OCR Fine Reader

Převod skenovaných stránek knih do znakové podoby wordovského souboru probíhá při použití OCR programu Fine Reader v následujících krocích:

Sken → **Převod na písmena (číst)** → **Převod do Wordu** → **Úprava ve Wordu**

Při tomto procesu dochází k různým nepřesnostem, které je pak třeba ručně odstraňovat ve výsledném dokumentu. Tyto nepřesnosti jsou dvojího druhu.

A. Do první kategorie patří nepřesná interpretace neostře vytištěných, slitých či zašpiněných znaků (interpret Fine Readru nerozpozná v bitmapě přesně příslušné písmeno nebo je interpretuje kupříkladu jako písmeno kurzivní místo písmeno v obyčejném řezu). S tímto druhem chyb se nedá moc dělat, pouze se můžete pokusit nastavit jiný jas či kontrast při skenování. Tyto chyby je pak třeba v textu jednu po druhé najít a ručně opravit.

B. Druhou kategorii tvoří chyby systematické. Do tohoto okruhu patří následující typy chyb:
   1. V českém textu (český jazyk) se chybně analyzují přehlásky (**ä, ü, ö**).
   2. Při převodu se systematicky díky zabudovanému slovníku chybně interpetují některá slova (např. slovenské slovo **ocko** se objeví jako **očko**).
   3. Jednotlivé verše se slijí do jednoho odstavce.
   4. Na konci některých řádků se bůhvíproč objeví ruční zalomení řádku
   5. Uvozovky a jednoduché uvozovky se převádí na horní rovné uvozovky, resp. čárku a apostrof.
   6. Pomlčky se interpretují jako diviz nebo naopak dlouhá pomlčka.
   7. Je-li pomlčka na začátku odstavce (někteří autoři tak uvozují přímou řeč), objeví se ve wordovském textu odstavec s odrážkou, nikoliv však samotný znak pomlčky.
   8. Trojtečka je interpretována jako 3 tečky.

Některé z uvedených systematických chyb lze odstranit volbou správného režimu převodu bitmapy na písmena (1. a 2.) nebo režimu převodu textu do Wordu (3. a 4.), jiné pak standardními hromadnými úpravami výsledného textu ve Wordu (4.–8.).

Dále naleznete tipy pro uvedených obtíží.

TIPY PRO SKENOVÁNÍ

ROZDĚLENÍ DVOUSTRAN (při skenování rozevřených knížek)

Nástroje – Možnosti – Skenovat/otevřít obrázek zaškrtnout ***Rozdělit dvojstrany***

SKENUJTE DOSTATEČNĚ SYTĚ A KONTRASTNĚ

Vyhnete se tak mnoha zbytečným nepřesnostem při následném rozpoznávání znaků.

RADĚJI SI VYZKOUŠEJTE CELÝ POSTUP skenu, rozpoznávání a převodu na několika málo stránkách

. Ušetříte si tak zbytečný nový sken nebo spoustu úmorné práce při následném ručním odstraňování zbytečných chyb. Sledujte při tom zejména:

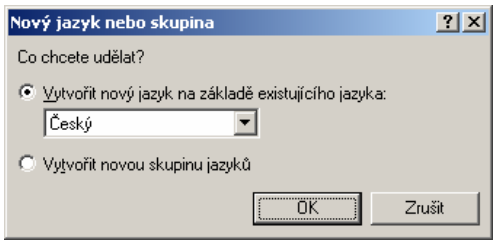
- Jak se převádějí akcentovaná písmena (zejména ú, ů, é, ě, ď, ť)
- Jak se převádějí přehlásky
- Zda se nekomolí slova
- Jak se zalamují řádky (případně verše)
- Jak se zachovává kurziva a tučný řez.

V případě nedostatečnosti konverze zkuste zvýšit jas či kontrast skenování, eventuálně nastavit jiný režim rozpoznávání a převodu (viz. dále).

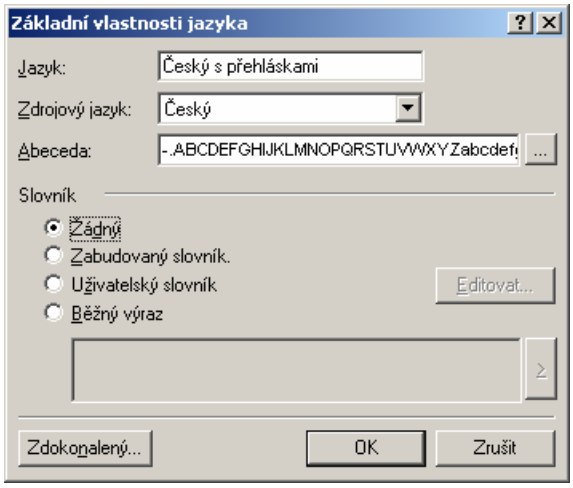
TIPY PRO ČTENÍ A PŘEVOD

PŘEHLÁSKY – VYTVOŘ NOVÝ JAZYK (ČESKÝ S PŘEHLÁSKAMI)

Nástroje – Možnosti – Rozpoznávání – Editovat jazyky – Nový – Vytvořit nový jazyk na základě existujícího jazyka – Český – OK



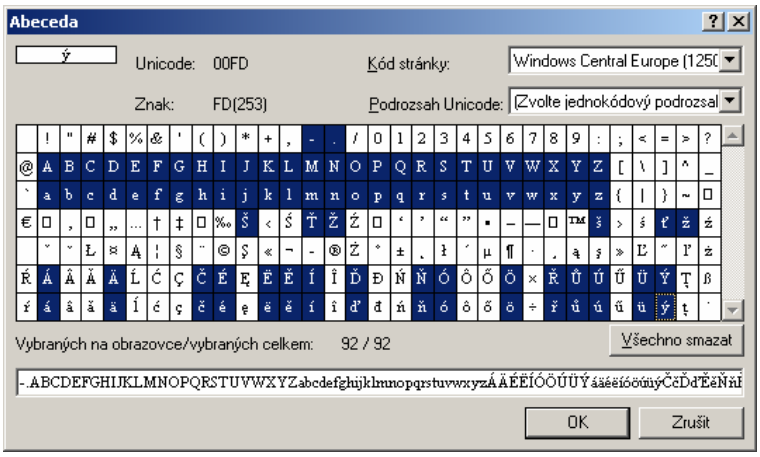
Zvol jméno nového jazyka (**Český s přehláskami).** Následně specifikuj novou sadu rozpoznávaných znaků:



Můžeš také zvolit, zda používat či nepoužívat slovník. Zvolíš-li **Slovník: Žádný**, neuplatní se při analýze slovník a Fine Reader si nebude vymýšlet a komolit slova. Nebude ale ani opravovat případné chyby.

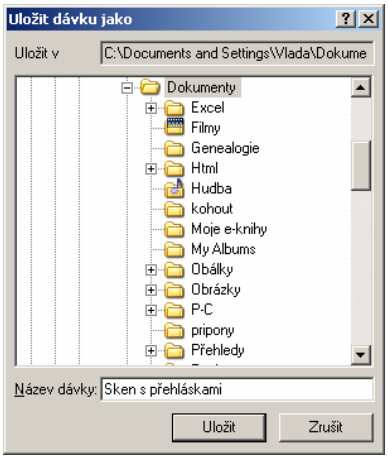
Klikni na tlačítko **…** na konci řádku **Abeceda**.

V okně s abecedou klikni na přehláskové znaky (tím je vybereš) a nakonec potvrď kliknutím na OK



Zavři klávesou OK i okno **Základní vlastnosti jazyka** a klávesou Ukončit i **Editor jazyka**

Definování nového jazyka platí pouze v aktuální dávce. Proto je třeba ji uložit, čímž se uloží i všechny její vlastnosti (a tedy i nový jazyk):

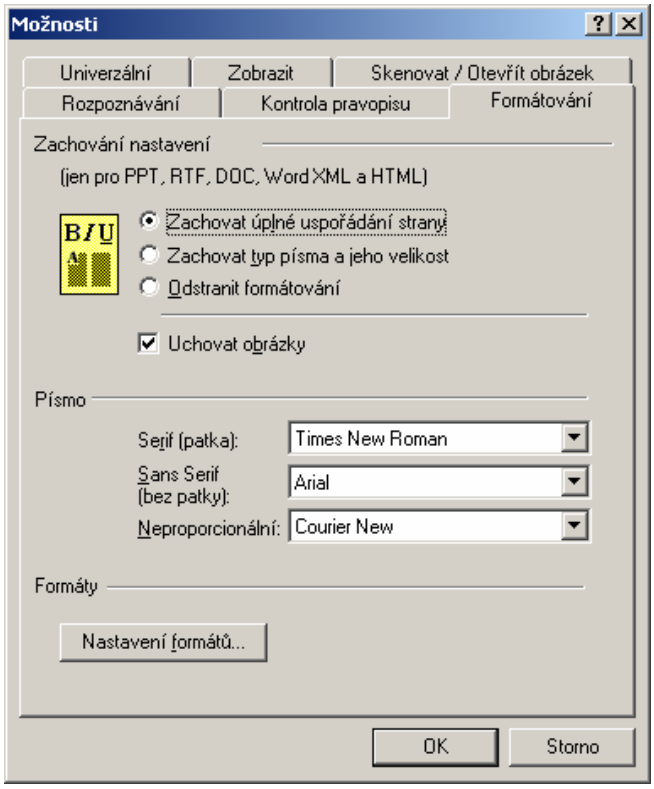


V poli volby jazyku (na Standardní liště) zvol nový jazyk (***Český s přehláskami***) a ulož dávku: ***Soubor – Uložit Dávku***

Kdykoliv v budoucnu budete potřebovat rozeznávat i přehlásky, otevřete uloženou dávku, zvolte příslušný jazyk (***Český s přehláskami***) a dál už pokračujte skenováním atd.

ZVOLTE SPRAVNÝ REŽIM FORÁTOVÁNÍ při převodu textu z Fine Readru do Wordu

- **Zachování formátování stránky** – zachovává všechny formátové atributy, vertikální vzdálenost mezi odstavci řeší jako proklad (nikoliv prázdný odstavec). Může však špatně řešit odrážky (jako odstavce s odrážkou a nikoliv samostatné znaménko) a pomlčky. Každá stránka je samostatný oddíl (s vlastním zrcadlem tisku, počtem sloupců atd.). Často bývá problém s verši, které spojí do jedné nudle. Občas z nejasného důvodu končí sadu řádek ručním zalomením řádku (**^l** = Shift+Enter). Tento režim bývá výhodný spíše pro skenování menšího počtu stránek bez veršů a pomlček na začátku odstavců.

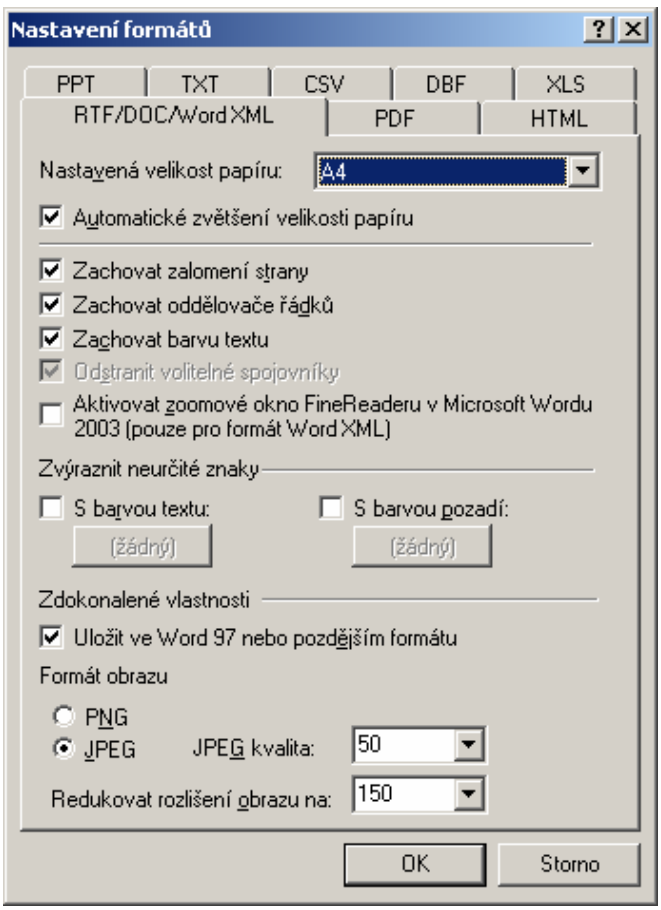


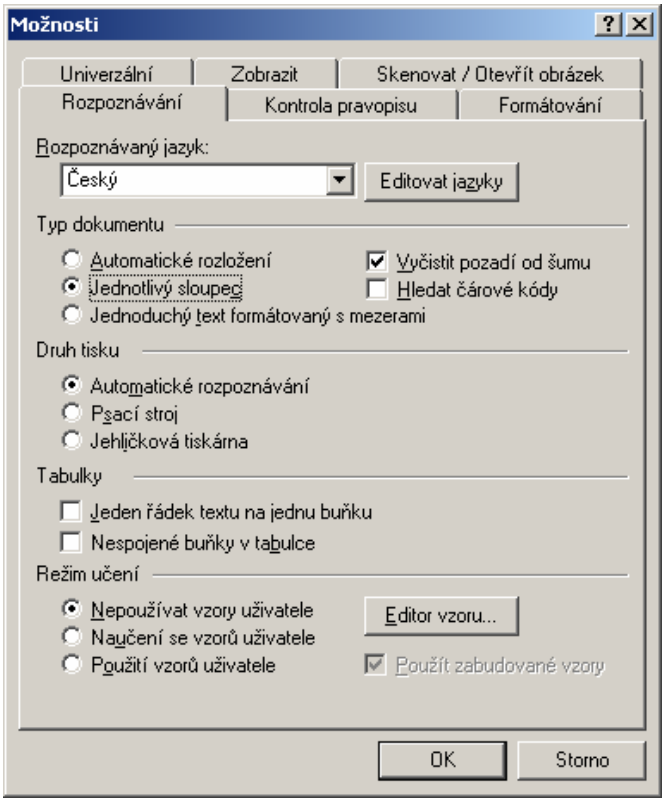
- **Zachování typu písma** – zachová atributy písma (tučné, kurziva…), ne však atributy odstavců. Stránky jsou spojené do jednoho oddílu (je při kontrole třeba hledat jejich konce). Celá dávka je transformována do jednoho oddílu. Tento režim zvolte, chcete-li použít naskenovaný text k novému zlomu!

- **Odstranit formátování –** zmizí veškeré atributy textu a jednotlivé řádky jsou ukončeny koncem odstavce. (Používej zejména pro verše: každý řádek zakončí koncem odstavce.)

ČTENÍ VERŠŮ

- **Nastav režim zachování oddělených řádků (případně i zalomení stran)**
  Ctrl+Shift+X (nebo **Nástroje – Nastavení formátů…** nebo **Možnosti – Formátování – Nastavení formátů…**) a zaškrtni příslušný režim **Zachovat oddělovače řádků** event. **Zachovat zalomení stránky**. Před následným čtením odstavců nezapomeň opět tento režim vypnout. Nejpříhodnější je patrně kombinace **Možnosti – Rozpoznávání – Jednotlivý sloupec** a **Nastavení formátů – Zachovat oddělovače řádků + Zachovat zalomení stránky**. V tomto případě budou jednotlivé verše končit povelem nové řádky (**^l**) a sloky (následované prázdným řádkem) znakem nového odstavce (**^p**). Následně lze doupravit ve Wordu (**^p→^p^p a ^l→^p**).





Jiná možnost, jak číst verše, je nastavit režimy:

- **Možnosti – Rozpoznávání – Typ dokumentu: Jednoduchý text formátovaný s mezerami**
- **Nastavení formátů – Zachovat oddělovače řádků + Zachovat zalomení stránky**

V tomto případě se nezachovají tučné a kursivní řezy. Zařídíte-li však, aby se bloky textu prostíraly přes celou stránku (a nejen přes jednotlivé sloky, přičemž by se vynechaly prázdné řádky), bude každá verš ukončen koncem odstavce a prázdné řádky budou prázdné odstavce.

Bloky textu se mohou shodně nastavit na všech stránkách takto:

- Máš-li už nastavené nějaké bloky textu, vymaž je: vyber příslušné stránky a pak: **Dávka – Vymazat bloky a text** (nebo **Ctrl+Del**)
- Nastav příslušný blok na jedné stránce.
- Ulož tento blok: **Obrázek – Uložit bloky** – a ulož si jej rozvržení bloků pod nějakým jménem.
- Zkopíruj blok do všech dalších vybraných stran: **Obrázek – Načíst bloky**

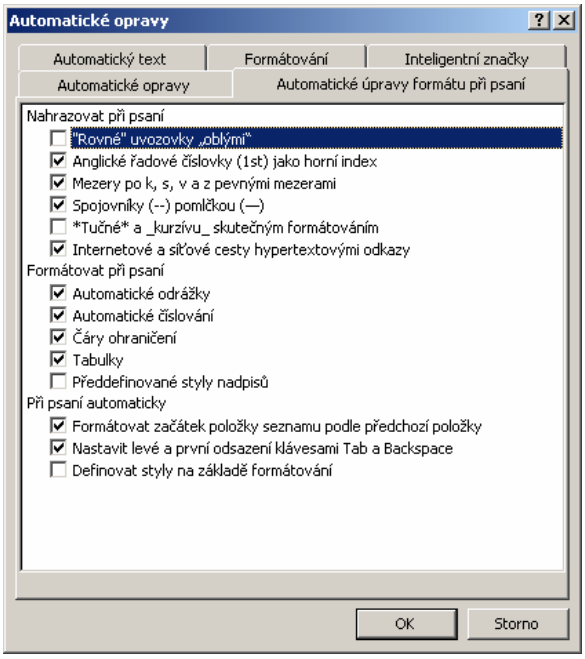
TIPY PRO ÚPRAVU VE WORDU

Chcete-li použít text pro nový zlom, odstraňte raději všechny pozůstatky dřívějšího formátování: různé velikosti písma, odsazení, zarovnání atd. (to vše se provede v rámci zlomu), odstraňte tabulátory, alespoň ze začátků odstavců. Vertikální odsazení odstavců řešte raději dodatečným ručním vložením prázdného odstavce. Proveďte proto následující akce:

- Sjednoť okraje stránek: spoj vše do jednoho oddílu tím, že nahradíš konce oddílu konci stránek: **^b → ^m** (důležité zejména v režimu ***Zachovat úplné uspořádání stran***)
- Sjednoť písmo (např. na Times 12•):
  - Vyber vše (***Ctrl+A***)
  - ***Formát – Písmo – Proložení znaků – Měřítko: 100%, Mezery: normální, Umístění: normální)***
- Sjednoť odstavce do stejné podoby (vyber vše ***Ctrl+A***, pak ***Formát – Odstavec – Řádkování: 1, Před:0, Za: 0***). Pak přesuň zarážky na měřítku tak, aby první řádek odstavce začínal asi o 5 mm vpravo a konec odstavce se kryl s koncem zrcadla.
- Odstraň tabulátory ze začátku odstavců: **^p^t → ^p**

Dále proveďte následující standardní literní záměny:

- Odstraň náhodně vzniklá ruční zalomení řádek: **^l → mezera**
- Odstraň mezery ze začátku odstavce: **^p mezera → ^p**
- Uvozovky: " → "
- Pravá jednoduchá uvozovky: **Alt+39 mezera → Alt+0145 mezera** (s vypnutou náhradou automatické náhrady rovných uvozovek oblými – ***Nástroje–Možnosti automatických oprav*** a dále viz. obrázek)
- Levá jednoduchá uvozovka: **mezera čárka → mezera Alt+0130** (s vypnutou náhradou automatické náhrady rovných uvozovek oblými – ***Nástroje–Možnosti automatických oprav*** a dále viz. obrázek)
- Apostrofy**: Alt+39 → Alt+0146** (s vypnutou náhradou automatické náhrady rovných uvozovek oblými – ***Nástroje–Možnosti automatických oprav*** a dále viz. obrázek)
- Trojtečky: **… → Alt+0133**
- Pomlčka: **mezera diviz mezera → mezera Alt+0150 mezera**
  (eventuálně **mezera diviz → mezera Alt+0150**)
  vyhoď pozůstatky po dělení na koncích řádků **diviz mezera**
- Zaměň dlouhou pomlčku na pomlčku: **Alt+0151 → Alt+150**
- Odstraň dvojité mezery (zaměň dvě mezery za jednu), opakuj.



Všechny popsané wordovské úpravy lze provést najednou pomocí maker obsažených v šabloně PUMA, kterou si můžete stáhnout spolu s návodem, jak ji instalovat a používat z internetové stránky www.paseka.cz/pistorius.