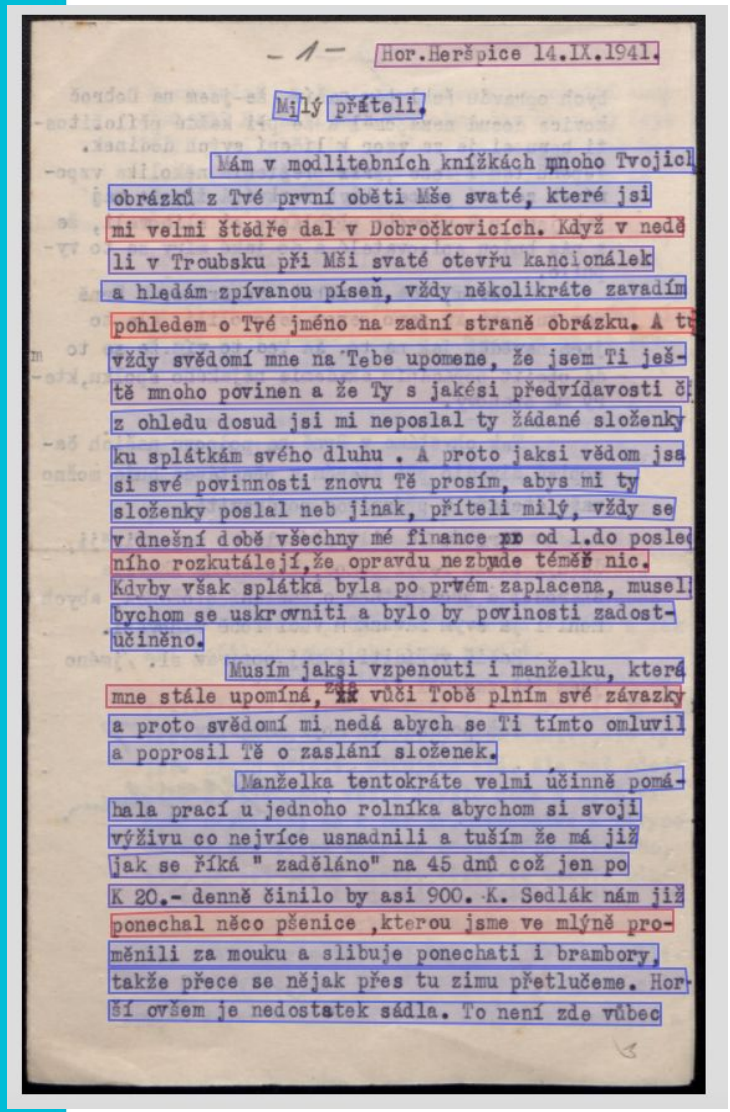
nakřičely

zence z posledních bouří. Co se nakřičely

reskcionářské listy, že socialisté svou hlučné



Strojopis



– 1 – Hor.Heršpice 14.IX.1941.

Milý příteli,

Mám v modlitebních knížkách mnoho Tvojich obrázků z Tvé první oběti Mše svaté, které jsi mi velmi štědře dal v Dobročkovicích. Když v neděli v Troubsku při Mši svaté otevřu kancionálek a hledám zpívanou píseň, vždy několikráte zavadím pohledem o Tvé jméno na zadní straně obrázku. A tu vždy svědomí mne na Tebe upomene, že jsem Ti ještě mnoho povinen a že Ty s jakési předvídavosti čili z ohledu dosud jsi mi neposlal ty žádané složenky ku splátkám svého dluhu . A proto jaksi vědom jsa si své povinnosti znovu Tě prosím, abys mi ty složenky poslal neb jinak, příteli milý, vždy se v dnešní době všechny mé finance xx od 1.do posled ního rozkutálejí,že opravdu nezbude téměř nic. Kdyby však splátka byla po prvém zaplacena, museli bychom se uskrovniti a bylo by povinosti zadostučiněno.

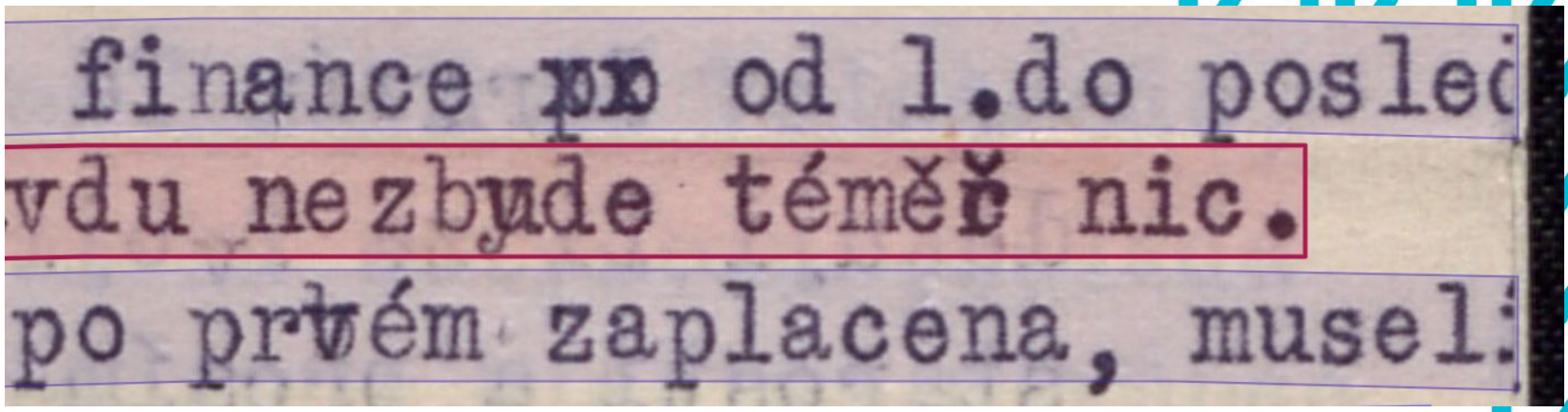
Musím jaksi vzpenouti i manželku, která mne stále upomíná, xxx vůči Tobě plním své závazky a proto svědomí mi nedá abych se Ti tímto omluvil a poprosil Tě o zaslání složenek.

Manželka tentokráte velmi účinně pomáhala prací u jednoho rolníka abychom si svoji výživu co nejvíce usnadnili a tuším že má již jak se říká " zaděláno" na 45 dnů což jen po K 20.- denně činilo by asi 900.-K. Sedlák nám již ponechal něco pšenice ,kterou jsme ve mlýně proměnili za mouku a slibuje ponechati i brambory, takže přece se nějak přes tu zimu přetlučeme. Horší ovšem je nedostatek sádla. To není zde vůbec

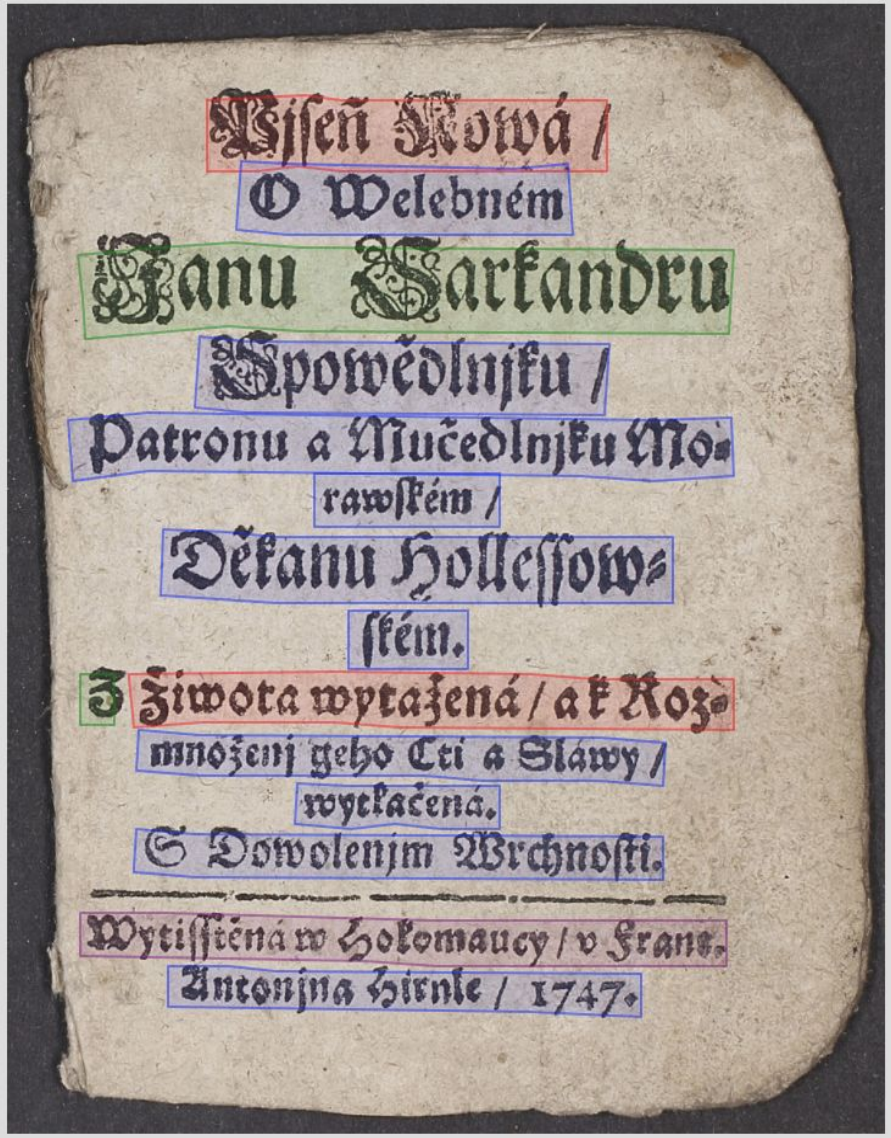
v dnešní době všechny mé finance od 1. do posled

ního rozkutálejí,že opravdu nezbude téměř nic.

Kdyby však splátka byla po prvém zaplacena, museli

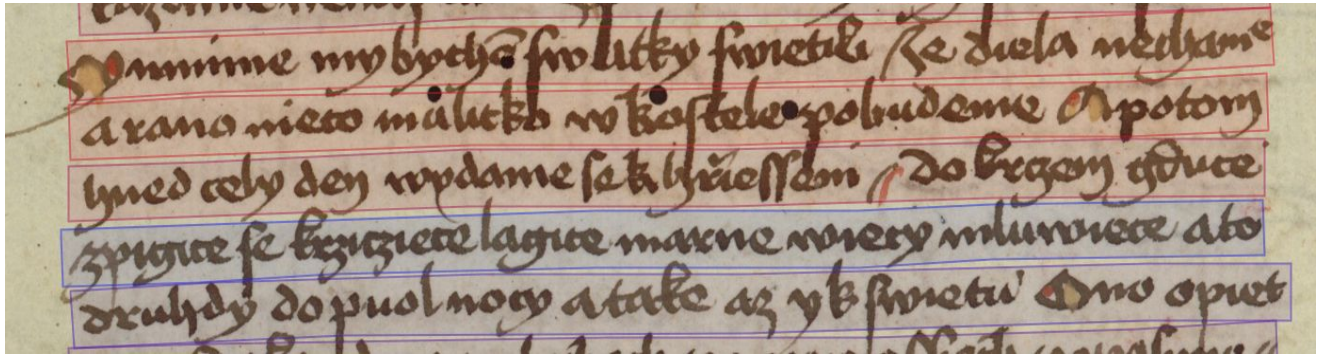


Kramářské tisky

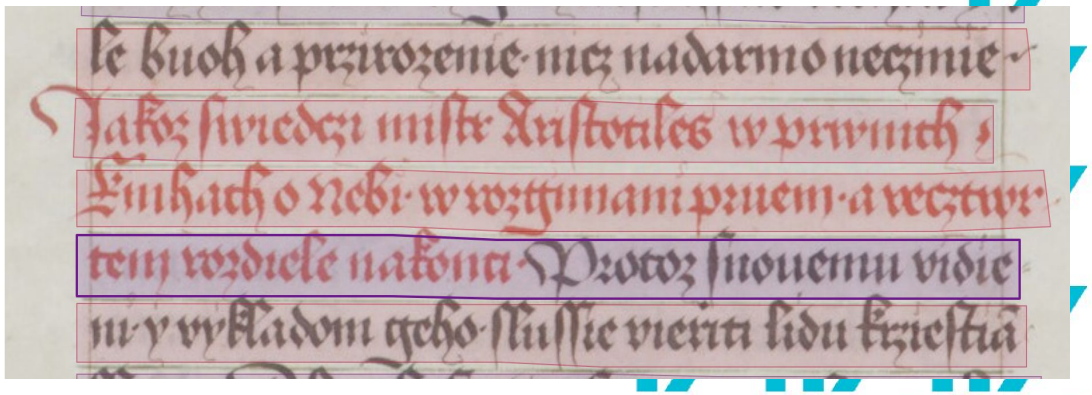


Pjſeň Nowá/
O Welebném
Janu Sarkandru
Spowědlnjku/
Patronu a Mučedlnjku Mo=
rawſkém/
Děkanu Holleſſow=
ſkém.
Z
žiwota wytažená/ a k Roz=
množenj geho Cti a Sláwy/
wytľačená.
S Dowolenjm Wrchnoſti.
Wytiſſtěná w Hoľomaucy/ v Frane.
Antonjna Hirnle/ 1747.

Jan Hus, Vavřinec z Březové

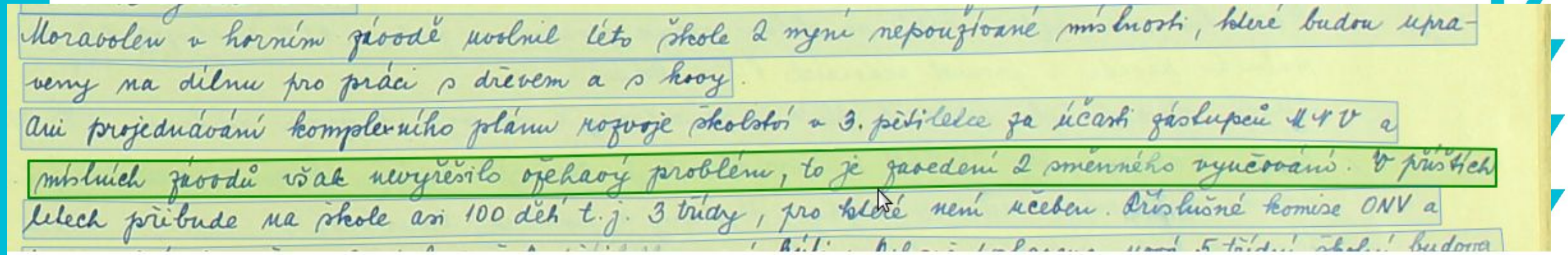


I mnime mybycha ſwatky ſwietili ſe diela nedani
a rano nieto malitko w koſtee pobudeme A potom
hned cely den wydame ſek hrieſſem Do krczem gduce
zpigice ſe krzicziece agite marne wiery mluwiece a to
druhdy do puo nocy a take az ys ſwietu Ono opiet



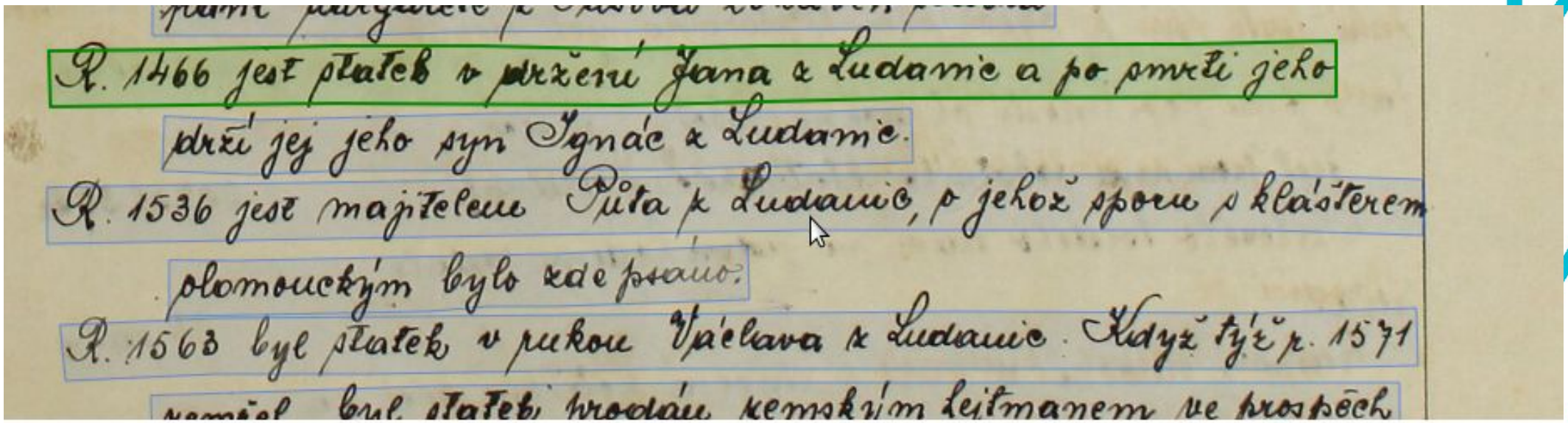
le buoh a przirozenie·niez nadarmo neczиie-
Jakoz ſwiedczi miſtr Ariſtocileś w prwnich
Eiihach o Nebi w rozgiani pruem·a vecztryr
tem rozdiele nakona Protoz ſnouemu vidie
ni·y vykladom geho ſluſſie vieriti lidu krzieſtia

České kroniky 20. století



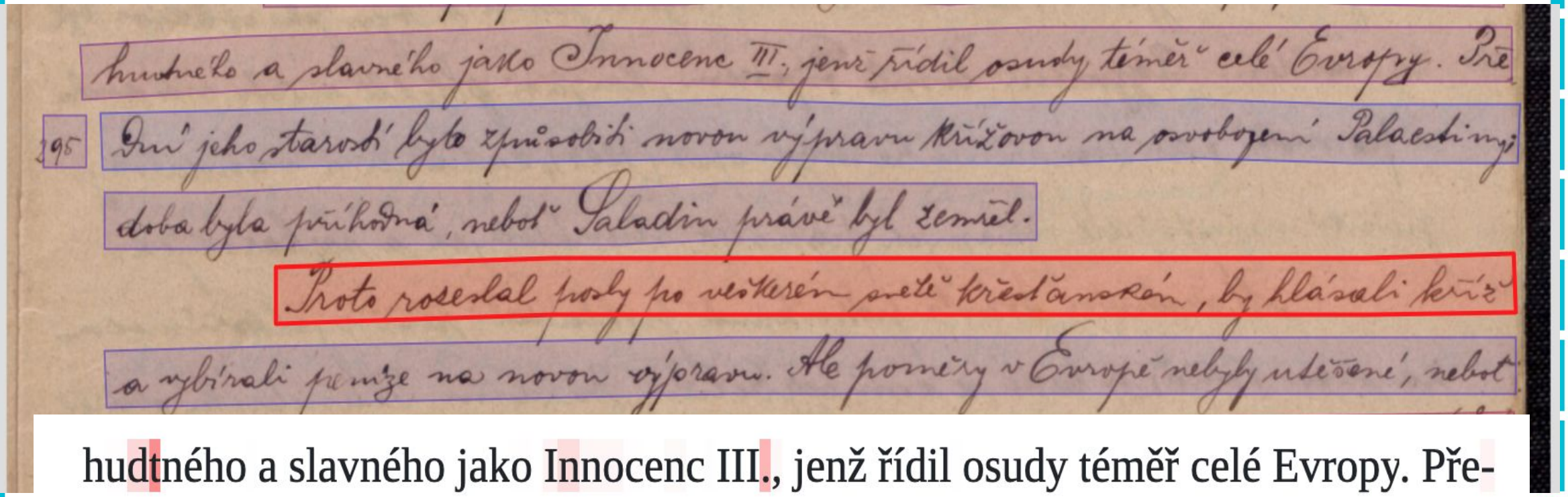
Moravolen v horním závodě uvolnil této škole 2 nyní nepoužívané místnosti, které budou upra-
veny na dílnu pro práci s dřevem a s kovy
ani projednávání komplexního plánu rozvoje školství v 3. pětiletce za účasti zástupců MNV a
místních závodů však nevyřešilo ožehavý problém, to je zavedení 2 směnného vyučování. V příštích
letech přibude na škole asi 100 dětí t. j. 3 třídy, pro které není učeben. Příslušné komise ONV a

České kroniky 20. století

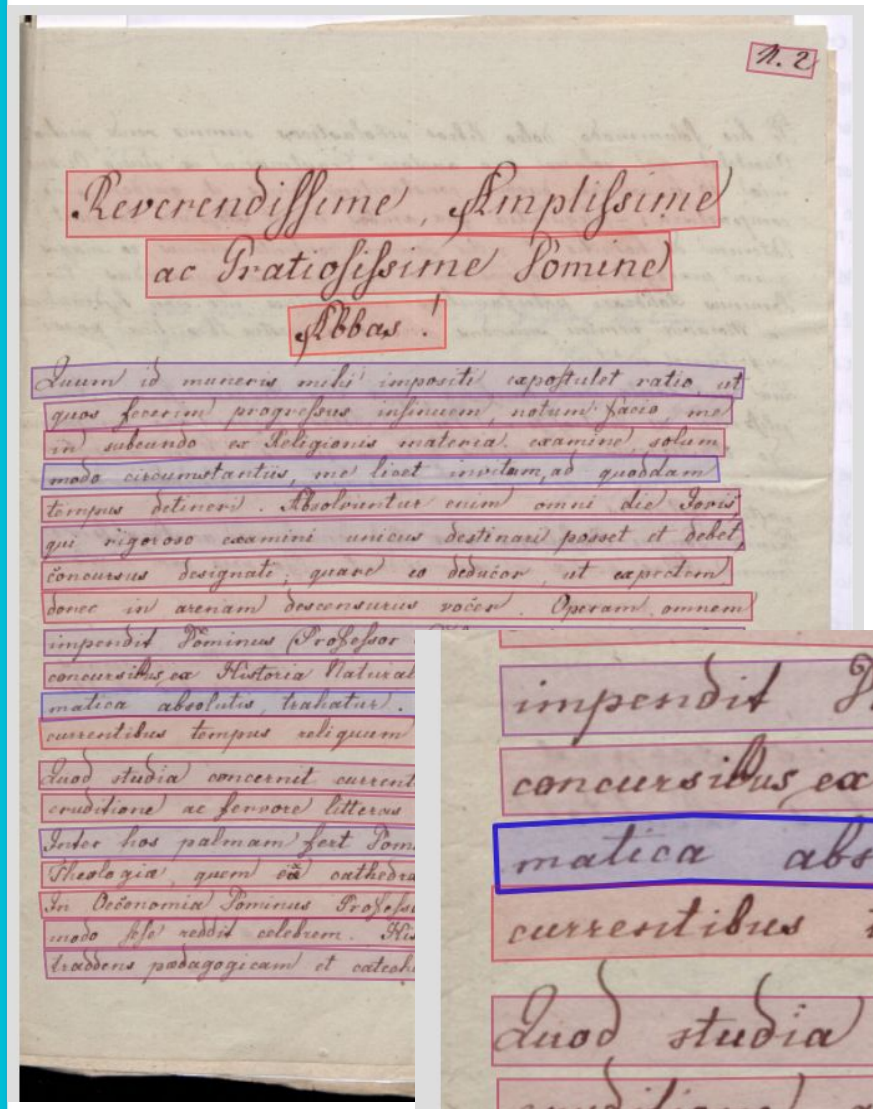


R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho
drží jej jeho syn Ignác z Ludanic.
R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášterem
olomouckým bylo zde psáno.
R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571

Rukopisy

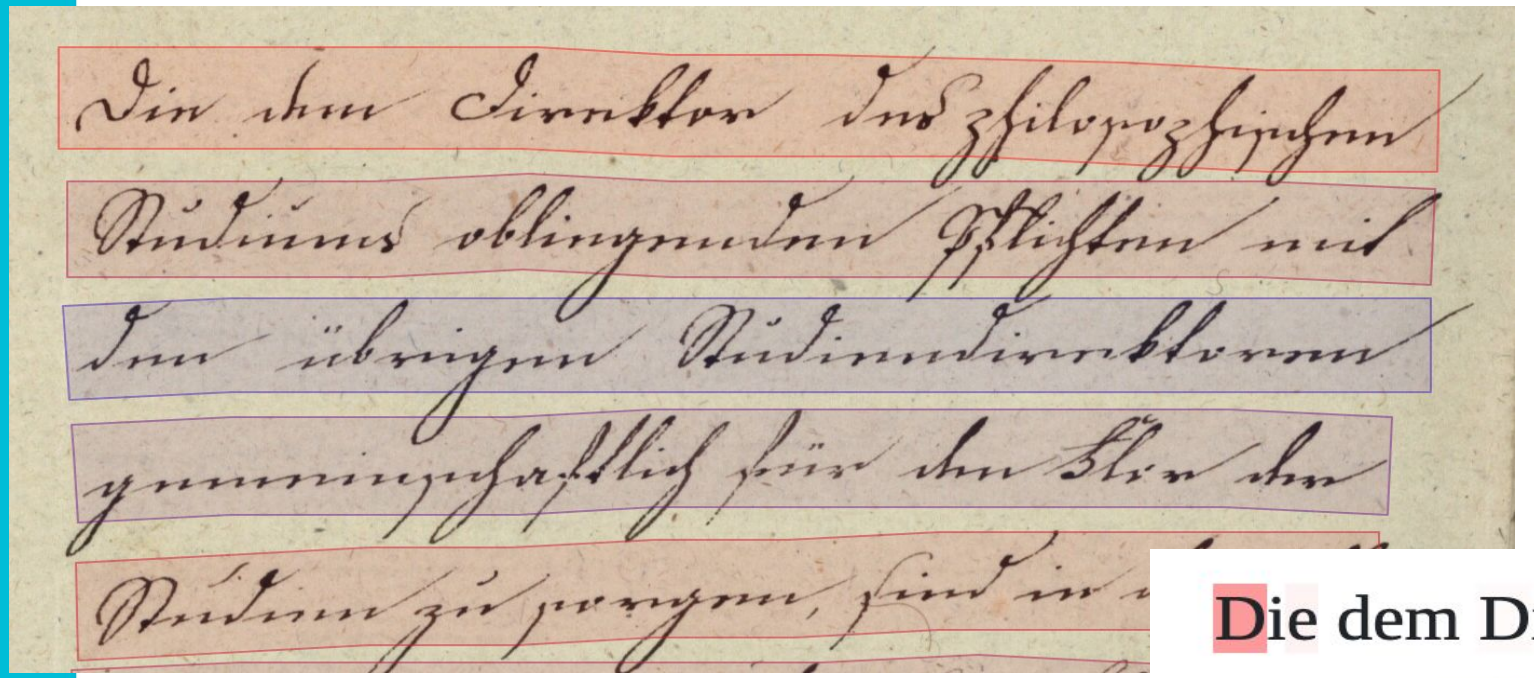

hudtného a slavného jako Innocenc III., jenž řídil osudy téměř celé Evropy. Pře-
dní jeho starostí bylo způsobiti novou výpravu křížovou na osvobození Palaestiny,
doba byla příhodná, neboť Saladin právě byl zemřel.
Proto rozeslal posly po veškerém světě křesťanském, by hlásali kříž
a vybírali peníze na novou výpravu. Ale poměry v Evropě nebyly utěšené, neboť

Rukopisy



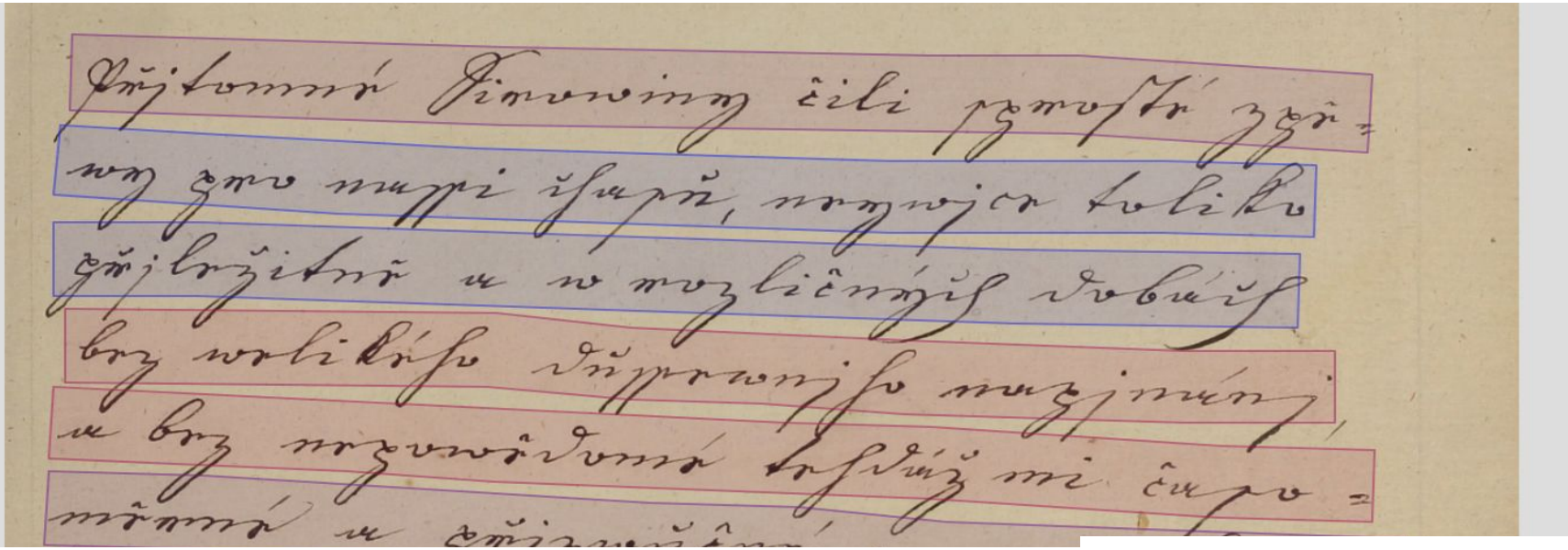
impendit Pominus Proſeſsor Löcher ne mora amplius,
concursilus ex Historia Naturali, Diplomatica, et Nemis,
matica absolutis, trahatur Interea omne a studis
currentibus tempus reliquum materia revolvenda conseoro.
Quod, studia concernit currentia, nactus sum dučès ingenio

Kurent



Die dem Direktor, des philoſophiſchen
Studiums obliegenden Pflichten mit
den übrigen Studiendirektoren
gemeinſchaftlich für den Fler der
Studien zu ſorgen, ſind in der all„

Kurent



Přjtomné Sirowiny čili ſproſté zpě„
wy pro naſſi chaſu, neywjce toliko
přjležitně a w rozličných dobách
bez welikého duſſewnjho napjmáej
a bez nepowědomé tehdáž mi čapo=

PERO - důležité odkazy

- Jádro OCR - pero-ocr python balíček https://github.com/DCGM/pero-ocr
- Webová aplikace pro kontrolu a opravy - pero_ocr_web
  - Běží na https://pero-ocr.fit.vutbr.cz
  - Zdrojové kódy https://github.com/DCGM/pero_ocr_web
- OCR API pro hromadné zpracování
  - https://pero-ocr.fit.vutbr.cz/api
- Informace o projektu - https://pero.fit.vutbr.cz/

Identifikace obrázků na stránce

- Cílem je automaticky detekovat pozici obrázků v naskenovaných dokumentech
- **Proč je to těžké?**

O PRÁCI VÝBORŮ ŽEN

Úspěchy a chyby libeňského výboru žen

Milena Majorová

Výbor žen při obvodním národním výboru v Praze-Libni jsme navštívili při jedné z jeho pravidelných čtrnáctidenních schůzí, kterých se také zúčastňuje předseda ONV soudruh Dohnal. Program schůzí je již předem určen pracovním plánem. Toto úterý je na pořadu církevní otázka, o které referoval soudruh Vacek z ONV. Soudružky sledovaly s velkým zájmem výklad soudruha Vacka o vatikánských příslušovačích kapitalismu a o jejich pletichách v našem státě. Ale naši věřící se ve své velké většině nedali zmást církevním panstvem. Příkladem toho je na příklad členka výboru žen, paní Holubičková. Je příslušnicí Čs. strany lidové a její práci předsedkyně výboru žen, soudružka Červenková, velmi kladně hodnotí.

Paní Holubičková dostala od výboru žen za úkol získat chlapce pro hornictví v rámci Lánské akce. Úspěch se dlouho nedostavoval. Paní Holubičková se však nevzdala. Chodila do škol, do rodin, rozmlouvala s chlapci, s učiteli i rodiči. Do akce zapojila i svou organizaci Čs. strany lidové. A výsledek? Šest hochů získala paní Holubičková sama a celkem Libeň vyslala 18 chlapců na hornictví. Těchto paní Holubičkové může být, že to byli chlapci s nejlepším prospěchem, a že mnozí z nich byli přijati do vyšší hornické školy.

Jinou úspěšnou akcí libeňského výboru žen byl nábor vdaných žen do těžkého průmyslu. Praha 8 měla za úkol získat celkem 241 žen. Soudružky z výboru žen utvořily šestatřicet dvojic a neúnavně chodily přesvědčovat ženy v domácnosti o nutnosti jejich vstupu do práce. Setkaly se s případy, že jim ani nebylo otevřeno, nebo se ženy dávaly zapřít. Ale soudružky se nedaly odradit. Dobrá, dnes nám zde neotevřeli, přijdeme tedy po druhé, po třetí, pokud to bude třeba, kolikrát bude třeba. Soudružky měly i takový případ, kdy se žena nezdráhala jít do zaměstnání, ale její muž, dokonce „uvědomělý" soudruh, jí bránil. Prý jeho žena půjde do zaměstnání tehdy, uzná-li to on za vhodné. Ale ani takové nepochopení nezneuchutilo soudružkám jejich úkol. Zhostily se ho s úspěchem. Získaly celkem 252 žen, které jsou vydatnou pomocí blízkým továrnám těžkého průmyslu.

Libeňský výbor žen má velkou zásluhu také o zdraví libeňských žen. Soudružka Dr Wiklická, lékařka z gynekologického oddělení nemocnice na Bulovce, již před třemi lety, jako členka Svazu žen v Libni začala provádět podle sovětského vzoru preventivní léčbu rakoviny u žen. Libeňský svaz žen a nyní i výbor žen propagoval mezi všemi libeňskými ženami od 35 do 60 let lékařskou prohlídku u soudružky Dr Wiklické. Tyto výzvy výboru žen spolu s přednáškami soudružky Wiklické o rakovině, které opět organisoval výbor žen a které byly velmi četně ženami navštěvovány, způsobily, že již asi 6.000 libeňských žen bylo vyšetřeno a v případě potřeby léčeno.

Výbor žen získává libeňské ženy také na zemědělské brigády do okolních státních statků. Ženy pracovaly při denních i nočních výmlatech, pomáhaly se sklizní řepy a zúčastnily se chmelové brigády na Žatecku. Významně pomohly zdolávat mandelinku bramborovou na bohnickém státním statku.

Je samozřejmé, že se libeňský výbor žen stará i o školy, školky a jesle ve svém obvodu. Ve spolupráci se školským referátem obvodního národního výboru odstraňuje jednu menší nedostatky, které zjistil. Ale necouvá ani před překážkami většími, jako je nedostatek místa ve školách. Ačkoliv tento problém není ještě zcela odstraněn, přece jen neúnavná práce výboru žen v tomto směru dává záruku, že bude nalezeno vyhovující řešení.

O tom, že výbor žen jíž v Libni dobře pracuje, svědčí také to, že besedy, které pořádá, bývají četně navštěvovány. Dokonce přišly ženy na besedu i tehdy, kdy byl vypjat elektrický proud. Byla to právě beseda o mírovém kongresu.

„Nejsme však ještě se svou prací zcela spokojeny," říká předsedkyně soudružka Anežka Červenková. „Pro nás není a nemůže být měřítkem, že snad některé výbory žen v Praze pracují hůře než my. My musíme svou práci ještě v mnohém zlepšit. Domníváme se, že nám v tom velmi pomůže široký s b o r a k t i v i s t e k, který začínáme budovat. Znamená to, že každá členka výboru žen získá asi 20 až 30 žen, které nám pomohou s agitační prací mezi ostatními ženami. Máme prostě nejlepší snahu stát se skutečně dobrým výborem žen a jsme přesvědčeny, že při obětavosti našich soudružek se nám to jistě podaří."

Soudruh předseda obvodního národního výboru Dohnal se soudružkou předsedkyní výboru žen Červenkovou a dvěma členkami výboru žen při projednávání úkolů lidosprávy.

Věříme, že se to libeňským soudružkám skutečně podaří. Ale musíme přece jen poukázat na jednu chybu, které se dopouštějí a to nejen ony, ale i soudruzi z obvodního národního výboru. Členky výboru žen se totiž staly současně členkami stálých komisí při ONV. To jistě není správné, neboť práce v komisích je nutně zdržuje od práce ve výboru žen. Soudruzi jim dají, podle jejich názoru, třeba jen malý úkol: roznést několik dosítek pozvánek. Přihlédneme-li k tomu, že většina žen z výboru je zaměstnána, vyžádá si i tento malý úkol dosti času. A to ani nehovoříme o větších úkolech, které jsou soudružky jako členky stálých komisí povinny vykonávat.

Soudružky z výboru žen si musí uvědomit, že není účelem, aby se ony zapojily do lidosprávy, ale že je účelem, zapojit do lidosprávy jiné, schopné ženy, které při besedách nebo jiném styku s nimi poznají. Nejde o to, aby několik žen bylo zaplněch a každá z nich měla několik funkcí, ale jde o to, aby i z těch, zatím brzejmenných, rosty nové, uvědomělé kádry.

Snad se libeňské soudružky kolem sebe porozhlédnou a vyberou místo sebe do komisí při obvodním národním výboru nové, schopné ženy. A že takových má Libeň dost a dost, o tom jistě nebudeme pochybovat.

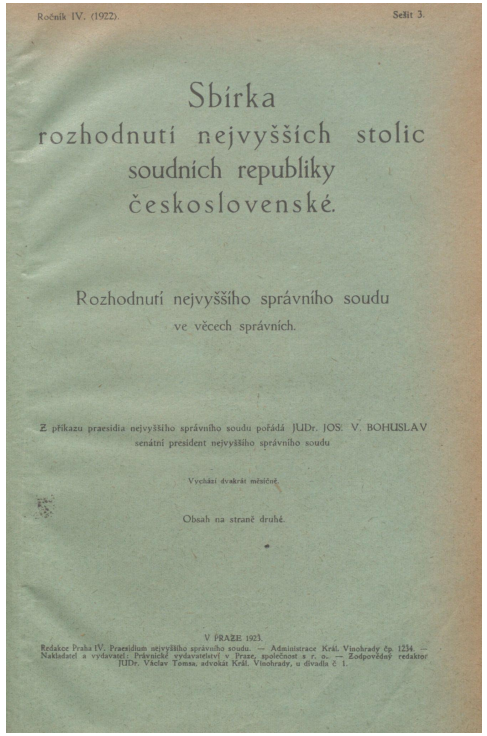
Předsedkyně soudružka Červenková se svými nejbližšími spolupracovnicemi. Foto Novák

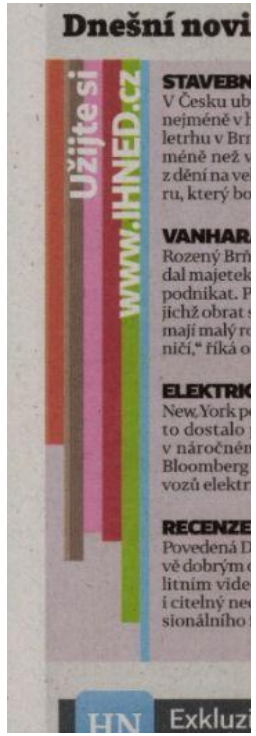
Členky výboru žen si dělají poznámky k referátu soudruha Vacka o církevních otázkách, aby mohly v diskusích se ženami správně argumentovat.

10

Co není text ještě nemusí být obrázek

- Grafické elementy, artefakty, pozadí

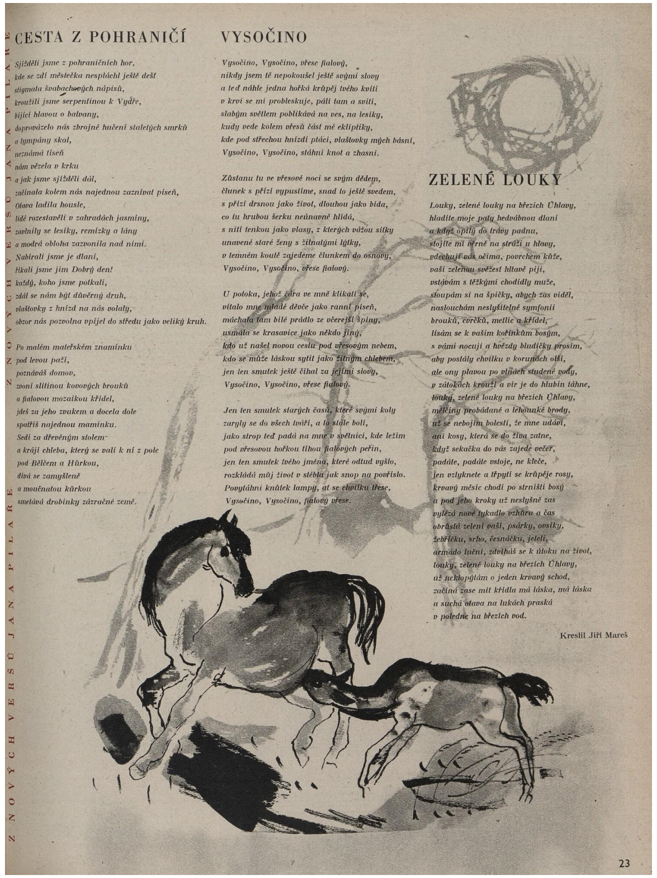




Co obsahuje text stále může být obrázek

- text v obrázcích, překryv

Obrázky nemusí být ohraničeny



Z NOVÝCH VERŠŮ JANA PILAŘE

CESTA Z POHRANIČÍ

Sjížděli jsme z pohraničních hor,
kde se zdí městečka nespláchl ještě dešť
stigmata švabachových nápisů,
kroužili jsme serpentinou k Vydře,
bijící hlavou o balvany,
doprovázelo nás zbrojné hučení staletých smrků
a tympány skal,
neznámá tiseň
nám vězela v krku
a jak jsme sjížděli dál,
začínala kolem nás najednou zaznívat píseň,
Otava ladila housle,
lidé rozestavěli v zahradách jasmíny,
zavlnily se lesíky, remízky a lány
a modrá obloha zazvonila nad nimi.
Nabírali jsme je dlaní,
říkali jsme jim Dobrý den!
každý, koho jsme potkali,
zdál se nám být důvěrný druh,
vlaštovky z hnízd na nás volaly,
obzor nás pozvolna vpíjel do středu jako veliký kruh.

Po malém mateřském znamínku
pod levou paží,
poznáváš domov,
zvoní slitinou kovových brouků
a fialovou mozaikou křídel,
jdeš za jeho zvukem a docela dole
spatříš najednou maminku.
Sedí za dřevěným stolem
a krájí chleba, který se valí k ní z pole
pod Bělčem a Hůrkou,
dívá se zamyšleně
a moučnatou kůrkou
smetává drobinky zázračné země.

VYSOČINO

Vysočino, Vysočino, vřese fialový,
nikdy jsem tě nepokoušel ještě svými slovy
a teď náhle jedna hořká krůpěj tvého kvítí
v krvi se mi probleskuje, pálí tam a svítí,
slabým světlem poblikává na ves, na lesíky,
kudy vede kolem vřesů část mé ekliptiky,
kde pod střechou hnízdí ptáci, vlaštovky mých básní,
Vysočino, Vysočino, stáhni knot a zhasni.

Zůstanu tu ve vřesové noci se svým dědem,
člunek s přízí vypustíme, snad to ještě svedem,
s přízí drsnou jako život, dlouhou jako bída,
co tu hrubou šerku neúnavně hlídá,
s nití tenkou jako vlasy, z kterých vážou sítky
unavené staré ženy s žilnatými lýtky,
v temném koutě zajedeme člunkem do osnovy,
Vysočino, Vysočino, vřese fialový.

U potoka, jehož čára ve mně klikatí se,
vítalo mne mladé děvče jako ranní píseň,
máchala tam bílé prádlo ze včerejší špíny,
usmála se krasavice jako někdo jiný,
kdo už našel novou cestu pod vřesovým nebem,
kdo se může láskou sytit jako žitným chlebem,
jen ten smutek ještě číhal za jejími slovy,
Vysočino, Vysočino, vřese fialový.

Jen ten smutek starých časů, které svými koly
zaryly se do všech tváří, a to stále bolí,
jako strop teď padá na mne v světnici, kde ležím
pod vřesovou hořkou tíhou fialových peřin,
jen ten smutek tvého jména, které odtud vyšlo,
rozkládá můj život v stébla jak snop na povříslo.
Povytáhni knůtek lampy, ať se chvilku třese,
Vysočino, Vysočino, fialový vřese.

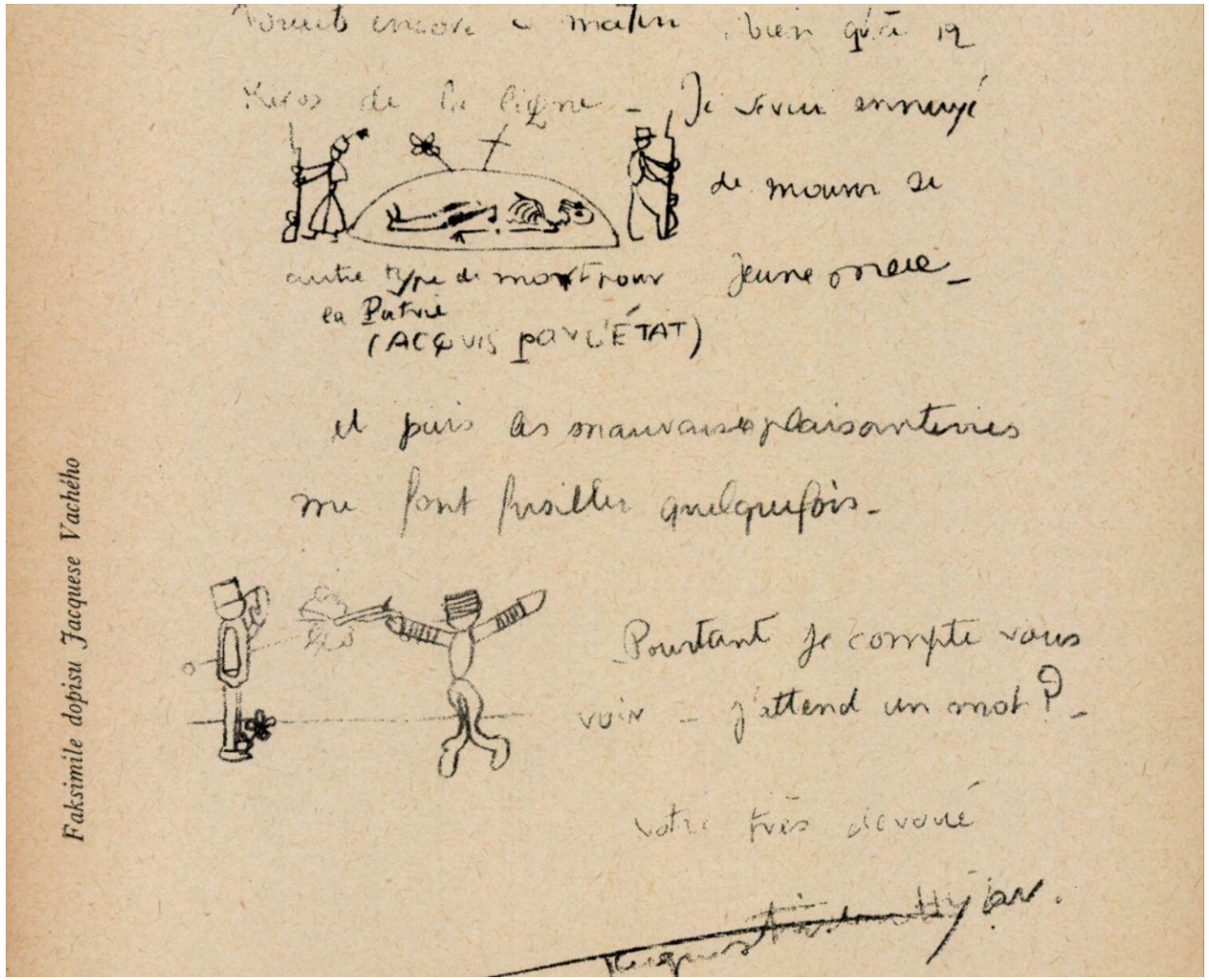
ZELENÉ LOUKY

Louky, zelené louky na březích Úhlavy,
hladíte moje paty hedvábnou dlaní
a když opilý do trávy padnu,
stojíte mi věrně na stráži u hlavy,
vdechuji vás očima, povrchem kůže,
vaši zelenou svěžest hltavě piji,
vstávám s těžkými chodidly muže,
stoupám si na špičky, abych zas viděl,
naslouchám neslyšitelné symfonii
brouků, cvrčků, metlic a křídel,
lísám se k vašim kořínkům bosým,
s vámi nocuji a hvězdy bludičky prosím,
aby postály chvilku v korunách olší,
ale ony plavou po vlnách studené vody,
v zátokách krouží a vír je do hlubin táhne,
louky, zelené louky na březích Úhlavy,
mělčiny probádané a lehounké brody,
už se nebojím bolesti, že mne udáví,
ani kosy, která se do živa zatne,
když sekačka do vás zajede večer,
padáte, padáte vstoje, ne kleče,
jen vzlyknete a třpytí se krůpěje rosy,
krvavý měsíc chodí po strništi bosý
a pod jeho kroky už neslyšně zas
vylézá nově tykadlo vzhůru a čas
obrůstá zelení vaší, psárky, ovsíky,
žebříčku, srho, česnáčku, jeteli,
armádo luční, zdviháš se k útoku na život,
louky, zelené louky na březích Úhlavy,
už neklopýtám o jeden krvavý schod,
začíná zase mít křídla má láska, má láska
a suchá otava na lukách praská
v poledne na březích vod.

Kreslil Jiří Mareš

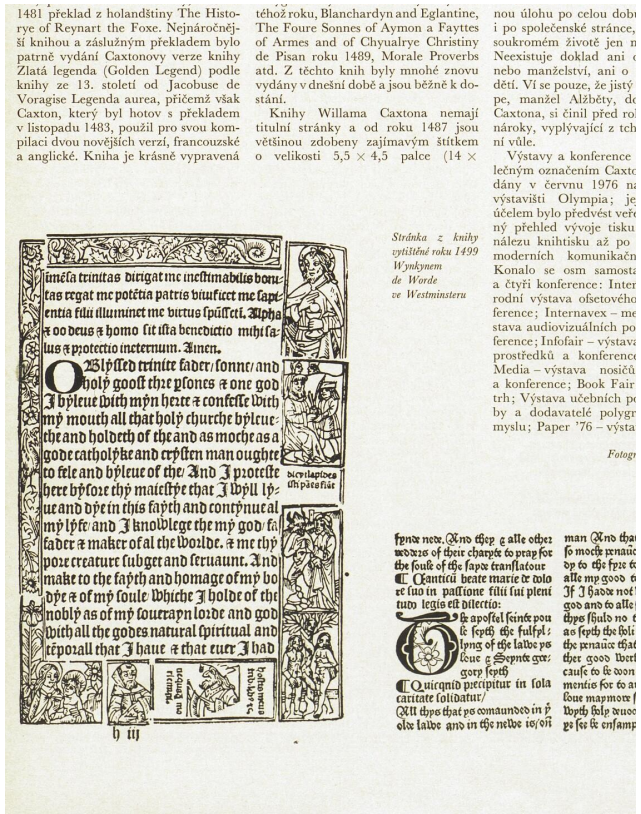
23

Obrázek a text mohou být vizuálně podobné



Kategorie nejsou jasné ani lidem

- Sken textu, typografie, grafy a tabulky



1481 překlad z holandštiny The Historye of Reynart the Foxe. Nejnáročnější knihou a záslužným překladem bylo patrně vydání Caxtonovy verze knihy Zlatá legenda (Golden Legend) podle knihy ze 13. století od Jacobuse de Voragise Legenda aurea, přičemž však Caxton, který byl hotov s překladem v listopadu 1483, použil pro svou kompilaci dvou novějších verzí, francouzské a anglické. Kniha je krásně vypravená

téhož roku, Blanchardyn and Eglantine, The Foure Sonnes of Aymon a Fayttes of Armes and of Chyualrye Christiny de Pisan roku 1489, Morale Proverbs atd. Z těchto knih byly mnohé znovu vydány v dnešní době a jsou běžně k dostání.

Knihy Willama Caxtona nemají titulní stránky a od roku 1487 jsou většinou zdobeny zajímavým štítkem o velikosti 5,5 × 4,5 palce (14 ×

nou úlohu po celou dobu i po společenské stránce, soukromém životě jen n Neexistuje doklad ani nebo manželství, ani o dětí. Ví se pouze, že jistý pe, manžel Alžběty, d Caxtona, si činil před ro nároky, vyplývající z tch ní vůle.

Výstavy a konference lečným označením Caxto dány v červnu 1976 n výstavišti Olympia; je účelem bylo předvést veř ný přehled vývoje tisku nálezu knihtisku až po moderních komunikačn Konalo se osm samost a čtyři konference: Inter rodní výstava ofsetového ference; Internavex – me stava audiovizuálních po ference; Infofair – výstav prostředků a konferenc Media – výstava nosičů a konference; Book Fair trh; Výstava učebních p by a dodavatelé polygr myslu; Paper '76 – výsta

Fotogr

Stránka z knihy vytištěné roku 1499 Wynkynem de Worde ve Westminsteru

TABU

o čem se nemluví

HLAVNÍ TÉMA: RODINNÉ VZTAHY BY MĚLY BÝT TY NEJPEVNĚJŠÍ. JENŽE MÍSTO TOHO JE RODINA ČASTO MÍSTEM NEPOCHOPENÍ.

Ten dům je naše prokletí

ADRIANA (45): „ASI DOJDE AŽ K SOUDU."

CO TOMU ŘÍKÁTE:

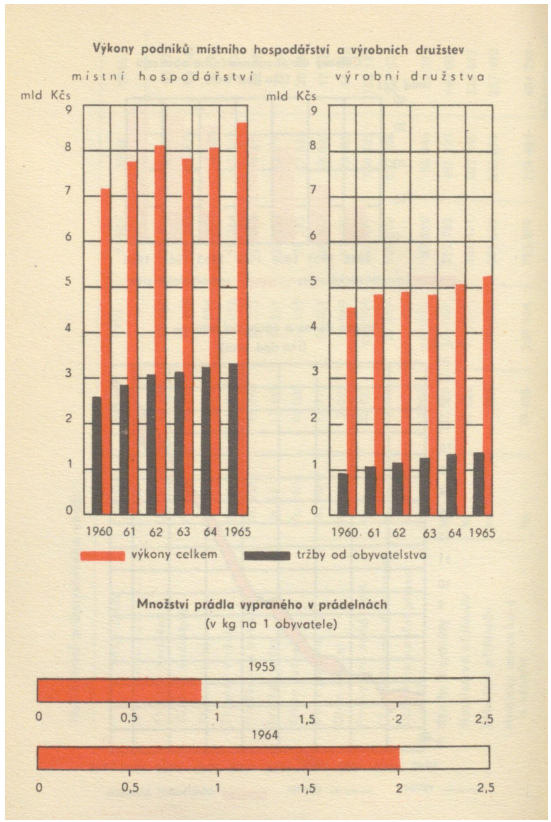
Věra (54)

Adrianu moc lituje. Protože taky bydlí na vesnici, umí si představit, kolik práce a starostí se všim má. Jedině řešení by asi bylo najmout si nějakého právníka, který by uměl poradit, jestli platí závěť, nebo ne. Ale vztahy se sourozenci se už asi nikdy nenapraví.

Naše rodina nikdy nebyla vzorem spokojenosti, kde by si všichni nezištně pomáhali a byli si oporou. Nevím, jak se to stalo, asi to byla nějaká výchovná chyba rodičů nebo co, ale zkrátka já, sestra a bratr nejsme žádní dokonalí sourozenci. Mezi mnou a sestrou je rozdíl pět let, bratr je ještě o dva roky mladší než Pepina. Jako malí jsme si spolu samozřejmě hráli, ale jakmile jsme šli do puberty, přestali jsme si rozumět. Každý si našel kamarády, a i když jsme se nikdy nějak zvlášť nesnášeli, začali jsme si být cizí. Navíc sestra na mě začala žárlit, možná proto, že jsem starší. Já jsem se taky brzo vdala a měla děti a jsem s mužem dodnes šťastná, zatímco ona se sice taky brzo vdala, ale stejně brzo se rozvedla. Od té doby, a to už je přes dvacet let, se plácá mezi mužskýma, ale žádný s ní nevydrží víc než půl roku. Zdá se mi taky zapšklá, takže kdo by s ní byl… Bratr zase nemá žádný cíl, živí se tím, co se zrovna namane, a taky dost pije. Já se svým spořádaným životem jim možná piju krev, možná jsem vzor toho, co oni by sice rádi, ale nedokážou to. Rozhodně, jak jsem se dozvěděla, pro mě nemají moc dobrých slov, dokonce prý sestra říkala naším známým, že mi je manžel nevěrný a já se tvářím, že o tom nevím. Nejhorší situace nastala po smrti našich rodičů. Táta umřel už dávno, ale maminka ani ne před půlrokem. Potíž je v tom, že my jsme bydleli s mužem a dětmi s nimi v jejich původním domě. Ne že bych o to nějak zvlášť stála, ale Pepina se odstěhovala se svým mužem do města a pak už tam jenom střídala bydliště. A bratr zmizel v osmnácti... Nikdy nevím, kdy se náhodou objeví. Takže naše rodina zůstala v domě, což sice zní jako výhra, ale kdo někdy měl něco takového jako vesnický dům, ví, že to je hrozná práce a stojí to spoustu peněz. Těch hodin, co manžel strávil při opravách! V životě jsme nebyli na pořádné dovolené, protože všechno, co jsme ušetřili, šlo na opravy domu, topení a tak. Maminka říkala, že ona nám z důchodu přispívat nemůže, ale že by si přála, aby za naši práci a investice a také za to, že jsme se o ni starali, když už byla nemocná, případ dům nám. Proto taky napsala závěť, ve které to píše. Potom umřela a začalo peklo. Sestra si najednou vzpomněla, kolik asi takový dům stojí a že my nemáme právo si na něj dělat nároky. Podle ní by se měl dům prodat a peníze by se měly rozdělit. Nebo, když my tam chceme bydlet, bychom měli ji i bratra vyplatit. Bratr s ní samozřejmě souhlasí, přestože nevrazil do opravy ani korunu. A že by se třeba starali jeden nebo druhý o mámu, když už nemohla, chudák, ani sama dojít na záchod, o tom ani nemluvím.

Jsem nešťastná. Prodat dům, to si neumím představit, protože kde bychom bydleli? A kde bychom vzali peníze na to, abych sourozence vyplatila, taky nevím, protože všechny peníze, co jsme měli navíc, šly do materiálu a na dělníky. Sestra říká, že jestli okamžitě nenavrhnu nějaké řešení, dá vše k soudu. Ještě jsme se ani nesnažili zjistit, jak to právně všechno je. Vím jenom, že mě mrzí nejenom to, jak se teď sourozenci handrkují o něco, na čem jim nikdy nezáleželo, ale hlavně to, že se o maminku pořádně nezajímali, když byla ještě na světě. Podle mě se musí v hrobě obracet. Manžel mě uklidňuje, že vše nějak dopadne, ale samozřejmě nemá pro švagra se švagrovou dobré slovo. Tuhle, když jsem plakala, se rozčilil a řval, že už nikdy nepřekročí náš práh. Copak takhle se chovají lidé v rodině? Je mi jasné, že takovéhle spory se už nikdy nenapraví, protože jsme už příliš rozhádaní.

12 VLASTA



Dostupná řešení - OCR od ABBYY (v ALTO)

- Dataset: 500 náhodných, manuálne oanotovaných stránek
- Výsledek:
  - Celkové IOU*: 0.69
  - Sensitivity**: 0.69
  - Precission**: 0.36
  - 18% obrázků není detekováno vůbec
  - 57% detekovaných obrázků jsou falešná pozitiva***

OCR od ABBYY Ruční anotace

* Intersection over union
**Vypočteno pro práh citlivosti IOU=0.5
***Neobsahují obrázek, nebo obsahují obrázek který už byl detekován, tzn. obrázky se překrývají.

Detekce pomoci vlastního modelu strojového učení

Řešení: Segmentace pomocí konvoluční neuronové sítě

Proč konvoluční sítě?

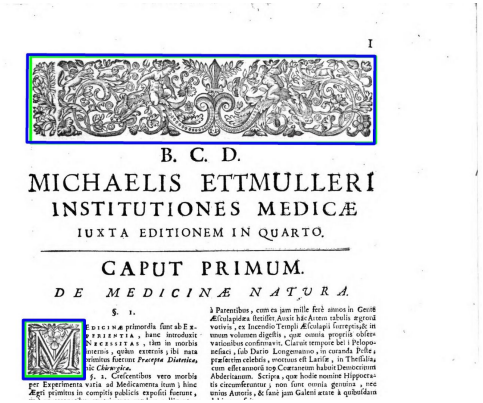
- Translační invariance: poloha obrázku na stránce není důležitá
- Množství volně dostupných implementací state of the art modelů (napr. ResNet, AlexNet, VGG,...)
- Možnost použít váhy předtrénované na velkém datasetu.

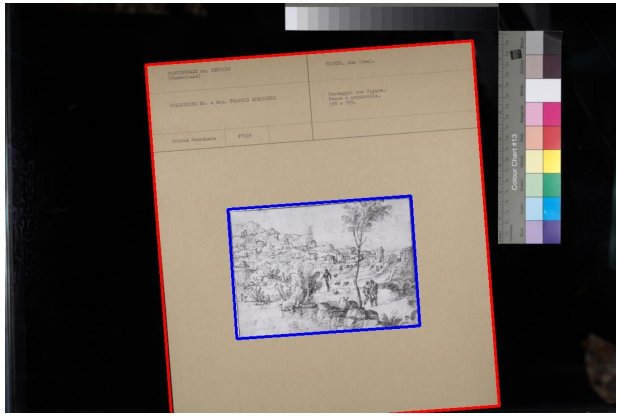
Slepé uličky

- **Newspaper Navigator model**
  - Vyvinut v Library of Congress na detekci obrázků, nadpisů a dalších objektů ve starých novinách.
  - Založen na síti Faster-RCNN od Facebooku.
  - Při evaluaci (bez našeho trénování) měl mnohem horší výkon než OCR od ABBYY.
  - Domníváme se, že model je náchylný k overfittingu a špatně generalizuje na náš dataset.
- **Natrénování UNet sítě (bez předtrénovaných vah)**
  - Architektura využívaná k segmentaci v medicíně a při zpracování satelitních snímků.
  - Nepodařilo se nám dosáhnout dostatečné přesnosti, náš dataset byl zřejmě příliš malý na to, aby se model naučil všechny potřebné vizuální znaky (features).

Neuronová síť dhSegment

- Vyvinuta v 2018 specificky k zpracování skenovaných historických dokumentů
- Použita k detekci ornamentů a fotek na naskenovaných stránkách
- Inspirována segmentační sítí UNet
- Možnost využít předtrénované váhy z ResNet-50
- Implementována v Pythonu pomocí Tensorflow

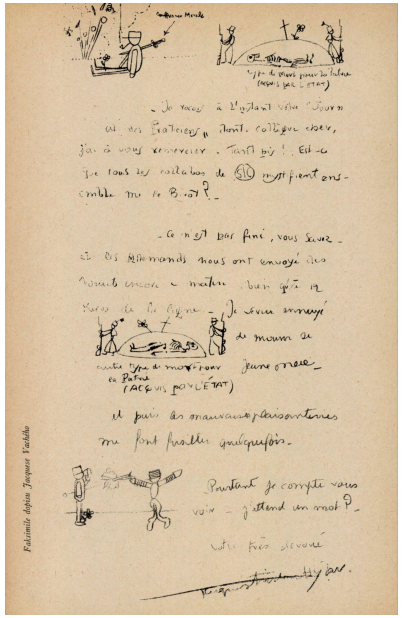




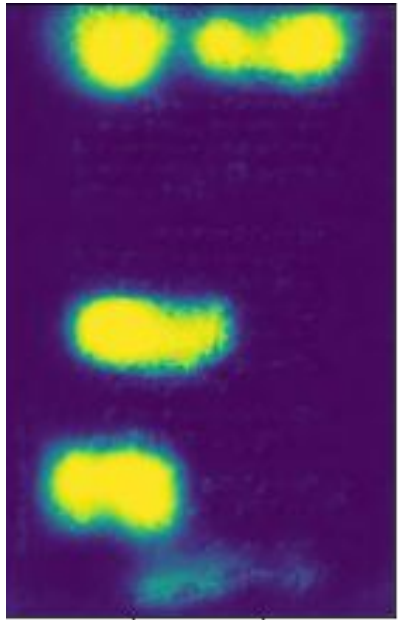
Neuronová síť dhSegment

- Natrénována na ~1000 manuálně anotovaných stránkách na notebooku bez GPU

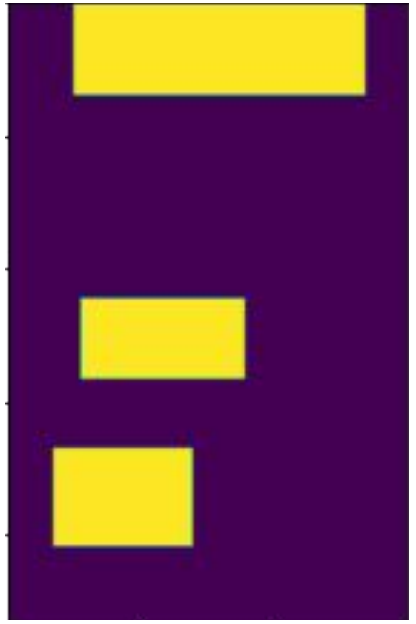
zpracování sítí → binarizace → výstup



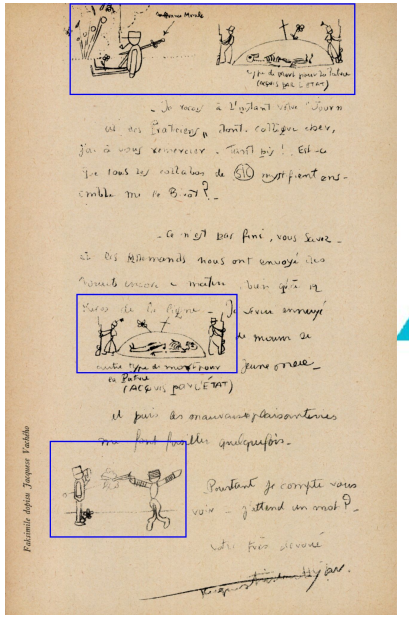
vstup



pravděpodobnost pro každý pixel



binární maska pro každý pixel



poloha obrázků

Neuronová síť dhSegment

- Výsledek:
  - Celkové IOU: 0.65
  - Senzitivity*: 0.65
  - Precission*: 0.4
  - 24% obrázků není detekováno vůbec
  - 53% detekovaných obrázků jsou falešná pozitiva**

- Rychlost (notebook bez GPU):
  - 2.87s na stránku
  - 33 dní na milion stránek

- **Při vynaložení relativně malého množství práce je kvalita prakticky identická s OCR od ABBYY**

*Vypočteno pro práh citlivosti IOU=0.5
**Neobsahují obrázek, nebo obsahují obrázek který už byl detekován

Odkazy

- dhSegment
  - Github https://github.com/dhlab-epfl/dhSegment
  - Publikace: Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, preprint at https://arxiv.org/abs/1804.10371
- NewspaperNavigator projekt: https://github.com/LibraryOfCongress/newspaper-navigator
- OCR od ABBYY: https://www.abbyy.com/finereader-server/
- Resnet: https://arxiv.org/pdf/1512.03385.pdf
- UNet: https://arxiv.org/pdf/1505.04597.pdf

Vyhledávání obrázků podle podobnosti - VISE

- Vyhledávání příbuzných obrázků podle prostorové podobnosti (výřez)
- Periodika, staré ilustrace, grafiky, loga...
- Přesný i pro malé rozlišení (testováno na 1024x1024)
- Možnost kombinace daty identifikujícími obrázky na stránce
- 13 000 obrázků zaindexováno za 7 hodin

Datová sada obrázků (jpg, png)

Indexace / trénování vizuálních deskriptorů..

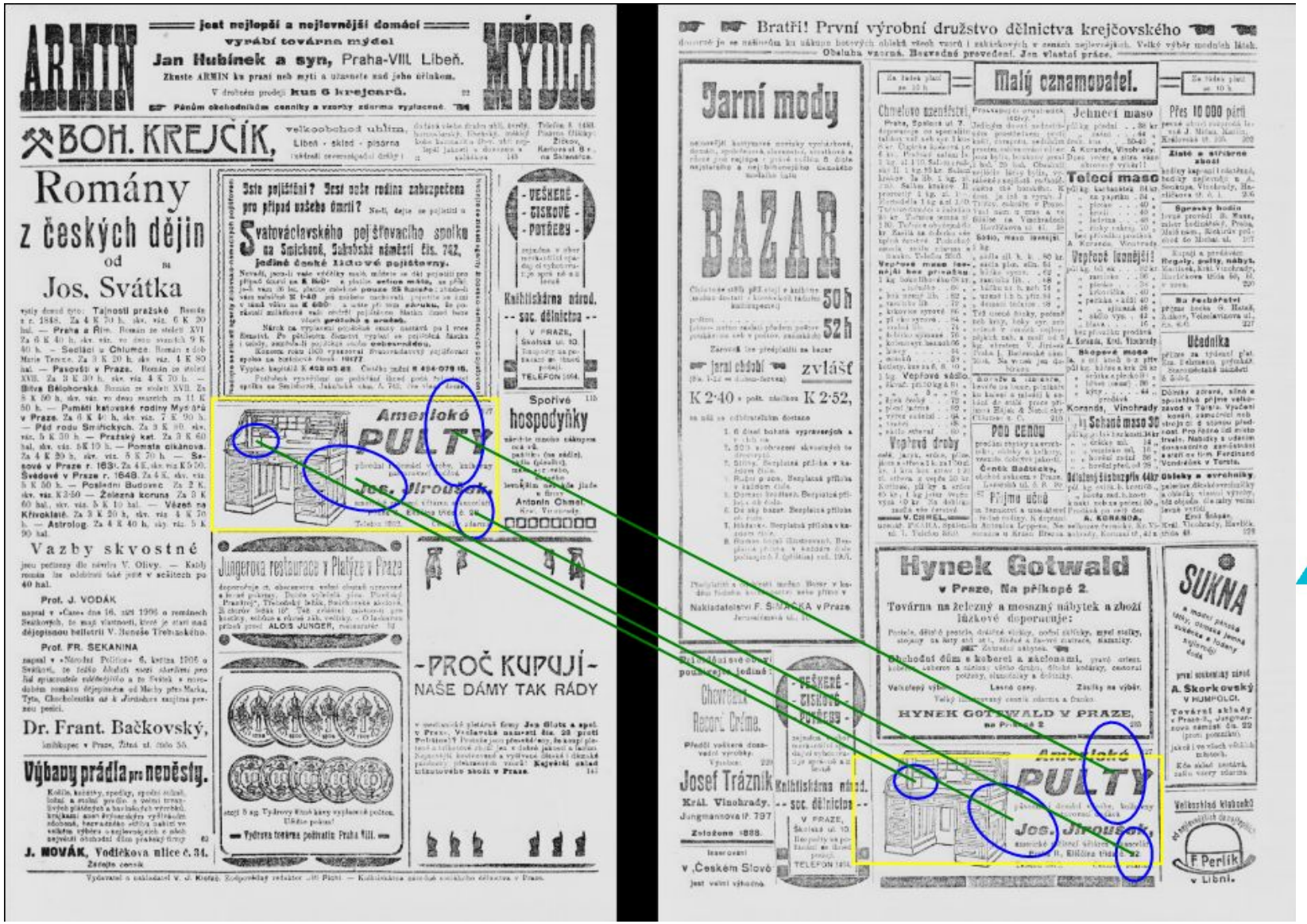
Zaindexovaná datová sada

Search ready

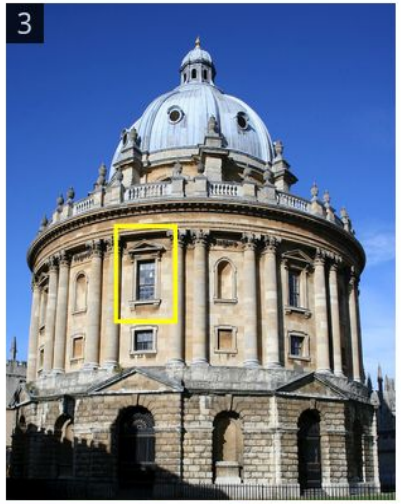
Demonstrace na inzerátech deníku Svobodné slovo

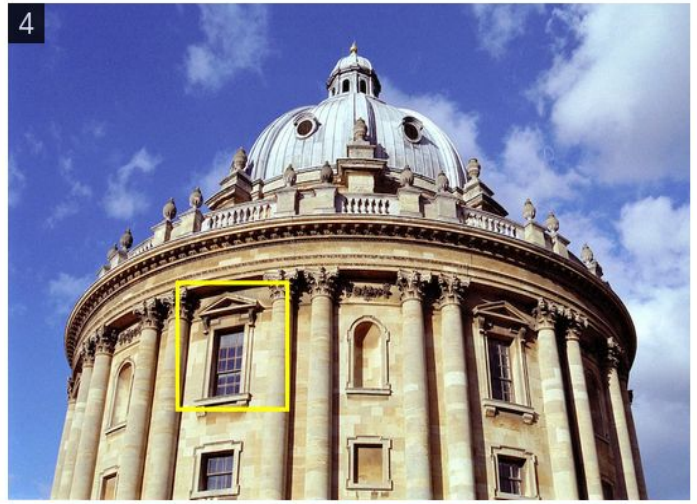
Hledat

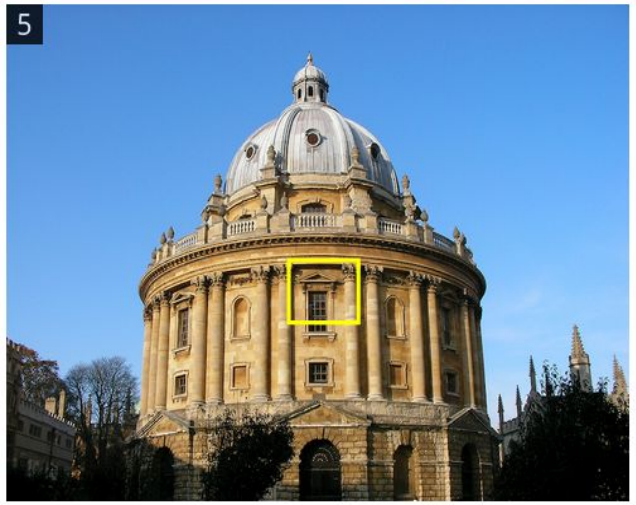
Porovnání obrázků inzerce deníku Svobodné slovo

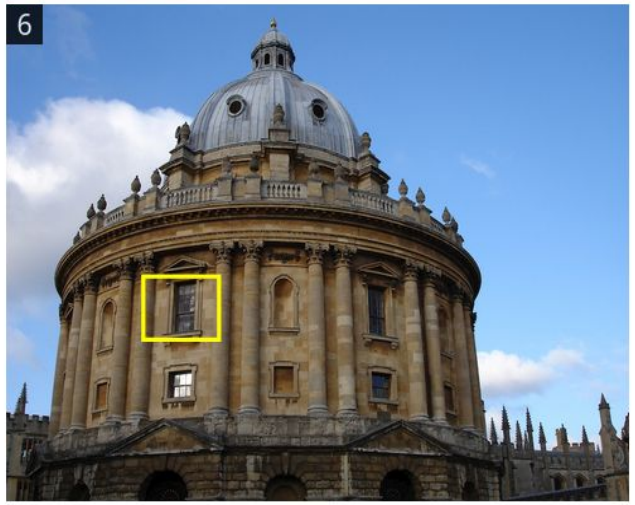


Funkční i pro 3D prostorové obrázky









Výhody VISE

- Velmi dobrá přesnost i pro obrázky s malým rozlišením
- Vyhledávání je rychlé
- Není příliš mnoho konkurenčních systémů
- Není potřeba grafické karty
- Systém je dále rozvíjen

Nevýhody

- Nefunguje pro běžné bloky textů, vhodné spíše pro obrázky, ilustrace nebo větší texty jako tituly, nadpisy atd.
- Může nastat nepřesnost ve vyhledávání, např. pokud je výřez málo detailní nebo je špatná trénovací sada
- Nelze použít obrázky s vysokým rozlišením
- Nelze přidávat nové obrázky již k natrénované datové sadě
- Nelze použít JPEG2000

Odkazy

- Abhishek Dutta, Relja Arandjelović, and Andrew Zisserman. 2021. VGG Image Search Engine. from https://www.robots.ox.ac.uk/~vgg/software/vise/
- Gitlab: https://gitlab.com/vgg/vise
- Oxford Visual geometry group: https://www.robots.ox.ac.uk/~vgg/

Co dál?

- Existuje řada zajímavých projektů, aplikací, modelů
- Není snadné najít hotové řešení
- Specifika reálných datových sad
  - rozsah, variabilita
- Velký prostor pro další rozvoj

Děkuji za pozornost!

Dotazy?

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně