Automatizace přepisu staročeských textů

pro potřebu lingvistických výzkumů

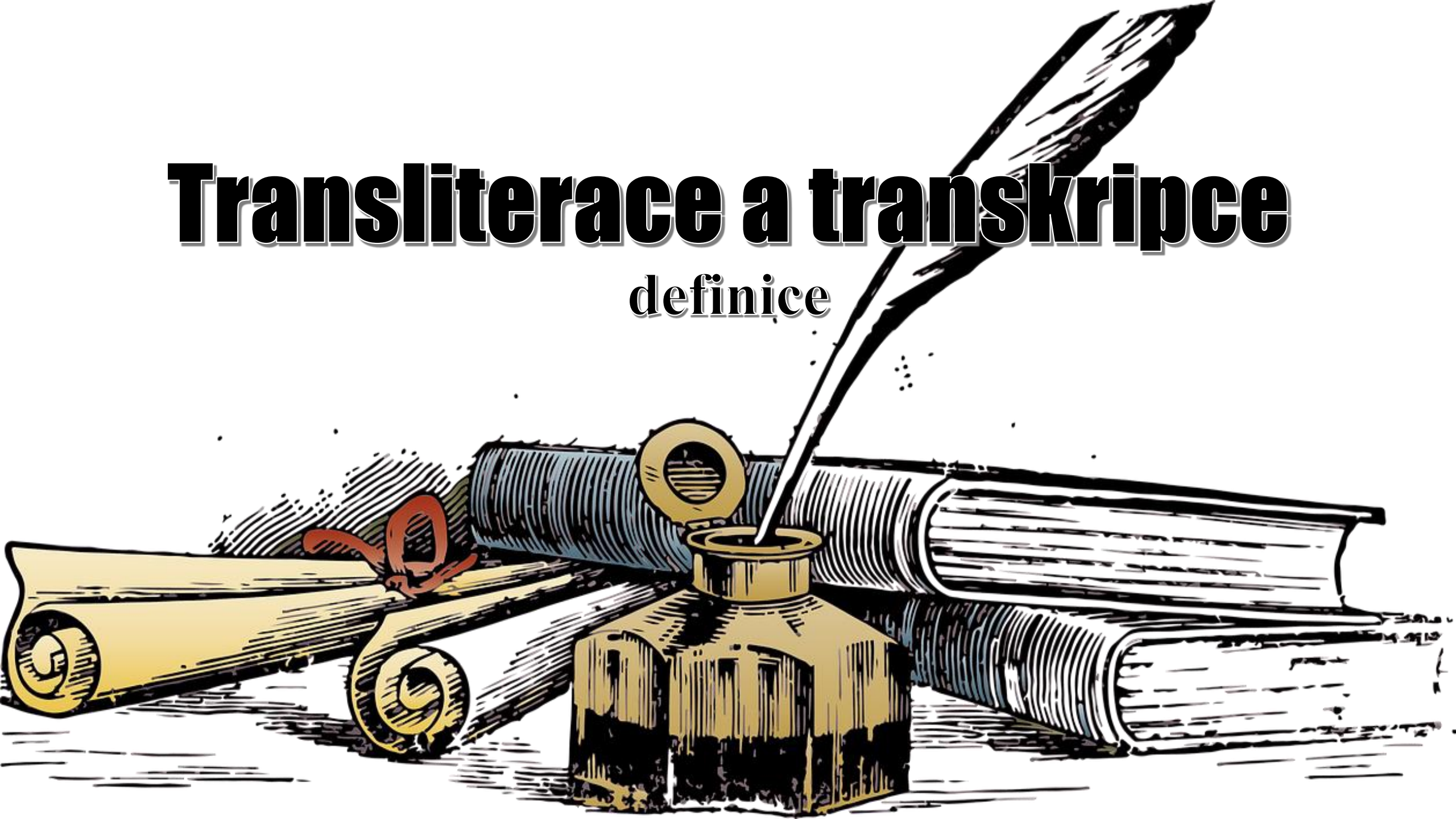
Eliška Pěnkavová

Otázky

1. Jak zrychlit dlouhé přepisování starých písem?
2. Jaká je chybovost při automatickém přepisování textů?
3. Jak tyto chyby zredukovat? Dají se úplně odstranit?
4. Jak těžit data z velkého příjmu textů?
5. Lze zautomatizovat transkripci?
6. Lze vytvořit univerzální digitální nástroj pro přepis staré češtiny?

Transliterace a transkripce

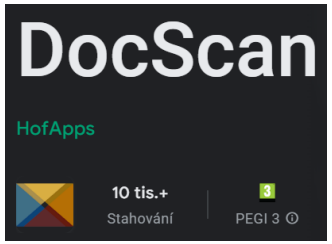
definice



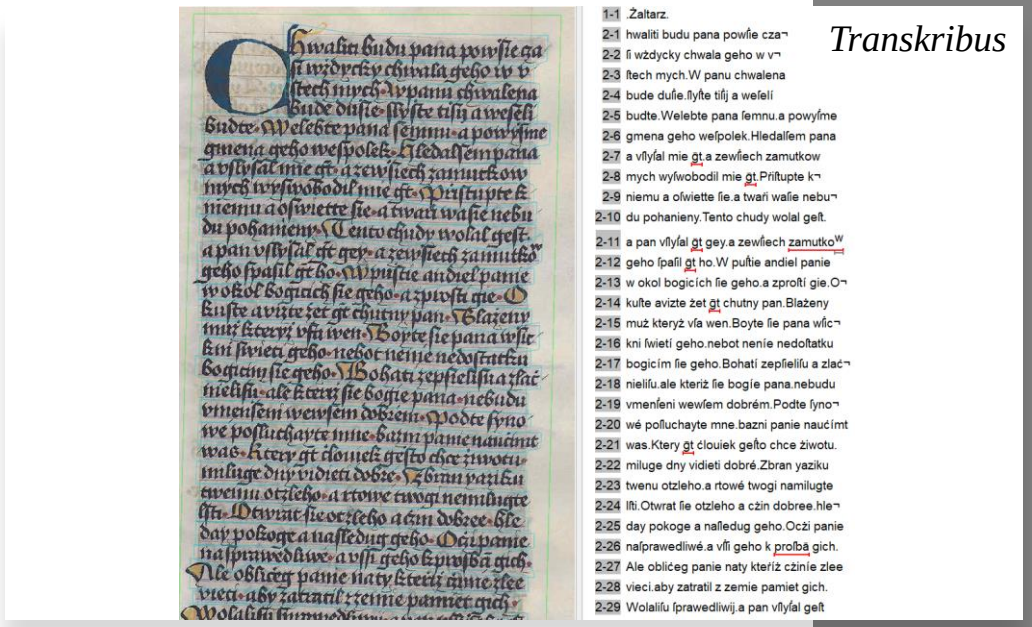
Automatizace přepisu

Aplikace:

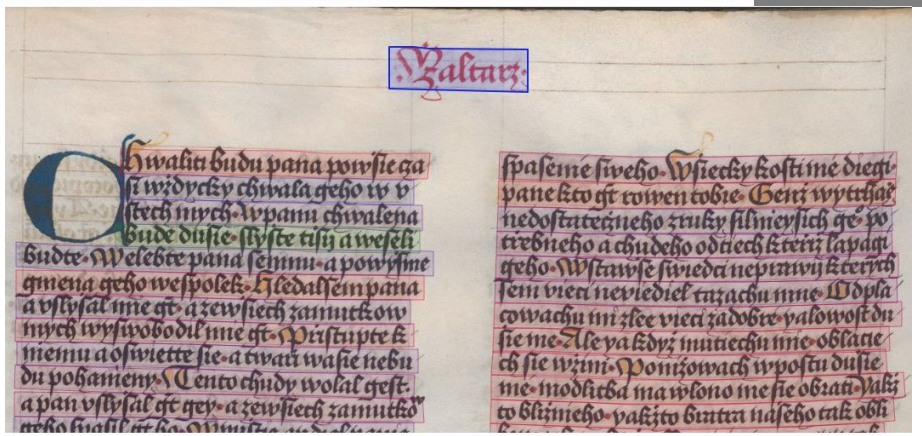
- Transkribus (HTR)
- Projekt Pero (OCR)
- Diptychon
- eScriptorium (OCR/HTR)
- OCR4ALL (OCR)
- Rescribe.xyz (OCR)



mobilní aplikace sloužící k digitalizace dokumentů



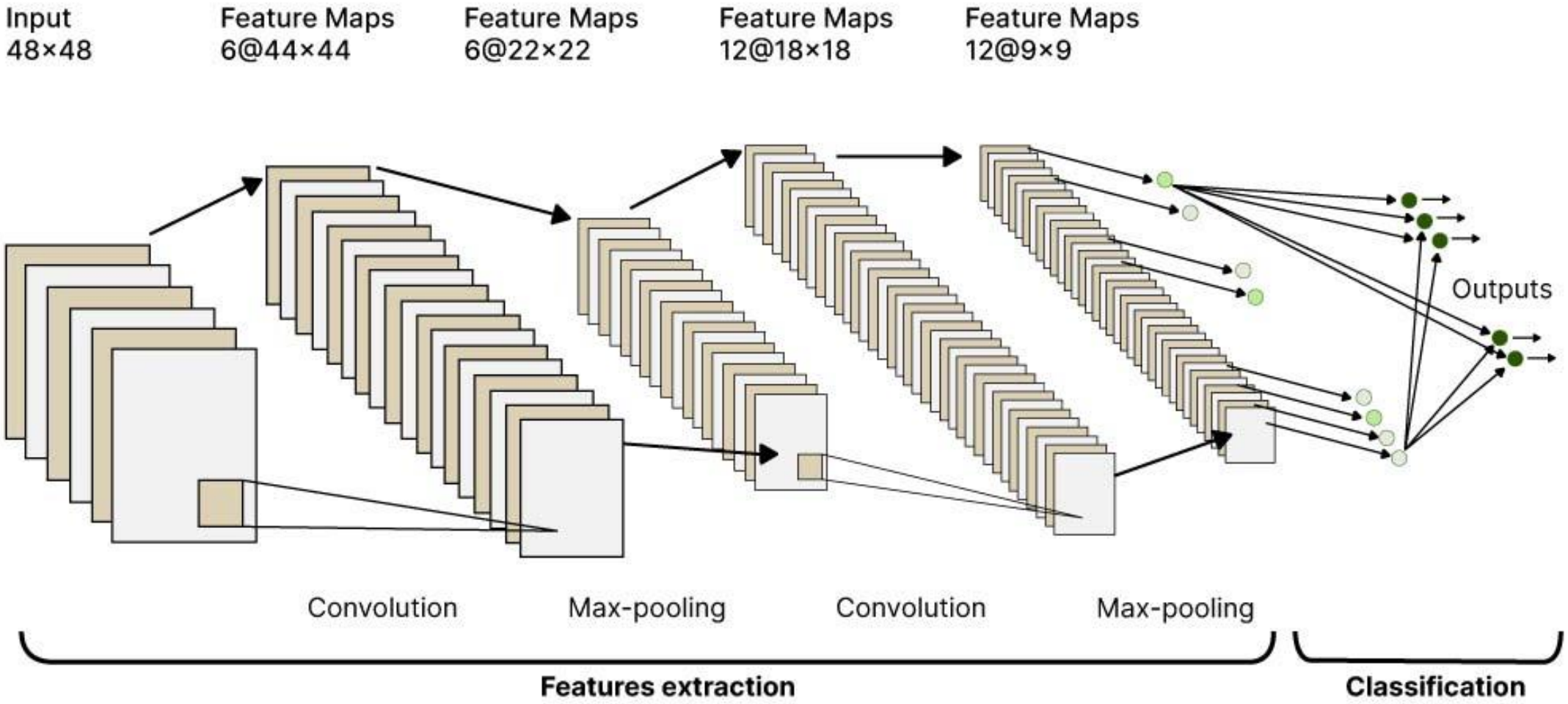
Transkribus



Projekt Pero

HTR x OCR

rozdíl, využití, výhody a nevýhody



struktura konvoluční sítě (HTR)

Transkripční pravidla

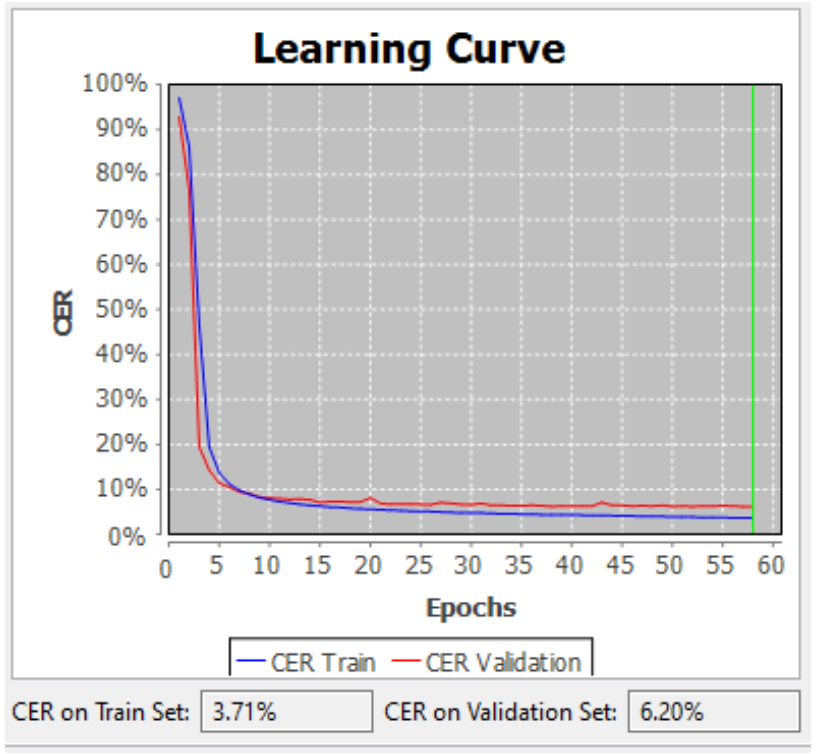
Publikace zabývající se pravopisem a transkripčními pravidly:

- Jiří Daňhelka, Směrnice pro vydávání starších českých textů (Husitský Tábor 8, 1985, s. 285–301)
- Jiří Daňhelka, Obecné zásady ediční a poučení o starém jazyce českém (in: Výbor z české literatury od počátků po dobu Husovu. Praha, Nakladatelství Československé akademie věd 1957, s. 25–35)
- Jiří Daňhelka, Obecné zásady ediční a poučení o češtině 15. století (in: Výbor z české literatury doby husitské. Svazek první. Praha, Nakladatelství Československé akademie věd 1963, s. 31–41)
- Josef Vintr, Zásady transkripce českých textů z barokní doby (Listy filologické 121, 1998, s. 341–346)

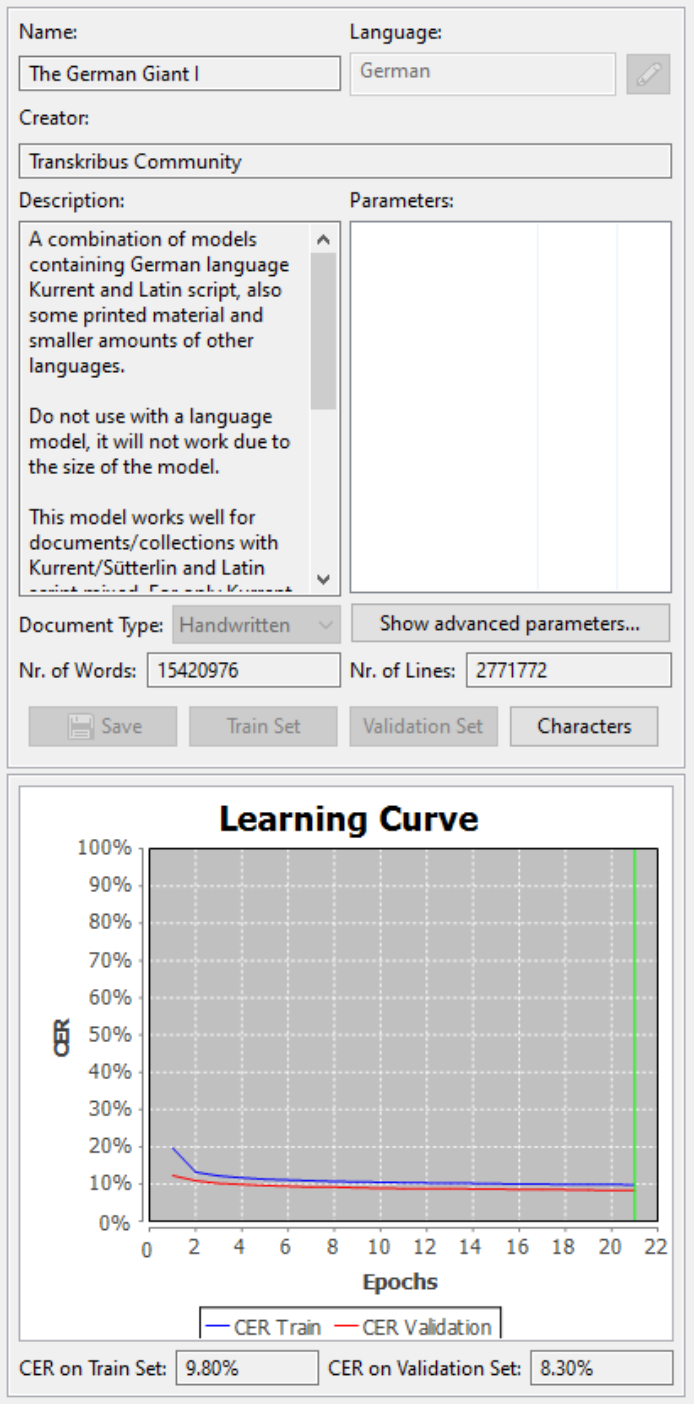
Lze vytvořit univerzální digitální nástroj pro přepis staré češtiny?

- Transliterace x Transkripce
- Character Error Rate (CER) – chybovost

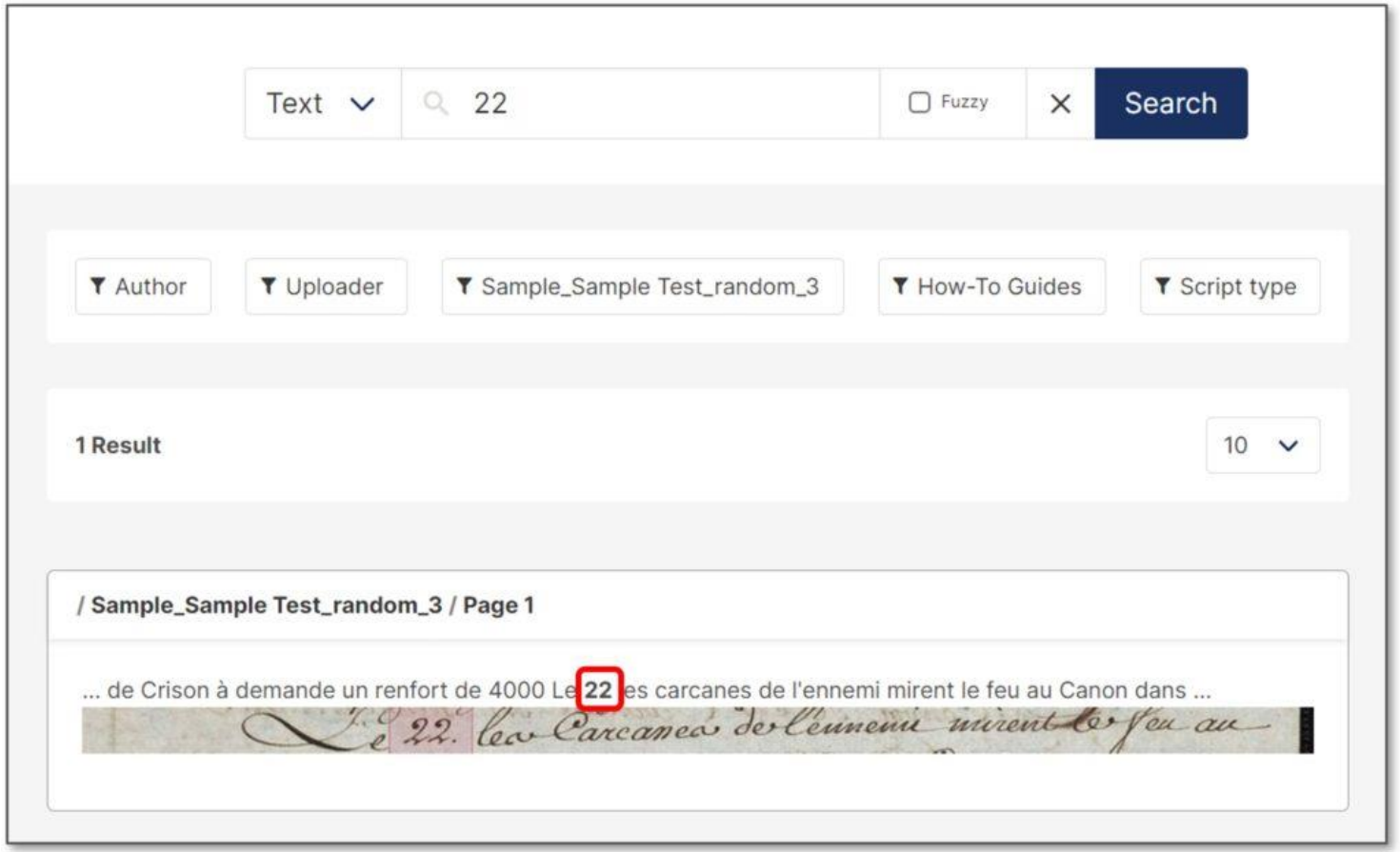
(lidská x automatická)



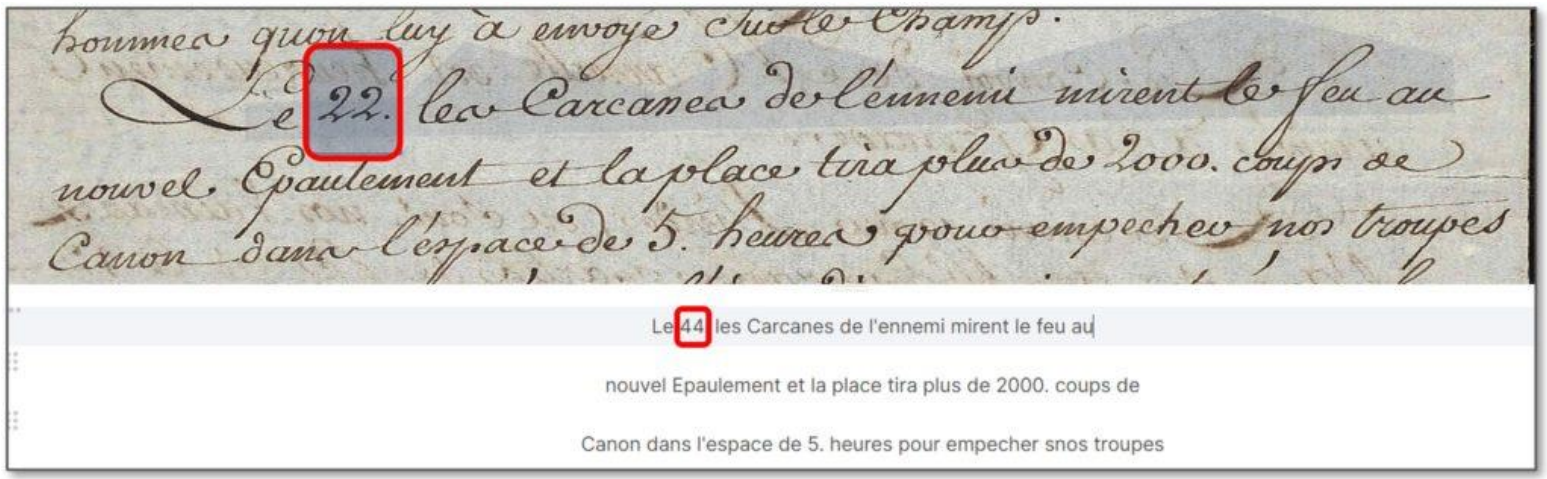
Old Czech Handwriting (without spaces)



- Smart Search



vylepšené fulltextové vyhledávání i s možností hledání při chybovosti modelu

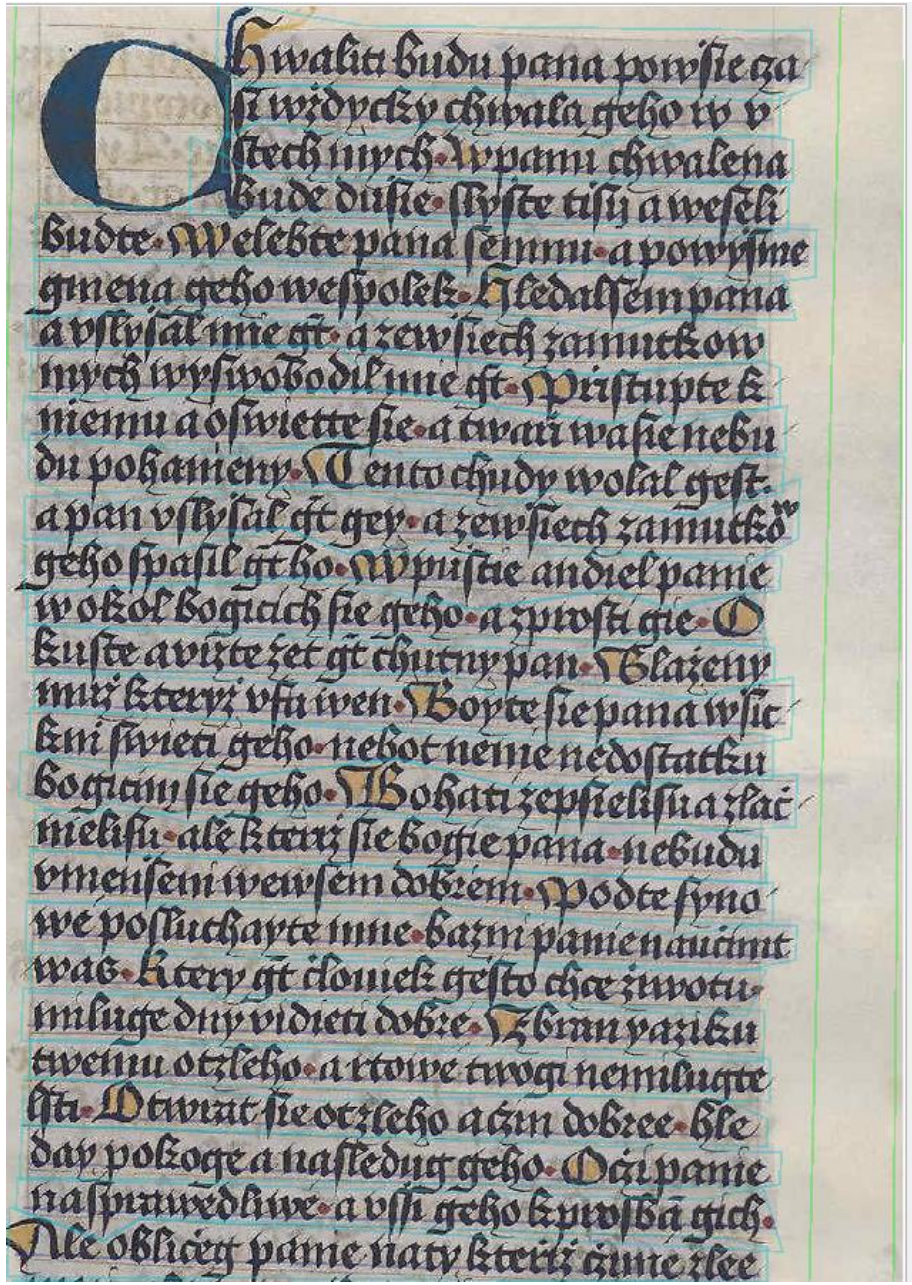


Jak zrychlit dlouhé přepisování starých písem?

- transliterační aplikace (Transkribus aj.)
- transkripční modely a šablony:

  - Word šablona
  - Transkribus – transkribovaný model
  - Transkribus – transliterovaný model se zkratkami
  - aplikace vytvořená na míru

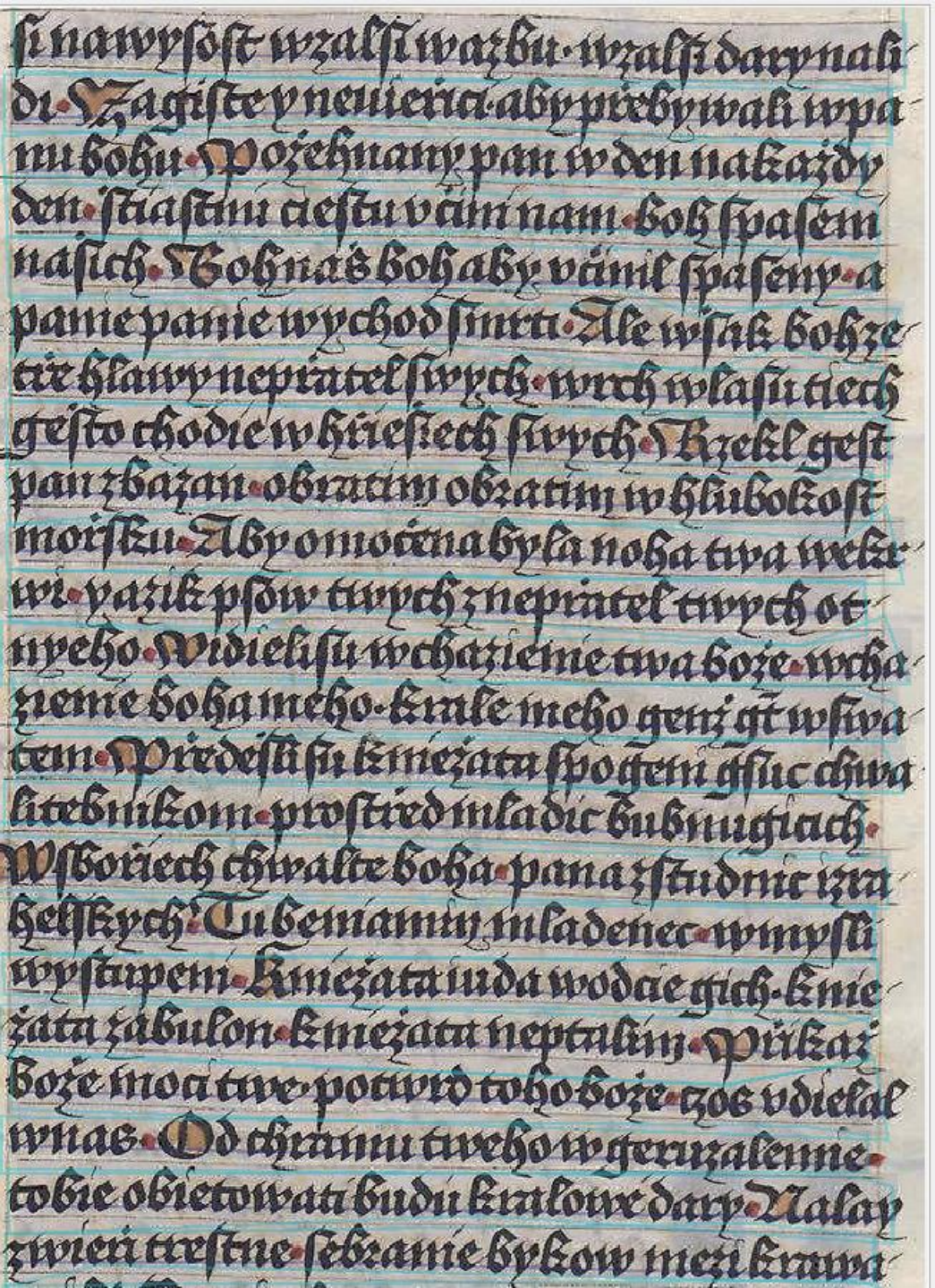
2-1 hwaliti budu pana powſie cza¬
2-2 ſi wżdycky chwala geho w v¬
2-3 ſtech mych.W panu chwalena
2-4 bude duſie.ſlyſte tiſij a weſelí
2-5 budte.Welebte pana ſemnu.a powyſme
2-6 gmena geho weſpolek.Hledalſem pana
2-7 a vſlyſal mie gt.a zewſiech zamutkow
2-8 mych wyſwobodil mie gt.Přiſtupte k¬
2-9 niemu a oſwiette ſie.a twaři waſie nebu¬
2-10 du pohanieny.Tento chudy wolal geſt.
2-11 a pan vſlyſal gt gey.a zewſiech zamutkow
2-12 geho ſpaſil gt ho.W puſtie andiel panie
2-13 w okol bogicích ſie geho.a zproſtí gie.O¬
2-14 kuſte avizte żet gt chutny pan.Blażeny
2-15 muż kteryż vſa wen.Boyte ſie pana wſic¬
2-16 kni ſwietí geho.nebot neníe nedoſtatku
2-17 bogicím ſie geho.Bohatí zepſieliſu a zlać¬
2-18 nieliſu.ale kteriż ſie bogíe pana.nebudu
2-19 vmenſeni wewſem dobrém.Podte ſyno¬
2-20 wé poſluchayte mne.bazni panie naućímt
2-21 was.Ktery gt ćlouiek geſto chce żiwotu.
2-22 miluge dny vidieti dobré.Zbran yaziku
2-23 twenu otzleho.a rtowé twogi namilugte
2-24 lſti.Otwrat ſie otzleho a cżin dobree.hle¬



2-1 hvaliti budu pana povše ča¬
2-2 si vždycky chvala jeho v v¬
2-3 stech mych.V panu chvalena
2-4 bude duše.slyšte tiší a veselí
2-5 budte.Velebte pana semnu.a povyšme
2-6 jmena jeho vespolek.Hledalsem pana
2-7 a vslyšal me jest.a zevšech zamutkov
2-8 mych vysvobodil me jest.Přistupte k¬
2-9 nemu a osvette se.a tvaři vase nebu¬
2-10 du pohaneny.Tento chudy volal jest.
2-11 a pan vslyšal jest jej.a zevšech zamutkov
2-12 jeho spasil jest ho.V pušte andel pane
2-13 v okol bojicích se jeho.a zprostí je.O¬
2-14 kuste avizte žet jest chutny pan.Blaženy
2-15 muž kteryž vsa ven.Bojte se pana všic¬
2-16 kni svetí jeho.nebot není nedostatku
2-17 bojicím se jeho.Bohatí zepselisu a zlač¬
2-18 nelisu.ale kteriž se bojí pana.nebudu
2-19 vmenšeni vevsem dobrém.Podte syno¬
2-20 vé posluchajte mne.bazni pane naučímt
2-21 vas.Ktery jest človek ješto chce životu.
2-22 miluje dny videti dobré.Zbran jaziku
2-23 tvenu otzleho.a rtové tvoji namiluite

ruční přepis transliterace

ruční přepis transkripce – 1. neoficiální a nefinální verze modelu



si navysost vzalsi vazbu.vzalsi darynali
di.Zajiste y neveřiciabypřebyvali vpa¬
nu bohu.Požehnany pan v den nakaždy
den.stastnu cestu včíní nam.boh spasení
našich.Bohnaš boh aby včinil spaseny.a
paní pane vychod smrti.Ale však boh ze¬
tře hlavynepřatel svych.vrch vlasu tech
ješto chode v hřešech svych.Rzekl jest
pan zbazan.obratim obratim v hlubokost
mořsku.Aby omočena byla noha tva vekr¬
vi.jazik psov tvych znepřatel tvych ot¬
nyeho.Vidílisu v chazení tva bože.vcha¬
zení boha ního.krale meho jenž jest v sva¬
tem.Předešlisu knížata spojení jsuc chva
litebníkom.prostředmladic bubnujicich
Vsbořích chvalte boha.pana zstudnic izra¬
helskych.Ču benamín mladenec.vmysli
vystupení.Knížata vida vodce jich.kní
žata zabulon.knížata neptalim.Přikaz
boze mačti tve patvčd toho boj čjs vdela.
vnas.Od chramu tvehov jeruzalemí.
tobí obetovati budú kralove dary.Aalaj
zveři trestne.sebraní bykov mezi krava¬

transkripce jiného listu

Zdroje

- Marie Krčmová (2017), Transkripce. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. URL: https://www.czechency.org/slovnik/TRANSKRIPCE (poslední přístup: 4. 10. 2023)
- Marie Krčmová (2017), Transliterace. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. URL: https://www.czechency.org/slovnik/TRANSLITERACE (poslední přístup: 4. 10. 2023)
- Ivan Šťovíček, Zásady vydávání novověkých historických pramenů z období od počátku 16. století do současnosti: příprava vědeckých edic dokumentů ze 16.-20. století pro potřeby historiografie. Praha: Archivní správa Ministerstva vnitra ČR, 2002. ISBN 80-86466-00-0.
- Internetová jazyková příručka [online] (2008–2023). Praha: Ústav pro jazyk český AV ČR, v. v. i. Cit. 4. 10. 2023. <https://prirucka.ujc.cas.cz/>.
- Vokabulář webový: webové hnízdo pramenů k poznání historické češtiny [online]. Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. © 2006–2023. Verze dat 1.1.23 [cit. 4. 10. 2023]. Dostupné z: https://vokabular.ujc.cas.cz
- https://readcoop.eu/insights/ocr-vs-htr/
- Ondřej Tomiška, Automatizovaný přepis rukopisných historických dokumentů a jejich využití pomocí moderních IT [online]. Hradec Králové, 2021 [cit. 2023-10-08]. Dostupné z: https://theses.cz/id/9lbwy8/. Diplomová práce. Univerzita Hradec Králové, Filozofická fakulta. Vedoucí práce doc. RNDr. Štěpán Hubálovský, Ph.D.
- Jiří Daňhelka, Směrnice pro vydávání starších českých textů (Husitský Tábor 8, 1985, s. 285–301)
- Jiří Daňhelka, Obecné zásady ediční a poučení o starém jazyce českém (in: Výbor z české literatury od počátků po dobu Husovu. Praha, Nakladatelství Československé akademie věd 1957, s. 25–35)
- Jiří Daňhelka, Obecné zásady ediční a poučení o češtině 15. století (in: Výbor z české literatury doby husitské. Svazek první. Praha, Nakladatelství Československé akademie věd 1963, s. 31–41)
- Josef Vintr, Zásady transkripce českých textů z barokní doby (Listy filologické 121, 1998, s. 341–346)
- Michal Wanner, Slovník současné archivní terminologie. Praha: Odbor archivní správy a spisové služby MV, 2022. ISBN 978-80-7616-118-4.
- Pavlína Kuldanová, Český jazyk v minulosti. Ostrava: Ostravská univerzita, 2013. ISBN 978-80-7464-479-5.
- Transkribus: AI powered Handwritten Text Recognition [online]. READ-COOP SCE, 2019 [cit. 8. 10. 2023]. Dostupné z: https://readcoop.eu/transkribus
- Projekt PERO [online]. Vysoké učení technické v Brně, Moravská zemská knihovna [cit. 8. 10. 2023]. Dostupné z: https://pero-ocr.fit.vutbr.cz/
- Obrázky: pixabay.com