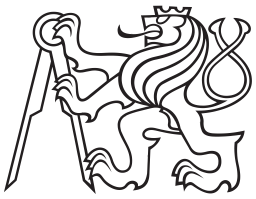


Středoškolská technika 2016

Setkání a prezentace prací středoškolských studentů na ČVUT

Nástroj pro korekci interpunkce na Wikipedii

Martin Scheubrein

Gymnázium Třebíč
Masarykovo nám. 9/116

Abstrakt

Tato práce si klade za cíl vytvoření nástroje pro automatizované vyhledávání chybně užitých interpunkčních znamének, jmenovitě pomlčky a spojovníku, v české sekci internetové encyklopedie Wikipedia. Výsledný počítačový program by měl usnadnit a výrazně urychlit jejich nalezení a za asistence korektora provádět opravné editace. Práce je koncipována tak, aby zároveň byla vzorem pro tvorbu podobných nástrojů užívajících rozhraní MediaWiki.

Klíčová slova: Wikipedie; MediaWiki; typografie; pomlčka; spojovník.

Abstract

The purpose of this thesis is to create a tool automatically searching for incorrectly used punctuation marks, namely dash and hyphen, on the Czech edition of Wikipedia, the internet encyclopædia. The resultant computer program is supposed to facilitate and speed up their retrieval and to commit the corrective editations with the assistance of a human corrector. The project may also serve as an exemplar for similar tools utilizing the MediaWiki API.

Keywords: Wikipedia; MediaWiki; typography; dash; hyphen.

Title translation: A punctuation-correcting tool for Wikipedia

Obsah

1 Úvod 4
2 Typografické pozadí 5
2.1 Spojovník 5
2.2 Pomlčka 6
2.3 Minus 6
3 MediaWiki 7
3.1 Wiki markup 7
3.2 Webové rozhraní 8
3.3 MediaWiki API 8
3.3.1 Přihlášení 8
3.3.2 Seznam náhodných článků 9
3.3.3 Stažení článku 9
3.3.4 Editace 10
4 Technologie a datové struktury 11
4.1 Pthreads 11
4.2 Knihovna readline 12
4.3 Knihovna libcurl 12
4.4 Parser jsmn 12
4.5 Regulární výrazy 12
4.6 Fronta 13
4.7 Halda 13
5 Architektura nástroje 15
5.1 Náhodné články 15
5.2 Stažení článku 16
5.3 Zablokování zakázaných úseků 16
5.4 Hledání chyb 17
5.5 Schválení uživatelem 18
5.6 Aktualizace chyby 19
5.7 Oprava textu 19
5.8 Odeslání změn 19
6 Zhodnocení výsledků 21
6.1 Typy článků 21
6.2 Vylepšení regulárních výrazů 22
6.3 Povaha chyb 23
6.4 Přínos pro Wikipedii 24
7 Závěr 26
Reference 27

Kapitola 1
Úvod

Wikipedie je svobodná internetová encyklopedie založená roku 2001 Jimmym Walesem.[1] Už za krátkou dobu své existence si získala oblibu veřejnosti právě díky své dostupnosti a volně přístupnému obsahu.

Slovo *svobodná* v jejím podnázvu ovšem neznamená jen zdarma stažitelný obsah. Význam tohoto slova – jak vysvětluje Free Software Foundation,[2] nadace podporující vývoj svobodného softwaru, s nímž Wikipedie sdílí základní myšlenky – se dotýká nejen pasivního užívání, ale vyjadřuje i svobodu veškerý obsah modifikovat a při zachování těchto svobod ho dále šířit. Takový princip se nazývá *copyleft*.[3]

To, že Wikipedii může kdokoli měnit a upravovat, bylo klíčovým faktorem jejího úspěchu. Získala tím desetitisíce aktivních přispěvatelů,[4] díky čemuž rychle pokryla velké množství encyklopedických hesel a stala se v praxi dobře použitelnou.

Kvůli neorganizovaným editacím ale nelze provádět systematickou kontrolu textu a obzvlášť ojediněle navštěvovaná hesla trpí faktickými nesrovnalostmi, nekonzistentním formátováním, překlepy a typografickými chybami. Právě posledními zmiňovanými se zabývá tato práce.

Častou chybou přispěvatelů Wikipedie je nerozlišování pomlčky (–) a spojovníku (-), interpunkčních znamének souhrnně označovaných francouzským výrazem *tiret*, jež jsou kvůli vzájemné grafické podobnosti mnohdy zaměňovány; a chybné mezerování okolo nich.

Smyslem této práce je navrhnout a realizovat nástroj, který by tyto jevy na Wikipedii sám vyhledával a usnadnil editorovi jejich opravu. Pomocí něj by pak bylo možné vymýtit alespoň znatelnou část těchto typografických omylů a tím učinit encyklopedii o něco více oku lahodící. Zároveň se při korekcích bude sbírat statistika o tomto druhu chyb, která poslouží i jako zpětná vazba a program bude na základě výsledků dále vylepšován.

V neposlední řadě má posloužit i jako vodítko či vzor při psaní podobných nástrojů, neboť řeší jak prohledávání textu a nalézání chyb, tak i efektivní softwarové řešení a komunikaci se servery Wikipedie přes rozhraní MediaWiki.[5]

[1] `en.wikipedia.org/wiki/Wikipedia` [cit. 4. 1. 2016]
[2] `www.gnu.org/philosophy/free-sw.html` [cit. 17. 11. 2015]
[3] `www.gnu.org/copyleft/copyleft.html` [cit. 17. 11. 2015]
[4] `en.wikipedia.org/wiki/Wikipedia:Wikipedians#Number_of_editors` [cit. 17. 11. 2015]
[5] MediaWiki – viz kapitolu 3 na straně 7.

Kapitola 2
Typografické pozadí

Pravidla počítačové sazby dokumentů čerpají ze tří hlavních pramenů.

Prvním jsou Pravidla českého pravopisu, jež definují závazná pravidla z pohledu lingvisty. Popisují rozdíly ve vztazích mezi slovy a větami, které by se měly v psaném textu vyznačit správnou interpunkcí.

Technické zhotovení řeší nezávazná, avšak obecně doporučovaná a dodržovaná norma ČSN 01 6910 *Úprava dokumentů zpracovaných textovými procesory* [1]. Ta kodifikuje použití konkrétních symbolů v textu, jejich vzhled a chování v závislosti na okolí.

Poslední neméně důležitou sadou pravidel jsou tradiční typografické zvyklosti. Typografie jako disciplína má od počátků svůj jasně vymezený účel: předat informaci obsaženou v textu jeho grafickou reprezentací tak, aby byl čtenář sazbou co nejméně rušen a mohl v maximální míře vnímat obsah textu.

Výše zmíněné tři prameny často nejsou úplné nebo jsou ve vzájemném rozporu. Příkladem budiž sekce normy [1] o nezlomitelné mezeře, již vyžaduje neroztažitelnou, ačkoli ve znakové sadě Unicode je tato mezera roztažitelná stejně jako obyčejná mezera mezislovní. Roztažitelná nezlomitelná mezera v sazbě esteticky mnohem lépe působí a neruší, což je v souladu s tradičními typografickými zásadami.

Pro tuto práci jsou nejzajímavější pravidla pro psaní znaků, jež lze souhrnně pojmenovat francouzským pojmem *tiret* – zahrnuje spojovník a pomlčku – a znaku minus. Všechny tři mají podobu různě vyhlížející vodorovné čárky, na klávesnicích počítačů však najdeme jediný – spojovník. Z toho důvodu je často sázen chybně na místa, kam patří pomlčka nebo minus.

Optimální průnik výše zmíněných souborů pravidel pro tyto znaky bude shrnut v této kapitole.

2.1 Spojovník

Spojovník je graficky znázorněn krátkou vodorovnou čárkou (-) sázenou zhruba v polovině výšky minusek[1] písma. Jeho tloušťka se blíží tloušťce hlavních tahů písma.

Jeho funkcí je – jak název napovídá – spojovat významově úzce svázané celky. Tomu napomáhá i jeho poměrně malá délka, díky níž celky těsně spojuje i graficky. Obyčejně se tedy užívá na úrovni jednotlivých slov.

Používá se například při kombinaci vlastností (*česko-německý slovník, modro-zelený dres*[2]), bližším určení (*kuchař-číšník*, *Brno-střed*), při spojení jmen (*Rakousko-Uhersko*)

[1] *Minusky* – malá písmena; opakem jsou *verzálky* – velká písmena.
[2] Ve významu modrý a zelený. Lze psát i *modrozelený* ve významu odstín mezi modrou a zelenou

nebo úzce souvisejících slov (*propan-butan*) naznačuje rozdělení slova (*jedno- a dvouděložné rostliny*), připojuje spojku -li (*Budeš-li mlčet, vyvázneš živ.*). Podrobné vymezení možností užití i s méně častými výskyty spojovníku lze dohledat např. v jazykové příručce Ústavu pro jazyk český [2].

Kolem spojovníku se nikdy nepíší mezery.

2.2 Pomlčka

Pomlčka má podobu delší vodorovné čárky. Je většinou tenčí než spojovník, a tak, ačkoli bývá sázena do podobné výšky, se opticky jeví být výš. V češtině rozlišujeme dvě podoby pomlčky. Pomlčku dlouhou (—) o délce jednoho čtverčíku[3] a pomlčku krátkou (–), která je dlouhá pouze půl čtverčíku. V češtině se téměř výhradně používá pomlčka krátká, s dlouhou pomlčkou se lze nejčastěji setkat v beletrii při značení aposiopese, funkčně však mezi nimi není žádný rozdíl a jejich použití závisí jen na vkusu a vůli sazeče.

S pomlčkou se nejběžněji setkáváme při oddělování vsuvky namísto interpunkčních čárek (*Toto tvrzení – jakkoli se zdá být absurdním – je pravdivé.*) nebo při vysvětlování (*nejvyšší bod Norska – Galdhøpiggen*). Při oddělování celých větných celků je žádoucí a zároveň nutné pomlčku oddělovat mezerami z obou stran. Graficky to výpověď člení na logické celky.

Další funkcí pomlčky je vyjadřovat rozsah nebo protikladnost. V takovém případě ale mezery kolem pomlčky nepíšeme, neboť má spíše spojovací než členicí funkci. V takovéto podobě se s ní setkáme u letopočtů (*mezi lety 1918–1938*), výčtu stran (*viz s. 49–53*), dopravních tras (*rychlík Znojmo–Jihlava*), sportovních střetů (*utkání Sparta–Slavia*) a jiných.

Pro výčet možností použití viz opět příručku [2]. Obecně ale platí, že mezery kolem pomlčky píšeme tehdy, plní-li ve větě funkci členicí, a bez mezer ji použijeme tam, kde nahrazuje spojku nebo předložku. V případě víceslovných pojmenování bývá vhodné uchýlit se k variantě s mezerami i tam, kde by se jinak vynechaly – nedochází tak k optickému spojení samostatných slov (srovnej *Náměstí Republiky–Černý Most* a *Náměstí Republiky – Černý Most*).

2.3 Minus

Znaménko minus (−) vzhledem připomíná krátkou pomlčku, ale má několik na první pohled nepatrných specifik. Kvůli uniformitě vzhledu matematických výrazů je vyžadováno, aby mělo stejnou délku jako znaménko plus. Navíc není sázeno jako běžný text na účaří,[4] ale na matematickou osu, do jiné výšky.

Při použití jako matematický operátor musí být od obou operandů odděleno vhodnou mezerou. Užívá-li se k zápisu záporných čísel, například teploty, přiléhá k číslu bez mezery.

[3] *Čtverčík* je typografická míra o velikosti výšky kuželky, tj. vertikální velikosi písma.

[4] *Účaří* je základní písmová linka, na které je položený řádek. Některá písmena, např. j, g, y, ji svým spodním dotahem přesahují.

Kapitola 3
MediaWiki

MediaWiki je software vytvořený původně pro účely Wikipedie a jiných projektů nadace Wikimedia (Wikibooks, Wiktionary...), ale posléze rozšířený i na další stránky, kupříkladu WikiLeaks nebo Uncyclopedia. Jeho účelem je implementovat *wiki*, tedy webovou stránku, kterou mohou sami uživatelé editovat pomocí webového rozhraní.

3.1 Wiki markup

Editace stránky neprobíhá přímo na úrovni zdrojového kódu. Články jsou psány značkovacím jazykem *wiki markup*, z nějž je až při požadavku na zobrazení stránky uživateli vygenerován HTML kód. Wiki markup je člověku srozumitelný, i laik se rychle naučí používat jeho základní funkce.

Klíčovým prvkem Wikipedie jsou odkazy. Jimi jsou články hustě provázány a umožňují čtenáři uživatelsky příjemnou formou zpřístupnit související témata nebo vysvětlit použitou terminologii. Zpočátku Mediawiki podporovala pouze jednoslovné odkazy,[1] v současnosti však disponuje syntaxí, která umožňuje zapisovat nejen odkazy víceslovné, ale dokonce i odkazy na článek, jehož název se liší od textu odkazu:

```
Zde je víceslovný odkaz na stránku [[Wikipedie]].
A tady je odkaz na [[Wikipedie|Wikipedii]] ve správném pádě.
```

Podobnou syntaxí – dvojicemi hranatých či složených závorek – lze do stránky vkládat obrázky, šablony nebo tabulky:

```
Obrázek lachtana: [[File:lachtan.png]]
V současnosti je na Wikipedii {{NUMBEROFARTICLES}} článků.
```

V jazyce wiki markup lze velmi snadno vyznačovat názvy sekcí a podsekcí, které se ve finálním dokumentu zobrazí příslušně naformátované, a měnit řezy písma:

```
== Sekce ==
=== Podsekce ===
Antikva, ''kurzíva'', '''tučný text'''.
```

Další nesčetné způsoby modifikace textu nabízí XML tagy, například:

```
<code> Citace zdrojových kódů </code>
<nowiki> Text, jenž zůstává netknut parserem jazyka wiki markup </nowiki>
Matematické prostředí: <math> x^2 + 2 = 0 </math>
```

Úplnou specifikaci jazyka si lze přečíst na jeho manuálové stránce [3].

[1] `en.wikipedia.org/wiki/MediaWiki#Markup` [cit. 28. 10. 2015]

3.2 Webové rozhraní

Uživatelsky přístupným způsobem úpravy článku je editace přes webové rozhraní. Z každého článku na Wikipedii vede odkaz přímo na editační pole, ve kterém uživatel může upravovat zdrojový wikitext. Rozhraní usnadňuje práci nabídkou vložení speciálních znaků (mezi nimi i pomlčky), často používaných konstrukcí (dvojité hranaté závorky), umožňuje zobrazit náhled stránky a zároveň vyznačit provedené změny ještě před odesláním na server.

V roce 2013 byla zavedena experimentální možnost upravovat stránku intuitivnějším způsobem ve WYSIWYG editoru.[2] Odpadá v něm nutnost ovládat jazyk wiki markup, a editace tak zvládnou i počítačově méně zdatní uživatelé, obtížně však zvládá pokročilejší editace, zejména speciálních prvků, jako jsou tabulky, infoboxy nebo šablony. Až dosud (listopad 2015) nebyla vydána stabilní verze.

3.3 MediaWiki API

MediaWiki poskytuje vedle webového rozhraní i aplikační rozhraní vhodné pro počítačem prováděné úpravy. Pro českou jazykovou mutaci Wikipedie je toto rozhraní dostupné na adrese `cs.wikipedia.org/w/api.php`. Na tuto adresu lze v přesně definovaném formátu odesílat požadavky a tím provádět na Wikipedii veškeré úkony od přihlášení uživatele přes stažení stránky až po její editaci.

3.3.1 Přihlášení

Registrace není sice nutným předpokladem k provádění editací, v případě automatizovaných editací je však velmi vhodná jednak k udržení vnitřního přehledu o tom, co program způsobil, jednak pro ostatní editory, aby mohli své připomínky k editacím cílit na konkrétného uživatele. V případě plně automatizovaných nástrojů registraci vyžadují pravidla pro provoz botů na Wikipedii [4].

V průběhu přihlašování se na straně klienta ukládají cookies dokládající, že je uživatel přihlášen. Je-li – stejně jako v tomto nástroji – použita knihovna libcurl,[3] o jejich správu se stará automaticky. Nutné je jen tuto volbu v programu zapnout.

Přihlášení do systému probíhá ve dvou fázích. V první si program zažádá o přihlášení a zasílá uživatelské jméno a heslo. Tato data je nutné odeslat metodou POST. Jednak proto, aby nezůstala v historii prohlížeče ani nebyla cachována, jednak kvůli tomu, že to MediaWiki vyžaduje. Struktura tohoto požadavku – kde `jmeno` a `heslo` jsou přihlašovací údaje uživatele – je:

```
 GET: action=login & format=json
POST: lgname=jmeno & lgpassword=heslo
```

[2] *WYSIWYG* je zkratkou anglického *what you see is what you get* – uživatel v reálném čase vidí finální podobu editovaného díla bez nutnosti kompilovat zdrojový kód.

[3] Libcurl – viz kapitolu 4.3 na straně 12.

MediaWiki odpoví textem obsahujícím tzv. *přihlašovací token*. Jde o řetězec znaků, který poslouží ve druhé fázi přihlašování. V té program provede druhý požadavek, v němž spolu s údaji z předešlého požadavku zašle i přihlašovací token na místě `token`:

```
 GET: action=login & format=json
POST: lgname=jmeno & lgpassword=heslo & lgtoken=token
```

Poté by mělo být přihlášení úspěšně dokončeno, což MediaWiki potvrdí v odpovědi.

3.3.2 Seznam náhodných článků

MediaWiki dokáže vygenerovat seznam názvů náhodně vybraných článků. Žádný článek se navíc nevyskytne v odpovědích vícekrát ani při opakovaném použití.[4] Požadavek vypadá takto:

```
GET: action=query & list=random & utf8 & rnnamespace=0 & rnlimit=pocet
```

Volba `rnamespace=0` znamená, že máme zájem pouze o články z hlavního jmenného prostoru, tedy o encyklopedické články, nikoli například o diskusní stránky. Na místě `pocet` je číselná hodnota vyjadřující, kolik názvů chceme zaslat. Pro standardního uživatele je maximální počet 10, uživatel s atributem *bot* může zažádat až o 20 názvů najednou.

3.3.3 Stažení článku

Jelikož články hodláme nejen číst, ale i upravovat, je rozumné naráz zažádat server jak o obsah článku, tak o jeho časové razítko.[5] Provedeme dotaz ve formátu:

```
GET: action=query & prop=revisions & format=json & utf8
     & rvprop=content|timestamp & redirects & titles=clanek
```

Za `clanek` dosadíme název článku, který chceme stáhnout. Obsahuje-li název nějaké speciální znaky, které mají v URL speciální význam, například mezeru, ampersand nebo lomítko, je nutné zakódovat název do URL kódování.[6]

Časové razítko má formu textového řetězce, který nemusí být dále upravován.

Textový obsah sice získáváme v kódování UTF-8, které je pro další zpracování vyhovující, obsahuje ale nerozkódované *escape sekvence*. To je zápis speciálních znaků, skládající se ze zpětného lomítka a běžného znaku nebo jeho oktalového zápisu. V programu, který jim rozumí, se tato dvojice přemění v jediný znak s jiným významem. Kupříkladu konec řádku zapsaný `\n`, tedy znak zpětného lomítka (hexadecimální hodnota `5C`) následovaný znakem malého písmene en (`6C`) se přečtou jako jediný znak konce řádku (`0A`). MediaWiki API tyto escape sekvence zasílá nerozkódované – tedy jako dva samostatné znaky. Aby se správně zobrazovaly, je nutné jejich výskyty nahradit dekódovanou variantou.

[4] `www.mediawiki.org/wiki/API:Random` [cit. 22. 12. 2015]
[5] Časové razítko – viz kapitolu 5.2 na straně 16.
[6] `www.w3schools.com/tags/ref_urlencode.asp` [cit. 4. 11. 2015]

3.3.4 Editace

K editaci – stejně jako k přihlašování – je nutné nejprve obdržet *editační token*. Ten lze získat jednoduchým požadavkem:

```
GET: action=query & meta=tokens & format=json
```

Token má podobu hexadecimálního čísla zakončeného znaky `+\`. Aby se těmito znaky nenarušila struktura URL, musí se zakódovat, tedy být nahrazeny sekvencí `%2B%5C`. Zbytek tokenu není třeba kódovat – vyskytují se v něm pouze bezpečné znaky ASCII.

Editační token zůstává platný po celou dobu přihlášení. Není třeba jej obnovovat před každou editací – stačí o něj požádat po přihlášení a po celou dobu běhu programu používat stále tentýž.

Při samotné editaci se kromě nového textu článku zasílá i krátké shrnutí změn, časové razítko získané při stažení stránky a editační token. Program edituje takto:

```
 GET: action=edit & format=json & minor & title=nazev
POST: summary=shrnuti & text=obsah & basetimestamp=razitko & token=etoken
```

Shrnutí editace se odesílá na místě `shrnuti`, samotný text článku na místě `obsah`, časové razítko získané současně s článkem samotným se dosazuje za `razitko` a editační token za `etoken`. Všechny tyto řetězce musí být kódovány.

Parametr `minor` označuje editaci jako tzv. malou změnu – takovou, která žádným způsobem nemění význam textu. Používá se při opravě typografie, překlepů nebo změně formátování.[7]

V takovéto konfiguraci se provádí editace celého článku najednou. Pokud by bylo žádoucí editovat jen určitou sekci článku, přibyl by parametr `section=sekce`, kde `sekce` je číslo označující editovanou sekci. Úvodní sekce je označena číslem nula, každá další postupně čísly od jedné výš.

[7] `cs.wikipedia.org/wiki/Wikipedie:Malá_editace` [cit. 23. 12. 2015]

Kapitola 4
Technologie a datové struktury

Plně automatizované nástroje operující na Wikipedii (tzv. *boti*) bývají nejčastěji řešeny jako webová služba. Jsou psány v některém z jazyků webu, například PHP nebo Perl. Vlastní program běží na vzdáleném serveru a ovládá se přes webové rozhraní. Jelikož ke svému běhu nepotřebují asistenci člověka, bývá ovládání omezeno na instrukce *spustit* a *vypnout*.

Oproti tomu zde popisovaný nástroj je koncipován jako počítačová aplikace. Běží přímo u uživatele a ten má nad ním plnou kontrolu. Proto je psán v jazyce C. Tato volba s sebou nese výhody i nevýhody. Jako desktopová aplikace se snáz ladí, je robustnější a uživatel je schopen velmi rychle reagovat na nepředvídatelné situace, které by mohly za jeho běhu nastat. Na druhou stranu pro jazyk C neexistují žádné specializované knihovny pro práci s MediaWiki[1] a jelikož je jazykem s nízkou úrovní abstrakce, práce s textem v něm není triviální.

Neznamená to ale, že je nutné řešit všechny technické detaily od základu. Pro jazyk C existuje celá řada knihoven zpřístupňujících i složité funkce.

4.1 Pthreads

Pthreads je POSIXový[2] standard procesových vláken. Vlákna se v mnohém podobají procesům – tedy jednotlivým instancím programů. Každé vlákno je nezávisle řízeno plánovačem procesů a chová se jako samostatná výpočetní jednotka. Vlákna jsou ale procesům podřízena – každé vlákno běží pod nějakým procesem. Vlákna běžící pod stejným procesem mohou sdílet určité části paměti a mnohem snáz tak spolu mohou komunikovat a synchronizovat svou činnost.

Pthreads definuje funkce, které vlákna vytváří a manipulují s nimi, stejně jako synchronizační mechanismy, především pak mutexy a podmínkové proměnné. V případě Pthreads základy operací s vlákny zajišťuje jádro[3] a funkce definované standardem jich využívají skrze systémová volání.

[1] MediaWiki – viz kapitolu 3 na straně 7.

[2] *POSIX* je rodina standardů, která se snaží o kodifikaci jednotného aplikačního rozhraní a zajistit tak kompatibilitu mezi operačními systémy, zejména klony systému Unix, mezi něž patří mj. GNU/Linux, OS X nebo Solaris.

[3] *Jádro* neboli *kernel* je základní součást operačního systému zodpovědná za řízení všech procesů a přidělování systémových prostředků.

4.2 Knihovna readline

Readline je knihovna poskytující funkce k pohodlnému zadávání příkazů. Kromě vkládání textu umožňuje i jeho mazání, přesuny kurzoru a další obvyklé funkce, na které jsou uživatelé zvyklí, ale nejsou při programování samozřejmostí. Přes klávesové zkratky je možné používat i pokročilé editační funkce.

Mimo rozsáhlou implementaci editačního rozhraní je součástí knihovny i sada funkcí pro správu historie příkazů zadaných v minulosti a jejich opětovnému vyvolání.

Nejznámějším softwarem používajícím tuto knihovnu je Bash – oblíbený unixový shell, tedy textové rozhraní provádějící příkazy zadané uživatelem. V této podobě se knihovna rozšířila na většinu operačních systémů GNU/Linux.

4.3 Knihovna libcurl

Knihovna libcurl je svobodná knihovna pro přenos dat skrze protokoly HTTP, FTP, SMTP a mnoho dalších. Používá standardní URL syntaxi známou ze zápisu webových adres a umožňuje snadný vývoj klientských programů. Mezi funkce významné pro nástroj, jímž se zabývá tato práce, patří HTTP GET a POST požadavky a zpracování cookies.

4.4 Parser jsmn

MediaWiki vrací odpovědi na požadavky ve formátu JSON. Ten zapisuje prostorovou hierarchii objektů do člověkem i strojem snadno čitelného textového řetězce. K získání hodnot je třeba nejprve řetězec rozparsovat do původní hierarchické struktury a v té pak požadované hodnoty vyhledat.

Z množství parserů formátu JSON v jazyce C byl zvolen parser jsmn. Je pojat minimalisticky, zkompilovaný je velký pouze 6 KB a neváže se na žádné, dokonce ani standardní knihovny. Vyniká svou rychlostí a jednoduchým použitím.

4.5 Regulární výrazy

Regulární výrazy jsou mocným nástrojem pro vyhledávání v textu. Regulární výraz popisuje konečný automat, jenž přijímá nějakou množinu řetězců. Jde o jakýsi vzor, jemuž odpovídají vybrané části řetězce. Ty pak lze například nahradit jiným řetězcem nebo statisticky vyšetřit jejich počet.

Nejjednodušším regulárním výrazem je řetězec znaků, jemuž odpovídá ten samý řetězec znaků. Například regulárnímu výrazu `abc` odpovídá řetězec *abc*. Regulární výraz ale může obsahovat i speciální znaky.[4] Například znak `.` (tečka) zastupuje libovolný znak; potom výrazu `a.c` odpovídají například řetězce *abc* nebo *afc*. Kvantifikátor `*` (asterisk)

[4] Zde uváděný způsob zápisu odpovídá v praxi používané syntaxi, nikoli formálním značkám operací s regulárními výrazy.

zase vyjadřuje libovolný počet opakování předchozího znaku nebo skupiny znaků. Výrazu `ab*c` tedy odpovídají řetězce *ac*, *abc* i *abbbc*. Výrazu `.*` odpovídá libovolný řetězec.

Speciálních znaků existuje samozřejmě mnohem více, konkrétní význam se napříč implementacemi liší. Významnou syntaxí regulárních výrazů je ta definovaná standardem POSIX,[5] rozšířená nejen v Unixových systémech, pro něž byl standard psán, ale i v mnoha programovacích jazycích.

Vzorovou implementaci standardu poskytuje glibc – standardní knihovna jazyka C projektu GNU. K implementaci nástroje byla použita právě ona. Dokáže totiž označit začátek i velikost úseku, jenž odpovídá regulárnímu výrazu, takže lze chybný úsek textu snadno vyjmout a nahradit správným.

4.6 Fronta

Fronta je abstraktní datová struktura typu FIFO. To pochází z anglického *first-in, first-out* a znamená, že se z ní prvky odebírají ve stejném pořadí, v jakém byly vloženy. Základními operacemi nad frontou jsou tedy vložení prvku na jeden její konec a vyjmutí prvku z konce opačného.

Frontou lze snadno realizovat komunikaci mezi asynchronně běžícími vlákny programu. Když chce jedno vlákno předat data druhému bez použití fronty, nastaví globální vlajku, která signalizuje, že jsou data připravena k přenosu. První vlákno musí čekat, než si druhé vlákno vlajky všimne, což způsobí nepředvítadelně dlouhé pozastavení činnosti prvního vlákna. Teprve až bude druhé vlákno připraveno ke čtení, může dojít k přenosu.

S pomocí fronty je komunikace podstatně plynulejší. Odesílatel data vloží do globální fronty a pokračuje v činnosti. Maximální počet prvků fronty – pokud není uměle omezen programátorem – není stanoven; i když tedy první vlákno zapisuje rychleji, než je druhé vlákno ochotné/schopné číst, data se neztrácí. Příjemce je vyzvedává z fronty takovou rychlostí, jaká jemu vyhovuje. Za cenu vyšší spotřeby paměti tak dostaneme plynulý běh programu.

Fronta je v nástroji realizována obvyklým způsobem – dvojitým spojovým seznamem. Fronta obsahuje ukazatel[6] na první a poslední prvek a každý prvek pak kromě dat obsahuje také ukazatele na prvek předchozí a následující. Díky této provázanosti lze prvky přidávat i odebírat v konstantním čase. Operace s frontou jsou tedy velmi rychlé.

4.7 Halda

Halda je stromová datová struktura, ve které má každý vrchol vyšší hodnotu než všichni jeho potomci (pak hovoříme o *max-heap*) nebo ve které má každý vrchol hodnotu nižší než všichni jeho potomci (*min-heap*). Základními operacemi nad haldou jsou přidání libovolného prvku do haldy, zjištění hodnoty maximálního/minimálního prvku a vyjmutí tohoto prvku.

[5] `http://www.regular-expressions.info/posix.html` [cit. 15. 11. 2015]

[6] *Ukazatel*, též *pointer*, je proměnná obsahující adresu nějakých dat v paměti, jinak řečeno ukazuje na ně.

Vkládání prvku do haldy začíná vložením listu do nejkratší větve stromu.[7] To ale může porušit haldu v místě mezi novým prvkem a jeho otcem. Pokud je halda porušená, tyto dva prvky vyměníme. Tím se problém přesune o úroveň výš. Vyměňujeme takto prvky, dokud porušují haldu nebo dokud se nedostaneme do kořene. V nejhorším případě musíme provést tolik výměn, jaká je výška stromu, která je vzhledem k počtu jejích prvků logaritmicky velká.

Nalezení maximálního/minimálního prvku je triviální, neboť z definice haldy vyplývá, že se vždy nachází v kořeni.

Při vyjímání prvku z kořene se na jeho pozici přesune poslední list z nejnižší úrovně stromu. To haldu mezi kořenem a jeho potomky poruší. Opraví se tím, že podobně jako v případě vkládání vyměníme prvek s potomkem (tím, jehož hodnota je rozdílnější od hodnoty otce) a přesuneme tím porušení haldy o úroveň níže. Takto „probubláváme“ haldou do doby, než je obnovena. Stejně jako v případě vkládání se v nejhorším případě provede logaritmicky mnoho výměn.

Díky tomu, že vkládání prvku do haldy a vyjmutí nejmenšího/největšího prvku probíha v logaritmickém čase, je halda základem řadícího algoritmu *heapsort*, který běží s optimální asymptotickou časovou složitostí $\mathcal{O}(N \log N)$. Nástroj opravující interpunkci na Wikipedii haldu používá právě k řazení podezřelých úseků textu.

[7] Najít takové místo není těžké, v praxi se strom haldy většinou mapuje do pole a nový prvek se přidává na konec tohoto pole.

Kapitola 5
Architektura nástroje

Nástroj sestává ze tří téměř nezávislých komponent. Každá z nich běží jako samostatné vlákno.[1] V hlavním vlákně běží uživatelské prostředí a další dvě hostí vstupní a výstupní rutinu. Ty byly z hlavního vlákna vyčleněny kvůli tomu, že zasílají a přijímají data přes internet. Internetová komunikace je vždy časově náročná a mírně nepředvídatelná, navíc pravidla Wikipedie pro provoz automatizovaných nástrojů [4] vyžadují, aby mezi každými dvěma dotazy na server byla alespoň desetisekundová prodleva. Aby komunikace s Wikipedií neblokovala práci uživateli, provádí se nezávisle na uživatelském vlákně a předává si s ním data přes globální fronty.[2] Jelikož do těchto front přistupují různá vlákna asynchronně, musí být přístup k nim chráněn mutexem tak, aby s frontou mohlo v jeden okamžik manipulovat jen jedno vlákno.

Stránky Wikipedie jsou v programu reprezentované datovým typem s těmito prvky:

- Název článku – jednoznačně identifikuje stránku
- Časové razítko – k zamezení editačních konfliktů
- Textový obsah článku
- Pole atributů textu – vymezují v obsahu zóny, ve kterých se nemají hledat chyby
- Halda chyb nalezených programem
- Fronta chyb schválených uživatelem

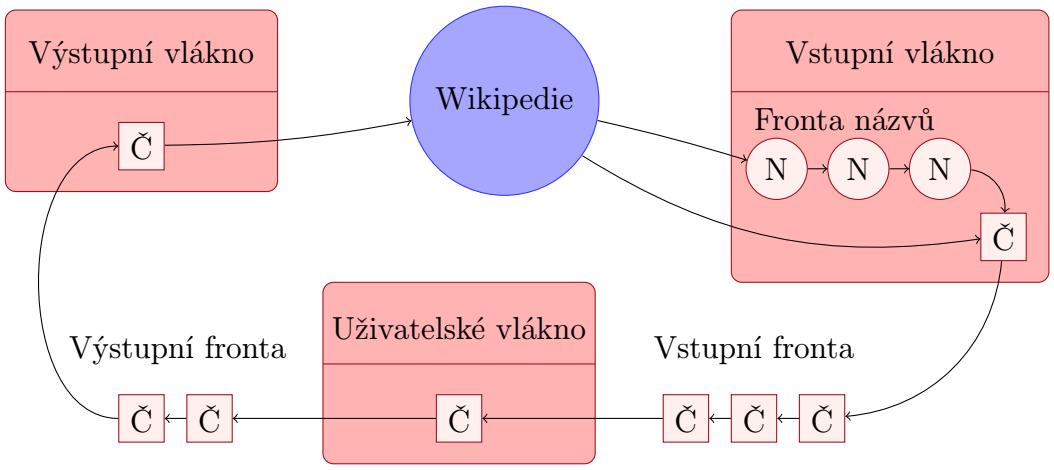
Každá taková struktura vzniká ve vstupní rutině při stažení stránky. Prochází postupně procesem automatického hledání chyb, jejich schválení uživatelem a zaslání opraveného textu Wikipedii ve výstupní rutině, kde svou existenci končí (obrázek 1).

5.1 Náhodné články

Aby mohl program nějaký článek stáhnout, musí znát jeho přesný název. MediaWiki poskytuje způsob, jak získat seznam náhodných, ale neopakujících se názvů článků[3]. Na jeden požadavek zašle MediaWiki seznam čítající jednotky až desítky názvů.

Funkce zajišťující stažení článku potřebuje vždy právě jeden název. Proto je seznam náhodných názvů rozparsován a jednotlivé názvy vloženy do fronty, ze které mohou být postupně odebírány. K zaslání požadavku na nový seznam a doplnění fronty ale dochází až těsně před jejím vyčerpáním – mít ve frontě názvy „do zásoby“ nemá význam, neboť stahování obsahu článku musí počkat na stažení seznamu v obou případech.

[1] Vlákno – viz kapitolu 4.1 na straně 11.
[2] Fronta – viz kapitolu 4.6 na straně 13.
[3] Seznam náhodných článků – viz kapitolu 3.3.2 na straně 9.



Obrázek 1. Schéma průchodu článků (Č) programem. Fronta názvů (N) je popsána v kapitole 5.1.

Fronta je lokální – je na ni přistupováno z jediného vlákna, a není proto třeba ji zabezpečovat synchronizačními primitivy. Dokonce je přístupná jen funkci stahující článek a jejím potomkům. Proces získání náhodného názvu má tedy formu nezávislého uzavřeného modulu.

5.2 Stažení článku

Stahování článků probíhá ve vlákně vstupní rutiny. Články se stahují po jednom a spolu s vlastním obsahem se získává i jejich časové razítko. To bude později sloužit ke zjištění, zdali během průchodu článku programem neprovedl jiný uživatel na stejném článku vlastní editaci, čímž by došlo k editačnímu konfliktu – změny provedené jedním uživatelem by se přepsaly změnami druhého uživatele.

Po úspěšném stažení se v datové struktuře reprezentující článek zinicializují i všechny prvky, které zatím nejsou naplněny – pole atributů textu a fronta a halda typografických chyb.

5.3 Zablokování zakázaných úseků

Zdrojový wikikód článku může obsahovat rozličné typy speciálních objektů.[4] Uvnitř mnoha z nich není žádoucí typografické chyby opravovat.

Mezi takové patří jednoznačně cíle odkazů, šablony, citace zdrojových kódů, předformátovaný text, vlastní speciální konstrukce jazyka wiki markup, matematické prostředí a mnoho dalších. V nich může mít znak spojovníku speciální význam a jeho změna by mohla mít fatální následky – může dojít k rozbití hypertextového odkazu nebo dokonce k porušení syntaxe kódu.

Naopak názvy sekcí, seznamy či úseky textu psané jiným než standardním řezem či velikostí písma pro účely korektury zajímavé jsou. Jejich odlišný vzhled není důvodem k tolerování typografických chyb.

[4] Wiki markup – viz kapitolu 3.1 na straně 7.

Většina objektů má ve wiki markup pevně danou strukturu – začínají a končí specifickou skupinou znaků nebo jsou ohraničeny XML tagy. K nalezení objektu, uvnitř kterého si nepřejeme provádět úpravy, tak stačí najít ve zdrojovém kódu uvozující skupinu znaků a k ní příslušnou skupinu ukončujících znaků. V článku (v poli atributů textu) pak můžeme označit celý úsek mezi touto dvojicí příslušnou značkou.

Je třeba dát si pozor na to, že objekty do sebe mohou být vnořeny a uvnitř jednoho objektu může nějaký jiný ztratit platnost. Například znak <, normálně uvozující XML tag, získává v matematickém prostředí význam operátoru *menší než*. Jednotlivé objekty je tak třeba hledat jednotlivě a v hierarchickém pořadí.

5.4 Hledání chyb

Každou typografickou chybu v článku reprezentuje datový typ obsahující tři prvky:

- Poloha začátku chybného úseku
- Délka tohoto úseku
- Náhrada, která se má vložit místo chybného úseku

První generace těchto struktur – chyb nalezených programem – vzniká na základě prohledání regulárními výrazy.[5] Každý výraz najde polohu všech výskytů jednoho druhu typografických chyb.

Nástroj si klade za cíl opravit zejména typografické chyby[6] tohoto druhu:

- Letopočty a rozsahy – vyskytne-li se spojovník, ať už s mezerami kolem něj nebo bez nich, mezi dvěma skupinami číslic, jde pravděpodobně o omyl a má být nahrazen pomlčkou bez mezer. Taktéž pomlčka oddělená od dvou skupin číslic má být nahrazena pomlčkou těsně přiléhající k číslům.
- Běžné interpunkční pomlčky – spojovník v běžném textu nikdy nemůže být oddělen mezerami od okolních slov. Dojde-li k tomu, pravděpodobně došlo k záměně a spojovník má být nahrazen pomlčkou.
- Záporná čísla – předchází-li spojovník číslici a z levé strany je oddělen mezerou, zřejmě má jít o vyjádření záporného čísla a spojovník má být nahrazen znakem minus.

Regulární výrazy pokrývající tyto případy jsou poměrně jednoduché. Jejich přesnou podobou včetně implementovaných vylepšení se zabývá kapitola 6.2.

Z každého výskytu chyby program vygeneruje výše popsanou strukturu a zařadí ji do haldy nalezených chyb. Díky vlastnostem haldy[7] se nalezené chyby setřídí podle polohy v textu, což bude později velmi výhodné jak z hlediska kontroly člověkem, tak z hlediska efektivity zpracování počítačem.

Po nalezení všech chyb vstupní vlákno vkládá článek do vstupní fronty, čímž jej předává uživatelskému vláknu. Vstupní vlákno pokračuje se zpracováním dalšího článku – jeho stažením, příp. vygenerováním náhodných názvů článků.

[5] Regulární výrazy – viz kapitolu 4.5 na straně 12.
[6] Správná typografie – viz kapitolu 2 na straně 5.
[7] Halda – viz kapitolu 4.7 na straně 13.

5.5 Schválení uživatelem

Jediným úkolem uživatelského vlákna je poskytovat uživateli rozhraní pro ovládání programu a schvalování oprav automaticky nalezených chyb.

Vlákno vyzvedne článek ze vstupní fronty. Pro každou z chyb nalezených v článku zobrazí uživateli její navrhovanou korekci spolu s okolím chyby, aby ji mohl uživatel posuzovat na základě širšího kontextu.

Jelikož některé terminály[8] nemají graficky rozlišenou pomlčku od spojovníku, v uživatelském rozhraní se pro pomlčky a minusy používá zápis podobný zápisu v programu TEX, tedy zdvojený spojovník `--` pro krátkou pomlčku a spojovník mezi dolary `$-$` pro minus. Příslušné znaky jsou nahrazeny až těsně před zobrazením uživateli a ihned po dokončení editace jsou zase převedeny zpět – interně program s těmito znaky pracuje v kódování UTF-8.

Zdá-li se mu navrhovaná oprava přijatelná, schválí ji stisknutím klávesy `Enter`. Je-li programem navržená oprava chybná, může uživatel text opravit za použití většiny možností editace nabízených knihovnou readline[9].

```
* 2007--2010: [[Univerzita Jana Amose Komenského Praha]], bakalářský obor ''Sociální a
* 2010--2012: [[Univerzita Jana Amose Komenského Praha]], navazující magisterský obor

=== Keylogger ===
>> errors found: 2

'''PS2 keylogger''' -- tento keylogger už v současnosti nepatří mezi příliš používané,
'''USB keylogger''' -- patří mezi nejpoužívanější keyloggery.

=== Tenis ===
>> errors found: 1

* [[tartan]] -- tvrdý povrch

=== Nigerokonžské jazyky ===
>> errors found: 8

* [[Atlantské jazyky]] (např. fulbština -- [[Burkina Faso]], [[Gambie]], [[Mali]],
Faso]], [[Gambie]], [[Mali]], [[Guinea-Bissau]], [[Senegal]]; wolofština -- [[Gambie]],
[[Jazyky benue-kongo]] (nejrozšířenější skupina jazyků -- patří sem [[bantuské jazyky]],
** [[Kvaské jazyky]] (neboli jazyk kwa -- jorubština, ibština, akanština, eweština, ve
```

Obrázek 2. Uživatelské rozhraní za běhu. Aktuálně se zpracovává čtvrtá chyba z osmi v článku „Nigerokonžské jazyky“, na posledním řádku.

Uživatel má možnost danou chybu ignorovat a neopravovat. Využije ji, když si například není jistý správným pravopisem – pak je lepší nechat text nezměněný. Učiní tak současným stiskem kláves `Ctrl-I`. Klávesovou zkratkou `Ctrl-P` pak ignoruje celý článek a neprovede v něm žádné změny – ani ty, které již byly schváleny.

V rámci aktuálně zpracovávaného článku se lze vracet k již schváleným chybám klávesami `Ctrl-U`. Tento příkaz přesune uživatele zpět na předchozí chybu. Historií lze listovat

[8] *Terminál* je zařízení skládající se z klávesnice a obrazovky schopné zobrazit pouze prostý text. V moderních operačních systémech se používají jejich emulátory jakožto jednoduchý, avšak robustní způsob zadávání příkazů počítači.

[9] Readline – viz kapitolu 4.2 na straně 12.

až k první chybě v článku. Zpět na místo, kde přestal, se uživatel dostane opakovaným stiskem klávesy `Enter`, přičemž chybám, které již jednou upravil, zůstává tato upravená podoba.

Program se ukončí klávesovou zkratkou `Ctrl-X`. Před ukončením počká na dokončení operací zahrnujících komunikaci s MediaWiki a uloží články ve vstupní a výstupní frontě, aby nebyla ztracena nedokončená práce.

5.6 Aktualizace chyby

Jelikož uživatel mohl chybu opravit jinak, než jak mu navrhoval program, a změnit tím datový typ, jenž chybu reprezentuje, je nutné jej na základě uživatelova rozhodnutí upravit. Porovná se proto úsek textu zobrazený uživateli ve své původní podobě před opravou a tentýž úsek textu, avšak tak, jak jej schválil uživatel. Najde se první a poslední znak, v němž se texty liší, a nová struktura chyby se vytvoří podle této rozdílné části.

Původní struktura vyjmutá z haldy nalezených chyb se může uvolnit a do fronty chyb schválených se vloží tato nově vytvořená. (Pokud uživatel navrhovanou opravu bez dalších zásahů rovnou schválí, obě struktury budou identické.)

Až uživatel přes rozhraní odsouhlasí všechny opravy – přesune je do fronty schválených chyb – není dále třeba článkem uživatele zatěžovat. Uživatelské vlákno jej vloží do výstupní fronty a začne zpracovávat další článek.

5.7 Oprava textu

Úlohou výstupního vlákna je vyzvedávat články z výstupní fronty, provést na nich opravy schválené uživatelem a odeslat opravené texty zpět Wikipedii.

Jelikož jsou v každém článku chyby a jejich opravy seřazeny podle místa výskytu, celý text lze opravit v jediném průchodu. Do výsledného řetězce se střídavě kopírují úseky původního textu článku a náhrady za chybné úseky. Na závěr tato nová, typograficky korektnější verze textu nahradí původní chybnou.

Po tomto kroku (nebo rovnou v jeho průběhu) lze z paměti uvolnit i všechny struktury představující chyby. Nejsou už třeba – opravenou informaci nyní nese samotný text článku.

5.8 Odeslání změn

Posledním krokem je odeslání opraveného textu Wikipedii.[10] Nenastane-li systémová chyba – ať už na straně MediaWiki nebo v editačním nástroji – může tento úkon selhat jedině v případě editačního konfliktu. Ten lze očekávat zejména v případech, kdy byla činnost programu pozastavena po stažení článku, ale ještě před jeho odesláním, a nebyl

[10] Editace – viz kapitolu 3.3.4 na straně 10.

delší dobu používán. V takovém případě je uživatel o chybě informován a může na situaci adekvátně reagovat.

Ze statistických dat,[11] a sice poměru průměrného stáří článku a průměrného počtu editací na článek, vyplývá, že průměrná prodleva mezi dvěma editacemi je 2,7 měsíce. Tento údaj je jistě značně nepřesný, ale praxe ukazuje, že se od skutečnosti podstatně neliší. Šance, že někdo provede změnu článku dříve než člověk používající zde popisovaný nástroj, tedy není vysoká.

Ve většině případů proto editace proběhne bez problémů a během několika sekund se změny projeví ve vlastním obsahu Wikipedie. Taktéž přibude záznam o editaci v historii článku a na účtu uživatele, pod nímž je program přihlášen.

Tím končí jednak činnost výstupního vlákna programu na tomto článku, jednak i existence samotné struktury článku. Výstupní vlákno může pokračovat další editací.

[11] `http://stats.wikimedia.org/CS/TablesWikipediaCS.htm` [cit. 15. 11. 2015]

Kapitola 6
Zhodnocení výsledků

Za dobu své existence prohledal nástroj přes 17 000 článků – zhruba 5 % celého objemu české Wikipedie. To je dostatečně velké množství ke statistickému zpracování. Na textech byly vypozorovány některé společné znaky a podle reálných situací byly značně zefektivněny i některé části programu.

Program si vede záznam o všech článcích, které prohledal, a záznam o okolí všech schválených typografických chyb při nálezu, při předložení opravené verze uživateli a po jejím schválení uživatelem. Všechna statistická data uvedená v této kapitole jsou extrahována z těchto záznamů.

6.1 Typy článků

Mezi zpracovávanými články nelze přehlédnout některá nebývale četná témata, specifická druhem chyb, jenž v nich převládá.

- Rozcestníky – speciální články, jejichž účelem je odkázat čtenáře z nejednoznačného hesla na konkrétnější. Jejich obsahem často nebývá nic jiného než dvojice *heslo* – *vysvětlení* oddělené pomlčkou. Velmi snadno se opravují.
- Osobnosti – Každý článek věnující se nějaké historické osobnosti začíná vymezením dat (a často i místa) narození a úmrtí. V samotném těle článku převažují chyby v rozsazích letopočtů.
- Sportovní tabulky – záznamy výsledků soutěží nebo sportovních výkonů týmu či sportovce. Vyskytují se v nich chybně zapsaná záporná čísla, ale i další číselné údaje. Bohužel jich často obsahují desítky, spíše stovky, takže se jejich oprava z časového hlediska nevyplatí. V tabulkách jazyka wiki markup se navíc špatně orientuje, vidí-li uživatel pouze jeden řádek zdrojového kódu.
- Hudební skupiny – často obsahují výčet svých členů ve formátu *jméno – nástroj*, někdy i diskografii se stopáží. Stejně jako rozcestníky se snadno opravují, neboť uživatel nemusí při kontrole číst souvislý text.

Lze najít i další významné kategorie, ty jsou však řidčeji zastoupeny a/nebo chyby, které se v nich vyskytují, nesdílí do takové míry společné rysy. Jsou jimi například články přírodovědné (rozměry, počty), články o obcích (letopočty, k nim přiřazené události) nebo texty o silnicích a železničních tratích (body propojené danou tratí).

Chyby ve všech výše zmíněných typech článků samozřejmě nejsou vymezeny striktně, články mohou obsahovat i pestrou plejádu dalších druhů typografických omylů, zejména chybně zapsaných mezivětných pomlček v běžném textu.

6.2 Vylepšení regulárních výrazů

Nejvýraznějších změn se program dočkal v oblasti vyhledávacích regulárních výrazů.[1] Při prvním spuštění disponoval třemi výrazy pro úpravu pomlček mezi čísly, dvěma pro mezislovní pomlčky a jedním pro minus u záporných čísel.

Všechny znaky - v následujících regulárních výrazech jsou znakem spojovníku a úsek odpovídající závorce je při opravě nahrazen samotným znakem pomlčky, není-li uvedeno jinak. Závorky v regulárních výrazech ohraničují chybný úsek, jehož polohu si program zapamatuje, pokud celý výraz přijme část textu článku. Právě ten úsek textu totiž bude později nahrazen jiným tak, aby byla typografie opravena.

Toto je úplný výčet regulárních výrazů použitých při prvním ostrém spuštění nástroje:

```
[[:digit:]]+(␣-␣)[[:digit:]]+
[[:digit:]]+(␣-␣)[[:digit:]]+  (znak - je pomlčka)
[^-[:digit:]][[:digit:]]+(-)[[:digit:]]+[^-[:digit:]]

[[:upper:]][[:alpha:]]+(␣-␣)[[:upper:]][[:alpha:]]+
[[:graph:]]+␣(-)␣[[:graph:]]+

␣(-)[[:digit:]]+  (nahradit znakem minus)
```

Třetí výraz má komplikovanější podobu kvůli tomu, aby měnil spojovník na pomlčku u rozsahu dvou čísel, ale už ne, pokud je spojovníkem spojeno více skupin čísel, neboť pak jde zřejmě o jakési kódové označení (například ISBN).

Čtvrtý výraz má za účel spojovat vlastní jména (slova začínající velkým písmenem) do konstrukcí typu *spoj Hodonín–Břeclav*. Praxe ale ukázala, že mezi vlastními jmény převažují víceslovné názvy (*Hodonín – Uherské Hradiště*), ve kterých je vhodnější použít variantu s mezerami. Navíc se vlastní jména příliš často vyskytují v okolí pomlčky v pozici přístavku (*hlavní město Dánska – Kodaň*), kde jsou mezery nutné. Proto už tento regulární výraz není v současnosti používán.

Poměrně rychle vyšlo najevo, že velké množství letopočtů je zároveň odkazem. Některý nebo oba z číselných údajů jsou tedy obklopeny dvojicemi hranatých závorek. Proto v prvních třech výrazech přibyly sekvence povolující číselné odkazy.[2] Taktéž přibyla varianta prvního regulárního výrazu pro rozsah řadových číslovek (*15.–16. století*), ve výčtu na druhé pozici:

```
[[:digit:]]+\]{2}?(␣-␣)\[{2}?[[:digit:]]+
[[:digit:]]+\.\]{2}?(␣-␣)\[{2}?[[:digit:]]+\.
[[:digit:]]+\]{2}?(␣-␣)\[{2}?[[:digit:]]+  (znak - je pomlčka)
[^-[:digit:]][[:digit:]]+\.?\]{2}?(-)\[{2}?[[:digit:]]+[^-[:digit:]]
```

Výše vypsaná čtveřice má jednu vadu – články o historických osobnostech uvádějí v hlavičce dobu života s rokem i datem narození (*3. března 1952 – 11. května 2001*).

[1] Regulární výrazy – viz kapitolu 4.5 na straně 12.

[2] V zápisu regulárního výrazu mají znaky [,] a . speciální význam, proto jim předchází zpětné lomítko. Ve zdrojovém kódu programu se s výrazem pracuje jako s řetězcem a je nutné všechna zpětná lomítka zdvojit, neboť v kontextu řetězců jazyka C mají samy zvláštní význam.

Regulární výrazy zaměřené na čísla vyhodnotí rok narození a datum úmrtí jako rozsah *1952–11*. Z toho důvodu byl přidán zvláštní výraz pro tato data beroucí v úvahu typický počet číslic roku a dne v měsíci. Zároveň bylo nutné přidat i výraz odpovídající správně napsanému datu, který nenahrazuje nic, pouze označí dané místo jako opravené, čímž zabrání dalším výrazům v jeho mylném opravování.

```
[[:digit:]]{4}\]{2}?( - )\[{2}?[[:digit:]]{1,2}\.  (nahradit pomlčkou ␣-␣)
[[:digit:]]{4}\]{2}?( - )\[{2}?[[:digit:]]{1,2}\.  (znak - je pomlčka,
                                                    nahradit pomlčkou ␣-␣)
```

Regulární výraz k nalezení chybného spojovníku mezi slovy prošel jen minimální změnou, a sice přijímá i čárku, která se někdy píše za druhou pomlčkou oddělující vloženou větu (*Přiznal se – ač nerad –, že udělal chybu.*).

```
[[:graph:]]+ (-),? [[:graph:]]+
```

Výraz pro záporná čísla byl vylepšen tak, aby přijímal i záporná čísla uvedená v tabulce nebo závorce – před číslem se může nacházet nejen mezera, ale i interpunkční znak (*interpunkční* ve smyslu klasifikace znaků podle jazyka C). V hlavičce bodových sportovních tabulek bývá často sloupec označený symboly +/−, který obsahuje kladný či záporný bodový zisk. Aby hlavička zůstala konzistentní s daty (a byla zapsaná korektně), byl pro tento případ přidán další regulární výraz.

```
[␣[:punct:]](-)[[:digit:]]+  (nahradit znakem minus)
\+/(-)  (nahradit znakem minus)
```

6.3 Povaha chyb

K datu 13. března 2016 bylo pomocí nástroje prohledáno 17 735 článků. Z toho ve 6 129 článcích byly nalezeny typografické chyby – to je každý třetí. To poukazuje na nepoučenost o typografické kultuře i u jinak vzdělaných lidí tvořících obsah Wikipedie.

Z nalezených chybných článků bylo 374 ignorováno – obyčejně šlo o články s desítkami chyb, které byly přeskočeny z důvodu neefektivity takových oprav. Proto, ačkoli bylo ignorováno jen zhruba 6 % článků, byla tím vynechána téměř třetina všech nalezených chyb. Převážně šlo o sportovní tabulky a dlouhé seznamy, u kterých se stejně nepředpokládá souvislá četba.

Průměrný článek obsahuje 2,3 chyby. Bereme-li ohled jen na články s alespoň jednou chybou, průměrně každý z nich obsahuje 6,6 chyb. A konečně započítáme-li pouze články, které nebyly ignorovány, průměrně budou v každém 4,9 chyby.

V tabulce 1 je uvedeno, kolikrát byly nalezeny jednotlivé druhy chyb. Každý typ odpovídá jednomu regulárnímu výrazu, ale pro lepší čitelnost jsou tyto chyby zastoupeny příklady chybného tvaru a jejich korektního protějšku. Spolu s absolutním počtem nálezů je uvedeno i procento chyb přijatých uživatelem v nezměněné podobě. Tento údaj vypovídá o vhodnosti a kvalitě implementace daného regulárního výrazu. Je však mírně zkreslen opravami konanými nástrojem, které ale nemají přímou souvislost s nalezenou chybou – například do očí bijících pravopisných chyb v blízkosti opravy.

Chyba	Oprava	Výskyt	Přijato
září 2014 - 27. ledna	září 2014 – 27. ledna	1980	94,5 %
John Lennon - zpěvák	John Lennon – zpěvák	16099	97,9 %
plodí 10 - 15 vajec	plodí 10–15 vajec	749	90,7 %
13. - 14. století	13.–14. století	168	97,0 %
v letech 1926-1931	v letech 1926–1931	5537	98,3 %
délka 8 – 12 mm	délka 8–12 mm	423	99,1 %
teplota -5 °C	teplota −5 °C	1140	98,4 %
+/-	+/−	164	61,6 %

Tabulka 1. Srovnání počtu nálezů jednotlivých druhů chyb, zastoupených příklady chybného a správného tvaru. Sloupec *Přijato* udává, kolikrát byla navržená oprava beze změny schválena.

Počtu nalezených chyb dominují spojovníky na místě pomlček mezi slovy. To je dáno množstvím vysvětlovacích konstrukcí a různých seznamů, kterých je na Wikipedii hojně. Dle očekávání jsou na druhém místě rozsahy čísel, zejména letopočtů. Malý podíl zaujímají minusy a rozsahy plných dat.

Při podrobnějším rozlišení chyb v rozsazích čísel výrazně převládá prostá záměna pomlčky za spojovník přiléhající těsně k oběma číslům. Tento druh chyby se vyskytnul 5537×, ostatní tři dohromady pouze 1340×. Lze usuzovat, že přispěvatelé cítí spjatost mezi dvěma číselnými údaji ve významu *od–do* a neoddělují proto spojovník mezerami. Převážně si tedy jen nejsou vědomi rozdílu mezi grafickou podobou pomlčky a spojovníku. To dokládá i fakt, že když už uživatel napíše pomlčku, málokdy ji chybně použije (třetí řádek odspodu v tabulce 1).

Co se úspěšnosti přijetí jednotlivých regulárních výrazů týče, vyčnívají dva. Výraz opravující znaky +/− má nízké procento přijetí proto, že společně s náhradou spojovníku za minus bývá nutné rozšířit sloupec dané tabulky, aby se do něj širší znak minus vešel. Sloupce mívají často šířku pevně nastavenou na 20 px, ale symboly plus/minus potřebují alespoň 30 px. Druhým výrazem s nižší úspěšností přijetí je spojovník mezi čísly s mezerami po obou stranách. Příčinou jeho selhání v některých případech jsou pravděpodobně věty, v nichž je namísto interpunkční pomlčky nesprávně použit spojovník a pouze náhodou se před i po něm vyskytují číselné údaje. Pro zatížení chybami nelze objektivně srovnat úspěšnost přijetí ostatních regulárních výrazů, jelikož podle dat vykazují minimální rozdíly.

Poukázat lze na to, že nástroj založený na tak jednoduchém principu, který nebere zřetel na sémantickou stránku jazyka, nemůže pracovat na podobných úkolech samostatně – jeho chybovost vždy bude nezanedbatelně vysoká. Zároveň je ale jeho úspěšnost dostatečná k provozu poloautomatického nástroje, který výrazně usnadní práci přispěvatelům – přes 97 % oprav chyb bylo schváleno v podobě navržené programem a k jejich opravě tedy nebylo nutné udělat nic víc, nežli stisknout klávesu `Enter`.

6.4 Přínos pro Wikipedii

Za dobu své existence nástroj prohledal 17 735 článků, tedy 5,1 % české Wikipedie. Ačkoli se tento počet vzhledem k obrovskému množství článků na Wikipedii zdá být nízký, svou

stopu na zkvalitnění encyklopedie zanechal – opravil celkem 29 825 typografických chyb a bude v opravách pokračovat i v budoucnu. Kdyby takové opravy prováděl člověk ručne, strávil by nad stejným množstvím neporovnatelně více času. S nástrojem trvá oprava jedné chyby v průměru čtyři sekundy, a to včetně nalezení chyby v článku, které je nejpomalejší fází při editaci obvyklým způsobem.

Kvantitativní průzkum prováděný společně s opravami poukázal na to, jak málo je současná populace vzdělaná v oblasti typografie. I tomu má nástroj napomoci tím, že čtenáři budou častěji potkávat správně použitá interpunkční znaménka, a pasivně se tak typografickým zásadám naučí. Jeden člověk i s pomocným nástrojem takovou práci sám nezastane. Může ale svým jednáním vzbudit iniciativu, která podnítí další uživatele k drobným typografickým opravám.

Dvanáct aktivních uživatelů Wikipedie si všimlo práce provedené nástrojem a zaslali mu *poděkování* – formu odměny za editaci konkrétního článku. Prvního ledna byla soustavná práce na typografii Wikipedie oceněna *Wikikytičkou* – uživatelským vyznamenáním II. třídy udělovaným za menší, avšak užitečnou nebo nečekaně vydařenou práci. I drobné úpravy Wikipedie se cení.[3]

Zdrojový kód programu je zveřejněn pod svobodnou licencí *2-clause BSD* na serveru GitHub na adrese `github.com/m-sche/tiret` společně se stručným návodem ke kompilaci a ovládání. Nejenže tak nástroj může kdokoli použít a čistit s ním českou Wikipedii od chybných spojovníků, ale může program i upravit a použít ho například pro jiný druh chyb. Míře modifikace se ale meze nekladou, zdrojový kód může být použit způsobem téměř libovolným.

[3] `cs.wikipedia.org/wiki/Wikipedie:Editujte_s_odvahou` [cit. 15. 12. 2015]

Kapitola 7
Závěr

Nerozlišování interpunkčních znamének je prohřeškem přetrvávajícím z dob počátků počítačové typografie, kdy z technických důvodů nebyly jednotlivé znaky graficky odlišeny. S nástupem jemných obrazovek a velkokapacitních paměťových médií už ale není pro ústupky v typografii omluvy. Dnes, kdy se dokumenty sází v počítači a mnohdy se ani netisknou na papír, tak typografie znova nabývá na významu.

Internetová encyklopedie Wikipedie je výborným vzorkem současné internetové literární tvorby a projevují se na ní všechny současné trendy nejen jazykového, ale i typografického barbarství. Zároveň je médiem velmi populárním a svým obsahem ovlivňuje širokou veřejnost.

Tato práce ověřila, že velká část editorů Wikipedie zcela neovládá zásady sazby textu v počítači a chybně používá interpunkční znaménka. Ta nejsou rozlišována samoúčelně – podle konkrétní formy pomáhají čtenáři sjednotit nebo rozdělit části výpovědi. Díky studiu chyb, které editoři nejčastěji dělají, se povedlo vytvořit nástroj schopný tyto chyby vyhledávat a s vysokou úspěšností navrhnout jejich opravu. Díky němu může být Wikipedie opravována mnohonásobně rychleji a s vynaložením podstatně menšího úsilí.

Pomocí nástroje bylo prohledáno přes 17 000 článků a opraveno více než 29 000 chyb. V poměru k celkové velikosti Wikipedie to sice je jen malá část, jde ale o zdařilou demonstraci toho, že i bez sémantické analýzy textu lze velmi přesně odhalovat chybně napsané pomlčky a minusy. Tento poznatek lze jistě aplikovat i na další interpunkci a rozšířit o ni funkčnost programu. Zároveň byl program spolu s krátkým návodem k použití zpřístupněn na adrese `github.com/m-sche/tiret`, aby mohli zájemci sami pomocí nástroje Wikipedii editovat.

Práce dále funguje jako ucelený, byť stručný návod k práci s MediaWiki API, užitečný zejména ve spojení s oficiálním návodem, který je podrobnější, zato roztříštěný. Použitá trojvláknová architektura se při používání nástroje osvědčila a taktéž může být inspirací při vytváření programů komunikujících s internetem.

Tato práce spolu s aktivitami na Wikipedii si tak nakonec kladou za cíl i oslovit alespoň několik jedinců a přimět je nebýt lhostejnými k počítačové sazbě. A to nejen na Wikipedii, ale i ve veškerých dokumentech, jež budou na počítači psát.

Reference

[1] ČSN 01 6910. *Úprava dokumentů zpracovaných textovými procesory*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

[2] Ústav pro jazyk český AV ČR. *Internetová jazyková příručka*. `www.ujc.cas.cz`.

[3] *Wikipedia help page: Wiki markup*. `en.wikipedia.org/wiki/Help:Wiki_markup`.

[4] *Wikipedia: Bot policy*. `en.wikipedia.org/wiki/Wikipedia:Bot_policy`.