

# Automatizace přepisu staročeských textů pro potřebu lingvistických výzkumů



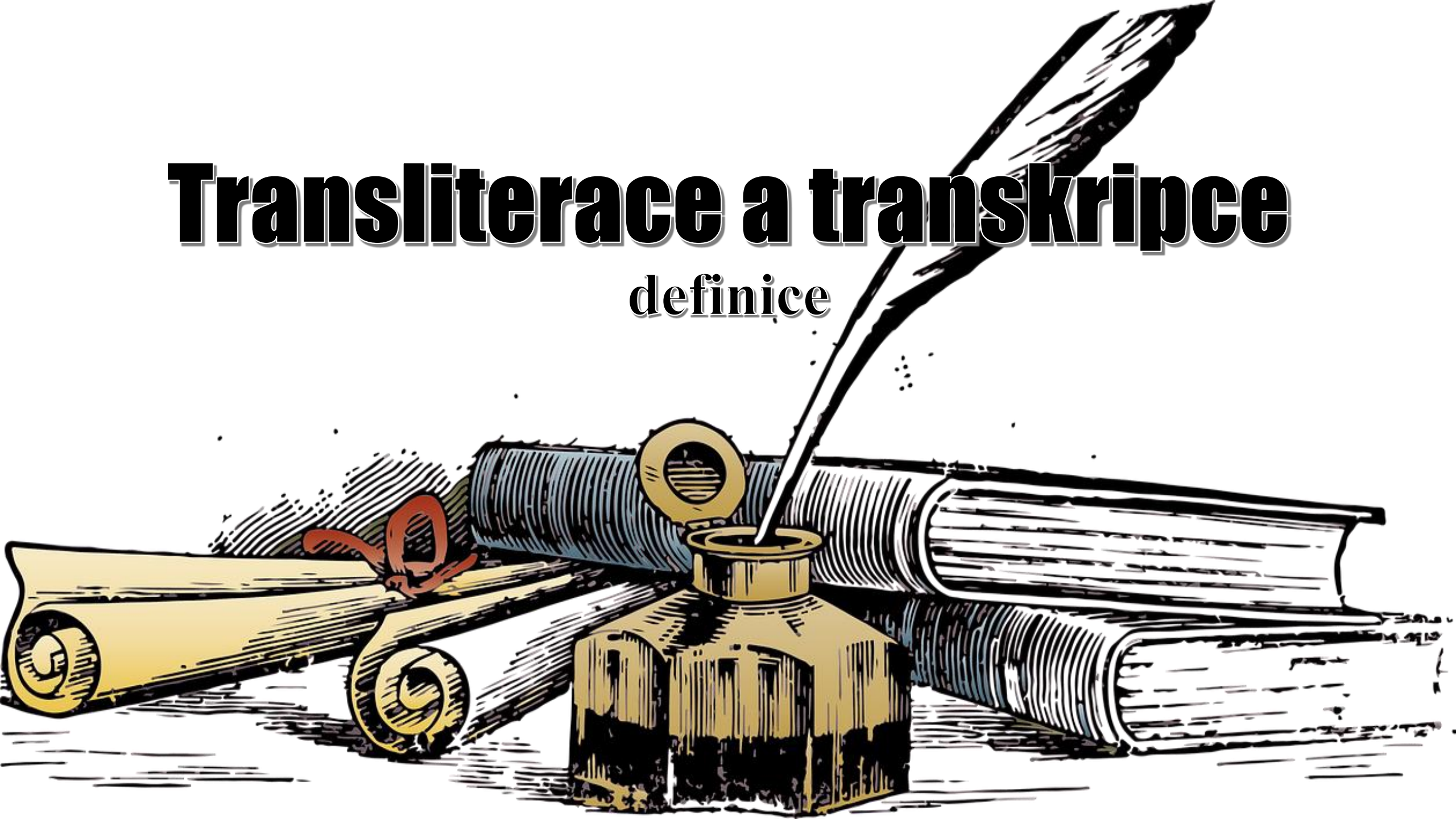
# Otázky

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

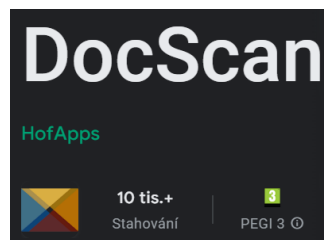
## definice



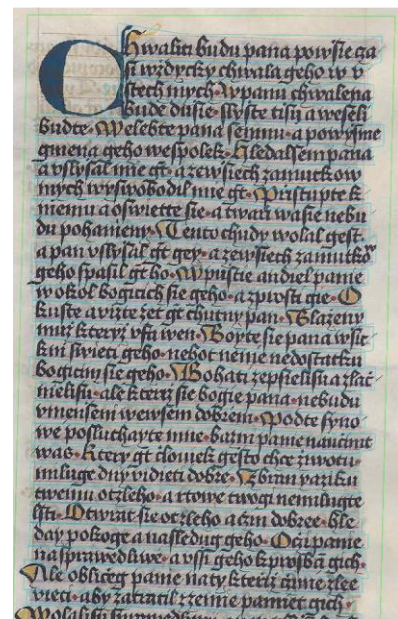
# Automatizace přepisu

## Aplikace:

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

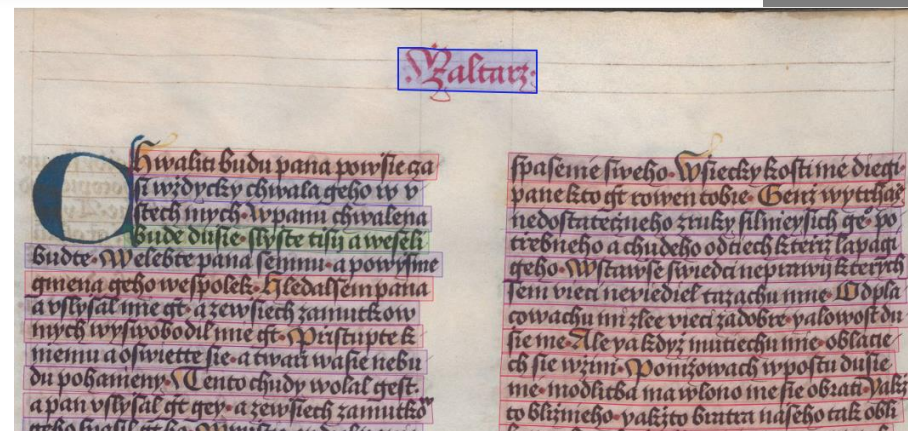


*mobilní aplikace sloužící  
k digitalizaci dokumentů*



1-1 Zaltar.  
2-1 hwaliti budu pana powlie cza-  
2-2 fi wždycky chwala gehu w v-  
2-3 tech mych. W panu chwalena  
2-4 bude dusie. Ilyste tifi a wesele  
2-5 bude. Welebbe pana semnu. a powymie  
2-6 gmena gehu wespolek. Hledalsem pana  
2-7 a wlyfal mie gt. a zewsiech zamutkow  
2-8 mych wyswobodil mie gt. Wristu pte k  
2-9 niemu a ofswiette sie. a twari wafie nebu-  
2-10 du pohanieny. Tento chudy wolal gest.  
2-11 a pan wlyfal gt. ge. a zewsiech zamutko-  
2-12 gehu spafil gt. ho. W wysie andiel panie  
2-13 w okol bogicich fie gehu. a zprofti gie. O-  
2-14 kuftie awizte zet gt. chutny pan. Blazeny  
2-15 muz kteryz wla wen. Boyte fie pana wlic-  
2-16 kni wtielti gehu. nebot nenie nedotatku  
2-17 bogicim fie gehu. Bohati zepfeliu a zlac-  
2-18 niellu. ale kteriz fie bogie pana. nebudu  
2-19 wmenfeni wewsem dobrem. Podte lyno-  
2-20 we posluchajte mne. bati panie naučit  
2-21 was. Ktery gt. clouiek gehu chce ziwotu.  
2-22 miluge dny widieti dobre. Zbran yaziku  
2-23 twenu otzleho. a rtowé twogi namilugte  
2-24 iti. Otwart fie otzleho a czin dobre. hie-  
2-25 day pokoge a nalleduq gehu. Ocz panie  
2-26 nalprawedliwé. a wlii gehu k proba gich.  
2-27 Ale oblicég panie naty kteriz czinie zlee  
2-28 weci. aby zatratil z zemie pamiet gich.  
2-29 Wolalifu sprawedliw. a pan wlyfal gest.

*Transkribus*



Zaltar

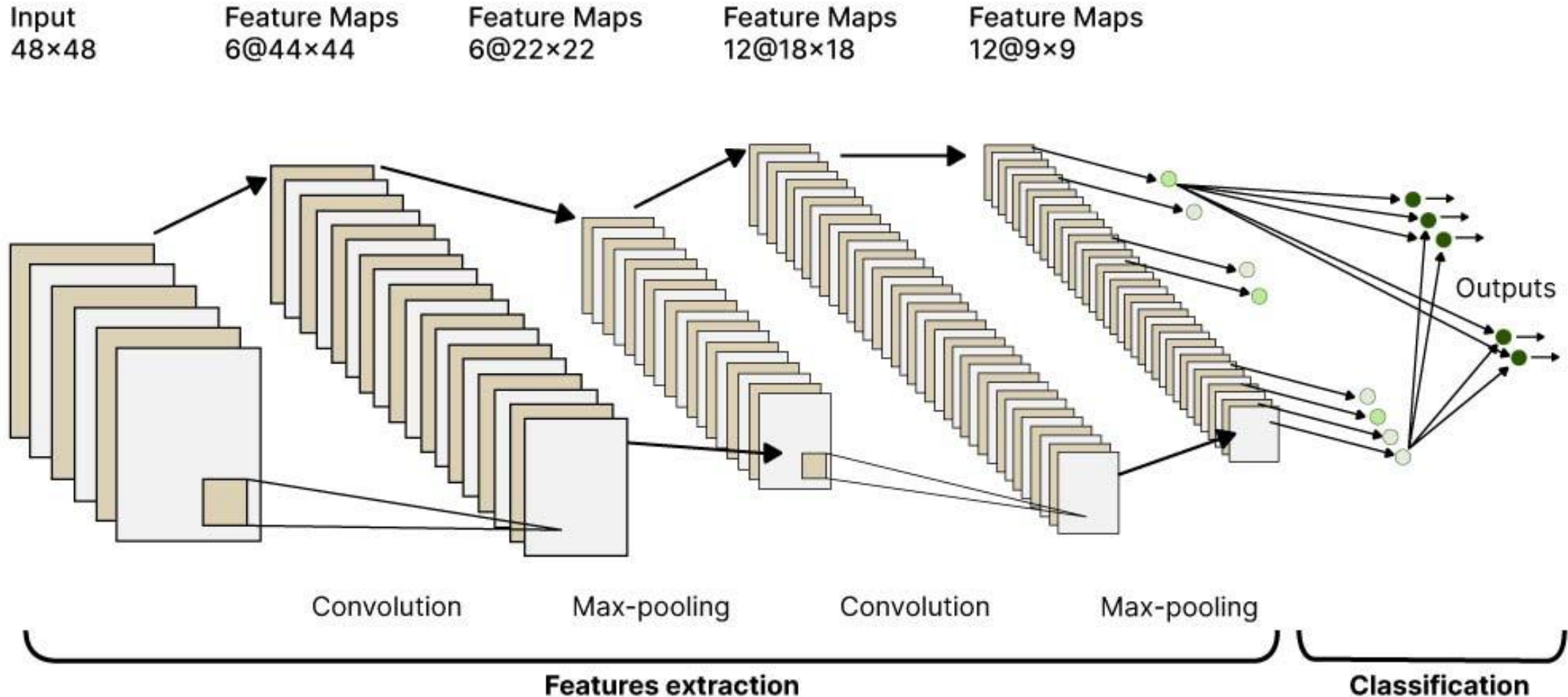
hwaliti budu pana powlie cza-  
fi wždycky chwala gehu w v  
tech mych. W panu chwalena  
bude dusie. Ilyste tifi a wesele

*Projekt Pero*



# HTR x OCR

## rozdíl, využití, výhody a nevýhody



*struktura konvoluční sítě (HTR)*

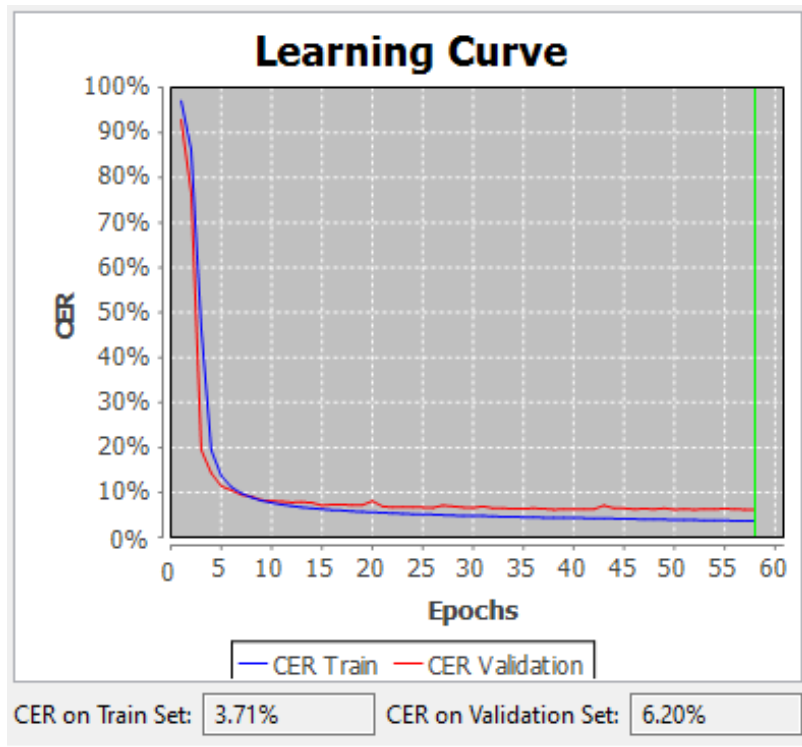
# Transkripční pravidla

*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 Vintř, 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)*

Name: The German Giant I Language: German

Creator: Transkribus Community

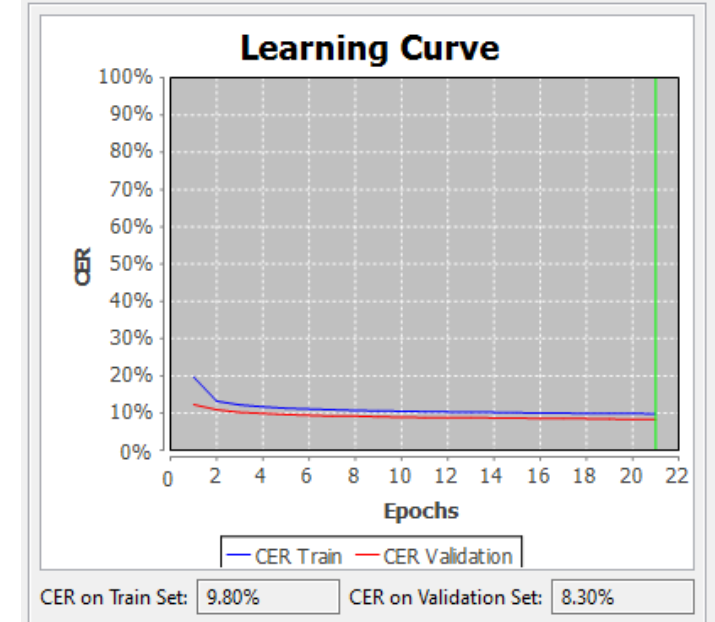
Description: A combination of models containing German language Kurrent and Latin script, also some printed material and smaller amounts of other languages. Do not use with a language model, it will not work due to the size of the model. This model works well for documents/collections with Kurrent/Sütterlin and Latin

Parameters:

Document Type: Handwritten Show advanced parameters...

Nr. of Words: 15420976 Nr. of Lines: 2771772

Save Train Set Validation Set Characters



- Smart Search

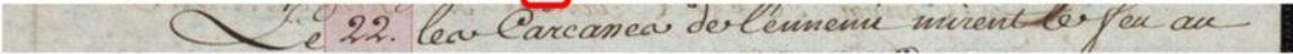
Text   Fuzzy

▼ Author   ▼ Uploader   ▼ Sample\_Sample Test\_random\_3   ▼ How-To Guides   ▼ Script type

1 Result 10 ▼

/ Sample\_Sample Test\_random\_3 / Page 1

... de Crison à demande un renfort de 4000 Le **22** les carcanes de l'ennemi mirent le feu au Canon dans ...



vylepšené fulltextové  
vyhledávání i s možností  
hledání při chybovosti modelu



Le **22**. les Carcanes de l'ennemi mirent le feu au  
nouvel Epaulement et la place tira plus de 2000. coups de  
Canon dans l'espace de 5. heures pour empêcher nos troupes

Le **44** les Carcanes de l'ennemi mirent le feu au  
nouvel Epaulement et la place tira plus de 2000. coups de  
Canon dans l'espace de 5. heures pour empêcher nos troupes



# 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 powšie cza<sup>7</sup>  
2-2 si wždycky chwala geho w v<sup>7</sup>  
2-3 ſtech mych. W panu chwalena  
2-4 bude dušie. ſlyšte tiſij a weſelí  
2-5 budte. Welebte pana ſemnu. a powyſme  
2-6 gmena geho weſpolek. Hledalſem pana  
2-7 a vſlyſal mie g̃t. a zewſiech zamutkow  
2-8 mych wyſwobodil mie g̃t. Priſtupte k<sup>7</sup>  
2-9 niemu a ofwiette ſie. a twari waſie nebu<sup>7</sup>  
2-10 du pohanieny. Tento chudy wolal geſt.  
2-11 a pan vſlyſal g̃t gey. a zewſiech zamutko<sup>w</sup>  
2-12 geho ſpaſil g̃t ho. W puſtie andiel panie  
2-13 w okol bogicích ſie geho. a zproſti gie. O<sup>7</sup>  
2-14 kuſte avizte zet g̃t chutny pan. Blažený  
2-15 muž kterýž vſa wen. Boyte ſie pana wſic<sup>7</sup>  
2-16 kni ſwieti geho. nebot nenie nedostatku  
2-17 bogicím ſie geho. Bohatí zepſielifū a zlač<sup>7</sup>  
2-18 nielifū. ale kteriž ſie bogie pana. nebudu  
2-19 vmenſeni wewſem dobrém. Podte ſyno<sup>7</sup>  
2-20 wé poſluchayte mne. bazni panie naučímt  
2-21 was. Ktery g̃t člouiek geſto chce žiwotu.  
2-22 miluge dny widieti dobré. Zbran yaziku  
2-23 twenu otzleho. a rtowé twogi namilugte  
2-24 lſti. Otwrac ſie otzleho a czin dobree. hle<sup>7</sup>

**O** hwaliti budu pana powſie ga  
si wždycky chwala geho w v  
stech mych. W panu chwalena  
bude dušie. ſlyšte tiſij a weſelí  
budte. Welebte pana ſemnu. a powyſme  
gmena geho weſpolek. Hledalſem pana  
a vſlyſal mie g̃t. a zewſiech zamutkow  
mych wyſwobodil mie g̃t. Priſtupte k  
niemu a ofwiette ſie. a twari waſie nebu  
du pohanieny. Tento chudy wolal geſt.  
a pan vſlyſal g̃t gey. a zewſiech zamutko<sup>w</sup>  
geho ſpaſil g̃t ho. W puſtie andiel panie  
w okol bogicích ſie geho. a zproſti gie. O  
kuſte avizte zet g̃t chutny pan. Blažený  
muž kterýž vſa wen. Boyte ſie pana wſic<sup>7</sup>  
kni ſwieti geho. nebot nenie nedostatku  
bogicím ſie geho. Bohatí zepſielifū a zlač<sup>7</sup>  
nielifū. ale kteriž ſie bogie pana. nebudu  
vmenſeni wewſem dobrém. Podte ſyno  
we poſluchayte mne. bazni panie naučímt  
was. Ktery g̃t člouiek geſto chce žiwotu.  
miluge dny widieti dobré. Zbran yaziku  
twenu otzleho. a rtowé twogi nemilugte  
lſti. Otwrac ſie otzleho a czin dobree. hle  
dap poſkože a naſledug geho. Oči panie  
na ſpravedliwe. a vſi geho kypſbá gich.  
Ale obſlizej panie naty kteriž čime rlee

2-1 hwaliti budu pana powſe ča<sup>7</sup>  
2-2 si vždycky chwala jeho v v<sup>7</sup>  
2-3 stech mych. V panu chwalena  
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<sup>7</sup>  
2-9 nemu a osvette se. a tvaři vase nebu<sup>7</sup>  
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<sup>7</sup>  
2-14 kuste avizte žet jest chutny pan. Blažený  
2-15 muž kterýž vsa ven. Bojte se pana všic<sup>7</sup>  
2-16 kni swetí jeho. nebot není nedostatku  
2-17 bojicím se jeho. Bohatí zepselisu a zlač<sup>7</sup>  
2-18 nelisu. ale kteriž se bojí pana. nebudu  
2-19 vmenšeni ve všem dobrém. Podte syno<sup>7</sup>  
2-20 vé posluchajte mne. bazni pane naučímt  
2-21 vas. Ktery jest člověk ješto chce životu.  
2-22 miluje dny videti dobré. Zbran jaziku  
2-23 tvenu otzleho. a rtové tvoji namiluite



si nawysost wzalsi wazbu. wzalsi darynali  
di. Zajiste y nevericiaby přebywali wpa  
nu bohu. Požehnany pan w den nakaždy  
den. stastnu cestu učiní nam. Boh spasení  
našich. Boh naš boh aby učinil spaseny. a  
pane pane wychod smrti. Ale však boh ze  
tře hlavy nepřatel svých. vrch vlasu tech  
gesto chodie w hřešech svých. Rzekl gest  
pan z bazan. obratim obratim w hlubokost  
mořsku. Aby omočena byla noha tva wekr  
wi. y azik psow tvych z nepřatel tvych ot  
nyeho. Vidielisu w chazení tva bože. w cha  
zení boha ního. krale meho jenž jest w swa  
tem. Předěšlisu knížata spojení jsuc chwa  
litebníkom. prostřed mladíc bubnujících.  
W sbořích chwalte boha. pana z studnic izra  
helských. Ču benamín mladeneč. wmysli  
wystupení. Knížata vida wodce jich. kní  
žata zabulon. knížata neptalim. Příkaz  
bože mačti tve patvěd toho boj čjs v dela.  
wnas. Od chramu tveho w geruzalemie.  
tobie obietowati budu kralowe dary. A alaj  
z wieri trestne. sebranie bykow mezi krawa

1-2 si navysost vzalsi vazbu. vzalsi darynali  
1-3 di. Zajiste y nevericiaby přebywali wpa  
1-4 nu bohu. Požehnany pan v den nakaždy  
1-5 den. stastnu cestu učiní nam. boh spasení  
1-6 našich. Boh naš boh aby učinil spaseny. a  
1-7 paní pane vychod smrti. Ale však boh ze  
1-8 tři hlavy nepřatel svých. vrch vlasu tech  
1-9 ješto chode v hřešech svých. Rzekl jest  
1-10 pan z bazan. obratim obratim v hlubokost  
1-11 mořsku. Aby omočena byla noha tva wekr  
1-12 vi. jazyk psův tvých z nepřatel tvých ot  
1-13 nyeho. Vidělisu v chazení tva bože. v cha  
1-14 zení boha ního. krale meho jenž jest v swa  
1-15 tem. Předěšlisu knížata spojení jsuc chva  
1-16 litebníkom. prostřed mladíc bubnujících  
1-17 Vsbořích chwalte boha. pana z studnic izra  
1-18 helských. Ču benamín mladeneč. vmysli  
1-19 vystupení. Knížata vida vodce jich. kní  
1-20 žata zabulon. knížata neptalim. Příkaz  
1-21 bože mačti tve patvěd toho boj čjs v dela.  
1-22 vnas. Od chramu tveho w geruzalemí.  
1-23 tobí obetovati budou kralove dary. A alaj  
1-24 zveři trestne. sebrání bykův mezi krawa

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/slovník/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/slovník/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