Masarykova univerzita

Filozofická fakulta

Ústav českého jazyka

Český jazyk se specializací počítačová lingvistika

Markéta Masopustová

Analýza srozumitelnosti textu

Diplomová práce

Vedoucí práce: RNDr. Vojtěch Kovář, Ph.D.

2018

Prohlašuji, že jsem diplomovou práci vypracovala

samostatně s využitím uvedených pramenů a literatury.

……

Ráda bych poděkovala RNDr. Vojtěchovi Kovářovi, Ph.D., za vedení práce a cenné rady. Mgr. Haně Žižkové děkuji za nápady a konzultace a Matyášovi Antonovi za korekturu práce. Dále děkuji všem, kteří vydrželi mé nářky během posledního tři čtvrtě roku práce a vyjádřili mi podporu. Děkuji také svému příteli, bez kterého bych v posledním měsíci práce pravděpodobně umřela hlady. A nesmím zapomenout na poděkování rodině, která mi celé studium na vysoké škole umožnila.

Shrnutí

Každý uživatel jazyka alespoň občas pochybuje o pravopisu. Málokdy se ovšem stane, že pisatelé pochybují o stylistických prostředcích a srozumitelnosti textu. Cílem diplomové práce je navrhnout lingvisticky orientovaná pravidla, která popíší stylistické prohřešky ve větách. Teoretická část rozebírá teorii stylistiky, současné nástroje pro automatickou kontrolu textu a analýzu lingvistických problémů. Na ni navazuje praktická část, která se věnuje popisu pravidel pro syntaktický analyzátor SET a vyhodnocení pravidel na korpusu, který byl pro potřeby práce vytvořen.

Klíčová slova

Stylistika, syntaktická analýza, srozumitelnost, SET, grammar checker, stylistics, syntactic analysis, intelligibility.

Použité zkratky

CZPJ FI MU	Centrum zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity
KČG	Kontrola české gramatiky
NESČ	Nový encyklopedický slovník češtiny
PMČ	Příruční mluvnice češtiny
SSČ	Slovník spisovné češtiny
SSJČ	Slovník spisovného jazyka českého
ÚFAL MFF UK	Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy
*	Hvězdička označuje negramatické věty nebo tvary slov.

OBSAH

ÚVOD

Pomyslnou nejvyšší metou při psaní textů je stylistická správnost. Gramatické či pravopisné chyby totiž většinou nebrání porozumění textu, kdežto špatně postavené věty nebo nekorektní vazby sloves mohou vést k nepochopení a ke ztrátě předávané informace. V rámci stylistické korektury se kontroluje také správný slovosled a časté opakování slov. V diplomové práci se zaměříme na vybrané chyby, které se objevují v psaných textech, a to jak v tištěných publikacích, tak na internetu. Cílem práce je vytvořit způsob, jak tyto chyby detekovat.

V první kapitole nastíníme problematiku stylistiky jako vědy. Nejprve definujeme, co se pod termínem styl, stylistika a stylizace skrývá, poté představíme stylotvorné faktory a podrobně rozebereme jednotlivé funkční styly. V závěru kapitoly popíšeme otázku prvků knižních, neutrálních a hovorových v psaném a mluveném jazyce a vztah stylistiky a srozumitelnosti textu.

Současné situaci na poli automatických oprav pravopisu či gramatiky a stylistiky se budeme věnovat v druhé kapitole. Zmíníme různé typy korektorů a podrobně se zaměříme na Kontrolu české gramatiky. Nevynecháme ani další nástroje, které vznikly ať už na akademické půdě, nebo v komerční sféře.

Třetí kapitola rozebírá lingvistické aspekty, kterým jsme se v rámci práce věnovali. Popíšeme problém slovosledu v češtině, opakování slov, pleonasmů, hovorových slov a další prohřešky.

Čtvrtá kapitola se soustředí na popis nástrojů, které jsme využili. Uvádíme zde tokenizaci, morfologickou analýzu a věnujeme se také syntaktické analýze. V této části vysvětlíme, jakým způsobem se tvoří pravidla pro syntaktický analyzátor SET.

V páté kapitole přichází na řadu praktická část a její popis. Pro každou kategorii, kterou jsme popsali ve třetí kapitole, popíšeme vytvořená pravidla, včetně příkladů a případných obtíží, na něž jsme v rámci tvorby přišli.

Šestá a poslední kapitola rozebírá korpus, který jsme pro testování pravidel vytvořili. Také jsme popsali, jakým způsobem se měří úspěšnost automatických nástrojů pro analýzu textu. Stěžejním bodem závěrečné kapitoly je vyhodnocení pravidel na vytvořeném korpusu včetně popisu problémů, které jsme během testování objevili.

1 ÚVOD DO STYLISTIKY

Jazykověda se skládá z několika disciplín, které popisují jazyk od nejmenších jednotek až po největší celky. Pomyslným nejvyšším stupněm je stylistika – tedy přiřazení textu k určitému stylu a volba správných morfologických, syntaktických a lexikálních prostředků.

V úvodní kapitole této práce přiblížíme, co znamená pojem styl a stylistika, popíšeme stylotvorné faktory a také jednotlivé funkční styly. V závěru nastíníme využití hovorových, neutrálních a knižních výrazů. Vysvětlíme také důležitost srozumitelnosti textu.

1.1 Definice

Na úvod bychom měli definovat, co se rozumí pod pojmem styl. Marie Čechová v *Současné stylistice* uvádí, že styl „v jazykovědě označuje určitý ráz verbálního komunikátu, zpravidla cílevědomě volený a uspořádaný tak, aby obsahem i formou vyhovoval komunikačnímu záměru autora“ (Čechová, 2008, str. 16). V *Novém encyklopedickém slovníku češtiny* (dále NESČ) můžeme najít definici od Michala Křístka, že jazykový styl „je v českém kontextu tradičně chápán jako záměrný výběr a organizace výrazových prostředků v textu, který se uplatňuje při jeho genezi“ (Křístek, 2017). V *Příruční mluvnici češtiny* (dále PMČ) Milan Jelínek definuje styl jako „výsledek výběru jazykových prostředků z množin prostředků konkurenčních,“ (Jelínek, 2012, str. 699) a zároveň upozorňuje, že styl je „záležitost řeči, nikoli jazyka“ (Jelínek, 2012).

Z výše uvedených definic je patrné, že stylistika není omezena hranicí slova, ale ani hranicí věty. Je to disciplína, která spojuje několik různých oborů a která využívá jejich poznatků. (Čechová, 2008, str. 20)

V naší práci se budeme zabývat stylistikou jazykovědnou, v některých publikacích označovanou tzv. lingvostylistikou. Čechová (2008) uvádí, že „lingvostylistika se věnuje deskripci stylu a studiu jeho účinku. (...) Jsou v ní analyzovány jednotlivé jazykové projevy z hlediska jejich kompozice a stylizace, jsou srovnávány projevy týchž i různých autorů po této stránce a zobecňovány stylové a stylizační vlastnosti týchž projevů i jazykových prostředků.“. Jelínek a Krčmová v NESČ uvádí, že „lingvostylistika se soustřeďuje na výrazovou složku jakýchkoli textů/komunikátů v celé její proměnlivosti, i když v novějším pojetí neodmítá ani přihlédnutí k tematice“. (Jelínek, 2017a)

Posledním pojmem, který v této části definujeme je stylizace. Jak uvádí Čechová, jedná se o „pojem užší (od stylistiky), týkající se stavby větné i nevětných (jmenných) konstrukcí, volby pojmenování aj. dílčích procesů při tvorbě komunikátu“. (Čechová, 2008, str. 21) V PMČ Jelínek označuje stylizaci jako stylotvorný proces. (Jelínek, 2012, str. 699)

Styly se v rámci stylistiky dělí do několika skupin na styly obecné, objektivní a subjektivní, které se ještě dále podrobněji dělí. Styly jsou vždy definovány různými faktory. Česká stylistika se již od V. Mathesia a B. Havránka zaměřuje nejvíce na funkci jazyka. „Funkcí ve stylistice se rozumí záměr, který sleduje autor, popř. účel, kterému projev slouží, cíl, který se projevem sleduje, event. zahrnuje i prostředek.“ (Čechová, 2008, str. 28)

1.2 Stylotvorné faktory

Každá komunikační situace si žádá dvě strany, autora a adresáta, a vždy je ovlivněna řadou faktorů, od způsobu komunikace po podmínky, ve kterých se komunikace odehrává. Kupříkladu jinak bude reagovat student na otázku „*Jak jdou zkoušky?*“ položenou rodiči u nedělního oběda a kamarádem v hospodě. „Všem okolnostem a vlivům, které usměrňují výběr výrazových prostředků a ovlivňují výsledný styl komunikátu, říkáme stylotvorné faktory neboli stylotvorní činitelé.“ (Čechová, 2008, str. 76) Ty se dále dělí na stylotvorné faktory objektivní a stylotvorné faktory subjektivní.

Mezi subjektivní faktory můžeme zařadit vše, co se týká subjektu, tedy autora komunikátu, např. znalosti o tématu, sociální prostředí, pohlaví či povahové vlastnosti.

Objektivní faktory naopak vystihují vše, co se netýká autora, ale komunikátu samotného, např. funkce komunikátu, forma (mluvenost × psanost) či soukromost × oficiálnost. Na základně jednotlivých funkcí se vydělují funkční styly.

1.3 Funkční styly

Funkční styly patří do skupiny stylů objektivních. Rozdělené jsou na základě zobecnění jednotlivých způsobů komunikace.

Pro následující popis funkčních stylů využijeme *Současnou stylistiku* od autorek Marie Čechové, Marie Krčmové a Evy Minářové (2008), která nabízí ucelený popis jednotlivých stylů.

1.3.1 Funkční styl prostěsdělovací

Jak již název stylu vypovídá, jedná se o často nepřipravené, spontánní jazykové projevy běžné denní komunikace. Mezi ně můžeme počítat například každodenní rozhovory s rodinou a přáteli, vzkazy na lednici či zamilované statusy na sociálních sítích.

Tento styl má dvě hlavní funkce, a to komunikační (sdělovací) a fatickou (kontaktovou). Stává se, že funkce fatická je důležitější: můžeme se setkat s tím, že rozhovor probíhá jen z důvodu osamělosti člověka, přičemž v tu chvíli nemá žádný sdělovací efekt. Jak jsme již zmínili, patří sem projevy spontánní, především mluvené, určené osobě, kterou mluvčí zná a která je velmi často v přímém kontaktu s příjemcem.

Základem prostěsdělovacího stylu je „běžně mluvený jazyk/běžná mluva, funkční podoba národního jazyka, která může být z hlediska strukturního naplněna kteroukoli varietou (spisovný jazyk, interdialekt, obecná čeština jako interdialekt, který se dostává do pozice nižšího standardu, dialekt)“ (Čechová, 2008, str. 201).

Z hlediska kompoziční stavby můžeme o textech tohoto stylu prohlásit, že jejich struktura je volná, mohou se v nich opakovat stejná oznámení (pleonasmy) a vyjádření nemusí být dokončeno. Syntaktická stavba je jednoduchá. Do sdělení často bývají zapojeny emoce, komunikát může být eliptický, dochází k míšení vazeb, a to především v mluveném projevu. Setkáváme se s hovorovými prostředky a frazémy, často na pokraji spisovnosti, a neologismy, které fungují jen v té dané situaci.

1.3.2 Funkční styl odborný

U odborného stylu je kladen důraz na přesnost, jasnost a úplnost. Jsou to předem připravené texty či projevy, se kterými se běžný uživatel jazyka nejčastěji setká ve škole či na univerzitě. V tomto typu textů je cílem předání informace, a proto je mu podřízena celková výstavba – od volby slov přes morfologii a syntax po kompozici. Můžeme sem zařadit studii, úvahu, esej, referát, kritiku, rešerši, poster či přednášku.

Funkce stylu je především odborněsdělná a základní formou je monolog autora. Texty odborného stylu bývají většinou situačně nezakotvené. „I když vznikají v jistém čase, a odrážejí tedy stav poznání i stav jazyka doby vzniku, směřují ve skutečnosti k nadčasovosti: autor formuluje poznatky tak, aby byly jednoznačně vnímatelné a pochopitelné i po letech.“ (Čechová, 2008, str. 211)

Z kompozičního hlediska je důležitá jasnost a jednoznačnost. Pro tyto texty je typická promyšlená kompozice. Pokud se na odborný text podíváme jako na celek, zaznamenáme nápadné vertikální členění – poznámky, citace, rejstříky, členění do odstavců, mezititulky a další. Z jazykových prostředků jsou vybírány výrazy neutrální (bez stylového a emočního zabarvení) a spisovné. Větná stavba je často složitá a s dlouhými souvětími. Je kladen důraz na aktuální členění, cílem odborného textu je objektivnost, proto se v něm réma umisťuje na konec výpovědi.

Jak jsme již zmínili, tento styl z morfologické stránky definuje především spisovnost. Čechová (2008) uvádí, že je v této oblasti význačné zastoupení genitivu obou čísel.

Pro odborný styl jsou typické termíny a odborné názvy, přičemž je snaha používat spíš termíny internacionální. Velké zastoupení mají také slova přejatá.

1.3.3 Funkční styl administrativní

S rozrůstajícím se množstvím byrokracie se ustálil administrativní styl. Patří sem texty úřední (úřední dopis, protokol, formulář) a také texty administrativně-právní (rozsudek, vyhláška). „Dřívější stylistiky nevyčleňovaly samostatný funkční styl administrativní; v rámci odborného stylu rozlišovaly oblast praktickou a teoretickou, přičemž k praktické oblasti přičleňovaly styl jednací.“ (Čechová, 2008, str. 231)

Funkcí má tento styl více, mezi nejdůležitější patří funkce sdělovací, regulativní a operativní. Vyjádření jsou strohá a věcná a náchylná na formálnost a dodržování norem (př. grafická úprava písemnosti i formát papíru). Na autora není brán zřetel, setkáváme se s prvkem tzv. právnické osoby, která může zastupovat autora textu, adresáta i oba.

Texty administrativního stylu mají tendenci být unifikované a textová výstavba je u velkého množství z nich podobná, ne-li stejná. Z hlediska syntaxe se projevuje úsilí o strohost a jednoznačnost, které vedou k ustáleným obratům, např. děkujeme za kladné vyřízení. Typické je časté využívání jmenných a pasivních konstrukcí.

Užívat by se měla neutrální slovní zásoba a spisovné sufixy (odchylky jsou zpravidla dány neznalostí kodifikačních pravidel). Frekventované jsou číslovky a kombinované výrazy číslicového a slovního typu. Administrativní styl má vlastní terminologii, ale využívá i výrazů z různých oborů, podle typu textu. Na první pohled jsou v textech patrné zkratky, zkratková slova a značky.

1.3.4 Funkční styl publicistický

Tvorba v rámci publicistického stylu je především v rukou novinářů. Jedná se o styl rozsáhlý a mnohotvárný, z útvarů do něj patří například zpráva, reportáž, glosa, recenze, polemika či interview. Můžeme sem zařadit také reklamu a propagandu.

Hlavní funkcí je funkce informativní a vedle ní také funkce persvazivní, působicí a ovlivňovací. Cílem je rychle a výstižně informovat čtenáře bez ohledu na věk a vzdělání, textům tohoto stylu by měli rozumět všichni bez rozdílu. I proto je důležitá sémantická jednoznačnost a maximální srozumitelnost. „Výrazové publicistické prostředky odrážejí politicko-ekonomickou situaci. (...) Texty nejsou charakterizovány nadneseností, obraty a vazbami výrazně knižními, běžně se v nich objevují prvky jiných stylů. Na jedné straně přibývá odborného a profesního vyjadřování, na druhé straně jsou hovorovější, zvláště ve zpravodajských útvarech a v komunikátech rozhlasových a televizních.“ (Čechová, 2008, str. 246)

Je potřeba si uvědomit, že ač jsou na texty a projevy tohoto stylu kladeny požadavky na dobrou úroveň obsahu i formality, vznikají v časové tísni. Výhodou je však jistá modelovost, která usnadňuje práci (např. denní tisk má stále stejné uspořádání a rubriky). S publicistickými texty jsou spjaté dva pojmy, a to aktualizace a automatizace. Aktualizace je „záměrná odchylka od standardního užití jazykových výrazových prostředků“ (Krčmová, 2017a), zatímco automatizace je její opak, tedy „užití výrazového prostředku ve shodě se stylovou normou“ (Krčmová, 2017a). Časté je užití obrazných vyjádření a metafor, které však nejsou složité, a také přejatých slov z různých oborů.

Z hlediska syntaxe je typická sevřená stavba vět a využití konstrukcí s nepůvodními předložkami, nepravé věty vedlejší či hromadění genitivních konstrukcí. Využity by měly být především prostředky neutrální spisovné vrstvy a použití nespisovné češtiny je záměrné (např. kvůli autentickému podání situace).

1.3.5 Funkční styl umělecký

S uměleckým stylem se setkáváme běžně například v beletristických knihách. Vnitřně se rozděluje na tři literární druhy, a to epiku, lyriku a drama, přičemž všechny druhy se dále dělí.

Jádrem tohoto stylu je estetická funkce, která má ve čtenáři podněcovat představy a působit na city, v kombinaci s funkcí komunikační. Od ostatních stylů se ten umělecký výrazně liší, po všech stránkách je plně v rukou autora, jak bude výsledný text vypadat. Výrazné je využití přímé řeči a poetických prvků. Není zde kladen důraz na spisovnost a v textech můžeme objevit výrazy zastaralé, knižní, nářeční i slangové. Z lexikální stránky je kladen důraz na rozmanitost lexika.

Texty uměleckého stylu v naší práci záměrně vynecháme, protože jsou z hlediska morfologického, syntaktického a lexikálního velice náročné na zobecňující popis a každý autor má svůj vlastní styl, který je více či méně odlišný od ostatních autorů.

1.3.6 Funkční styl rétorický

Přestože se v práci budeme věnovat textům psaným, pro úplnost v krátkosti zmíníme také rétorický styl.

Cílem mluveného projevu je přesvědčit posluchače či ho informovat. Na rozdíl od psaných textů je adresát bezprostředně přítomen, rétor tedy může reagovat na různé podněty (například dovysvětlení, pokud adresát nerozumí). Svou roli hrají také neverbální prostředky a využití hlasu. Mluvčí je na svůj projev dobře připravený a vyjadřuje se spisovně.

1.4 Výrazy neutrální, knižní a hovorové

Neutrální výrazy, tzv. stylově bezpříznakové, může uživatel jazyka využít v jakékoli situaci bez ohledu na styl a způsob komunikace. Oproti tomu výrazy příznakové mohou pozměnit ráz či dokonce význam výpovědi. Příznakové prostředky lze stylově „seřadit na ose (archaický) – knižní – neutrální – hovorový – (nespisovný: obecný, regionální nářeční)“. (Jelínek, 2017b)

Čechová uvádí, že tato stylová příznakovost vzniká díky množství synonymních prostředků, ať už lexikálních, frazeologických a slovotvorných, nebo syntaktických, tvarových a hláskových, které umožňují existenci variantních prostředků. (Čechová, 2008, str. 131) Důležité je si uvědomit, že spolu s vývojem jazyka se hranice na vyznačené ose posouvají a mění, tedy například slova, která byla dříve nespisovná, dnes považujeme za neutrální a dále slova, která byla dříve neutrální, jsou dnes hodnocena jako knižní. Tento posun se týká nejen slov, ale také slovosledu.

V psaných textech se setkáme s výrazy knižními. Jelínek (2012) uvádí, že „jsou spjaty s tzv. vysokým stylem. Jsou provázeny odstínem jisté výrazové „vznešenosti“, a pokud neslouží cílům parodizačním nebo humorným, stávají se prvky prestižních stylů. (...) Jsou v současném jazykovém systému neupevněny a jejich užití musí být zvlášť motivováno.“ (Jelínek, 2012, str. 780)

Hovorové prostředky využíváme každý den v mluvených projevech. Jak uvádí Jelínek (2012), „jejich struktura obyčejně vypovídá spisovnému systému, posouvají se postupně mezi prostředky spisovné.“ (Jelínek, 2012, str. 779)

Prostředky slangové spadají pod prostředky hovorové a jsou využívány specifickou skupinou mluvčích (např. lékaři).

V naší práci se budeme zabývat psanými texty. Jak vyplývá z předchozího popisu, texty psaného rázu by měly využívat výrazy neutrální; výrazy knižní jsou považovány za známky „vyššího stylu“, avšak mohou snížit čitelnost textu, kdežto hovorových výrazů je vhodné se v psaných textech vyvarovat.

1.5 Stylistika a srozumitelnost

Srozumitelnost je jedním z faktorů, které jsou důležité pro všechny funkční styly. Není-li dodržena, může dojít k chybnému předání informace či úplnému nedorozumění. Srozumitelnému textu je možné dobře rozumět a čtenář nemá problém s jeho výkladem.

Jedním z nejdůležitějších faktorů srozumitelnosti je správný pravopis a gramatika, tedy správné psaní interpunkce, slov s *i–y*, *s–z*, velkých písmen apod. Se špatným pravopisem se pravděpodobně nejvíce setkáme u prostěsdělovacího stylu. Dalším důležitým faktorem je slovosled odpovídající aktuálnímu větnému členění a správné použití tématu a rématu (podrobněji dále v kapitole 3.1). Mezi další patří délka věty: v rámci uměleckého

stylu můžeme dlouhé a vzletné věty považovat za autorský styl, avšak u administrativního stylu, zvláště u právnických textů, sníží pochopitelnost. Nejen v odborných textech se často setkáme s pasivními a jmennými konstrukcemi. Jejich využití nemusí být vždy špatné, avšak jejich přemíra většinou snižuje srozumitelnost komunikátu. Pro neznalého čtenáře bude náročný na pochopení odborný text plný termínů, které se netýkají jeho oboru. Srozumitelnost a čitelnost textů snižují také knižní a archaické výrazy, ale opět musíme dodat, že v uměleckém stylu je jejich využití očekáváné, zatímco v administrativním stylu naprosto nevhodné.

2 SOUČASNÉ NÁSTROJE PRO AUTOMATICKOU OPRAVU TEXTU

Lidé každý den produkují velké množství psaných textů. Dříve to byly texty psané ručně, avšak se zaváděním počítačů do všech odvětví lidské činnosti a s postupem digitalizace přebírají první laťku texty psané strojově, na počítači. Nárok na gramaticky správné texty byl vždy, s psaním na počítači se k němu přidal i nárok na psaní bez překlepů. A právě počítač přináší možnost zjednodušit si tyto korektury, a to pomocí nástrojů pro automatickou opravu textu.

Než budeme pokračovat dále, je potřeba si uvědomit rozdíl mezi pravopisem a gramatikou, protože korektory se dělí podle těchto dvou zaměření. Pravopis neboli ortografie označuje souhrn pravidel o grafickém zaznamenávání jazykových projevů (definice podle SSČ). Oproti tomu gramatika neboli mluvnice je nauka o stavbě jazyka, o jazykových prostředcích spojujících slova ve věty a o výstavbě textu, zahrnující syntax a v některých jazycích morfologii (definice podle Akademického slovníku cizích slov).

V následující kapitole popíšeme typy korektorů a nástroje, které existují pro češtinu.

2.1 Typy korektorů

V NESČ najdeme rozdělení od Karla Paly na korektory pravopisné, gramatické a stylistické. Pravopisný korektor je „program, který umožňuje opravovat v českých textech pravopisné chyby a různé typy překlepů.“ (Pala, 2017a) Dále upozorňuje, že tento korektor dokáže odhalit pouze chyby v jednotlivých slovních tvarech, př. malý kočka neodhalí jako chybný tvar, protože obě slova jsou správně česky.

Gramatický korektor, na rozdíl od pravopisného, je „program, který kontroluje gramatickou správnost textu a upozorňuje autora textu na zpravidla syntaktické chyby, jichž se při psaní textu dopustil“. (Pala, 2017b) Tento typ korektoru je složen z morfologického analyzátoru a souboru pravidel pro strojové zpracování.

Jako třetí typ Pala představuje stylistický korektor jakožto „program, který se snaží v souvislém textu rozpoznávat stylistické chyby a upozorňovat na ně uživatele“. (Pala, 2017c)

Oproti tomu Petkevič (2014) uvádí, že nástroje jsou dvojího typu: systémy pro kontrolu pravopisu, které jsou označovány jako spell checkery, a systémy pro kontrolu gramatické správnosti, jež jsou označovány jako grammar checkery. Uvádí také, že vytvořit jednoduchou kontrolu pravopisu je snadné, zatímco vytvořit systém, který bude správně kontrolovat gramatiku, je poněkud „ambicióznější".

Pala (2017b) i Petkevič (2014) zdůrazňují, že u těchto nástrojů je kladen velký nárok na rychlost a správnost. Z hlediska rychlosti by měl být nástroj schopen víceméně okamžitě vyhodnotit text a zvýraznit chybu. Zároveň by se neměl mýlit – takzvaná chyba prvního druhu (angl. *false positive*, FP) je nepřípustná.

2.2 Kontrola české gramatiky

V rámci popisu Kontroly české gramatiky (dále KČG) budeme vycházet ze článku Vladimíra Petkeviče *Kontrola české gramatiky (český grammar checker)*, který vyšel v roce 2014 v časopise *Studie z aplikované lingvistiky*. Kontrola české gramatiky vznikla Ústavu pro jazyk český AV ČR, v. v. i., v letech 2004–2005 a slouží jako podklad pro kontrolu gramatiky v kancelářském balíku Microsoft Office.

Sledované chyby jsou rozděleny do tří kategorií, a to na chyby gramatické a pravopisné, chyby stylistické a chyby formální.

2.2.1 Chyby gramatické a pravopisné

Z hlediska gramatiky a pravopisu jsou chyby rozděleny na 13 typů.

Prvním typem je *chyba ve jmenné skupině*. KČG sem řadí chyby v atrakci[1] a chyby v rozvíjejících adjektivních přívlastcích.

Mezi *chyby týkající se sloves* KČG zařazuje chyby v tvorbě kondicionálu, chybějící reflexiva, negramatický vztah spojky a slovesa ve větě a slovesa s homonymním tvarem (př. *je*).

[1] „Označení syntaktické struktury vzniknuvší mechanickým přizpůsobením gramatického tvaru nějakého slova gramatickému tvaru slova sousedního." (Karlík, 2017)

Dalším typem je *chyba ve shodě u sloves*, tj. ve shodě ve víceslovném slovesném přísudku. Kontroluje se také vztah mezi finitním slovesem a přechodníkem.

Samozřejmostí je kontrola *chyb ve shodě přísudku s podmětem*.

Mezi *pravděpodobnou chybu ve shodě přísudku s podmětem* se počítají věty, u nichž může být více interpretací shody.

Typ *chyb týkajících se zájmen* se soustředí na chyby v užívání zájmenných tvarů, vztahy na předělu hlavní věty a věty závislé a chyby v homofonních podobách zájmena já a ona.

V rámci *chybného slovosledu* se KČG zabývá pozicí příklonek ve větě, tedy zda jsou v syntakticky druhé pozici.

Chyby ve valenci jsou velice složité a KČG se jim věnuje pouze ve vybraných případech. Petkevič uvádí, že zde je velký prostor pro rozšíření.

V rámci *chyb ve vokalizaci předložek* KČG kontroluje správnou podobu přeložky.

Také se hlídají *chyby v použití předložek*, a to z hlediska pádové rekce, když spolu nesouhlasí předložka se jménem.

Detekuje-li KČG *chybějící nebo přebývající čárku ve větě*, zobrazí uživateli upozornění, aby věnoval větě pozornost, protože se v ní vyskytují dvě finitní slovesa.

Do skupiny *chyb ve tvarech slova* patří některé překlepy, spřežky, chybné užití spojovníku, psaní *i-y, u-ů*, jednoho či dvou *n*, účelových adjektiv či specifické chyby při práci s OCR.

Mezi *ostatní chyby* patří chybějící čárka u vokativu.

2.2.2 Chyby stylistické

Stylistickým chybám se KČG příliš nevěnuje. Jsou rozděleny na dva typy, a to nadbytečná slova a ostatní stylové chyby. Petkevič (2014) uvádí tyto příklady:

(a) Nadbytečné formy:

Udělal to bez toho, aniž by se začervenal.

Koupil jsem to proto, protože jsem chtěl.

(b) Nespisovné slovesné tvary:

**Kolegové ti to poví.*

**Oni ví, oni tu pečínku sní.*

Za prohřešek proti stylu je v KČG hodnocen i opisný komparativ (spojení *více/méně* + pozitiv/komparativ). Je však nutné věnovat pozornost homonymii adverbia *více* s číslovkou *více*:

**Přiběhli více rychleji.*

Měli více lépe propracovaných řešení.

V první větě KČG správně hlásí prohřešek, avšak druhá věta je zcela správně, ale i tady KČG upozorňuje na chybu (tedy tzv. chyba prvního druhu). U složitých adjektiv, především kompozit, je možné stupňovat opisně, proto u nich KČG chybu nehlásí, například:

Odboj byl více protisrbský. (Petkevič, 2014)

2.2.3 Chyby formální

Mezi formální chyby KČG zařazuje hromadění interpunkce, malé písmeno na začátku věty, chybějící či přebývající mezery, přebývající mezery kolem interpunkce, přípony u čísel, chyby ve psaní zkratek, rozdíl mezi spojovníkem a rozdělovníkem a také příliš dlouhé věty (hranice je stanovena na 300 slov).

2.3 Nástroje společnosti Lingea

Nástroje společnosti Lingea jsou rovněž využity v kancelářském balíčku MS Office a navíc také v nástrojích společnosti Adobe (InDesing, Photoshop). Patří mezi ně *dělení slov*, *korektor překlepů* a oprava *oslovení v dopisech*. Nutno podotknout, že se nezaměřují pouze na češtinu, ale také na další slovanské, románské a jiné jazyky.

Korektor překlepů[2] patří do kategorie pravopisných korektorů. Jeho základem je detailní popis formální morfologie a systému vzorů. Z jednoho slovního základu (kmene) se generují odvozené tvary. Slovník pro každý jazyk se skládá ze dvou částí: slovníku

[2] Dostupný na webu https://korektor.lingea.cz/.

vzorů, který zahrnuje lingvistické základy tvoření tvarů, alternací v kořeni a popis gramatických kategorií, a slovníku kmenů, jehož obsahem jsou kmeny většiny slov jazyka. (Lingea, 2018b)

Nástroj pro dělení slov pracuje na základě vybrané množiny vzorů pro dělení slov. Podle webu Lingea zvládne správně rozdělit i termíny a cizí slova a dodržuje typografické zásady nevhodného dělení slov (př. dělení slova knihovna na dva řádky). (Lingea, 2018a)

Dříve společnost Lingea vyvíjela nástroj Grammaticon. K dispozici byl na CD jako samostatný program či jako doplněk pro MS Office, avšak dostupný je pouze pro operační systém Windows XP a nižší, přičemž tyto verze již nejsou firmou Microsoft podporovány.

Grammaticon se kromě pravopisných, gramatických a interpunkčních chyb, zabýval také stylistickými chybami. Ve specifikacích programu je uvedeno, že umožňoval přepínání mezi pěti styly: standardním, formálním, technickým, neformálním a dopisem. Nabízel také možnost definovat si vlastní styl. Dále specifikace uvádí, že Grammaticon opravuje stylistické prohřešky týkající se např. použití první osoby v technických textech, zmnožení větných členů, opakování slov ve větě, přítomnost slovesa ve větě, oslovení v dopise velkým písmenem apod. (Lingea Grammaticon, 2018)

2.4 Srovnání Grammaticonu a KČG

Srovnání obou zmíněných nástrojů existuje několik, například (Pala, 2005) a (Petkevič, 2014). Dále rozebereme druhé zmíněné srovnání od Vladimíra Petkeviče.

Petkevič (2014) uvádí, že „KČG dokáže odhalit cca 30–40 % chyb v českých textech“. KČG s rozumnou mírou falešných chyb zpracuje všechny hlavní typy gramatických chyb, Grammaticon oproti tomu hlásí poměrně velké množství neoprávněných chyb (např. u slovesných skupin, parazitických slov, ve shodě přísudku s podmětem i ve slovosledu). KČG oproti Grammaticonu lépe odhaluje chyby v čárkách. KČG upozorňuje vždy na nejzávažnější chybu v daném kontextu, Grammaticon často hlásí stejnou chybu vícekrát. Grammaticon často nerozpozná očividnou chybu, což se KČG stane výjimečně. Grammaticon se na rozdíl od KČG více věnuje stylu.

Také Pala (2005) upozorňuje na velké množství falešných chyb, které hlásí nástroj Grammaticon.

2.5 Další nástroje

Na internetu můžeme najít různé další nástroje na kontrolu textu. Jejich kvalita však není valná a zpravidla se jedná pouze o pravopisné korektory.

Například na webu Litéra[3] je k dispozici Korektor, který se však zaměřuje na typografii a pracuje na základě slovníku.

V rámci akademické sféry vzniklo také několik nástrojů jako bakalářské a diplomové práce. Zmiňme například nástroj Korektor Michala Richtera na ÚFAL MFF UK[4]. Cílem bylo vytvořit pokročilý spell checker. Tento nástroj využívá jazykové modely, lexikální morfologickou analýzu a statistickou analýzu. V rámci testování dosáhl F-míry 0,87. (Richter, 2012)

Podíváme-li se na zahraniční nástroje, jistě stojí za zmínku nástroj Grammarly. Jedná se o komplexní grammar checker pro angličtinu založený na umělé inteligenci. Ta je využita například k transformaci neformálních textů na formální. (Tetreault, 2018)

[3] Dostupný na webu https://www.liteera.cz.

[4] Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy.

3 ANALÝZA LINGVISTICKÝCH PROBLÉMŮ

Definovat stylistickou chybu není jednoduché. Nejedná se vždy o prohřešky proti stylu, jak jsme jej popsali v první kapitole. Pala v NESČ uvádí, že „nejčastěji spočívají v chybné volbě použitých výrazů nebo konstrukcí vzhledem k danému komunikačnímu záměru“. (Pala, 2017)

V této kapitole vysvětlíme lingvistický podklad k analyzovaným problémům.

3.1 Slovosled

Mnoho uživatelů jazyka považuje češtinu za jazyk s volným slovosledem, a také *Nový encyklopedický slovník češtiny* uvádí: „Slovosled v češtině se vyznačuje vysokou mírou slovosledné flexibility.“ (Uhlířová, 2017) Pravidla však v češtině platí i pro slovosled, jen si je ne vždy uvědomujeme. Například různý pořádek slov v jedné větě může dodat naprosto odlišný význam. Ukažme to konkrétně na následujících větách:

Hanka šla nakoupit s Petrem.

Nakoupit šla Hanka s Petrem.

S Petrem šla nakoupit Hanka.

Každá ze zmíněných vět má trochu odlišný význam. Věta první říká, s kým šla Hanka nakoupit. Věta druhá naopak specifikuje, že nakoupit šla Hanka s Petrem dohromady. A třetí věta dává do popředí to, kdo šel nakoupit s Petrem.

S tímto souvisí také aktuální členění větné a s ním i pojmy téma a réma. Téma označuje východisko, základ věty, a definuje se jako „to, o čem se mluví“ a „to, o čem se vypovídá“. Réma naopak označuje informaci novou pro danou komunikační situaci a je také označováno termíny jádro či ohnisko věty. (Hajičová, 2017)

Věty, které jsou postaveny špatně, mnohdy pozbývají svého smyslu a nepřináší příjemci žádnou informační hodnotu. Často mají tyto věty více významů a nedávají smysl nikomu jinému než autorovi. Jako příklad uveďme:

Děti rodí ženy.

Z ukázané věty není jasné, zda ženy jsou rodičkami, což by bylo přirozené, nebo naopak děti jsou rodiči žen. Dále rozebereme slovosledná pravidla, kterým se budeme věnovat v praktické části.

3.1.1 Příklonky

V rámci slovosledných pozic můžeme určit pozici iniciální, postiniciální, mediální a finální. Za iniciální pozici je považována první přízvučná pozice ve větě a může se v ní vyskytovat pouze jeden větný člen. Postiniciální pozice je nepřízvučná a nezáleží na počtu nepřízvučných slov či větných členů. Relativně přízvučná je pozice mediální. Nachází se uprostřed věty a může se v ní vyskytovat jeden i více větných členů. Finální pozice je poslední přízvučná pozice ve větě a pojme maximálně jeden větný člen. Tato pozice je na rozdíl od ostatních obligatorní.

Pokud je v češtině nějaké slovosledné pravidlo důsledně dodržováno, je jím postiniciální pozice příklonek.

Př. *Letos* si *pořídím nové kolo.* × **Letos pořídím* si *nové kolo.*

Př. *Pročetl* jsem *hodně, ale ještě* jsem se *nic nedozvěděl.* × **Pročetl* jsem *hodně, ale ještě nic* jsem se *nedozvěděl.*

Druhou pozicí ve větě nemusí vždy být druhé slovo, jak si můžeme všimnout v druhém příkladu. Uhlířová ve své publikaci *Knížka o slovosledu* (1987) k příklonkám vysvětluje: „Hlavním důvodem je jejich zvuková kvalita. Jsou to slova většinou jednoslabičná, bez vlastního přízvuku. Z tohoto důvodu se ve větě „přiklánějí“ k předcházejícímu přízvučnému slovu a tvoří s ním jeden přízvukový a rytmický celek. V důsledku své nepřízvučnosti nemohou stát na počátku věty, ale až za prvním přízvučným úsekem věty, tedy na místě „druhém“, chcete-li, v tak zvané postiniciální pozici.“ (Uhlířová, 1987, str. 83) A na stejném místě také vysvětluje, co může být ve větě považováno za první pozici: jediné slovo či holý větný člen; skupina slov, která tvoří rozvitý větný člen; větný člen, který je rozvitý vedlejší větou přívlastkovou; samotná vedlejší věta; spojka souřadicí nebo podřadicí; větný člen, po kterém následuje vsuvka. Zjednodušeně řečeno: první pozicí je myšlen první přízvučný úsek.

Podíváme-li se na příklonky z hlediska funkčních stylů, zjistíme, že například v uměleckém stylu, konkrétně poezii, nemusí být pravidlo druhé pozice dodrženo, protože poloha příklonek, stejně jako ostatních slov, se podřizuje rytmickému schématu verše. Toto pravidlo bývá porušeno i v mluvených projevech a svědčí o nepřipravenosti mluvčího, který netušil, co chce sdělit. V psaných jazykových projevech je tato pozice dů-

ležitá, protože slovosled přebírá intonační funkci. Druhá pozice příklonky napomáhá syntaktické interpretaci věty, což je důležité u rozsáhlých textů odborného stylu. (Uhlířová, 1987, str. 85–87)

Pokud je ve větě přítomno předklonek více, předchází zvratné *se*, *si* krátkým tvarům osobních zájmen (př. *se mi*). Ani toto však není pravidlo, které platí vždy. Pokud se jedna předklonka vztahuje k přísudkovému slovesu a druhá k infinitivu, není nutné, aby stály vedle sebe (př. *Reakcí na to je snaha ze sevření se osvobodit*.). (Uhlířová, 1987, str. 93)

Také Svoboda (1984) se zabývá pořadím slov v postiniciální pozici. Řadí je takto:

- subjektový exponent slovesa vyjádřený samostatným slovem (*jsem* v préteritu, *bych* v kondicionálu);
- subjekto-objektový (reflexívní a recipročí) exponent slovesa (*se, si*) a formálně totožné partikule bez výše uvedené sémantické náplně;
- nepřímý objekt (*mu, ti*);
- přímý objekt (*ho, tě*);
- „pronominální“ adverbiále místa (*tu, tam, zde, sem*), času (*teď, pak*), způsobu (*tak*) a dále předložkový pronominální objekt či adverbiále (*k němu, s tebou*). (Svoboda, 1984)

3.1.2 Předložky

Z hlediska slovosledu můžeme za stylisticky nevhodné vyhodnotit také postavení dvou předložek vedle sebe, protože snižují plynulost textu během čtení a čtenář se musí pozastavit nad významem věty.

Př. *Dospěl* k pro *něj těžké otázce.*

3.2 Ukazovací zájmena

Velice častým jevem, který se opakuje u mnoha uživatelů jazyka, je opakování ukazovacích zájmen. Svědčí o nedostatečně velké slovní zásobě a případně o nepřipravenosti projevu, kdy autor neví, co je záměrem komunikace.

Př. *Až* tu *funkci implementujeme, pro* toho *přihlášeného uživatele* to *bude podmínkou, odsouhlasit* ten *souhlas.*

3.3 Opakování stejných výrazů

Malá slovní zásoba vede i k dalšímu problému, který můžeme zařadit do stylistických prohřešků, a to opakování stejných výrazů.

Př. Však *to byla ona,* však *ji všichni poznali.*

Do této kategorie patří tzv. slovní vata (popř. slovní vycpávka, parazitické slovo). Objevují se především v rámci mluvených projevů. Typickým příkladem je foném [ə]. Avšak i v psaných textech se s nimi můžeme setkat, jedná se například o slova *prostě* a *vlastně*.

Př. *Je to* prostě *tak, tento notebook je vskutku vybaven 20" LCD s rozlišením nádherných 1920 × 1200 bodů.*

3.4 Pleonasmy

Akademický slovník cizích slov uvádí definici pleonasmu: „stylistická figura vznikající nahromaděním výrazů významově blízkých; nadbytečné hromadění synonym různého slovního druhu“ (Kraus, 2005). V některých případech mohou sloužit jako zdůraznění nebo upoutání pozornosti čtenáře či posluchače, ale z hlediska stylu je vhodnější se jim vyhnout.

Př. *Představ si, že k nákupu dávali* dárek zadarmo*!*

3.5 Nesmyslné superlativy

Uživatelé jazyka se občas nechají unést a chtějí vyjádřit, že zážitek byl výborný, jednání důležité, dovednost jednoduchá apod. Slova, která už sama o sobě jsou pomyslným vrcholem, se jim nezdají dostatečně výstižná, a proto u adjektiv a adverbií vznikají superlativy, které sice z morfologického hlediska dávají smysl, ale z logického už nikoli.

Př. *Až se překladatelé naučí* nejzákladnější *základy základní terminologie, s radostí jednoho přidám.*

3.6 *Jakýkoli(v)–kterýkoli(v)*

Mezi slovy *jakýkoli(v)* a *kterýkoli(v)* je rozdíl ve významu, avšak v textech je můžeme nalézt zaměněné. *Kterýkoli(v)* se používá ve významu výběru z prvků určité množiny, zatímco *jakýkoli(v)* označuje výběr určité vlastnosti či kvality.

Př. Kterýkoli *z uchazečů má šanci postoupit do druhého kola.*

Př. *Byl ochoten přijmout* jakoukoli *pomoc.*

Př. **Demisi mohl podat* jakýkoli *z členů vlády.*

3.7 Dvojité spojky

Často se stává, že si mluvčí v průběhu výpovědi rozmyslí, co chtěl sdělit. S větší pravděpodobností se s tímto problémem setkáme v mluvených projevech. V psaných textech se nedomyšlené výpovědi projevují většinou vyšinutím z větné stavby, nebo také například zapomenutými dvojitými spojkami *buď'–(a)nebo*, *sice-ale (však, ale, přesto, nicméně)*. Setkat se můžeme i s použitím spojek *jednak–*druhak*, přičemž **druhak* není spisovný tvar a správné spojení je *jednak–jednak*. Jak uvádí Pravdová, „výraz **druhak* je stylově příznakový, motivovaný snad snahou o určité ozvláštnění řeči. Dodatečně byl přitvořen ke slovu *jednak*, avšak nikoli analogicky, protože vychází z číslovky řadové (*druhý*), zatímco *jednak* je utvořeno z číslovky základní (*jeden*)." (Pravdová, 2012)

Př. *Jednak těm lidem lépe rozumím, protože mluví spisovnější a pro mě srozumitelnější češtinou, a* *druhak *miluju Olomouc, Znojmo, Mikulov.*

3.8 Vztažná zájmena *který* a *jenž*

Jako stylisticky nevhodné můžeme vyhodnotit také téměř nekonečný cyklus vztažných vět navázaných na sebe, viz následující příklad. Čtenář je nucen pamatovat si informaci o hospodě, následně o hospodě a domu, poté o hospodě, domu a zahradě atd., než dospěje k vysvětlení, proč je hospoda vůbec zmíněna.

Př. *Šel do hospody, která byla za domem, který měl velkou zahradu, kterou vlastnil starý děda, který poštvával psa na kočky, které tam lovily myši, a dal si tam pivo, které měl rád.*

U zájmena *který*, případně *jenž*, vznikají chyby také použitím špatného tvaru slova. U zájmena *který* se to týká především špatného užití u středního rodu, u zájmena *jenž* je problém komplexní, anžto jeho uživatelé neznají správné tvary.

Př. *Dal si pivo, **který *měl tak rád.*

Př. *Dal si pivo, **jenž *měl tak rád.*

3.9 Hovorové prvky

Jak jsme zmínili v závěru první kapitoly, hovorové prvky do psaného textu nepatří. Do této skupiny můžeme zařadit například adjektiva zakončená v nominativu na *-ej* vzniklé změnou z koncovky *-ý*. Do psaných komunikátů se dostávají především v rámci prostěsdělovacího stylu. V rámci sloves zmiňme například rozšířené tvary *můžu* a *můžou*, které kodifikační příručky uvádějí jako hovorový. Musíme ovšem podotknout, že u tohoto slovesa dochází k posunu na již zmíněné ose knižní – neutrální – hovorový a tvary *mohu* a *mohou* jsou považovány za stylově vyšší. (Čechová, 2008, str. 150). Je tedy pravděpodobné, že tvary *můžu* a *můžou* budou v příští kodifikační publikaci uznány jako neutrální, protože například PMČ (2012) je již jako neutrální uvádí. K nespisovným tvarům sloves patří dále vynechání koncového *-e* u slovesa *budeme*.

Př. Můžu *s tebou jít, ale ten film musí být fakt* dobrej.

3.10 Ostatní chyby

V této části rozebereme ostatní chyby, kterých se dopouštějí autoři textů.

Mezi jednu z nejčastějších patří špatný zápis slova *výjimka* – tedy špatně **vyjímka*. Toto slovo vzniklo odvozením od slovesa *vyjímat*, avšak došlo ke zkrácení kořene a prodloužení předpony. Změna se týká také adverbia *výjimečný*.

Př. *On byl vždy **vyjímečný.

Mezi další častou chybu patří chybné psaní slova *permanentka/permanentní*, ve kterém pisatelé zamění písmena *m* a *n*, takže vznikne **pernamentka/pernamentní*.

Př. *Mám **pernamentku *na plavání, heč!*

V psaných textech, v největší míře v odborném stylu, je potřeba odkazovat na tabulky, grafy, obrázky či jiné úseky textu. Využívá se k tomu rozkazovací způsob slovesa *vidět*, tedy tvar *viz*, nebo *vizte*. Není výjimkou, že pisatelé tvar viz považují za zkratku, a proto za ním píší tečku.

Př. *Detaily jsou na obrázku, *viz. strana 90.*

Problém je také v psaní číslovek. Tvary typu *14ti-procentní* řeší ve své bakalářské práci například (Michálek, 2017). My jsme se zaměřili na špatný tvar číslovek *dva* a *oba*. Pravděpodobně z důvodu hyperkorektnosti vznikají tvary **dvěmi/oběmi* a **dvoumi/oboumi* ve třetím a sedmém pádě namísto správných tvarů *dvěma/oběma*.

Př. *Před *dvěmi minutami jsem zjistil, že my dva máme „lepší" vkus.*

Čeština na rozdíl od například angličtiny využívá k oslovování vokativ. To znamená, že oslovení typu *pane Strako* je označováno za spisovné. Především v prostěsdělovacím stylu občas můžeme objevit oslovení *pane Straka*, tedy využití nominativu namísto vokativu. Takováto oslovení jsou hodnocena jako nespisovná.

Př. Pane Straka*, mohl bych vám položit otázku?*

U nominativu plurálu maskulin zakončených na *-t* a *-ta* se využívají dvě koncovky, a to *-i* a *-é*, přičemž tvary s -é jsou některými lingvisty označovány jako knižní. U jisté skupiny jmen je však ve spisovném jazyce vyžadováno *-i* (př. *advokáti*), což je pravděpodobně způsobeno vokalickou disharmonií (ve slově **advokáté* by byly dvě dlouhé hlásky za sebou). U jmen na *-nt* se hyperkorektně ve projevech vyskytuje koncovka *-é*, protože ji autoři považují za stylově vyšší, avšak spisovně je pouze varianta *-i* (př. *reprezentanti*). (Čechová, 2008, str. 139)

Př. *Naši *reprezentanté *letos opět nevyhráli.*

Poměrně často v mluvené řeči a poté i textech zaznívá špatný tvar slova *datum*. Jedná se o slovo latinského původu a podobně jako ostatní jména latinského a řeckého původu má odlišné skloňování. Tedy místo správných *tvarů data, datu, datem, dat, datům, datech, daty* se v textech objevují tvary *datumu, datumech, datumy, datumům, datumům*. Nutno podotknout, že v textech týkajících se informatiky lze tyto tvary akceptovat, aby nedošlo k nedorozumění (př. Převedu *data* k výše uvedenému *datu*.) (Konečná, 2014)

Př. *Na těchto* datumech *jsme se všichni shodli.*

Předložka *mimo* se podle SSČ pojí s akuzativem. Avšak především v administrativních a publicistických textech se zaměňuje se předložkou *kromě*, která se váže s genitivem. Vzniká tak vazba *mimo* + genitiv, která je nedodržením spisovné formy. (Čechová, 2008, str. 158)

Př. *Ještě že v autě vezeme* mimo *grilu i dvoje elektrické topidla.*

3.11 Dlouhé věty

Délka vět je často daná stylem, do kterého je komunikát zařazen. Například věty v textech odborného stylu budou delší než věty stylu prostěsdělovacího. Pala (2017c) uvádí, že „typickou stylistickou chybou jsou i příliš dlouhé věty či souvětí čítající zhruba více než 25 slov, důsledkem je pak zpravidla jejich nízká srozumitelnost“. (Pala, 2017c)

V rámci uměleckého stylu mohou být dlouhé a složité věty uměleckým záměrem. Jako příklad můžeme uvést knihu Obsluhoval jsem anglického krále od Bohumila Hrabala.

Př. *Zákon také ukládá povinnost všem poskytovatelům internetu zaznamenávat veškerý pohyb uživatelů na internetu (v jaký čas a kam vstoupili) a všem poskytovatelům webhostingů povinnost veškeré uložená data zálohovat několik měsíců zpětně pro pozdější dohledání*.

4 VYUŽITÉ NÁSTROJE

Tato kapitola se věnuje nástrojům, které jsme využili v rámci praktické části. Začneme u tokenizace, která je základem pro počítačové zpracování textu, poté si představíme morfologický analyzátor ajka a následně budeme pokračovat popisem syntaktického analyzátoru SET a způsobu tvorby pravidel pro tento nástroj.

4.1 Tokenizace

Mezi základní operace, které se provádí na textech, patří tokenizace. Jedná se o „automatický proces, který člení text složený z písmen, interpunkčních znamének a mezer na jednotlivé izolované tokeny, tj. na slovní tvary a interpunkční znaménka pro účely dalšího (obvykle počítačového) zpracování.“ (Petkevič, 2017b)

Pro účely naší práce využíváme unitok, který byl vytvořen v rámci Centra zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity (dále CZPJ FI MU).

Příklad výstupu z nástroje:

```
Pokračování
dvoudílné
fantasy
komedie
o
neschopném
čaroději
<g/>
,
jenž
se
stává
průvodcem
prvního
turisty
<g/>
,
který
kdy
navštívil
Zeměplochu
<g/>
.
```

4.2 Morfologický analyzátor ajka

Morfologická analýza je „automatický proces, při němž se v užším smyslu každému slovnímu tvaru v (korpusovém) textu přiřadí všechny morfologické údaje včetně slovního druhu v podobě značky (tagu), v širším (obvyklém) smyslu navíc všechna jeho lemmata procesem lemmatizace“ (Petkevič, 2017a) a je jedním ze vstupních prvků při počítačovém zpracování textu.

V rámci Centra zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity vznikl nástroj majka/ajka původně jako diplomová práce Radka Sedláčka v roce 1999. Zjednodušeně řečeno, ajka pro každé slovo vytvoří základní tvar (lemma), všechny tvary slova (př. pro substantiva tvary singuláru a plurálu všech pádů) a přiřadí mu morfologické značky (tagy), které vychází z brněnského atributového značkování. V případě homonym přiřadí slovu všechny možné tvary (tedy pro slovo *ženu* lemma *hnát* a *žena* a značky k5eAaImIp1nS a k1gFnSc4). Následně s pomocí nástroje desamb proběhne desambiguace (tj. zjednoznačnění), která rozhodne, které lemma a která značka se má použít.

S výstupem, který vytvoří nástroj ajka a desamb, dále pracuje syntaktický analyzátor SET, který popíšeme v následující části. Příklad výstupu z nástroje:

```
<s desamb="1">
Pokračování     pokračování     k1gNnSc1
dvoudílné       dvoudílný       k2eAgInPc1d1
fantasy         fantas          k1gInPc1
komedie         komedie         k1gFnSc2
o               o               k7c6
neschopném      schopný         k2eNgMnSc6d1
čaroději        čaroděj         k1gMnSc6
<g/>
,               ,               kIx,
jenž            jenž            k3xRgMnSc1
se              se              k3xPyFc4
stává           stávat          k5eAaImIp3nS
průvodcem       průvodce        k1gMnSc7
prvního         první          k4xOgMnSc2
turisty         turista         k1gMnSc2
<g/>
,               ,               kIx,
který           který           k3yRgMnSc1
kdy             kdy             k6eAd1
navštívil       navštívit       k5eAaPmAgMnS
Zeměplochu      Zeměploch       k1gInSc2
<g/>
.               .               kIx.
</s>
```

4.3 Syntaktický analyzátor SET

Syntaktická analýza je „rozklad jazykové jednotky mající za cíl reflektovat fakt, že za linearitou segmentů, z nichž je jednotka složena, se skrývá hierarchizovaná struktura." (Karlík, 2002) Pro češtinu je syntaktická analýza složitější, protože na rozdíl od jiných jazyků má poměrně volný slovosled.

V rámci syntaktické analýzy existují dva hlavní přístupy. Prvním je závislostní přístup, který „za hlavní člen věty je považován predikát (nejčastěji sloveso), který je rozvíjen dalšími, závislými členy (které mohou být rovněž rozvíjeny)". (Kučera, 2006) Tento přístup je rozvíjen v rámci Pražského závislostního korpusu. A druhým je složkový přístup, který oproti závislostnímu dokáže analyzovat větší jednotky než pouze slova. Výstupem jsou složkové stromy, které jednoznačně ukazují, které části věty jsou jmenné skupiny a které z nich se pojí se slovesem (Kovář, 2008). Se složkovým přístupem se setkáme v CZPJ FI MU u nástroje Synt. V rámci těchto přístupů můžeme vymezit metodu parciální analýzy, pro kterou není předmětem celý strom, ale pouze některé jeho části.

Syntaktický analyzátor SET (*Syntactic Engineering Tool*) je vyvíjen v CZPJ FI MU pod dohledem Vojtěcha Kováře. Založen je na postupné segmentaci věty, což je metoda kombinující jak závislostní, tak složkový přístup i parciální analýzu. Naše práce je založena na tomto nástroji, proto se budeme věnovat jeho popisu. Vycházíme z oficiální dokumentace dostupné na webu[5] a z disertační práce Vojtěcha Kováře *Automatic Syntactic Analysis for Real-World Applications*. (Kovář, 2014) Nebudeme uvádět kompletní popis nástroje, vyjmenujeme pouze prvky, které jsme využili v praktické části.

Pravidla mají danou pevnou strukturu, která se skládá z šablony a seznamu akcí. Šablona je část pravidla, která popisuje podmínky, za kterých dané pravidlo platí. Poznáme ji podle klíčového slova `TMPL:` a musí být vždy umístěna na jednom řádku. Pokud jsou podmínky splněny, provede se příkaz zapsaný v seznamu akcí. Tento seznam může být rozdělen na více řádků. Obecné schéma pro pravidlo vypadá následovně:

```
TMPL: (ŠABLONA) (SEZNAM AKCÍ)
```

[5] Dostupné na http://nlp.fi.muni.cz/trac/set/wiki/documentation.

4.3.1 Šablona

Jak jsme již zmínili, šablona specifikuje podmínky platnosti pravidla. Podmínky lze vyjádřit více způsoby. Prvním z nich je tzv. *jedna podmínka*, která je vyjádřena kombinací atributu (word, lemma či tag) a podmínky. Psané jsou do kulatých závorek a pro výběr z možností je možné využít disjunktivní znak |.

Konkrétní příklady z gramatiky:

- `(word buď)`: vyhledá všechny výskyty slova buď v daném tvaru,
- `(lemma který)`: vyhledá všechny výskyty lemmatu který,
- `(tag k7)`: vyhledá všechny výskyty značky k7[6], tedy všechny předložky.

V rámci jedné šablony je možné využít více podmínek.

Další možností, jak vyjádřit podmínku, je *pojmenovaná proměnná*. Můžeme ji využít k zápisu složitější podmínky a v případě, že chceme využít stejnou podmínku ve více pravidlech. Definuje se vždy na konci pravidla či gramatiky a zapisuje se pomocí znaku pro dolar, následuje název pojmenované proměnné, její atributy a poté omezení, za kterých platí. Zápis tedy vypadá `$JmenoPromenne(atribut)`.

Seznam atributů je stejný jako u předchozího pravidla s jednou změnou:

- `word`: označuje přesnou shodu ve tvaru slova,
- `lemma`: označuje lemma slova, které může být v různých tvarech,
- `tag`: označuje morfologickou značku, která musí odpovídat,
- pojmenovaná proměnná umožňuje využít i záporné varianty atributů.

V šabloně také můžeme řešit nutnost provázání pravidla se začátkem či koncem věty. K tomu jsou určeny aliasy `bound` pro začátek a `rbound` pro konec věty.

Během tvorby pravidel se můžeme dostat do situace, kdy potřebujeme vyhledat slova, která neleží přímo vedle sebe, ale nachází se mezi nimi řetězec jiných slov. V tomto případě využijeme značky `...` (tři tečky), která pojme jakkoli dlouhý řetězec.

[6] Značky vychází z brněnského atributového značkování.

4.3.2 Seznam akcí

Uvedli jsme, že v seznamu akcí jsou uvedeny příkazy, které se mají provést při splnění podmínek v šabloně. Jedná se o funkci, která podle podmínek vybere daná slova a provede akci. Vždy se skládá z klíčového slova a argumentu, přičemž každé pravidlo může mít neomezený počet akcí. Argumenty mohou například vyznačovat, na kterých místech se má akce provést. První podmínka má pozici nula. Počet argumentů je také libovolný, minimální počet je stanoven na jeden.

`MARK`

Tato akce je využívaná k označení slov v analyzovaném segmentu a je možné ji použít dvojím způsobem. Zaprvé jako přiřazení vybraných slov do složkového elementu, kdy je přímo za názvem akce vypsán seznam indexů, jež mají být zařazeny do složky. Zadruhé jako označení jediného prvku segmentu, jemuž je poté přiřazena závislost.

Příklad pravidla: `TMPL: (lemma verlyba) MARK 0 LABEL <velryba-nok>` – vybere slovo na první pozici, které vyhovuje podmínce, že lemmatem slova je **verlyba*.

`LABEL`

Akce `LABEL` přiřazuje slovu štítek, jak její název napovídá. Užitečná je například v případě, že vyžadujeme výstup v podobě závislostního stromu. Hraně nalezené akcí `MARK` přiřadí zvolené ohodnocení.

Příklad pravidla: `TMPL: (lemma verlyba) MARK 0 LABEL <velryba-nok>` – označí vybrané slovo štítkem „velryba-nok“.

`PROB`

`PROB` je akce, která mění váhu pravidla. Základní váha je nastavena na hodnotu 100, tedy pokud není váha specifikována, využije se tato základní hodnota. Argumentem akce vždy musí být kladné přirozené číslo (případně ve speciálních případech 0).

Příklad pravidla: `TMPL: (lemma buď) ... (lemma nebo|anebo) MARK 0 2 LABEL <bud-ok> PROB 1000` – zvýší váhu pravidla ze 100 na 1000.

5 POPIS PRAVIDEL

V předchozí kapitole jsme teoreticky nastínili práci s nástrojem SET. Nyní podrobně popíšeme, jak vypadají pravidla, která jsme napsali. Celou gramatiku, kterou jsme vytvořili, přikládáme v příloze A.

V rámci podmínkové části pravidel jsme ve výsledku využili všech tří možných atributů, tedy `word`, `lemma` i `tag`. V gramatice najdeme *jednu podmínku* i *pojmenovanou proměnnou*. Z možných akcí jsme využili `MARK`, `LABEL` a `PROB`, které jsme popsali v předchozí části.

Uvedeme vždy stručné odůvodnění pro dané pravidlo a jak jsme pravidlo vytvářeli. Poté představíme příklad pravidla z gramatiky, jeho výklad a příklad věty, kterou řeší. Zmíníme také případné obtíže.

V poslední části kapitoly představíme program v Pythonu, který jsme vytvořili po počítání slov ve větě.

5.1 Pravidla pro slovosled

Jak jsme popsali v analýze v kapitole 3.1, v rámci slovosledu se budeme věnovat příklonkám, protože jejich pozice je v české větě daná, a kombinaci dvou předložek vedle sebe, které snižují čitelnost.

5.1.1 Příklonky

V analýze jsme podrobně rozebrali pořadí příklonek ve větě, ale také mezi sebou. V praktické části jsme vytvořili pravidla, která detekují příklonky jak ve správné pozici a pořadí, tak ve špatné pozici či pořadí. Vzhledem k tomu, že postiniciální pozice ve větě není obligatorní a ani všechny příklonky se nemusí vyskytovat zároveň (například ve větě *Koupil jsem ho* jsou mezi *jsem* a *ho* dvě prázdné pozice pro příklonky), je sada pravidel rozsáhlejší. Pro jednotlivé pozice ve větě jsme definovali *pojmenované proměnné* `$PRVNI-POZICE`, `$DRUHA-POZICE`, `$TRETI-POZICE` a `$CTVRTA-POZICE` obsahující výpis slov, které se na dané pozici mohou vyskytnout, a také jsme vytvořili *proměnnou* `$FINITUM`, která pomáhá najít ve větě sloveso v určitém tvaru. Dále jsme využili aliasu `bound` pro začátek věty

a atributu word. Z akcí jsme kromě MARK a LABEL zapojili také PROB pro zvýšení váhy pravidel, jež nám umožnila diferenciaci mezi správným a špatným pořadím.

```
TMPL: bound $DRUHA-POZICE $FINITUM MARK 1 LABEL <priklonky-nok> PROB 101
```

Výklad: Pokud na začátku věty je slovo odpovídající proměnné $DRUHA-POZICE následované slovem odpovídajícím proměnné $FINITUM, označ slovo odpovídající proměnné $DRUHA-POZICE štítkem <priklonky-nok>. Toto pravidlo má váhu 101.

Příklad: Si *pořídím nové kolo.*

5.1.2 Předložky

Postavení dvou předložek vedle sebe je stylisticky neobratné a často kvůli němu dojde k roztržení vazeb ve větě. Vzhledem k tomu, že tento problém se může týkat téměř všech předložek, postavili jsme pravidlo pomocí dvou tagů k7.

```
TMPL: (tag k7) (tag k7) MARK 0 1 LABEL <predlozky-nok>
```

Výklad: Pokud se ve větě nachází vedle sebe dvě slova odpovídající tagu pro předložky k7, označ tyto dvě předložky štítkem <predlozky-nok>.

Příklad věty: *Valná většina (80 %) dotázaných označila internet věcí* za do *určité míry relevantní jejich podnikání.*

5.2 Pravidlo pro ukazovací zájmena

Opakování ukazovacích zájmen je poměrně častou chybou. Je však potřeba stanovit si hranici, kolik zájmen je ještě přijatelných a kolik už zasluhuje upozornění uživatele. V rámci pravidla jsme nastavili počet na dvě a více zájmen ve větě.

Autoři textů jsou ohledně ukazovacích zájmen poměrně vynalézaví a vytváří tvary ukazovacích zájmen, které nejsou spisovné, například **tohleto*. Ajka tato slova špatně rozpoznává a pro zmíněný tvar **tohleto* určí lemma **tohlet*. Vytvořili jsme *pojmenovanou proměnnou* $UKAZOVACI, do které jsme uložili lemmata všech ukazovacích zájmen spolu s lemmaty těchto nespisových výrazů.

```
TMPL: $UKAZOVACI ... $UKAZOVACI MARK 0 2 LABEL <ukazovaci-nok>
```

Výklad: Pokud jsou ve větě dvě a více slova, která odpovídají proměnné $UKAZOVACI, označ je štítkem <ukazovaci-nok>.

Příklad: *Až* tu *funkci implementujeme, pro* toho *přihlášeného uživatele* to *bude podmínkou, odsouhlasit souhlas.*

5.3 Pravidla pro stejné výrazy

V rámci opakování stejných slov jsme hledali výskyt dvou a více stejných výrazů v jedné větě. Zaměřili jsme se na ta následující slova: *proto, protože; však; tedy; jakoby; nicméně*. Z atributů jsme k vyhledávání použili word.

```
TMPL: (word Však|však) ... (word však) MARK 0 2 LABEL <opakovani-nok>
```

Výklad: Pokud se ve větě nachází slovo *však* dvakrát na různých místech, označ tyto výskyty štítkem <opakovani-nok>.

Příklad: *Kundis potáhl balon a vystřelil jeho střela byla však zblokována, k odraženému balonu se však dostal Mekki.*

Také jsme do této kategorie zařadili pravidla pro tzv. slovní vatu. Jedná se o jednoduchou detekci, jestli se slovo nachází ve větě.

```
TMPL: (word prostě) MARK 0 LABEL <vata-nok>
```

Výklad: Pokud se ve větě vyskytuje slovo *prostě*, označ ho štítkem <vata-nok>.

Příklad: *Recepce je rádoby moderně přestavěná, ale křivě postavené zdi* prostě *nezakryjete žádnou moderní technologií.*

5.4 Pravidlo pro pleonasmy

Pleonasmy jsou sousloví či fráze, které opakují stejný význam. Vznikají nejčastěji neznalostí významu jednoho z použitých slov. Pravidlo je založeno na atributu lemma.

```
TMPL: (lemma dárek|bonus) (lemma zadarmo|zdarma) MARK 0 1
      LABEL <pleonasmus>
```

Výklad: Pokud se ve větě nachází jakýkoli tvar slova *dárek*, nebo *bonus*, následovaný jakýmkoli tvarem slova *zadarmo*, nebo *zdarma*, označ tato dvě slova štítkem `<pleonasmus>`.

Příklad: *Představ si, že k nákupu dávali* dárek zadarmo*!*

Během práce jsme nasbírali velké množství pleonasmů. Jejich kompletní seznam uvádíme v příloze D. Přiložený výčet samozřejmě nepostihuje všechny pleonasmy v češtině, počítáme s jeho úpravou.

5.5 Pravidla pro nesmyslné superlativy

Superlativy u některých adjektiv a adverbií je sice možné vytvořit, ale ne vždy dávají smysl. Je tomu tak například u slov, která už sama o sobě jsou vrcholem, tím nejvyšším. Hledali jsme ta, která jsou nejčastější, a to: *nejoptimálnější*, *nejhlavnější*, *nejdokonalejší*, *nejideálnější*, *nejzákladnější*. Do *pojmenované proměnné* `$SUPERLATIV` jsme uložili všechny možné tvary těchto slov.

```
TMPL: $SUPERLATIV MARK 0 LABEL <superlativ-nok>
```

Výklad: Pokud se ve větě vyskytuje slovo uložené v proměnné `$SUPERLATIV`, označ ho štítkem `<superlativ-nok>`.

Příklad: *To jsou* *nejhlavnější *faktory neutralismu jako „nástroje“, které nám mohou pomoci z přítomné, pro nás neprospěšné situace.*

5.6 Pravidlo pro *jakýkoli*

V textech dochází k záměně slov *jakýkoli(v)* (výběr z vlastností či kvalit) a *kterýkoli(v)* (výběr z množiny prvků). Všimli jsme si, že je využívána vazba *jakýkoli(v) z* a postavili na ní pravidlo. V tomto případě jsme zvolili kombinaci atributu lemma a word.

```
TMPL: (lemma jakýkoli|jakýkoliv) (word z) MARK 0 1 LABEL <jakykoli-nok>
```

Výklad: Pokud se ve větě vyskytuje jakýkoli tvar slova *jakýkoli*, nebo *jakýkoliv* následovaný předložkou *z*, označ tato dvě slova štítkem `<jakykoli-nok>`.

Příklad: *Člověk si mohl vybrat* jakékoli ze *čtyřiadvaceti okének.*

Bohužel nedokážeme pomocí pravidel postihnout celý problém rozdílu mezi *jakýkoli(v)* a *kterýkoli(v)*, protože je záležitostí spíše sémantiky než syntaxe.

5.7 Pravidla pro dvojité spojky

Pisatelé občas v průběhu psaní zapomenou, jak začali větu, a proto v textech můžeme najít, respektive postrádat druhou z dvojitých spojek. Zaměřili jsme se na párové spojky *buď'–(a)nebo*, *sice–ale/přesto/však/nicméně* a chybné použití *jednak–*druhak*. Vytvořili jsme sadu pravidel, která tyto spojky vyhledá a náležitě je označí. Sice se jedná o nesklonné výrazy, takže se nabízí využít atributu `word`, nicméně slovo *buď'* je homonymní (spojka a rozkazovací způsob slovesa *být*), proto jsme zvolili atribut `lemma`. V tomto pravidle se nalézá také akce `PROB`, která zvyšuje váhu pravidla, a využita je pro diferenciaci správných a problémových vět.

```
TMPL: (lemma buď) ... (lemma nebo|anebo) MARK 0 2 LABEL <bud-ok> PROB 1000
TMPL: (lemma buď) MARK 0 LABEL <bud-nok>
```

Výklad: První pravidlo říká: pokud se ve větě nachází slovo *buď'* a slovo *nebo* či *anebo*, označ tato slova štítkem `<bud-ok>`. Druhé pravidlo označuje: pokud se ve větě nachází slovo buď, označ jej štítkem `<bud-nok>`.

Příklad: *Nové družstevní byty nechají pro své členy stavět* buď *družstva s dlouhou historií, v posledních letech pak v menší míře vznikají pod taktovkou developerů.*

5.8 Pravidla pro vztažná zájmena *který* a *jenž*

V textech také můžeme najít téměř nekonečná souvětí spojená vztažným zájmenem *který*, případně *jenž*. Takto navázané věty, zvlášť pokud jsou dlouhé, zhoršují čitelnost textu a je lepší souvětí rozdělit na více samostatných vět. V rámci této práce jsme se rozhodli, že budeme detekovat souvětí, která mají tři a více vztažných zájmen. Opět bylo potřeba vytvořit sadu pravidel. První z nich detekují, zda se nejedná o zdvojené zájmeno v rámci souřadně spojených vedlejších vět, protože v těchto případech nemusí být nutné se opakování vyvarovat. Druhé vyhledává tři a více výskytů zájmena. Využili jsme atribut pro `lemma`.

```
TMPL: (lemma který|jenž) ... (lemma který|jenž) ... (lemma který|jenž)
      MARK 0 2 4 LABEL <ktery-nok>
```

Výklad: Pokud se ve větě nachází tvar zájmena *který*, nebo *jenž* třikrát na různých místech, označ tato zájmena štítkem `<ktery-nok>`.

Příklad: *Šel do hospody,* která *byla za domem,* který *měl velkou zahradu,* kterou *vlastnil starý děda,* který *poštvával psa na kočky,* které *tam lovily myši, a dal si tam pivo.*

Další sada pravidel se netýká stylistiky, avšak pokládali jsme za užitečné ji implementovat, protože se v ní často chybuje. Také souvisí se vztažnými zájmeny *který* a *jenž* uvozujícími vedlejší věty. Jedná se o použití špatného rodu u zájmena: u zájmena *který* je nejčastějším problémem špatné užití středního rodu, respektive jeho nevyužití (př. *pivo, který jsem měl rád*); u zájmena *jenž* je problém komplexní, poněvadž autoři textů neznají jeho správné formy.

Sada pravidel nejprve detekuje rod jména a poté se podívá na rod vztažného zájmena za čárkou. Pro jednotlivé rody jsme definovali *pojmenované proměnné* `$KTERY-NOTF, $KTERY-NOTN, $KTERY-NOTM, $KTERY-NOTI`, do kterých jsme uložili tagy špatných tvarů zájmen.

```
TMPL: (tag k1gF.*|k2.*gF.*|k3.*gF.*|k4.*gF.*) (word ,) $KTERY-NOTF MARK 0 2
      LABEL <rodktery-nok>
```

Výklad: Pokud se ve větě nachází slovo odpovídající značce následované čárkou a výrazem z proměnné `$KTERY-NOTF`, označ nalezené slovo a výraz z proměnné štítkem `<rodktery-nok>`.

Příklad: *Neměl rád kočky, **který** lovily ptáky*.

5.9 Pravidlo pro hovorové prvky

Hovorové prvky pronikají do celé slovní zásoby a týkají se většiny slovních druhů. Díky práci s brněnským atributovým značkováním se nám práce zjednodušila na pouhé jedno pravidlo. Značky (tagy) totiž obsahují atribut stylistického příznaku, v němž je na výběr mimo jiné hovorovost.

```
TMPL: (tag .*wH) MARK 0 LABEL <hovorove>
```

Výklad: Pokud se ve větě vyskytuje slovo vyhovující tagu, označ jej štítkem `<hovorove>`.

Příklad: *Jinak jsme od Labské nížiny, tak* *můžem, *až tu organizaci* *omrknem, *uspořádat* *Labskej *okruh.*

5.10 Pravidla pro ostatní chyby

Mezi ostatní chyby jsme zařadili jednoslovné problémy, se kterými se můžeme setkat a je potřeba na ně pisatele upozornit.

První je špatný zápis slova *výjimka*, respektive **vyjímka*, a z něj odvozená adjektiva a adverbia. Použili jsme atribut `lemma` pro zachycení všech možných tvarů slova.

```
TMPL: (lemma vyjímka|vyjímečný|vyjímečně) MARK 0 LABEL <vyjimka-nok>
```

Výklad: Nachází-li se ve větě slovo *vyjímka*, *vyjímečný*, nebo *vyjímečně* v jakémkoli tvaru, označ je štítkem `<vyjimka-nok>`.

Příklad: *On byl ale* *vyjímka.*

Další chybou je metateze[7] ve slově *permanentka* a *velryba*, chybně psaná jako **pernamentka* a **verlyba*. Pravidla jsou podobná jako předchozí.

```
TMPL: (lemma pernamentka|pernamentní|pernamentně) MARK 0
LABEL <permanentka-nok>
```

Výklad: Nachází-li se ve větě slovo *pernamentka*, *pernamentní*, nebo *pernamentně* v jakémkoli tvaru, označ je štítkem `<permanentka-nok>`.

Příklad: *Mám* *pernamentku *na plavání, heč!*

[7] „Přeskupení hlásek ve slově.“ (Krčmová, 2017b)

K ostatním chybám jsme zařadili také tečku za rozkazovacím způsobem slova *vidět*. Slovo *viz* je často pokládáno za zkratku, a proto za ni autoři píší tečku. Tentokrát jsme použili atribut `word`, protože potřebujeme vyhledat přesný tvar slovesa.

```
TMPL: (word viz) (word .) MARK 0 1 LABEL <viz-nok>
```

Výklad: Pokud se ve větě vyskytuje slovo *viz* následované tečkou, označ tyto dvě pozice štítkem `<viz-nok>`.

Příklad: *Detaily jsou na obrázku, *viz. strana 90.*

Dále jsme se zabývali chybným tvarem číslovky *dva* a *oba*. Nespisovné tvary vznikají z většinou z důvodu hyperkorektnosti a z hovorového jazyka pronikají do psaného, především v rámci prostěsdělovacího stylu. Opět jsme využili atribut `word` pro přesnou podobu slova.

```
TMPL: (word dvěmi|dvoumi|oběmi|oboumi) MARK 0 LABEL <dve-nok>
```

Výklad: Pokud se ve větě nachází slovo *dvěmi, dvoumi, oběmi*, nebo *oboumi*, označ ho štítkem `<dve-nok>`.

Příklad: *Před *dvěmi *minutami jsem zjistil, že my dva máme „lepší“ vkus.*

Věnovali jsme se také oslovení mužů, které není ve vokativu. Na základě frekvenční analýzy v korpusu czTenTen jsme vybrali oslovení *pane, doktore, ministře, profesore, kolego, poslanče, starosto, místopředsedo, prezidente, premiére, kapitáne, mistře, primátore, redaktore, inženýre*. Tato oslovení jsme uložili do *pojmenované proměnné* `$OSLOVENI`.

```
TMPL: $OSLOVENI (tag k1gMnSc1) MARK 1 LABEL <osloveni-nok>
```

Výklad: Pokud se ve větě vyskytuje slovo uložené v proměnné `$OSLOVENI` následované slovem odpovídajícím tagu, označ ho štítkem `<osloveni-nok>`.

Příklad: *Tak vy jste si,* pane Janáček*, přinesl i fotografický přístroj.*

Dalším z příkladů hyperkorektnosti je koncovka *-é* u plurálu životných maskulin zakončených na *-nt* namísto koncovky *-i*. V tomto případě jsme využili *pojmenovanou proměnnou* `$HYPER-E`, do které jsme vypsali možné varianty slov. Slova jsme získali pomocí frekvenční analýzy na korpusech SYN2015 a czTenTen12.

```
TMPL: $HYPER-E MARK 0 LABEL <hyper-e>
```

Výklad: Pokud se ve větě vyskytuje slovo uložené v proměnné `$HYPER-E`, označ toto slovo štítkem `<hyper-e>`.

Příklad: *Přestože Švýcarsko bylo v tomto zápase favoritem, francouzští* reprezentanté *znepříjemňovali hru svému soupeři, jak se jen dalo.*

Z hlediska špatné deklinace jsme se zabývali také slovem *datum*, které slyšíme často špatně užité v mluveném jazyce, odkud se přenáší do psané formy. Konkrétně jsme opět vytvořili *pojmenovanou proměnnou* `$DATUM`, do které jsme uložili všechny špatné varianty.

```
TMPL: $DATUM MARK 0 LABEL <datum-nok>
```

Výklad: Pokud je ve větě slovo odpovídající proměnné `$DATUM`, označ toto slovo štítkem `<datum-nok>`.

Příklad: *Přesné* datumy *jsou vyvěšeny v občerstvení na nástěnce.*

Nekorektní užití se nevyhýbá ani předložkám. Kromě již zmíněného pořadí dvou předložek vedle sebe jsme se věnovali také špatné vazbě předložky *mimo* s genitivem. Vzniká kvůli záměně s předložkou *kromě*, která se právě s genitivem váže.

```
TMPL: (word mimo) (tag k1.*c2.*) MARK 1 LABEL <mimo-nok>
```

Výklad: Pokud se ve větě nachází slovo *mimo* následované slovem odpovídajícím tagu, označ druhé slovo štítkem `<mimo-nok>`.

Příklad: *Klause označuje za Cikána, v minulém díle tak častoval* mimo Čechů *i další národy východní Evropy.*

5.11 Řešení pro délku vět

Délku věty by bylo možné řešit i pomocí analyzátoru SET, avšak toto řešení by nebylo elegantní. Rozhodli jsme se proto navrhnout řešení v programovacím jazyce Python 2.7, které spočítá počet slov ve větě. Ukázka kódu je na obrázku 1 (str. 39), celý program přikládáme v příloze B. V rámci našeho kódu jsme hranici stanovili na čtyřiceti slovech ve větě.

Program počítá se stejným vstupem, jaký zpracovává SET, tedy s vertikálem. Využívá jeho struktury a řídí se znaky `<s>` a `</s>` pro začátek a konec věty. Nejprve se do proměnné `corpora` načtou data ze souboru. Program následně prochází soubor řádek po řádku; pro každý řádek, který se nachází mezi znaky `<s>` a `</s>` a zároveň neobsahuje tečku, čárku, středník ani znak `<g/>`, uloží slovo na řádku do proměnné `sentence` a zvedne číslo v proměnné `line_count` o jedna. Ve chvíli, kdy nalezne značku pro konec věty `</s>` a kdy je zároveň `line_count` delší než 40, uloží proměnnou `sentence` do pole `result` a vynuluje proměnné `sentence` a `line_count`. Pokud je počet menší jak 40, proměnná se do pole neuloží. Na konci programu se vypíší všechny věty zaznamenané do pole `result` spolu s počtem slov.

```
for line in corpora:
    if not count:
        if "<s>" or "<s desamb=\"1\">" in line:
            count = True
            continue
    if count:
        if not "</s>" in line:
            if not "<g/>" in line:
                if not line.startswith(".") and not line.startswith(",") \
                   and not line.startswith(";"):
                    sentence += line.split()[0]
                    sentence += ' '
                    line_count +=1
        else:
            if line_count > 40:
                result.append ([sentence, line_count])
            sentence = ''
            line_count = 0
            count = False
long_sentence = json.dumps(result, ensure_ascii=False)
print long_sentence
```

Obrázek 1. Ukázka kódu.

Příklad výstupu:

```
[["Principiálně probíhá financování buď ze soukromých zdrojů a veřejných
prostředků formou stálé ekonomické podpory vybraných subjektů a vypisováním
grantů v případech veřejných prostředků na základě přijatých zásad a speci-
fických kritérií ", 30]]
```

6 VYHODNOCENÍ

Zmínili jsme, jak se vytváří pravidla pro syntaktický analyzátor SET, popsali jsme, jaká pravidla jsme vytvořili, a v poslední části naší práce se budeme věnovat vyhodnocení. Nejprve uvedeme, jak vypadá testovací sada dat, a poté popíšeme, jakým způsobem jsme tuto sadu dat zpracovali. Zmíníme také, jakým způsobem se počítá úspěšnost. Následně analyzujeme výsledky.

6.1 Korpus

Webová příručka pro práci s korpusy Českého národního korpusu uvádí, že „jazykový korpus je rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednotném formátu tak, aby bylo možné v něm jednoduše vyhledávat různé jazykové jevy – zejména slova a slovní spojení (kolokace).“ (Cvrček, 2014) Velikost korpusu se odvíjí od počtu tokenů, tedy od nejmenších jednotek textu (zpravidla slova).

Během tvorby našeho testovacího korpusu jsme využili tří zdrojů. První z nich je korpus SYN2015, který spravuje pracoviště Českého národního korpusu na Filozofické fakultě Univerzity Karlovy. Jedná se o korpus psané češtiny, texty byly vytištěny a veřejně publikovány. Skládá se ze tří typů literatury: beletrie, oborové literatury a publicistiky. Celkově zahrnuje přes 120 milionů tokenů. Druhým zdrojem je korpus czTenTen12, který obsahuje texty stažené z internetu v letech 2010 a 2011. Dostupný je přes rozhraní Sketch Engine a obsahuje přes 5 miliard tokenů. Jako třetí zdroj je využito malé množství vět, které během své korektorské praxe zachytila autorka práce. Čtvrtým a posledním zdrojem sada náhodných článků ze serveru Novinky.cz.

Vznikly dva korpusy, jeden testovací s větami, které obsahují chyby, a druhý kontrolní, který by žádné chyby obsahovat neměl. Testovací korpus označený jako `data-nok` obsahuje celkem 13 626 tokenů. Skládá se z chybných vět, které jsme našli v obou zmíněných korpusech a ze zdrojů autorky práce.

Kontrolní korpus, který je pojmenovaný jako `data-ok`, obsahuje celkem 28 169 tokenů. Složili jsme ho z vět nalezených v korpusu, které neobsahují chyby, z opravených vět z testovací sady a několika desítek vět ze serveru Novinky.cz.

Oba korpusy přikládáme k práci v příloze C, v označkované i neoznačkované verzi.

6.2 Zpracování dat

Během přípravy vyhodnocení jsme pracovali na fakultním serveru `aurora`, protože jsou na něm k dispozici všechny tři nástroje, které jsme k práci potřebovali. Korpusy jsme nejdříve předali nástroji unitok, který věty rozdělil na jednotlivé tokeny. Následně jsme je pomocí nástroje ajka automaticky označkovali a pomocí nástroje desamb zjednoznačnili.

Příkazem `cat` jsme načetli data ze souboru `_data-ok.txt/_data-nok.txt` a následně zavolali nástroj unitok, kterému jsme nastavili jazyk na češtinu. Poté jsme zavolali nástroje ajka a desamb, přičemž výstup jsme přesměrovali do souboru `data-ok.txt/data-nok.txt`.

```
cat _data-ok.txt | /nlp/projekty/set/unitok.py --language=czech | /cor-
pora/programy/desamb.utf8.ajka.sh > data-ok.txt
```

Získali jsme dva soubory ve tvaru vertikálu, tedy souboru, který má na každém řádku jedno slovo či znak interpunkce. Díky majce jsme pro každé slovo získali také jeho lemma a značku. Nyní jsme měli připravená data pro syntaktický analyzátor SET. Zavolali jsme ho následujícím příkazem s přepínačem `--grammar`, který nám umožnil využít vlastní sadu pravidel `pravidla.txt`, a korpusem `data-ok.txt/data-nok.txt`. Výstup analýzy jsme uložili do souboru `vysledek-ok.txt`.

```
/nlp/projekty/set/set/set.py --grammar ./pravidla.txt data-ok.txt > vysle-
dek-ok.txt
```

Výsledný soubor uvádí ve čtvrtém sloupci štítek určený pravidly, pokud jim věta odpovídala.

Příklad výstupu:

```
0     Přeci        11    p     <hovorove>
1     jenom        12    p
2     však         12    p
3     má           12    p
4     toto         12    p
5     intermezzo   12    p
6     jeden        12    p
7     pozitivní    10    p     <pleonasmus>
8     klad         10    p     <pleonasmus>
9     .            12    p
10    <sentence>   -1    p
```

6.3 Výpočet úspěšnosti

K vyhodnocení úspěšnosti automatických nástrojů se nejen v informatice využívají dva vzorce: přesnost (*precision*) a výtěžnost pokrytí (*recall*). Přesnost se určuje jako procentuální poměr relevantních výsledků analýzy ke všem výsledkům analýzou získaným. Pokrytí je oproti tomu definováno jako poměr relevantních výsledků analýzy ke všem relevantním výskytům ve zkoumaném vzorku bez ohledu na to, zda byly analýzou identifikovány. (Cvrček, 2013)

K vyjádření vztahu mezi přesností a pokrytím se využívá F-míra (či F-skóre).

$$\text{Přesnost} = \frac{TP}{TP + FP}$$

$$\text{Pokrytí} = \frac{TP}{TP + FN}$$

$$\text{F-míra} = 2 \times \frac{\text{přesnost} \times \text{pokrytí}}{\text{přesnost} + \text{pokrytí}}$$

V těchto vzorcích TP znamená *True Positive*, tedy počet správně identifikovaných jednotek; FP znamená *False Positive* a vyjadřuje počet špatně identifikovaných jednotek; FN znamená *False Negative* a označuje počet jednotek, které nebyly rozpoznány.

6.4 Vyhodnocení

Podrobně jsme prošli oba výsledné soubory a spočítali jsme pro každé pravidlo, či sadu pravidel, výskyt TP a FP, ze kterých jsme následně vypočítali přesnost, pokrytí a F-míru. Celkové počty uvádíme v tabulce 1 (str. 43). Pokud bylo ve větě více chyb stejného druhu, tedy například několik označení `<priklonka-nok>`, počítali jsme ji vždy jen jednou.

Nejdřív jsme se věnovali vyhodnocení pravidel pro příklonky. Jak je možné si všimnout v tabulce, pravidla pro špatnou pozici příklonky mají velice malou přesnost, nalezla v sadě `data-ok` velké množství vět, ve kterých chyba ve skutečnosti nebyla. Pokrytí je naopak naopak vysoké, protože pravidla v sadě `data-nok` objevila všechny věty, ve kterých chyba byla. Pro porovnání jsme do tabulky uvedli výsledky pro pravidla `<priklonky-ok>`, který je oproti tomu mnohem lepší. Tato pravidla odhalila všechny věty, v nichž se příklonka vyskytovala na správné pozici, a nehlásila žádnou falešnou chybu. Pro další využití

pravidel se nabízí možnost zaměřit se spíše na pravidla, která příklonku ve správné pozici odhalí.

Pravidlo	TP	FP	Přesnost	Pokrytí	F-míra
<priklonky-ok>	136	0	1	1	1
<priklonky-nok>	41	309	0,117	1	0,210
<predlozky-nok>	35	5	0,875	0,921	0,897
<ukazovaci-nok>	59	1	0,983	0,983	0,983
<opakovani-nok> <vata-nok>	46	0	1	1	1
<pleonasmus>	131	0	1	0,885	0,939
<superlativ-nok>	11	0	1	1	1
<jakykoli-nok>	20	0	1	1	1
<bud-nok> <sice-nok> <jednak-nok>	54	19	0,740	0,885	0,806
<ktery-nok>	1	3	0,250	0,1	0,143
<rodktery-ok>	151	1	0,993	1	0,997
<rodktery-nok>	37	133	0,218	0,925	0,352
<hovorove>	16	8	0,667	0,941	0,780
ostatní[8]	117	0	1	1	1

Tabulka 1. Vyhodnocení pravidel.

Dále jsme pokračovali vyhodnocením pravidel pro předložky. Dosáhli jsme celkem vysoké přesnosti i pokrytí. Mezi špatně rozpoznané výrazy patřilo spojení *vzhledem k*. Pravděpodobně by se šlo této chyby vyvarovat vytvořením vlastního pravidla pro toto spojení a nastavením dostatečné váhy. Během tvorby pravidel jsme tento problém nezachytili, poněvadž v našem malé testovacím vzorku ajka označkovala *vzhledem* jako podstatné jméno, nikoli jako předložku.

[8] Pod tuto položku patří všechna pravidla popsaná v části 0.

Výsledek pravidla pro ukazovací zájmena vyšel rovněž dobrý. U pravidla jsme stanovili hranici na dvě a více ukazovacích zájmen v jedné větě, avšak je na diskuzi, zda nebýt do budoucna mírnější.

Pravidla pro opakování stejných slov a slovní vatu dopadla výborně. U pravidel pro slovní vatu jsme byli striktní, hledali jsme každý výskyt. Otevírá se tady možnost pro rozčlenění nástroje na jednotlivé styly – v prostěsdělovacím stylu je možné tato slova tolerovat, avšak v rámci odborného stylu nikoli.

Bez chyby dopadla také detekce pleonasmů. V sadě `data-nok` se ovšem vyskytlo 17 výrazů, které pravidla nezvládla najít, proto není pokrytí stoprocentní.

Výborně dopadla i další dvojice pravidel, a to pravidla pro detekci superlativů a špatné vazby slova *jakýkoli*.

Další sadou pravidel bylo hledání zapomenuté dvojice spojky *buď* (*nebo*), *sice* (*ale*) a chybného tvaru **druhak*. Dosáhli jsme uspokojivého výsledku. Většina domnělých chyb je způsobena velkou vzdáleností mezi oběma spojkami.

Špatný výsledek pravidla pro opakování zájmena *který* a *jenž* je dán překryvem s pravidlem pro detekci správného a špatného rodu zájmena *který* a *jenž*. Z deseti špatných vět byla označena pouze jedna a k tomu pravidlo označilo tři věty, které byly správně. Pravděpodobně by kolizi bylo možné vyřešit upravením váhy pravidel. Ve fázi tvorby pravidel jsme si tento překryv bohužel neuvědomili, protože jsme testovali každou kategorii zvlášť. Pro kontrolu jsme pravidla otestovali znovu samostatně (tedy bez přítomnosti ostatních pravidel v gramatice). Na tomto menším vzorku dat jsme dosáhli stoprocentní přesnosti a šedesátiprocentního pokrytí.

U sady pravidel, která rozhodují o správnosti zájmena *který* a *jenž*, jsme zaznamenali podobný výsledek jako u pravidel pro příklonky, tedy že pravidla pro správný rod mají vysokou úspěšnost, zatímco pravidla pro špatný rod mají malou úspěšnost. Částečně je to způsobeno tím, že ajka určila špatný rod zájmena během značkování, tudíž následně neodpovídalo pravidlům. Pokrytí je opět vysoké, protože pravidla odhalila většinu vět, ve kterých byl rod správně.

Uspokojivý je výsledek pro detekci hovorových slov či jejich tvarů. Mezi slovy určenými špatně jsme našli například tvary *hospoda*, *hospodě*, *jarmarku*. SSČ hodnotí slovo

jarmark jako obecné, u slova *hospoda* však žádný stylový příznak neuvádí, pouze v SSJČ je toto slovo hodnoceno jako obecné a expresivní.

Všechna zbylá pravidla jsme zahrnuli do kategorie ostatní. Jednalo se především o vyhledávání konkrétních slov či tvarů. U těchto pravidel jsme dosáhli stoprocentní úspěšnosti.

Shrneme-li vyhodnocení jako celek, tak jednoduchá pravidla dosáhla výborných výsledků. U složitějších pravidel jsou výsledky podle očekávání horší. U pravidel pro příklonky a shody zájmena *který* a *jenž* s rodem a pádem jména, ke kterému se váže, bude třeba vyzkoušet ještě jiný přístup. Před dalším využitím pravidel bude potřeba rozsáhlejší testování.

ZÁVĚR

Každý pisatel, tedy každý člověk, má svůj vlastní styl. A ať už je tento styl jakýkoli, měl by dodržovat určitá pravidla. Opravy stylistiky jsou značně subjektivní záležitost, avšak existují v češtině pravidla, která musí být dodržena nehledě na autorský záměr. Cílem diplomové práce bylo s pomocí syntaktického analyzátoru SET automaticky odhalit prohřešky, kterých se autoři textů dopouštějí. Práce je součástí většího celku, jehož cílem je vytvořit nový korektor češtiny.

V teoretické části práce jsme se věnovali popisu stylistiky a funkčních stylů. Dále jsme vysvětlili rozdíly mezi jednotlivými druhy automatických korektorů a zpracovali přehled současných nástrojů. Poslední částí teoretického podkladu byla analýza chyb, které můžeme v psaných komunikátech najít, včetně jejich lingvistického vysvětlení.

Teoretickou část jsme začali popisem nástrojů, které jsme využili pro práci, přičemž jsme kladli důraz na syntaktický analyzátor SET. Následně jsme podrobně rozebrali pravidla pro tento nástroj, která jsme vytvořili. Nezapomněli jsme ani na důkladné testování a vyhodnocení výsledků. Většina pravidel uspěla s výbornými výsledky.

V rámci práce jsme se věnovali stylům obecně, s vynecháním stylu rétorického a uměleckého. Tady se nabízí možnost rozšíření, kterou jsme naznačili již u vyhodnocení. V rámci některých funkčních stylů lze hledané prvky tolerovat, např. u prostěsdělovacího stylu můžeme povolit výskyt slov *prostě* a *vlastně*, u kterých máme momentálně nastavenou nulovou toleranci, jelikož v odborném stylu by se tato slova vyskytovat neměla. Také by například bylo vhodné detekovat deminutiva, která nepatří do odborného stylu. Co se týče hovorovosti, bylo by zajímavé propojení se SSČ, který obsahuje stylové příznaky pro jednotlivá slova. Například slovo s příznakem *expresivní* by se v textech odborného stylu vyskytovat nemělo.

Několikrát jsme v rámci práce zmiňovali posun slov na ose knižní – neutrální – hovorové. Tato změna se dotkne i našich pravidel, která bude potřeba upravovat podle aktuální kodifikace. Týkat se bude především pravidel pro hovorovost, ale také třeba pravidel pro pleonasmy. Pravidla pro pleonasmy bychom mohli rozdělit na dvě části: pleonasmy, které jsou nepřípustné, a pleonasmy, které už nejsou vnímány jako natolik chybné.

V programovacím jazyce Python jsme vytvořili program, který počítá slova ve větě. V rámci práce jsme nastavili hraniční počet na čtyřiceti slovech, avšak do budoucna bude vhodné upravit počet podle přiřazení ke stylu.

Současná pravidla jsou postavena tak, že řeší prohřešky v rámci jedné věty. Do budoucna by bylo dobré se zaměřit i na větší celky, odstavce i celý text, a podle toho upravit například pravidla pro detekci ukazovacích zájmen a opakování slov.

BIBLIOGRAFIE

CVRČEK, Václav, 2013. Precision a recall. In: *Příručka ČNK* [online]. Praha [cit. 2018-06-25]. Dostupné z: https://wiki.korpus.cz/doku.php/pojmy:precision

CVRČEK, Václav, 2014. Co je korpus?. In: *Příručka ČNK* [online]. Praha [cit. 2018-06-19]. Dostupné z: http://wiki.korpus.cz/doku.php#co_je_korpus

HAJIČOVÁ, Eva, 2017. Aktuální členění větné. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-02-04]. Dostupné z: https://www.czechency.org/slovnik/AKTU%C3%81LN%C3%8D%20%C4%8CLEN%C4%9AN%C3%8D%20V%C4%9ATN%C3%89

JELÍNEK, Milan, 2012. Stylistika. KARLÍK, Petr, ed., Marek NEKULA, ed. a Zdenka RUSÍNOVÁ, ed. *Příruční mluvnice češtiny*. Vyd. 2., opr. [i.e. 4. vyd.]. Praha: Nakladatelství Lidové noviny, s. 699-780. ISBN 9788071066248.

JELÍNEK, Milan a Marie KRČMOVÁ, 2017a. STYLISTIKA. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy: Nový encyklopedický slovník češtiny* [online]. [cit. 2018-04-29]. Dostupné z: https://www.czechency.org/slovnik/STYLISTIKA

JELÍNEK, Milan a Marie KRČMOVÁ, 2017b. Stylový příznak. In: KARLÍK, Petr, Marek NEKULA a Jana PLESKALOVÁ. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-10]. Dostupné z: https://www.czechency.org/slovnik/STYLOV%C3%9D%20P%C5%98%C3%8DZNAK

KARLÍK, Petr, 2002. Analýza syntaktická. KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *Encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny, s. 40. ISBN 807106484x.

KARLÍK, Petr, 2017. Atrakce. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-05-29]. Dostupné z: https://www.czechency.org/slovnik/ATRAKCE

KONEČNÁ, Hana, 2014. Podstatná jména, která mají před koncovým -um souhlásku. PRAVDOVÁ, Markéta, ed. a Ivana SVOBODOVÁ, ed. *Akademická příručka českého jazyka*. První. Praha: Academia, s. 303. ISBN 978-80-200-2327-8.

KOVÁŘ, Vojtěch, 2008. *Syntaktická analýza s využitím postupné segmantace věty*. Brno. Diplomová práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Aleš Horák.

KOVÁŘ, Vojtěch, 2014. *Automatic Syntactic Analysis for Real-World Applications* [online]. Brno [cit. 2018-06-14]. Dostupné z: https://is.muni.cz/th/iadwb. Disertační práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Aleš Horák.

KRAUS, Jiří, 2005. *Nový akademický slovník cizích slov A-Ž*. 1. Praha: Academia. ISBN 80-200-1351-2.

KRČMOVÁ, Marie, 2017a. Aktualizace. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-05-26]. Dostupné z: https://www.czechency.org/slovnik/AKTUALIZACE

KRČMOVÁ, Marie, 2017b. Metateze. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-15]. Dostupné z: https://www.czechency.org/slovnik/METATEZE

KŘÍSTEK, Michal, 2017. STYL. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-04-15]. Dostupné z: https://www.czechency.org/slovnik/STYL

KUČERA, Ondřej, 2006. Pražský závislostní korpus jako elektronická cvičebnice jazyka českého. In: *Proceedings of the 4th Student Research Competition in Informatics and Information Technologies (finalists papers)*. Praha, 41–47. Dostupné také z: http://ufal.mff.cuni.cz/styx/doc/styx-acm.pdf

LINGEA, 2018a. Dělení slov. *Lingea* [online]. Lingea [cit. 2018-05-30]. Dostupné z: https://www.lingea.cz/deleni-slov

LINGEA, 2018b. Korektor překlepů. *Lingea* [online]. Lingea [cit. 2018-05-30]. Dostupné z: https://www.lingea.cz/korektor-preklepu

Lingea Grammaticon, 2018. In: *SW.cz: Specialista na software* [online]. [cit. 2018-06-26]. Dostupné z: https://www.sw.centrum.cz/vyuka-a-vzdelavani/vyuka/lingea-grammaticon/

MICHÁLEK, Zbyněk, 2017. *Algoritmizace hromadných oprav vybraných typograficko-pravopisných jevů českého jazyka* [online]. Brno [cit. 2018-06-13]. Dostupné z:

https://is.muni.cz/th/tqyxh/. Bakalářská práce. Masarykova univerzita, Filozofická fakulta. Vedoucí práce Vít Baisa.

PALA, Karel, 2005. Pište dopisy bez chyb! – Český gramatický korektor pro Microsoft Office. *Computer*. **7**(13-14), 72.

PALA, Karel, 2017a. Pravopisný korektor. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-05-30]. Dostupné z: https://www.czechency.org/slovnik/PRAVOPISN%C3%9D%20KOREKTOR

PALA, Karel, 2017b. Gramatický korektor. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-14]. Dostupné z: https://www.czechency.org/slovnik/GRAMATICK%C3%9D%20KOREKTOR

PALA, Karel, 2017c. Stylistický korektor. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-14]. Dostupné z: https://www.czechency.org/slovnik/STYLISTICK%C3%9D%20KOREKTOR

PETKEVIČ, Vladimír, 2014. Kontrola české gramatiky (český grammar checker). *Studie z aplikované lingvistiky*. Univerzita Karlova v Praze, Filozofická fakulta, **5**(2), 48–86, 175 s. ISSN 2336-6702.

PETKEVIČ, Vladimír, 2017a. Morfologická analýza. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-15]. Dostupné z: https://www.czechency.org/slovnik/MORFOLOGICK%C3%81%20ANAL%C3%9DZA

PETKEVIČ, Vladimír, 2017b. Tokenizace. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-06-16]. Dostupné z: https://www.czechency.org/slovnik/TOKENIZACE

PRAVDOVÁ, Markéta, ed., 2012. *Jsme v češtině doma?*. První. Praha: Academia. Lingvistika (Academia). ISBN 978-802-0021-465.

RICHTER, Michal, Pavel STRAŇÁK a Alexandr ROSEN, 2012. Korektor – A System for Contextual Spell-checking and Diacritics Completion. In: KAY, Martin, ed. a Christian

BOITET, ed. *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, India: Coling 2012 Organizing Committee, 1–12.

SVOBODA, Aleš, 1984. České slovosledné pozice z pohledu aktuálního členění. *Slovo a slovesnost*. **45**(1), 22–34.

TETREAULT, Joel, 2018. Under the Hood at Grammarly: Transforming Writing Style with AI. In: *Grammarly blog* [online]. [cit. 2018-06-12]. Dostupné z: https://www.grammarly.com/blog/transforming-writing-style-with-ai/

UHLÍŘOVÁ, Ludmila a Ivona KUČEROVÁ, 2017. Slovosled. In: KARLÍK, Petr, ed., Marek NEKULA, ed. a Jana PLESKALOVÁ, ed. *CzechEncy - Nový encyklopedický slovník češtiny* [online]. [cit. 2018-02-04]. Dostupné z: https://www.czechency.org/slovnik/SLOVOSLED