Beta version

Opravidlo

Nový český webový korektor

Mgr. Hana Žižková, Ph.D

Ústav českého jazyka Filozofické fakulty Masarykovy univerzity

Tento projekt je financován se státní podporou Technologické agentury ČR v rámci Programu ÉTA.
Projekt TL02000146 Webový pravopisný, gramatický a typografický korektor pro český jazyk

www.Opravidlo.cz

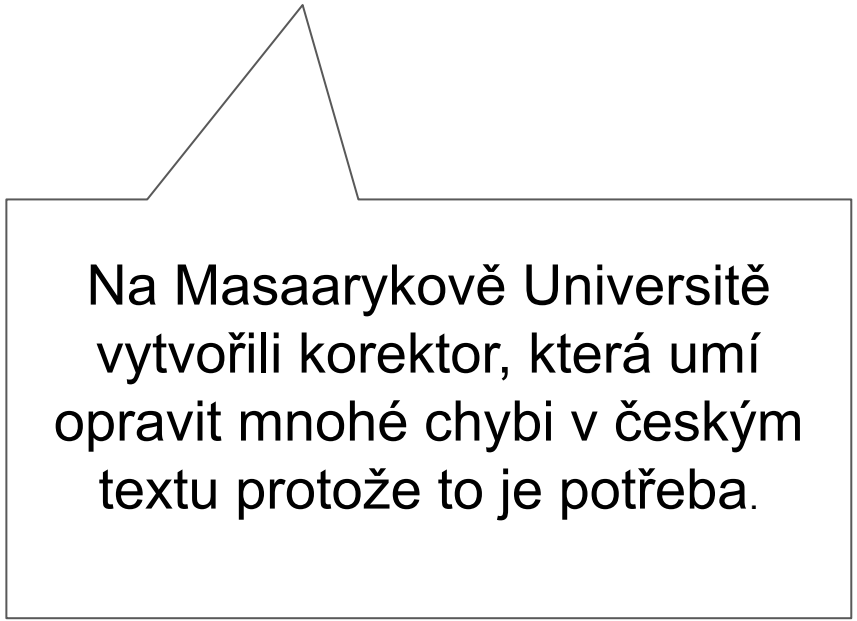
- webový korektor českých textů
- vznikl pod záštitou ÚČJ FF MU ve spolupráci s:
    - ÚJČ AV ČR
    - ÚTKL FF UK
    - Seznam.cz
    - Wikimedia
- založen na formálních pravidlech

Východiska pro strojové opravy

- Co je a co není chyba.
- Co opravovat bez zásahu autora textu a na co autora upozornit.
- Co lze strojově odhalit a opravit.
- Propojení oprav s výkladem v Internetové jazykové příručce.
- Pravidla českého pravopisu.
- ČSN 01 6910 Úprava dokumentů zpracovaných textovými procesory.

ZOBRAZIT AUTOMATICKÉ OPRAVY

Na Masaarykově Univerzitě vytvořili korektor, která umí opravit mnohé chybi v českým textu protože je to potřeba.



→ Návrhy oprav 6 OPRAVIT VŠE

~~Masaarykově~~ → Masarykově

Výraz Masaarykově není ve slovníku

~~Univerzitě~~ → univerzitě universitně

Výraz Univerzitě není ve slovníku

Zobrazit více

Chybný tvar zájmena

~~chybi~~ → chybí chyby chybu

Výraz chybi není ve slovníku

Zobrazit více

Chyba ve shodě přívlastku s řídícím větným členem

~~textu~~ → textu,

Chybějící větná čárka

SPUSTIT KOREKTOR KOPÍROVAT VYMAZAT OBSAH OPRAVA DIAKRITIKY

Co Opravidlo umí opravit?

- překlepy
- interpunkci
- gramatickou shodu
- negramatické struktury jako (např. zeugma)
- zájmena
- velká písmena
- typografické nedostatky

Opravidlo

Co korektor umí

Korektor Opravidlo umí kromě běžných překlepů opravit také řadu chyb z oblasti pravopisu, gramatiky a typografie. Jak můžete vidět v následujícím dokumentu, je jich opravdu hodně. Některé jazykové
vytvořit. Přesto doufáme, že Opravidlo může být cenným pomocníkem při psaní jakýchkoli textů.

Interpunkce

Doplnění interpunkce

- Chybějící čárka před spojovacím výrazem (např. *Myslel si* že už neprší.*);
- chybějící čárka u vložené vedlejší věty (např. *Vedle domu, kde bydlím* roste strom.*);
- chybějící čárka za přímou řečí (*„Měl jsem velká očekávání“* pravil Luboš.*);
- chybějící čárka mezi členy několikanásobného větného členu (např. *Maminka* tatínek a děti jeli na dovolenou.; O finále si zahrají Brazilci* Portugalci* Španělé a Francouzi.*);
- čárka oddělující oslovení (např. *Dobrý den* pane Nováku!*).

Odstranění interpunkce

- Nadbytečná čárka mezi vedlejšími větami v poměru slučovacím, pokud se v nich opakuje spojovací výraz (např. *Přála si, aby se o ni psalo*, a aby byla slavná.*);
- nadbytečná čárka před spojením a to v případech, kdy spojení neuvozujekonkretizující doplnění, ale odkazuje na obsah předchozí věty (např. *Přišli dlouho váhal*, a to se mu nevyplatilo.*);
- nadbytečná čárka před zkratkami typu *atd., apod., aj.* (např. *Prodávali jablka, hrušky*, apod.*);
- nadbytečná čárka v souslovích s co (případně dalšími relativy) (např. *Přijdu*, co nevidět.*);
- nadbytečná čárka před aneb, čili, neboli (např. *diabetes*, čili cukrovka*);
- nadbytečná čárka v infinitivních konstrukcích se slovesy být a mít (např. *Nemám*, co ztratit.*);
- nadbytečná čárka sloužící k oddělení části věty před přísudkem (čárka podle výdechové pauzy) (např. *Díky vysokému obsahu chlorafylu*, má obilná tráva silný alkalizační a detoxikační efekt na organ

Gramatická shoda

- Shoda přísudku s jednoduchým podmětem (např. *Psi *štěkaly.*);
- shoda přísudku s několikanásobným podmětem skládajícím se ze dvou substantiv spojených spojkami a, i, nebo, ani, či, & nebo předložkou s (např. *Psi a kočky se *vyhýbaly na sluníčku.*);
- shoda shodného přívlastku před podstatným jménem (např. *starý psi*);
- shoda přívlastku před podstatným jménem v 7. pádě (např. *barevnýma pastelkama a *čtyřmi očima*);
- shoda mezi slovesnými tvary (např. *jsme dělal, *byli připraven, *dělali jsem*);
- chybné tvary kondicionálu (např. *kdyby jste, *by jsme*).

Chybné větné konstrukce

Zeugma

- Vybrané případy zanedbání dvojí vazby sloves (např. *Apelují a prosím admina webu*).

Atrakce

- Vybrané případy přizpůsobení gramatické formy slova gramatické formě slova sousedního (např. *Před *sluncem východem jsme vyrazili na cestu. Na základě nových *nařízenich se musí dveře zamyka

Kontaminace

- Vybrané případy chybných předložkových vazev sloves (např. *Pomýšlel *o rezignaci. Nebudu přihlížet *na to, jak to tady upadá. Osočil se *na mě.*);
- chybné použití slovesa vyvarovat se se 3. pádem (např. *Vyvarujte se přímému slunci.*).

Velká písmena

- Oprava psaní velkých písmen u vybraných vlastních jmen (např. *Jel do *brna. Dali mu jméno *petr *novák. Narodil se ve znamení *blíženců.*).

Typografie (velká část oprav probíhá automaticky)

- Násobné mezery místo jedné (např. *Petr Novák*);
- nedělitelné mezery (zde pro ilustraci ~) po jednopísmenných předložkách, iniciálách, jednopísmenných zkratkách a řadových číslovkách vyjádřených číslicí (např. *a ~s, J. ~K. ~Rowlingová, k ~oknu, s. ~

MUNI ARTS

Typografická oblast

- Automatické opravy

  Nadbytečná mezera před tečkou, vykřičníkem, otazníkem: Proč ?; nezlomitelná mezera po jednopísmenných předložkách nebo oprava zápisu typu *12%ní na 12%; dvě tečky na jednu tečku: *Nepřišel.. nebo opravy mezer uvnitř závorek: *(slovo).

- Opravy, o kterých rozhodne uživatel:

  záměny spojovníku a pomlčky: *9-12 vs. 9–12; zápis peněžních částek: 20,– Kč vs. 20 Kč; případné mezery před znaky %, ‰ ad., všude tam, kde rozhoduje kontext a význam: o 10 % roztok této kyseliny nezdražil od roku 2004 vs. o 10% roztok této kyseliny neměli zájem.

Ortografická oblast

- Velká písmena: Navštívil *brno. Pozdravoval *markétu. Narodil se ve znamení *blížence̊.
- Stopy ho zavedly do písku. vs. Stopy ho zavedly do Písku.
- Část chyb detekována v porovnání se slovníkem: vidět vs. *vydět.
- Kolokace v případech homonymie: výr vs. vír; výr velký vs. vír velkoměsta.
- Hranice slov: vleže; nahoře; zpět; na modro vs. namodro; na rozdíl vs. *narozdíl; na shledanou vs. *nashledanou.
- Uživatel (vtom zahřmělo vs. v tom balíčku).
- Diakritika (Nemam cas. vs Nemám čas.; Zpival na zahrade vs. Zpíval na zahradě; ale ne páni vs. paní; vír vs. vir; krtiny vs. křtiny.

Lexikální oblast

- Označení slov hovorových (např. kafčo ad.), nářečních (čupnout, brundibál šulánky, čudla ad.) nebo slangových (pacoš, lajknout ad.)
- Upozornění na nesprávné konstrukce typu *více/méně jak vs. více/méně než; záměnu zájmen jakýkoli z a kterýkoli; pleonasmus: Dostal dárek zadarmo.
- Strojová detekce chyb není možná tam, kde je k posouzení významu potřeba širší kontext.

Morfologická oblast

- Porovnání hodnoceného tvaru se správným slovníkovým tvarem: *za zádama vs. za zády, *v konvy vs. v konvi; *s drahokami vs. drahokamy nebo *dvěmi vs. dvěma.
- Složitější případy: (jsou tam) lvi vs. (vidím) lvy.
- Na většinu chyb, které spadají do oblasti morfologie, korektor upozorní.

Syntaktická oblast

- Chybějící interpunkce před spojovacími výrazy: Řekl *že přijde.; u oslovení: Zavolám ti *Hanko* později.; částečně za vedlejší vloženou větou: Vlak, kterým měli přijet *měl zpoždění.; nadbytečná čárka: Samozřejmě*, že spal.
- Chyby ve shodě jednoduchého podmětu a přísudku (Studenti se *učily.), duálových tvarů (mezi čtyřma očima) nebo neshodující se slovesné tvary (*nesli jsem) nebo vybraných dalších složitějších případů shody podmětu a přísudku ad.

Sémantická oblast

- Limity všude tam, kde je potřeba kontext a význam.
- Kolokace výr velký vs. vír velkoměsta; kropicí konev vs. kropící zahradník.

Stylistická oblast

- Upozornění na opakování stejných slov nebo slov se stejným kořenem: Až tu funkci implementujeme, pro toho přihlášeného uživatele to bude podmínkou, odsouhlasit ten souhlas.
- Upozornění na užití nespisovných nebo hovorových výrazů nebo hromadění předložek: Přišel s pro něj nejlepším řešením.

Co Opravidlo trápí?

- Věty, kde je důležitý význam: Přijely autobusy. vs. Přijeli autobusy.
- Chybná data na vstupu: např. rozlišení nominativu a akuzativu: Poradily jim děti.; To je krize.
- Limit pravidlových oprav: např. předložkové vazby: Bydlí naproti břečtanem porostlému domu.
- Rychlost odezvy.

Jak Opravidlo funguje?

- Modulární struktura
- Několik použitých nástrojů
- Vlastní slovník
- Webové rozhraní
- Volně dostupný nástroj

Použité nástroje

1. tokenizace: UNITOK
2. morfologická analýza: MorphoDiTa a Majka
3. syntaktická analýza: SET
4. moduly

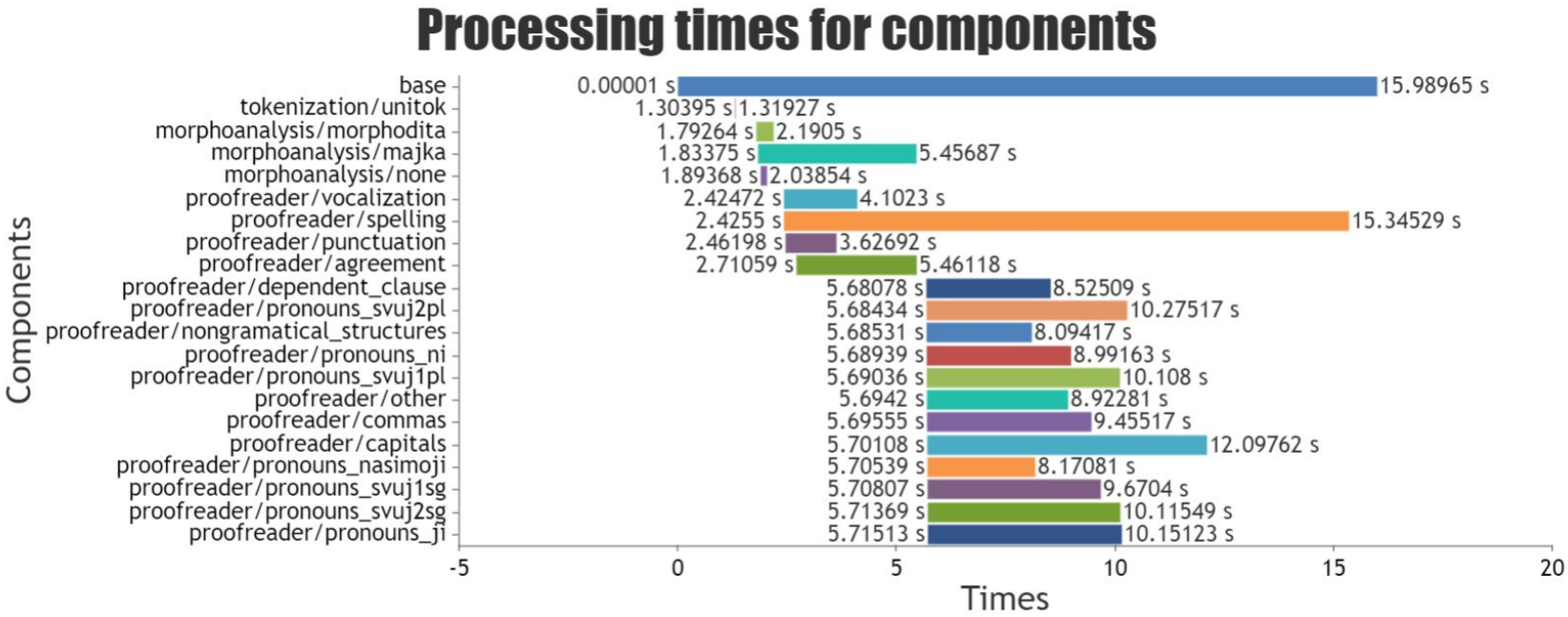
morfologický slovník: MorfFlex + IJP + frekvence

překlepy: SymSpell

Moduly

- interpunkce
- shoda podmětu s přísudkem, shoda přívlastku
- negramatické větné konstrukce (zeugma, kontaminace, atrakce)
- typografické chyby
- ostatní chyby
- kontrola překlepů (slovník)
- velká písmena
- doplňkové moduly dalších pravidel (chybné uvození vedlejších vět, konkurenční tvary zájmen)

Moduly vs. rychlost



MUNI ARTS

Slovníky

- **slovník korektních tvarů**
  - IJP, Ajka a MorfFlex – 23 mil. tvarů
  - IJP (*jiné je*, *lze i*), Ajka (zastaralé tvary) a MorfFlex
  - frekvenční seznam SYN v. 9 – 2,5 mil. tvarů
  - do slovníku se dostaly nekorektní tvary
- **našeptávač**
  - frekvenční seznam z korpusu CzechWeb2017
  - nabízí tři nejpodobnější tvary
  - vlastní slovník: MorfFlex 2.0 a slovníku z IJP (celkem 3 miliony tvarů)

MorphoDiTa vs. Majka

- moduly využívají výhodnější tagger pro danou oblast
  - některé moduly využívají i oba taggery
- rozdílné značkování způsobuje potíže
  - při formulaci CQL dotazů
  - při tvorbě pravidel
- rozdílný slovník
  - Majka: využívá kmeny, objevuje se přegenerování, neudržovaný
  - MorphoDiTa: MorfFlex, aktualizovaný

Formální jazyková pravidla

- více než 7500 pravidel
- Např.: Pravidlo (a) přikazuje vkládat čárku před vedlejší větu uvozenou spojkou že, pokud spojce že předchází ale (Nemyslím si ale, že má pravdu) a pokud předchozí věta není uvozená rovněž spojkou že (např. Musíš mě přesvědčit, že to nejsou jen plané sliby, ale že svá slova skutečně splníš), protože tím pádem by mezi větami vedlejšími vznikl poměr odporovací. Zmíněnou situaci řeší pravidlo (b). (J. Machura)

```
(a) myslim ale, ze
TMPL: $NEG ... (word ale) (word že)  MARK 3 <c>  PROB 500
    $NEG(word not): že

(b) VH, ze VV, ale ze VV
TMPL: (word že) ... (word ale) (word že)  MARK 2 <c>  HEAD 2
      PROB 4000
```


MUNI ARTS

Závěrem

- Možnosti a meze strojových oprav online webovým korektorem v českém textu v rámci jednotlivých jazykových rovin/oblastí.
- Typografie: jediná oblast automatických oprav některých zjištěných chyb.
- V oblasti ortografické, lexikální a morfologické velká část chyb detekována porovnáním existujících tvarů s daty ve slovníku korektoru.
- V oblasti syntaktické, sémantické a stylistické detekce chyb a jejich opravy více založeny na formálních pravidlech.

- **Opravidlo momentálně nejkomplexnější volně dostupný nástroj, který ocení každý, kdo píše česky.**

Děkuji za pozornost

MUNI ARTS

Věty pro Opravidlo

Myslel si že už neprší.

Vedle domu, kde bydlím roste strom.

„Měl jsem velká očekávání“ pravil Luboš.

Maminka tatínek a děti jeli na dovolenou.

O finále si zahrají Brazilci Portugalci Španělé a Francouzi.

Dobrý den pane Nováku!

Psi štěkaly.

Tudy by jsme nešli.

Apeluji a prosím admina webu.

Před sluncem východem jsme vyrazili na cestu.

[24] Jel do brna. Dali mu jméno petr novák. Narodil se ve znamení blíženců.

MUNI ARTS