Kontrola české gramatiky (český grammar checker)[1]

Vladimír Petkevič

ABSTRACT:

A Grammar Checker for Czech. The paper presents a detailed description of the linguistic software system *Kontrola české gramatiky* (*Grammar Checker for Czech*) that checks grammatical, orthographic and stylistic correctness of a text written in Czech, highlights the errors encountered and offers a user corrections of the errors committed. The Grammar checker for Czech has been integrated in the Microsoft Word software product within the system Microsoft Office™ version 2003 (since 2005) and subsequently 2007, 2010 and 2013. The paper focuses on the history of the project, presents ideas on which the Checker is based, describes how the Checker can be utilized including interactive error corrections, presents a detailed description of the errors detected and an algorithm of the Checker's processing an input text. It is not only external aspects of the grammar checking system that are depicted, i.e. how the Checker presents itself to the user and how the user should use it, but also the whole conceptual basis of the system: what components it consists of and how they cooperate. The core of the Checker is constituted by formalized grammatical (especially syntactic), orthographic and stylistic rules detecting mainly grammatical and also some orthographic a stylistic errors in Czech texts. The rules use primarily the results of automatic part-of-speech and morphological analysis and morphological disambiguation of Czech texts and also morphological synthesis. An error is conceived of as a violation of the grammar and orthography of contemporary standard written Czech, the grammar being expressed by the rule-based system. The Checker deals with errors of spelling only in special grammatically motivated cases and in those cases where the people usually make mistakes; a full-fledged spell checker had already been integrated in the Microsoft Office package before the emergence of the Checker and a new one was not needed. In the conclusion, the success rate of the Checker is briefly compared to the *Grammaticon* system, a survey of positive and negative aspects of the Checker is presented and future directions of its potential further development are shown.

KLÍČOVÁ SLOVA / KEY WORDS:

formalizovaná gramatická pravidla, formalizovaná pravopisná pravidla, formalizovaná stylistická pravidla, gramatická chyba, kontrola české gramatiky, lingvistický softwarový systém, morfologická analýza, morfologická disambiguace, morfologická syntéza, pravopisná chyba, stylistická chyba

formalized grammatical rules, formalized orthographic rules, formalized stylistic rules, grammar checker for Czech, grammatical errors, linguistic software system, morphological analysis, morphological disambiguation, morphological synthesis, spelling errors, stylistic errors

1 Tento příspěvek vychází s finanční podporou Programu rozvoje vědních oblastí na Univerzitě Karlově v Praze, č. P11, s názvem Český národní korpus, řešeného na Filozofické fakultě Univerzity Karlovy v Praze. Autor děkuje za veškeré připomínky Tomáši Jelínkovi, Pavlu Květoňovi, Karlu Palovi a Pavlu Straňákovi.

0. ÚVOD

Jednu z užitečných aplikací počítačové lingvistiky představují programy prověřující formální (ortografickou) a gramatickou správnost vstupního psaného textu podle pravidel platných v daném jazyce, například češtině, angličtině, francouzštině. Hlavními softwarovými systémy náležícími do této kategorie jsou:

— systémy pro kontrolu pravopisu (angl. spell(ing) checking, program se nazývá spell(ing) checker),
— systémy pro kontrolu gramatické správnosti (grammar checking, program se nazývá grammar checker).

Vyvinout program pro jednoduchou kontrolu pravopisu, který by měl odhalit neexistující slova v daném jazyce, je z lingvistického hlediska poměrně snadné: stačí mít k dispozici (i) soupis všech existujících slovních tvarů příslušného jazyka a efektivně jej uspořádat (je však třeba mít na paměti, že slovní zásoba každého jazyka se mění!) a (ii) tokenizační program (tokenizér), který dokáže rozdělit vstupní text na jednotlivé tokeny včetně interpunkčních znamének.

Kontrola pravopisu však může být podstatně ambicióznější. Místo pouhého soupisu platných slovních tvarů je možné použít morfologického analyzátoru používajícího slovní kořeny/kmeny a morfologická paradigmata (zejména pro flektivní jazyky, mezi něž čeština patří) a upravit takový analyzátor pro účely kontroly pravopisu. Kontrola může při své činnosti zkoumat také okolí zkoumaného slova a přitom nejen hlásit nesprávně zapsaná slova, ale také upozorňovat na lokální gramatické chyby dané negramatickými souvýskyty slov, která však sama v jazyce existují; obecně se takovým souvýskytům *n* slov říká negramatické n-gramy, nebo častěji negativní (gramatické) *n*-gramy, nejčastěji jsou to dvojice (bigramy) či trojice (trigramy), např. *do střeše*. Kontrola pravopisu se však nemusí spokojit s pouhým upozorněním na chybné slovo či slovní spojení: může navíc iniciativně navrhovat vhodnou opravu. Tak si počíná např. stochastický spell checker, který popisují Richter et al. (2012). Ten navrhuje opravu chybného slova na základě stochasticky vyhodnoceného kontextu, a blíží se tak kontrole gramatiky.

Ze softwarového hlediska je v každém případě nezbytné, aby spell checker byl natolik rychlý, jak to odpovídá lingvisticky relativně nenáročné úloze (v porovnání s netriviální prověrkou gramatičnosti textu): jakmile se uživateli objeví na obrazovce stránka textu, měl by program prakticky okamžitě zvýraznit chybné slovní tvary a nabídnout je k opravě, případně navrhnout, jak chybu opravit.

Prověřovat gramatickou správnost je úkol výrazně obtížnější, a to jak z hlediska lingvistického, tak softwarového. Toto tvrzení se pokusíme doložit v tomto článku, který se této problematice podrobně věnuje. Pro češtinu existuje několik programů pro kontrolu pravopisu a — pokud je nám známo — dva systémy určené výhradně pro prověrku české gramatiky:

— *Kontrola české gramatiky* vyvinutá v Ústavu pro jazyk český AV ČR, v. v. i.,
— systém Grammaticon vyvinutý brněnskou společností Lingea.

V tomto článku se budeme zabývat pouze systémem *Kontrola české gramatiky* (dále téměř vždy jen Kontrola), jen v závěru jej srovnáme se systémem Grammaticon. Kontrola bude popsána v těchto oddílech: 1. (pre)historie vzniku *Kontroly české gramatiky*, kolektiv jejích autorů; 2. metodologická východiska a zásady, na nichž je *Kontrola české gramatiky* založena; 3. používání *Kontroly české gramatiky*; 4. sledované gramatické, pravopisné, stylistické a formální chyby; 5. algoritmus zpracování vstupního textu; 6. úspěšnost *Kontroly české gramatiky*, srovnání se systémem Grammaticon; 7. shrnutí pozitivních a negativních rysů *Kontroly české gramatiky*; 8. možnosti dalšího rozvoje *Kontroly české gramatiky*.

1. (PRE)HISTORIE VZNIKU *KONTROLY ČESKÉ GRAMATIKY* A KOLEKTIV JEJÍCH AUTORŮ

Kontrola české gramatiky byla vyvinuta pro editor Microsoft Word a integrována do systému Microsoft Office™ (v dalším textu nebudeme značku ™ u společnosti Microsoft ani u jejích produktů nadále uvádět). Brněnská společnost Lingea vyvinula krátce po roce 2000 pro Microsoft Office program pro kontrolu pravopisu a usilovala o získání zakázky i na kontrolu (korektor) české gramatiky: hodlala uspět se svým systémem Grammaticon. Kontrakt se společností Microsoft však nebyl uzavřen, a tak se společnost Microsoft obrátila na Ústav pro jazyk český AV ČR, v. v. i., (dále ÚJČ AV ČR) a výsledkem byla smlouva obou subjektů. ÚJČ AV ČR najal na vývoj Kontroly externí spolupracovníky, kteří ji vyvíjeli od roku 2004 do května roku 2005. První verzi Kontroly dodal ÚJČ AV ČR společnosti Microsoft v podobě zakódovaného binárního souboru s definicí gramatických, pravopisných a stylistických pravidel a softwarovou knihovnou (DLL) pro jejich interpretaci a poté maďarská společnost Morphologics implementovala rozhraní mezi DLL a systémem Microsoft Office 2003. V tomto systému Kontrola funguje v editoru Microsoft Word, mimo něj ji nelze používat. Kontrola gramatiky byla v tomto systému v provozu od 1. července 2005, a kdo systém Microsoft Office 2003 v té době vlastnil, mohl si Kontrolu stáhnout z internetu zdarma jako rozšíření funkčních možností Microsoft Office 2003. Po první verzi Kontroly následovala v roce 2007 verze druhá, která je už začleněna do dalších verzí systému Microsoft Office: do verze 2007, 2010 a 2013. Uživatel tak má k dispozici dva systémy gramatické kontroly textu psaného v editoru Microsoft Word, které jsou naprosto nezávislé

— spell checker firmy Lingea,
— Kontrolu české gramatiky

a oba kontrolní systémy může zapínat nezávisle na sobě.

Autory Kontroly jsou pracovníci těchto institucí (s výjimkou níže uvedené maďarské společnosti uvádíme současné instituce autorů k 18. 5. 2014, nikoli instituce v době vzniku Kontroly):

— Ústav pro jazyk český AV ČR, v. v. i. (ÚJČ AV ČR),
— Ústav teoretické a komputační lingvistiky FF UK v Praze (ÚTKL FF UK v Praze),

— Ústav formální a aplikované lingvistiky MFF UK v Praze (ÚFAL MFF UK v Praze),
— IBM Česká Republika,
— maďarská firma Morphologics.

Níže uvádíme konkrétní autory spolu s náplní jejich práce na Kontrole (tituly autorů jsou dnešní): doc. RNDr. Karel Oliva, Dr., (ÚJČ AV ČR): institucionální podpora, styk se společností Microsoft, koncepce systému; RNDr. Pavel Květoň, Ph.D., (IBM Česká republika): koncepce systému, tvorba softwaru na podporu pravidel automatické morfologické disambiguace češtiny a pravidel pro kontrolu gramatiky, pravopisu a stylistiky; Mgr. Tomáš Jelínek, Ph.D., (ÚTKL FF UK v Praze): koncepce systému, vývoj lingvistických pravidel; doc. RNDr. Vladimír Petkevič, CSc., (ÚTKL FF UK v Praze): koncepce systému, vývoj lingvistických pravidel; RNDr. Milena Hnátková, CSc., (ÚTKL FF UK v Praze): správa morfologického slovníku češtiny, vytváření, správa a implementace slovních spojení; prof. RNDr. Jan Hajič, Ph.D., (ÚFAL MFF UK v Praze): vytvoření morfologického slovníku češtiny a morfologické analýzy češtiny; Maďarská firma Morphologics: implementace rozhraní mezi softwarovou knihovnou (DLL) a systémem Microsoft Office.

Kontrolu důkladně testovali: doc. RNDr. Vladislav Kuboň, Ph.D., (ÚFAL MFF UK v Praze); doc. PhDr. Karel Pala, CSc., (Centrum zpracování přirozeného jazyka, Fakulta informatiky Masarykovy univerzity), který navíc mimo testování Kontroly provedl její srovnání se systémem Grammaticon.

2. METODOLOGICKÁ VÝCHODISKA A ZÁSADY, NA NICHŽ JE *KONTROLA ČESKÉ GRAMATIKY* ZALOŽENA

Kontrola vychází z preskriptivní gramatiky a pravopisu spisovné češtiny (srov. MČ2; MČ3; PMČ; Havránek — Jedlička, 1981; Šmilauer, 1966; 1972; Pravidla českého pravopisu, 1993), v malé míře se zabývá i problematikou stylu. Snaží se v textu nalézt gramatické a pravopisné chyby, přičemž chyba je definována jako *prohřešek* vůči této gramatice a pravopisu. Například ve větě

(1) Přes obtíže, které vyvolává, je jeho řešení ve *filosofickému* oblasti jasné a přesvědčivé.

je dativní slovní tvar *filosofickému* gramaticky nesprávný, protože nevyhovuje ani akuzativní, ani lokálové rekci předložky *ve*. Takovéto chyby má Kontrola uživateli hlásit a pokud možno navrhnout uživateli náležitou opravu. Uveďme ještě jiný příklad:

(2) Kdyby demokracie užívala prostředky základně neslučitelné *se* mravními požadavky, pracovala by na své autodestrukci.

Ve větě (2) má předložka *se* vokalizovanou podobu, ač těsně za ní stojí slovní tvar, který vyžaduje užít výhradně podoby nevokalizované, tj. *s*. Je to tedy zjevná chyba a Kontrola by i v tomto případě měla tuto skutečnost ohlásit a navrhnout příslušnou opravu.

Kontrola se tedy zaměřuje hlavně na zjevné gramatické chyby a snaží se jemně oddělovat zrno (= gramaticky, pravopisně, příp. stylisticky správné věty) od plev (= gramaticky, pravopisně, příp. stylisticky nesprávné věty). To je ovšem obecně problematické, neboť existuje řada případů, kdy není jasné, zda jde o chybu, nebo ne. V takových případech (podrobněji níže) si Kontrola počíná spíše opatrně a permisivně: buď chybu vůbec neohlásí, nebo na její možnost upozorní, aniž by navrhovala nějakou opravu.

Kontrola se snaží pokrýt co nejvíce gramatických i pravopisných jevů v češtině, nehlásí však uživateli chyby v pravopisu jednotlivých slovních řetězců, pokud nejsou způsobeny gramatikou, čili např. nehlásí řetězce znaků, jež v češtině neexistují (to činí kontrola českého pravopisu, tedy spell checker společnosti Lingea; viz výše), i když je vnitřně samozřejmě odhalí v procesu morfologické analýzy dané věty (celý proces a algoritmus zpracování je podrobně popsán v oddílu 5 níže). Pokud se ovšem ve větě objeví slovní tvar, který v češtině neexistuje, zásadně to ztěžuje správnou činnost Kontroly: nachází-li se takové slovo mezi dvěma slovy spjatými nějakým gramatickým vztahem, nedokáže Kontrola takový vztah náležitě rozpoznat. Vypne-li si tedy uživatel funkci spell checkingu při práci s textem v programu Microsoft Word a objeví-li se v nějaké větě úsek slov obsahující neznámé slovo, nic se uživateli — alespoň co se týká tohoto úseku — neohlásí. Kontrola se tedy plně uplatní, když ve zkoumaném větném úseku rozpozná všechny slovní tvary. Aby tedy uživatel plně využil možností, jež Kontrola nabízí, měl by nejprve opravit všechny řetězce písmen, které v češtině neexistují, tedy většinou překlepy. Kontrola tak téměř veškerou kontrolu jednotlivých slov realizovanou nezávisle na kontextu přenechává spell checkeru společnosti Lingea a sama se zaměřuje na gramatické, formální (pravopisné) a v menší míře stylistické vztahy mezi existujícími českými slovy. Z povahy věci je tedy zřejmé, že její úkol je mnohem obtížnější než úkol spell checkeru, neboť se v co největší míře snaží postihnout syntagmatiku češtiny.

Vzhledem k požadavkům zadavatele, tj. společnosti Microsoft, volí Kontrola v poměru k chybám velmi konzervativní, opatrnický přístup. Hlásí uživateli jen vyložené chyby, a snaží se tak co nejméně hlásit chyby domnělé, konfigurace, jež sice mohou být chybné, ale nemusí. Není-li si jista, raději „mlčí", nebo případně oznámí uživateli, že danou konstrukci/větu lze interpretovat různým způsobem. Na druhé straně je však nedobré, když mlčí a nehlásí chyby tam, kde vskutku jsou: hranice tu bývá velmi úzká. Dovolíme si tu obraznou formulaci: Kontrola musí se ctí a bez úhony proplout mezi dvěma obludami v Messinské úžině: Skyllou, která Kontrolu nutí vidět všude chyby, tj. hlásit je i tam, kde nejsou, a Charybdou, jež Kontrole vnucuje shovívavou permisivitu. Stejně jako je Charybda méně nebezpečná než Skylla, je rovněž přehlédnutí chyby menším prohřeškem než hlášení chyby v gramaticky správné konstrukci: hlásit domnělé chyby (angl. *false flags*) v gramaticky správných větách je totiž to nejhorší, co se Kontrole může stát.[2]

2 Lze si ovšem představit, že opatrnost při přístupu k možným chybám by se dala parametrizovat: uživatel (typicky profesionální pracovník s jazykem) by si sám mohl určit, jaký přístup ke zkoumanému textu by měla Kontrola zvolit. Nemusel by třeba nutně preferovat přílišnou opatrnost, chtěl by naopak, aby Kontrola byla k textu přísnější a podezíravější a aby více riskovala, přičemž by mu nevadilo, že občas ohlásí chybu tam, kde není. Za tuto poznámku autor děkuje recenzentovi.

Dalším úskalím jsou konstrukce, které sice jsou sensu stricto gramatické, ale vzhledem k záměru autora textu gramaticky nesprávné: autor textu napsal z hlediska své intence gramaticky nesprávnou větu, ta je však ve skutečnosti (z jiných důvodů) gramatická; chyba, jíž se dopustil, tak může vést k odlišné, správné interpretaci. Například napíše-li uživatel větu

(3) Ženy *viděli*.

bude se třeba podivovat nad tím, že Kontrola nerozpozná ani tak jasnou chybu, jakou je chyba ve shodě přísudku s podmětem (podle něho je tu frapantní neshoda ve jmenném rodě!); takovou chybu je přece snadné odhalit, když se obě slova nacházejí těsně vedle sebe! Uživatel si však neuvědomí, že věta má jiný význam, než jaký uživatel chtěl světu sdělit, totiž

(3a) *Ženy*(Apl) viděli.

kde *ženy* jsou akuzativním předmětem přísudkového slovesného tvaru *viděli*, nikoli jeho podmětem (tj. je-li podmět lexikálně nevyjádřen).

Z tohoto příkladu je zřejmé, že Kontrola se řídí gramatickými pravidly platnými pro češtinu, v uvedeném příkladu např. jedním z několika pravidel o shodě přísudku s podmětem. Tato pravidla ve formalizované podobě tvoří jádro Kontroly. Jako další příklad uveďme větu:

(4) *Pak *od jí s* smutkem v duší odešel.

V této větě jsou hned tři chyby: (a) bezprostředně po předložce *od* nesmí stát osobní zájmeno začínající na *j* (správně *od ní*) (to platí až na speciální případy adjektivní skupiny zanořené do předložkové skupiny řízené předložkou, zde *od*); (b) předložka *s* není před slovem začínajícím na *s* (*smutkem*) vokalizována (správně *se smutkem*); a konečně (c) tvar *duší* je po předložce *v* nesprávný, neboť tato předložka vyžaduje jméno v akuzativu nebo lokálu (zde správně *duši*). Kontrola obsahující náležitě formalizovaná pravidla pro zachycení chyb typu (a)–(c) uživatele na tyto chyby upozorní a navrhne mu vhodné opravy věty, např.:

(4a) Pak od ní se smutkem v duši odešel.

Všechny sledované gramatické, pravopisné a stylistické chyby jsou podrobně rozebrány v oddíle 4.

3. POUŽÍVÁNÍ *KONTROLY ČESKÉ GRAMATIKY*

Velmi stručně nyní popíšeme, jak může uživatel s Kontrolou pracovat, a také představíme různé styly/mody Kontroly. *Kontrola české gramatiky* je podobně jako spell checker společnosti Lingea zabudována jako jeden z modulů do systému Microsoft

Office. Kontrola prověřuje text zapsaný v editoru Microsoft Word a hlásí zeleným vlnovkovitým podtržením chybných či podezřelých míst v textu na celé obrazovce prohřešky proti české gramatice/pravopisu/stylu. Kontrola se po spuštění editoru Word aktivuje takto:

> Nástroje → Jazyk → Nastavit jazyk → Čeština + nezaškrtnout volbu Neprovádět kontrolu pravopisu ani gramatiky

Kliknutím pravým tlačítkem myši na zeleně podtržený úsek lze zjistit podrobnosti o hlášené chybě, případně nastavit parametry, jež chce uživatel sledovat. Kontrola obecně nehlásí chyby v pravopisu jednotlivých slovních tvarů (činí tak jen v některých zvláštních případech na základě kontextu) — to je úkol souběžného a nezávisle fungujícího spell checkeru, který hlásí chyby červeným vlnovkovitým podtržením. U chybné věty Kontrola co nejpřesněji označí chybné či podezřelé místo, a klikne-li následně na toto místo uživatel pravým tlačítkem myši, pak Kontrola:

— Slovně popíše chybu.
— Většinou navrhne opravu.

Souhlasí-li uživatel s návrhem Kontroly, stačí na příslušný návrh kliknout a oprava se ihned provede. Pokud uživatel text nějak opraví — ať už podle návrhu Kontroly, nebo jinak —, danou větu Kontrola opět prověří a při nalezení další chyby opět podtrhne chybný úsek.

Když uživatel klikne na zeleně podtržený úsek, objeví se mu spolu s doporučením, jak chybu opravit, i volba *Gramatika*. Když na ni klikne, objeví se okno s podezřelým úsekem textu, přičemž v záhlaví okna je uveden obecný typ chyby (např. *Chyba ve shodě přísudku s podmětem*). Okno obsahuje i doporučení opravy a také nabídku podrobnějšího vysvětlení chyby (klikne-li se na volbu *Vysvětlit*). Mimoto se objeví i volba *Možnosti*. Klikne-li na ni uživatel, objeví se vpravo nové okno s podokněm *Styl dokumentu*, v němž si uživatel může vybrat jeden ze tří psacích stylů/modů:

— *Gramatika* — provede se kontrola gramatiky,
— *Gramatika+Styl* — provede se kontrola gramatiky i stylu,
— *Vlastní* — provede se kontrola jevů definovaných uživatelem.

Volba *Nastavení*, jež se uživateli také nabídne, mu po rozkliknutí umožní zjistit, jaké jevy Kontrola v závislosti na příslušném modu prověřuje. Celkem Kontrola umožňuje sledovat jevy / chybové kategorie, jež jsou podrobně popsány v oddílu 4, a standardně je také uživateli u obou prvních modů nabízí. Uživatel si navíc může u každého z modů sám zvolit jen ty kategorie, které chce sledovat.

4. SLEDOVANÉ GRAMATICKÉ, PRAVOPISNÉ, STYLISTICKÉ A FORMÁLNÍ CHYBY

Kontrola české gramatiky sleduje tyto tři kategorie chyb:

— chyby gramatické a pravopisné,
— chyby stylistické,
— chyby formální.

Zatím jsme se zmiňovali jen o chybách gramatických, pravopisných a stylistických, na konci tohoto oddílu popíšeme i chyby, jež nazýváme formální (lze je ovšem zařadit mezi (počítačově) pravopisné).

Podrobně nyní popíšeme chyby, jež Kontrola sleduje.

4.1 CHYBY GRAMATICKÉ A PRAVOPISNÉ

V modu *Gramatika* a *Gramatika+styl* Kontrola umožňuje odhalovat následující gramatické a pravopisné chyby:

— chyba ve jmenné skupině (shoda aj.),
— chyba týkající se sloves,
— chyba ve shodě u slovesa (ve shodě komponent víceslovného slovesného přísudku),
— chyba ve shodě přísudku s podmětem,
— pravděpodobná chyba ve shodě přísudku s podmětem,
— chyba týkající se zájmen,
— chybný slovosled,
— chyba ve valenci — obligatornost, pád (jen ve velmi omezené míře),
— chyba ve vokalizaci předložek,
— chyba v použití předložek (zejména pádová rekce),
— chybějící nebo přebývající čárka ve větě,
— chyba ve tvarech slova: některé překlepy, opakování téže formy, spřežky…,
— ostatní chyby.

Následuje charakteristika jednotlivých typů sledovaných chyb.

4.1.1 CHYBA VE JMENNÉ SKUPINĚ

Do této skupiny chyb patří zejména chyby:

— ve shodě komponent jmenné skupiny: rozvíjející syntakticky adjektivní přívlastky se musí shodovat se svým řídícím jménem ve jmenném rodě, čísle a pádě, jinak jde o chybu.
— v atrakci: odhaluje se naopak chybná přítomnost shody (i tvarové, nejen morfologické).

Uveďme některé příklady porušení pravidla o shodě komponent jmenné skupiny:

(5) *V otci, který se po letech vrátil domů, ušlechtilém, *báječnému*, mimořádném člověku, viděl mladý virtuos svůj vzor.

Adjektivum *báječnému* v dativu figuruje v přístavkové lokálové jmenné skupině *ušlechtilém, báječnému, mimořádném člověku*, což je zjevně gramatická chyba, neboť je porušeno pravidlo o shodě řídícího jména *člověku* s jeho adjektivními přívlastky v pádě. Kontrola navrhne opravit *báječnému* na *báječném*.

Uveďme ještě další typický příklad z této kategorie chyb:

(6) **Naši nový* kolegové koupili dobrou vodu.

Jelikož výchozí gramatikou Kontroly je gramatika spisovné češtiny, je slovo *nový* odhaleno jako chybné (jakožto atribut se neshoduje se svým řídícím substantivem *kolegové* v čísle) a je navržena náležitá oprava (*nový* → *noví*).

V textech se občas vyskytují také chyby v atrakci, kdy je nějaký tvar v chybném pádě, neboť tento pád negramaticky atrahoval od jiného slova:

(7) *Američané v řadě *případech* dokázali svou vstřícnost.
(8) *Pozice odborů je ve většině *zemích* historicky zakořeněná.

Kurzívou zvýrazněná slova jsou chybně v lokálu, neboť atrahují svůj lokálový tvar od předchozího slova *řadě*, resp. *většině*, které je ovšem náležitě v lokálu. Kontrola nabídne uživateli opravu v podobě správných genitivních tvarů *případů*, resp. *zemí*.

(9) *Domy kupců se nacházely poblíž velkých *náměstích*.

V této větě je tvar *náměstích* chybně v lokálu; jde patrně o chybnou tvarovou atrakci koncovky adjektiva *velkých* v Gpl. Kontrola chybu odhalí, neboť před tvarem *náměstích* není ve větě předložka s lokálovou rekcí, a nabídne náležitou opravu: *náměstí*. Poznamenejme, že větu lze samozřejmě opravit i jinak, než jak nabízí Kontrola, např. na

(9a) Domy kupců se nacházely *na* velkých náměstích.

čili ponechat jmennou skupinu *velkých náměstích* beze změny a doporučit náhradu jiného slova; zde by se předložka *poblíž* nahradila předložkou *na*. Ve svých doporučeních Kontrola nabízí co nejkonzervativnější opravu, zpravidla morfologickou, tj. jiný tvar téhož lexému.

4.1.2 CHYBA TÝKAJÍCÍ SE SLOVES

Do této skupiny chyb spadají především chyby při tvoření kondicionálu (včetně spojek *aby*, *kdyby*) v souvislosti s příklonnými reflexivy *se* a *si* a tvary slovesa *být*. Dále se

tu zpracovávají chyby v užívání reflexiv tantum (chybějící reflexivní částice), chybějící finitní slovesný tvar v klauzi uvozené vztažným zájmenem, negramatický vztah spojky a tvaru finitního slovesa v klauzi a chyby týkající se některých homonymních, velmi frekventovaných slovesných tvarů (*je*, *má* a další).

Chyby při tvoření kondicionálu se týkají kondicionálových tvarů:

(10) *My *by jsme udělali*.

Kontrola zvýrazní celý víceslovný přísudek *by jsme udělali*, jehož podoba nevyhovuje pravidlům gramatiky spisovné češtiny; správně má být *bychom udělali*.

Další typy odhalovaných chyb se týkají nesprávného užití reflexiva *se* a *si* v souvislosti s kondicionálovými tvary *by*, *bys*, … a spojkami *aby*, *abys*… a *kdyby*, *kdybys*…:

(11a) *Myslím, že *bys se* měl zastydět.
(11b) *Prosím, *abys si* to rozmyslel.
(11c) **Kdybys si* to rozmyslel, bylo by to lepší.

Kontrola náležitě navrhuje konstrukce, v nichž se příklonné *-s* odpoutá od tvaru *by*, *aby*, *kdyby* a připojí se k reflexivu; vznikne tak správné *by ses*, *aby sis*, *kdyby sis*. Podobně se Kontrola staví ke konfiguracím *jsi se*, resp. *jsi si* (správně *ses*, resp. *sis*).

Užitečné je upozorňovat uživatele na mylné vynechání povinného reflexiva u sloves a také u deverbativních adjektiv reflexiv tantum. Například ve větě

(12) *Snažil ho porazit v šachu.

schází reflexivní částice *se*, ač je u slovesa *snažit* povinná, a Kontrola tuto částici rovněž vyžaduje.

Pokud Kontrola identifikuje vedlejší větu vztažnou uvozenou zájmeny *který*, *jaký*, *čí*, *jenž*, vyžaduje v takovéto klauzi finitní slovesný tvar:

(13) *Byl to pracovník, který všechno vědět.

Kontrola si rovněž všímá vztahu některých spojek a tvarů sloves v klauzi, kterou uvozuje:

(14) *Chtěl, aby to *vypadá* pěkně.

U takovéto věty Kontrola navrhne nahradit prézentní tvar minulým příčestím téhož lexému.

Do kategorie *Chyba týkající se sloves* patří i chyby typu „dvě finitní slovesa v klauzi“. Na takovéto chyby je velmi užitečné upozorňovat, neboť se tak mj. odhalí mylné vynechání čárky či jiného interpunkčního znaménka tvořícího předěl mezi různými klauzemi. Chyby v následujících větách jsou zcela typické:

(15) *Protože *přijal* korunu bez dalších podmínek *je* evidentně hloupý.

Ve větě (15) schází čárka někde mezi slovesnými tvary *přijal* a *je*, velmi pravděpodobně těsně před *je*. Mylné vynechání čárky Kontrola na základě přítomnosti dvou finitních sloves, totiž *přijal* a *je*, v téže klauzi odhalí, a to i přesto, že tvar *je* je obecně i tvarem zájmenným. Kontrola však na základě poměrně rafinovaných úvah rozpozná, že *je* nemůže být ve větě (15) zájmenným tvarem, a následná kolize se slovesem *přijal* vede k detekci chyby. Kontrola ohlásí: „Věnujte pozornost slovu *je*" a uživatel se snad již dovtípí, kde je chyba. Jakmile umístí čárku na správné místo, tj. za tvar *podmínek*, Kontrola takovou opravu vzápětí vyhodnotí jako náležitou a mlčí.

4.1.3 CHYBA VE SHODĚ U SLOVESA (VE SHODĚ KOMPONENT VÍCESLOVNÉHO SLOVESNÉHO PŘÍSUDKU)

Shoda ve víceslovném slovesném přísudku se týká shody ve složených slovesných tvarech v čísle, popř. i jmenném rodě: tedy minulého a předminulého času 1. a 2. osoby, přítomného a minulého kondicionálu a opisného pasiva s výjimkou pasivního infinitivu. Rovněž se hlídá shoda finitního slovesa s přechodníkem v čísle a rodě. Následující příkladové věty osvětlí činnost kontrolního systému, narazí-li Kontrola na jev neshody:

(16a) **Byl* po všech těch zkušenostech opravdu *připraveni* na nejhorší.
(16b) **Byli jsme* po všech těch zkušenostech opravdu *připraven* na nejhorší.
(16c) *Pak *bys* teprve *koukaly*, co o vás vím.

U těchto vět Kontrola hlásí chybu ve shodě slovesných tvarů tvořících víceslovný přísudek, přičemž si povšimněme, že mezi těmito slovesnými tvary je poměrně velká vzdálenost. Kontrole to nevadí, neboť je schopna identifikovat slova patřící k téže klauzi nezávisle na slovosledné vzdálenosti mezi jednotlivými komponentami přísudku, pokud klauze nebo její část je souvislá, tj. slovosledně nepřerušená jinou, vnořenou klauzí.

Některé chyby ve shodě ve víceslovném přísudku hlásí Kontrola jako „Chybějící nebo přebývající čárka ve větě" a zároveň ovšem navrhuje správná nahrazení chybných slov. Při průzkumu věty

(17) *Pak *jsme* opravdu *spatřil* toho člověka.

tedy navrhuje nahradit buď tvar *jsme* tvarem *jsem*, nebo tvar *spatřil* tvarem *spatřili* či *spatřily*.

4.1.4 CHYBA VE SHODĚ PŘÍSUDKU S PODMĚTEM

Kontrola tu prověřuje shodu přísudku se substantivním/zájmenným podmětem ve jmenném rodě, čísle a osobě, jako např. v těchto příkladech:

(18) *Ten muž se *objevili* na pracovišti.
(19) *Chlapec *odešla*.
(20) *Žena se konečně potom *upravil* a vyšla na ulici.

(21a) *Partneři se znechuceně po dlouhé hádce *rozešly*.
(21b) **Rozešly* se po dlouhé hádce partneři.

Kontrola se vypořádá s oběma slovosledными variantami podmětu a přísudku; opět jí nevadí netěsné vzájemné postavení podmětu a přísudku. Kontrola však dokáže správně vyhodnotit i vztah některých jednodušších druhů podmětu tvořeného koordinovanou konstrukcí, jako například ve větách

(22a) *Žena a muž se *rozešly*.
(22b) *Žena, muž a dívka se *rozešly*.

kde navrhne opravit tvar *rozešly* na *rozešli*, jak bychom očekávali. Tyto případy jsou zvlášť problematické, neboť Kontrola musí rozpoznat význam spojky *a* a čárky: jak spojka, tak čárka jsou obecně homonymní, neboť mohou mj. oddělovat klauze, nebo naopak sdružovat jednotlivé větné členy v jedinou koordinaci. Kontrola tedy musí rozlišit, v jakém syntaktickém významu se spojky, resp. čárky užívá, tj. musí mít v sobě zabudovánu schopnost jemně identifikovat klauze (přesněji jejich spojité části) a jejich oddělovače (tj. spojky a interpunkční znaménka). To je v řadě případů nesmírně obtížné, v nemalém počtu případů je však Kontrola na základě hlubších syntaktických úvah při rozpoznávání úspěšná. Uvažme například gramaticky správnou větu:

(23) Muž seděl ve stínu a své zálibě v kouření doutníků se poživačně oddával.

Disambiguační podsystém (viz podrobněji oddíl 5) musí rozhodnout, co vlastně koordinuje spojka *a*: zda (i) koordinuje klauze

Klauze1: *Muž seděl ve stínu*
a
Klauze2: *své zálibě v kouření doutníků se poživačně oddával*

nebo zda (ii) koordinuje dvě jmenné skupiny řízené předložkou *ve*

JmSkupina1: *stínu*
a
JmSkupina2: *své zálibě*

neboť jak *stínu*, tak *své zálibě* mohou být v dativu nebo v lokálu (*stínu* navíc v genitivu), tj. mohou se v pádě i lišit, a v tomto případě ovšem nemohou tvořit jednu koordinovanou jmennou skupinu. Ve větě (23) však spojka *a* odděluje klauze: kdyby platila varianta (ii), nacházely by se v téže větě dva finitní slovesné tvary (l-ová participia považujeme pro účely Kontroly za finitní tvary), totiž *seděl* a *oddával*, a to je v psaném jazyce negramatické.

4.1.5 PRAVDĚPODOBNÁ CHYBA VE SHODĚ PŘÍSUDKU S PODMĚTEM

Do této kategorie patří konstrukce již výše zmíněné v oddílu 2. Jde o konstrukce gramaticky správné, uživatel Kontroly je však může vnímat při jiném chápání jako chybné. Mimo výše uvedenou větu (3) lze uvést pro uživatele záludnou konstrukci typu

(24) Autobusy *přijeli*.[3]

která se neškolenému uživateli může — obdobně jako věta (3) — jevit gramaticky defektní: je tu opět zjevná neshoda přísudku s podmětem ve jmenném rodě, interpretuje-li uživatel tvar *autobusy* jako podmět. Chápe-li se však tento tvar jako prostředek v interpretaci

(24a) *Autobusy*(Ipl) přijeli.

s možným pokračováním

(24b) *Autobusy* přijeli [výletníci] a tramvajemi zase odjeli.

kde je podmět lexikálně nevyjádřen, je věta plně gramatická. Celý problém tkví v pádové homonymii tvaru *autobusy* (Npl, Apl, Vpl, Ipl). Mělce uvažujícího uživatele může mlčenlivá reakce Kontroly znechutit natolik, že Kontrolu už nadále nebude používat. Autoři takového kontrolního systému musí tedy dobře zvažovat i psychologická hlediska související s tím, že uživatel obvykle není gramatický expert a neocení jemnosti gramatického systému češtiny, zejména si neuvědomí fakt homonymie českých slov a celých konstrukcí: jak tvar *ženy* ve větě (3), tak tvar *autobusy* ve větě (24) jsou v důsledku pádového synkretismu vícenásobně homonymní. Dokonce ani v případě, že podmět je sémanticky jasný, nemusí Kontrola nutně zjistit, že se neshoduje s přísudkovým slovesem, protože sémantické zřetele až na výjimky nebere v úvahu.

Při zpracování uvedených konstrukcí, při jejichž interpretaci na sebe mohou narazit odbornost autora Kontroly a jazyková bezelstnost neškoleného uživatele, je patrně vhodné, aby na možnost vícenásobné interpretace Kontrola uživatele upozornila. Proto Kontrola hlásí *pravděpodobnou chybu ve shodě přísudku s podmětem*. Podobných konstrukcí je více, věta (24) je však typická.

Je-li nicméně daná věta nehomonymní, jako třeba věta (25) či (26)

(25) **Jinoch* se *obrátila* na vedoucího.
(26) **Výletníci* včera za vydatného deště *přijely*.

bylo by namístě považovat případnou mlčenlivost Kontroly za vážný nedostatek: ve větách (25) a (26) je povinná shoda přísudku s podmětem ve jmenném rodě a čísle jasně porušena, přičemž slova *jinoch* a *výletníci* jsou jednoznačně v nominativu a mají v uvedených větách funkci jednoznačného podmětu. Takovéto chyby musí Kontrola

3 Za tento příklad autor děkuje kolegovi a spoluautorovi Kontroly Tomáši Jelínkovi.

hlásit, jelikož se tu nemůže vymlouvat na homonymii (která ovšem obecně představuje problém zásadní důležitosti, viz blíže v oddílu 5).

4.1.6 CHYBA TÝKAJÍCÍ SE ZÁJMEN

U tohoto typu chyb Kontrola zejména (a) upozorňuje na specifické chyby v užívání zájmenných tvarů tam, kde v češtině existují kontextem podmíněné varianty, (b) prověřuje vztahy zájmen na předělu hlavní klauze a závislé relativní klauze, zvláště pak shodu relativního zájmena uvozujícího relativní klauzi a jeho antecedentu v nadřazené klauzi, (c) odhaluje v některých případech časté chyby v homofonních podobách zájmena *mě* vs. *mně*, a také *ji* vs. *jí*.

Ad (a): V češtině existují dvě podoby některých osobních zájmen: (i) v komplementární opozici: tvary stojící po předložkách oproti tvarům, které po předložkách stát nesmějí: *mně* vs. *mi*, *tobě* vs. *ti*, *tebe* vs. *tě*, *němu* vs. *mu/jemu*, *něho* vs. *ho/jeho*, *ní* vs. *jí*, *ni* vs. *ji*, *nich* vs. *jich*, *nim* vs. *jim*, *ně* vs. *je*, *nimi* vs. *jimi*; (ii) příklonné tvary jsou v některých strukturách na rozdíl od tvarů nepříklonných negramatické (například příklonný tvar nemůže být konjunktem v koordinaci): *mi* vs. *mně*, *ti* vs. *tobě*, *tě* vs. *tebe*, *mu* vs. *jemu/němu*, *ho* vs. *jeho/něho*.

Kontrola odhalí použití nesprávné zájmenné varianty typu (i) v předložkových skupinách typu

(27) *Nikdo *do tě* nevidí.
(28) *Přistoupil *k jí*.

přičemž vydá hlášení: „Chyba týkající se zájmen“ a na požádání podrobněji vysvětlí, že: „Po předložce nemůže stát příklonka“, a uvede vysvětlující příklad.

Kontrola však dokáže odhalit i nesprávné slovosledné postavení klitických zájmen, ale takovéto chyby hlásí jako chyby ve slovosledu (viz následující oddíl 4.1.7).

Ad (b): Jako *Chybu týkající se zájmen* hlásí Kontrola i neshodu relativního zájmena *který*, *jaký*, *čí*, *jenž* s jeho nominálním antecedentem v čísle a jmenném rodě:

(29) *Muž, *kterou* jsem včera zahlédl, se opět objevil na rohu ulice.

Problém automatické identifikace antecedentu relativního zájmena je obecně velmi složitý, např. v nadřazené větě může být vhodný antecedent slovosledně dost vzdálen od relativa; lexémy *který*, *jaký*, *čí* jsou navíc homonymní: mohou totiž také uvozovat nepřímou otázku atd. Kontrola tedy hlásí chybu jen v jednoznačných případech jako ve větě (29), nikoli však např. ve větě

(30) Nakonec rozhodl otec, která z nich se načne.

kde vedlejší klauze je nepřímou otázkou, což je dáno valenční povahou slovesného lexému *rozhodl* v nadřazené větě, který patří do skupiny pár set slovesných lexémů, jež mohou být rozvity nepřímou otázkou. Tato složitá problematika ovšem do tvorby Kontroly zahrnuta nebyla, jelikož je to problematika valence a ta se pro svou složi-

tost „vešla“ do Kontroly jen okrajově. A tak si Kontrola, zkoumajíc větu (30), není jista, zda *otec* je rozvit relativní klauzí, nebo zda vedlejší věta není třeba nepřímou otázkou, a proto raději chybu nehlásí. To je u věty (30) jistě správné, nikoli však např. u věty

(30a) *O této záležitosti rozhodl otec, *která* někdy bývá velmi autoritativní.

kde klauze *která někdy bývá velmi autoritativní* není nepřímou otázkou, nýbrž vedlejší vztažnou klauzí rozvíjející (nesprávným tvarem *kterou*, správně má být *který*) substantivum *otec*. Kdyby totiž byla nepřímou otázkou, pak by sloveso *rozhodnout* mělo místo pro přímý předmět ve svém rámci obsazeno dvěma členy: (i) předložkovou skupinou *o této záležitosti* a (ii) nepřímou otázkou *která někdy bývá velmi autoritativní*, což ovšem valenční rámec slovesa *rozhodnout* nepřipouští.

Takto složitých úvah opírajících se o valenční rámce sloves Kontrola až na výjimky není schopna, a tudíž alibisticky mlčí. U věty (29) si však jista je, neboť klauze *kterou jsem včera zahlédl* nemůže být nepřímou otázkou. Rozdíl mezi větou (29) a (30) tkví právě v absenci/přítomnosti slovesa jistých vlastností v hlavní klauzi.

Ad (c): Vzhledem k homofonii grafémicky odlišných slov uživatelé často chybují ve tvarech *mě* vs. *mně* a *ji* vs. *jí*, jejichž syntaktická distribuce i význam jsou odlišné. Správné užívání těchto tvarů je ovšem dáno — až na případ, kdy se tato zájmena objevují v předložkových skupinách — především slovesnou valencí a ta je, jak už bylo uvedeno výše, v Kontrole zpracována jen velice okrajově, např. v neosobních konstrukcích typu

(31) *Mrzelo *mně*, že jsem se mu nestihl omluvit.
(32) *Potěšilo by *jí*, kdyby sis na ni vzpomněl.

kde Kontrola upozorní uživatele na chybu v pádě ve slově *mně* a nabídne mu tvar *mě*, resp. ve slově *jí* (nabídne správný tvar *ji*).

4.1.7 CHYBNÝ SLOVOSLED

V této oblasti se Kontrola zaměřuje zvláště na některé případy povinného slovosledu, což se týká v prvé řadě příklonek (zájmen a pomocných slovesných tvarů). Kontrola hlídá pořadí příklonek na syntakticky druhé, tzv. Wackernagelově pozici. Níže je uvedeno několik vět, v nichž je v příklonkovém trsu pořadí příklonek negramatické, což je hrubá gramatická chyba:

(33) *Ale za týden *mu se* ozvu.
(34) *Minulý týden *mu jsem* ještě nic neřekl.
(35) *Především *tě jsem* prosil, abys na to zapomněl.
(36) *Zavolají *ho si* za chvíli.
(37) *Bylo opravdu Valentina *mi* líto.

Ve větách (33) až (36) Kontrola ohlásí chybu a uživateli navrhne přehodit tvary vyznačené kurzívou. Věta (37) je složitější, ale i s ní se Kontrola vypořádá se ctí: navrhne uživateli správnou opravu:

(37a) Bylo mi opravdu Valentina líto.

Další pravidlo o příklonkách praví, že příklonka nemůže ve spisovné češtině stát na začátku věty/klauze, srov.:

(38) **Jsem* ti neřekl, že nemám čas.
(39) **Byste* to měli vědět.

Zde Kontrola uživatele upozorní na chybný slovosled, není však natolik iniciativní, aby neohroženě navrhla opravu.

Je-li Kontrole předložena gramatická věta, v níž se příklonky nacházejí v gramatickém kanonickém pořadí, Kontrola chybu správně nehlásí (obligatorní a fakultativní příklonky jsou níže vyznačeny kurzívou):

(40) Opravdu doporučil Petrovi vrátit *ho* Lídě.
(41) Myslím, že vážně měl chuť *tě* co nejdřív vyhodit.
(42) Poslal *jsem vám ho*.
(43) Poslal *vám ho*.

4.1.8 CHYBA VE VALENCI

Problematika valence je nesmírně složitá a do Kontroly byly valenční zřetele vyjádřené pravidly zahrnuty jen velmi výběrově. Právě v této oblasti jsou ovšem skryty velké možnosti, jak Kontrolu výrazně zlepšit a dále rozvíjet, a to zvláště v jemném rozpoznávání slovesných, adjektivních, substantivních a adverbiálních obligatorních doplnění v náležitém pádě, příp. stupni. Například Kontrola by měla upozornit uživatele na chybu ve valenci ve větě

(44) *Všiml si *tu dívku*.

kde objektové doplnění slovesného lexému musí být v češtině (na rozdíl třeba od slovenštiny) v genitivu; zatím zde však Kontrola chybu nehlásí.

Nicméně některé typy valence alespoň částečně zpracovány jsou. Vychází se přitom například z následujících netriviálních faktů. (a) Ve větě s jedním určitým slovesem majícím ve svém rámci jediné doplnění s nepředložkovou akuzativní valencí může stát pouze jeden nepředložkový akuzativ ve funkci předmětu. Případné ostatní slovní tvary v akuzativu mají nutně jinou funkci: jsou řídícími prvky akuzativní předložkové skupiny, nebo jsou časovými či měrovými substantivy (*chvíli*, *hodiny*, *kilometr*…), příp. figurují ve frazémech (*ruku v ruce*). Pokud tyto funkce akuzativních tvarů Kontrola „ve svých úvahách" bezpečně vyloučí, může uživateli pomoci. Například u věty

(45a) *Tu *dívku* pokládal schopnou *kandidátku*.

kde patrně vypadla předložka *za* před slovem *schopnou*, Kontrola ohlásí chybu ve valenci a upozorní uživatele na problematickou dvojici substantiv v akuzativu: *dívku* a *kandidátku*,

protože sloveso *pokládat* nemůže být rozvito dvěma akuzativními předměty a jinou větněčlenskou funkci tato slova ve větě (45a) mít nemohou. Předložíme-li však Kontrole větu

(45b) Tu dívku pokládal dlouhou chvíli za schopnou kandidátku.

správně nic nehlásí, jelikož ve větě odhalí časové substantivum *chvíli*.

(b) Významnou oblastí povrchové valence je problematika reflexiv *se* a *si* a jejich vlastnosti, a to zejména pro vliv, který má jejich přítomnost na syntaktické okolí. Kontrola využívá mj. toho, že je-li v klauzi reflexivní sloveso či adjektivum typu *se*, není v klauzi předmětné doplnění tohoto slovesa/adjektiva v akuzativu. Toto tvrzení neplatí bezvýhradně, neboť existuje nečetná množina slovesných lexémů reflexiv tantum typu *se*, která vyžadují svůj přímý předmět v akuzativu (*učit se fyziku*, *pomodlit se pět zdrávasů*). Pokud však v klauzi není sloveso z této skupiny, Kontrola se může s úspěchem uplatnit, jako je tomu např. u věty

(46) *Usmál *se tu dívku*.

kde nejspíše vypadla předložka *na* za reflexivem *se*. Kontrola si klade otázku, k jakému základovému slovu se reflexivum váže, a zjistí, že jediným kandidátem tu je slovesný tvar *usmál*. Jelikož tento lexém nemůže mít akuzativní předmět, což Kontrola dobře ví, neboť má k dispozici výše zmíněný soupis reflexiv tantum typu *se*, hlásí chybu (ovšem po prověření, zda *dívku* nemůže být časové či měrové substantivum). Zato u věty

(47) *Dověděl se* tu nepříjemnou *zprávu*.

chybu nehlásí, protože sloveso reflexivum tantum *dovědět se* má akuzativní předmět ve svém rámci.

Je-li ve větě sloveso reflexivum tantum typu *se*, které nemá ve svém rámci akuzativní předmět, a zároveň je v ní jiné sloveso, které akuzativní předmět ve svém rámci má, Kontrola správně nic nehlásí, jako třeba u věty:

(48) Snažil se přečíst novou zprávu.

Chybu však hlásí u chybné věty

(48a) **Snažil se* novou *zprávu*.

kde další sloveso mimo *snažit se* schází.

Kontrola si mimo uvedené, jen zčásti zpracované oblasti všímá i dalších vybraných jevů, jako například (c1) a (c2):

(c1) Syntaktická nekompatibilita slovesného rodu a genitivu. Jde o konstrukce typu:

(49) **Oběda* je uvařen.
(50) **Měst* bylo obnoveno.

(c2) Dva nominativy v klauzi. Tato problematika je obecně velmi složitá: Kontrola se snaží hlásit nesprávnou přítomnost dvou nominativních podmětů v klauzi:

(51) *Chlapec hledá *dívka*.
(52) *Dívka jde pomalu *dívka*.

Takovýchto chyb se uživatel vskutku může dopustit, u věty (52) například tak, že chce přesunout slovo *dívka* na začátek věty, přitom však slovo nepřesune, nýbrž zkopíruje. Kontrola musí ovšem rozpoznat, že nominativy jsou skutečně formami vyjadřujícími podmět, nikoli nějakou jinou větněčlenskou funkci, srov. gramaticky zcela správnou větu:

(53) Alena si prý říká prezidentka světa.

Dlužno dodat, že v oblasti valence toho zbývá ještě mnoho vykonat, zvláště v podrobném zpracování základních pádových valencí. Výše uvedené úvahy alespoň naznačily směr, jakým by se rozvoj Kontroly v oblasti zpracování valence mohl ubírat.

4.1.9 CHYBA VE VOKALIZACI PŘEDLOŽEK

Při psaní textu se může snadno přihodit, že pisatel správně napíše nevokalizovanou předložku bezprostředně následovanou slovem, které nevokalizovanou podobu předložky buď vyžaduje, nebo připouští, a pak v průběhu úprav textu slovo po předložce změní na jiné slovo, které naopak vyžaduje vokalizovanou podobu předložky. Tento problém se týká těchto jedenácti předložek: *bez/beze*, *k/ke/ku*, *nad/nade*, *od/ode*, *pod/pode*, *před/přede*, *přes/přese*, *s/se*, *skrz/skrze*, *v/ve*, *z/ze*. Pisatel například původně napsal

(54a) Se *starými* lidmi se odmítal bavit.

a poté pozměnil tvar *starými* na *postaršími*, přičemž nezměnil podobu předložky:

(54b) **Se postaršími* lidmi se odmítal bavit.

V tomto případě Kontrola hlídá správnou podobu předložky a navrhuje její náležitý tvar, totiž *s* před slovy nezačínajícími na sykavku a na specifické souhláskové shluky.

Podobně se pisatel může dopustit chyby opačné: správně nejprve napíše nevokalizovanou podobu předložky bezprostředně následovanou slovem, které nevokalizovanou podobu předložky buď vyžaduje, nebo připouští, a pak slovo po předložce změní na jiné slovo, které naopak vyžaduje vokalizovanou podobu předložky. Například původně správně napsal

(55a) *S postaršími* lidmi se odmítal bavit.

a poté pozměnil tvar *postaršími* na *starými*:

(55b) **S starými* lidmi se odmítal bavit.

Kontrola zde ohlásí chybu v podobě předložky *s* a navrhne uživateli její vokalizovanou podobu *se*.

U výše uvedených předložek se také prověřuje vokalizovaná/nevokalizovaná podoba předložky v postavení těsně před zájmenným tvarem *mě*/*mne*, např.:

(56) *Odešel *od mě* znechucen.

Není-li předložka v tomto postavení vokalizována, Kontrola hlásí chybu.

4.1.10 CHYBA V POUŽITÍ PŘEDLOŽEK

U předložek Kontrola sleduje hlavně chyby v pádové rekci, kdy po předložce nenásleduje jméno v pádě, který předložka vyžaduje. Následující příklady jsou charakteristické:

(57) *Přijeli poslové *ze zprávou*.
(58) *Neřekl jsi mi, *z čím* mám počítat.
(59) **Nad vám* i visí pohroma.
(60) **Za mně* se neschová nikdo.
(61) *Myslel si, *ze se* mu to nepovede.

V češtině ovšem může ve speciálním a dost výjimečném případě stát těsně po předložce vnořená adjektivní skupina, v níž na řídícím adjektivu Adj závisí syntaktické substantivum SyntSubst1, které mu slovosledně předchází a nesplňuje rekční požadavek předložky, neboť s ní syntakticky nesouvisí. Adj však typicky rozvíjí jiné substantivum, SyntSubst2, jako jeho shodný přívlastek, přičemž právě pád syntaktického substantiva SyntSubst2 je určen předložkovou rekcí. Takový případ nastává např. v takovýchto větách

(62) *Z vámi připraveného dokumentu* jsem to nevyčetl.

kde Kontrola chybu nehlásí, což je správné: *vámi* těsně po předložce *z* je tu SyntSubst1 v instrumentálu, které závisí na Adj *připraveného*, a toto Adj v genitivu atributivně rozvíjí slovní tvar *dokumentu* (= SyntSubst2) rovněž v genitivu, přičemž až tento tvar, nikoli tvar *vámi*, splňuje genitivní rekční požadavek předložky *z*. Na tomto příkladu ukazujeme, že Kontrola je založena na hlubších, nikoli jen prvoplánových syntaktických úvahách.

Do této skupiny chyb jsou zahrnuta i opomenutí potenciálně lokálové předložky (*o*, *na*, *po*, *při*, *v*) před jménem v lokálu

(63) [O] **těch* lidech nikdo nemluvil.

a dále chyby tohoto typu: předložka před nominativem, vokativem, slovesným tvarem, spojkou; absence tvaru, který předložka vyžaduje v náležitém pádě; více než dvě předložky za sebou a další.

4.1.11 CHYBĚJÍCÍ NEBO PŘEBÝVAJÍCÍ ČÁRKA VE VĚTĚ

Tento druh chyby se v textech vyskytuje snad nejčastěji, lidé totiž většinou neznají dost složitá pravidla o kladení čárek v češtině. Častěji se stává, že v textu je čárka spíše opomenuta, než že by přebývala, a to zvláště mezi klauzemi. Následují typické příklady hrubých pravopisných chyb

(64) *Dneska jsem *první který* půjde spát.
(65) *Hlupák *jsi kam* zase jdeš?
(66) *Nebylo příliš *jasné proč* nás odvážejí z Prahy.
(67) *Nevěděli jsme *však kam* nás vezou.

kde čárka schází mezi slovy zvýrazněnými kurzívou. Kontrola mimo oznámení chyby nabídne, kam čárku umístit.

Méně je na první pohled zřejmé, kam čárku umístit, předchází-li vedlejší věta větu hlavní, jako třeba ve větě:

(68) Že se opravdu *mýlí* mu *nedocházelo*.

Kontrola pokyne uživateli, aby věnoval pozornost oběma slovesům, jelikož v téže klauzi nemůže být více než jedno finitní sloveso (včetně l-ových participií); gramaticky správný víceslovný predikát, jako je třeba 1. a 2. osoba minulého času, Kontrolu samozřejmě nepřekvapí. Jakmile uživatel umístí čárku mezi obě slovesa, je Kontrola spokojena. Kontrola je tedy bohužel uspokojena i v případě, že uživatel umístí čárku za slovo *mu*, neboť se nepozastaví nad negramatickým dativním doplněním slovesa *mýlit se*. Odhalit tuto negramatičnost by znamenalo, že Kontrola by musela pořádně „umět" valenci.

Pravidla o nepřípustnosti dvou finitních sloves v téže klauzi Kontrola velmi úspěšně využívá a dva finitní slovesné tvary interpretuje jako opomenutou čárku, musí si ovšem být jista, že oba tvary jsou vskutku slovesné (rozhodování Kontroly se problematizuje v případě homonymie některého ze slovesných tvarů).

Pisatelé chybují i v kladení čárek v těsné blízkosti spojek, jak je vidět z následujících příkladů:

(69) *Franta tam přišel *a, když* uviděl Marii, zase odešel.
(70) **A, protože* byl dobrý, věnoval mi svůj čas.
(71) *Franta tam přišel, *ale, když* uviděl Marii, zase odešel.

Kontrola zde uživateli oznámí chybu a doporučí čárku umístit před spojku.

Pisatel někdy změní slovosled a nepřesune přitom pomocné sloveso, nýbrž ho zkopíruje, takže v klauzi se pak týž pomocný tvar nachází dvakrát:

(72) Opravdu *jsem* ti nechtěl *jsem* ublížit.

Při průzkumu této věty Kontrola vyzve uživatele, aby se věnoval oběma výskytům tvaru *jsem*. Rozdíl v hlášeních chyby ve větě (72) a ve větě (15), poprvé uvedené v oddílu 4.1.2,

(15) *Protože* přijal korunu bez dalších podmínek *je* evidentně hloupý.

spočívá v absenci homonymie ve větě (72) (u věty (15) obsahující homonymní slovní tvar *je* se hlásí „Chyba týkající se sloves").

Kontrola ve speciálních případech hlídá i shodu dvou koordinovaných slovesných predikátů v osobě, např. ve větě:

(73) *Našel jsem klíč, který ležel pod rohožkou a otevřel *jsem* dveře.

V této větě má vztažná klauze podmět *který* a jeho přísudek je ve 3. osobě (*ležel*). V klauzi *otevřel jsem dveře* není podmětem *který*, neboť nevyjádřený podmět je v 1. osobě. Z toho ovšem vyplývá, že tato klauze není součástí vztažné klauze, a že se tedy nekoordinuje s klauzí *který ležel pod rohožkou*. Proto Kontrola hlásí chybu v kladení čárky a navrhne doplnit čárku za slovo *rohožkou*.

Pravidlo se bohužel neuplatní na nesprávnou větu

(74) *Vrátil se ještě pro kolegu, který zpanikařil a *oheň* mu zatarasil cestu.

jež je podobně jako věta (73) také chybná, neboť buď má být za slovem *zpanikařil* čárka, nebo se před slovo *oheň* má doplnit *jemuž* a zároveň odstranit *mu*. Kontrola zde nic nehlásí: buď si neuvědomí, že *oheň* je podmětem v klauzi *oheň mu zatarasil cestu*, nebo slovo *oheň* sice identifikuje jako podmět, ale pak nezpracuje případ dvou odlišných podmětů v úseku *který zpanikařil a oheň mu zatarasil cestu*: jedním podmětem je *který* odkazující na *kolegu*, druhým pak *oheň*. Negramatičnost koordinované konstrukce spočívá v tom, že v české vztažné větě tvořené koordinací není (až snad na výjimky) možné, aby jeden z podmětů byl vyjádřen vztažným zájmenem a druhý substantivem za nepřítomnosti dalšího vztažného zájmena (nutno doplnit: a *jemuž* oheň… nebo případně: a *jehož* oheň, aby věta byla alespoň gramatická).

4.1.12 CHYBA VE TVARECH SLOVA: NĚKTERÉ PŘEKLEPY, OPAKOVÁNÍ TÉŽE FORMY, SPŘEŽKY…

Do této skupiny chyb odhalovaných Kontrolou patří pravopisné chyby v některých obtížných slovech, v nichž se často chybuje (soupis takových slov, jež má Kontrola k dispozici, ale zdaleka není vyčerpávající), Kontrola zde ovšem v žádném případě nenahrazuje obyčejný spell checker, který by upozorňoval na každý překlep. Uveďme několik skupin typických chyb, které Kontrola odhaluje, přičemž chyba je buď absolutní, tj. nezávislá na kontextu, nebo relativní, tj. dané slovo je chybné pouze v daném kontextu.

(a) Chyba v užívání spojovníku:

(75) *Lepra *nebo-li* malomocenství sužuje dodnes asi tři milióny lidí.
(76) **Jest-li* na mě počkáte, něco vám ukážu.

Má být samozřejmě *neboli*, resp. *jestli*.

Častěji je chyba způsobena homofonií: uživatel nesprávně napíše homofonní slovo. Kontrola tu upozorňuje na slova, která v češtině sice existují, ale v daném kontextu jsou nesprávně užita. Takových typů je více, uveďme jen některé klasické školské typy s příklady:

(b) Chyba v psaní *s*/*z*:

(77a) *nucená/státní *zpráva*,
(77b) **zpravit* náladu,
(78) **shlédnutí* představení,
(79) *vojenský *sběh*.

(c) Chyba v psaní *i*/*y*:

(80) *větrný *výr*,
(81a) **víska*t radostí,
(81b) **výskala* mu vlasy,
(82) *v *Břeclavy*,
(83) *Firma *vytypuje* vhodné kandidáty.

(d) Chyba v psaní *ú*/*ů*:

(84a) *Žloutky ušleháme s citrónovou *kúrou*.
(84b) **pomerančová/zemská *kúra*,
(84c) *odtučňovací *kůra*.

(e) Chyba v psaní jednoho či dvou *n*:

(85) **cenou* remízu,
(86a) **obranou* linii,
(86b) **obraný* val,
(87) **ochranou* známku,
(88) **ranná* tvorba,
(89) *stará *pana*,
(90) **varhaním* koncertem,
(91) **jeskyních* maleb,
(92) *veřejně *činí*.

(f) Chybné psaní slov *cenně*, *rodinně*, *sezónně*, *příčinně* apod. po předložce:

(93a) *K *záchranně* tonoucích lidí nedošlo.

(g) Chyba v psaní účelových adjektiv. Často se chybuje v užívání účelových adjektiv, která si pisatelé zhusta pletou s dějovými:

(94) *Koupil *balící* papír.
(95) *Stála tam *kropící* konev.
(96) *Nepoužil *čistícího* prostředku.
(97) *Máme doma dva *honící* psy.

(h) Chyby při rozpoznávání pomocí OCR[4] u naskenovaných dokumentů. Některé specifické chyby vznikají zvláště při rozpoznávání naskenovaných dokumentů pomocí OCR; při jejich prověrce se Kontrola uplatní velmi dobře:

(98) *Dozvěděl jsem se, *ze* Martin slib splnil.
(99) *Potřebuju něco pro *zahřáti*.
(100) *Neuznáváme to *rozhodnuti*.
(101) *Byli to ti, na *než* měl spadeno.

Jak vidno, často se nesprávně zpracovává diakritika, jež je zdrojem podivných chyb, ty však Kontrola často s úspěchem odhalí.

4.1.13 NĚKTERÉ JINÉ CHYBY

Do této zbytkové skupiny patří například absence čárky ve větě s vokativní nominální skupinou

(102) *Zavři už konečně *dveře Jardo*!

kde mezi slovy *dveře* a *Jardo* schází čárka.

4.2 CHYBY STYLISTICKÉ

Mezi nemnohé stylistické jevy, které Kontrola sleduje a jejichž sledování se dá zapnout v modu *Gramatika+Styl*, patří:

— nadbytečná slova,
— ostatní stylové chyby (hlášené jako Stylová chyba).

4 Optical Character Recognition.

Uveďme některé charakteristické příklady:

(a) Nadbytečné formy:

(103) Udělal to *bez toho, aniž* by se začervenal.
(104) Koupil jsem to *proto, protože* jsem chtěl.

(b) Nespisovné slovesné tvary:

(105) *Kolegové ti to *poví*.
(106) *Oni *ví*, oni tu pečínku *sní*.

Za stylovou chybu se považuje i opisný komparativ, tedy spojení *více*, *méně* + pozitiv/komparativ adjektiva či adverbia. Je tu však nutno dát pozor na homonymii adverbia *více* s číslovkou *více*:

(107) *Přiběhli *více rychleji*.

Zde Kontrola správně hlásí (přinejmenším) stylistický nedostatek, v následující gramaticky zcela správné větě však bohužel hlásí chybu také (je to typický *false flag*):

(108) Měli *více lépe propracovaných* řešení.

Ve větě (108) je totiž slova *více* užito ve významu číslovky (*více … řešení*), nikoli jako (nesprávné) formy opisného stupňování komparativu *lépe*.

Některá složitá adjektiva, hlavně kompozita, se mohou stupňovat opisně a chybu ve stupňování pak Kontrola nehlásí:

(109) Odboj *byl více protisrbský*.

4.3 CHYBY FORMÁLNÍ

Kontrola sleduje i některé ryze formální pravopisné jevy týkající se zkratek, interpunkčních pravidel, ba dokonce počtu mezer za sebou, a také ohlašuje příliš dlouhou větu (= delší než tři sta slov), na jejíž zpracování v první verzi Kontroly (v Microsoft Office 2003) rezignuje, ve verzích dalších však takové věty zpracovává. Uživatel může po Kontrole chtít, aby sledovala tyto formální pravopisné jevy:

— hromadění interpunkce,
— malé písmeno na začátku věty,
— chybějící mezera,
— přebývající mezera kolem interpunkce,
— hromadění mezer,
— přípony u čísel,
— chyba v psaní zkratek,

— spojovník a rozdělovník,[5]
— příliš dlouhá věta (delší než 300 slov).

Popíšeme nyní velmi stručně jednotlivé formálně-pravopisné chyby, jež Kontrola hlásí.

4.3.1 HROMADĚNÍ INTERPUNKCE

Hromadění interpunkce je chyba, kterou Kontrola signalizuje, odhalí-li nepravopisnou konfiguraci více interpunkčních znamének v těsném sledu, například dvě čárky za sebou:

(110) *To *nijak,, nesouvisí* s tím, jak to dělají.

4.3.2 MALÉ PÍSMENO NA ZAČÁTKU VĚTY

Díky dobré větné segmentaci Kontrola ve valné většině případů správně rozpozná, že věta začíná malým písmenem:

(111) Přešel k jinému *tématu. týkalo* se nových obilovin.

4.3.3 CHYBĚJÍCÍ MEZERA

Za čárkou a středníkem se píše mezera, takže věta (112) je psána chybně, neboť čárka a *že* nejsou odděleny mezerou:

(112) *Je tedy *vidět,že* chyb se nedopouštějí jen staří mazáci.[6]

4.3.4 PŘEBÝVAJÍCÍ MEZERA KOLEM INTERPUNKCE

Naopak před interpunkční znaménko, přesněji čárku, tečku, dvojtečku, středník, vykřičník, otazník, se mezera neklade:

(113) *Vše se řítí k jakémusi *poslednímu , osudnému* a stále bližšímu okamžiku.
(114) *Má zvláštní pohled na svou *práci : podstata*, tj. chemické složení farmak ho nezajímá.

5 Tak je nesprávně uvedeno v současné verzi Kontroly, správně má být: Spojovník a pomlčka (podobně dále).

6 Existuje však užití čárky v jiném významu, totiž ve významu počáteční jednoduché uvozovky, zvláště v textu, v němž se dále vyskytuje koncová jednoduchá uvozovka ‘: „Proto se také termínu labyrint začalo používat ve významu ‚bludiště‘.“ Za toto upozornění děkuje autor recenzentovi.

4.3.5 HROMADĚNÍ MEZER

Hromadění mezer nesouvisí s jazykem ani s „počítačovým" pravopisem, je však jakýmsi nepsaným pravidlem, že mezi slovy by nemělo být více mezer než jedna (znak * u věty (115) upozorňuje pouze na tento prohřešek v jazykově správné větě):

(115) *Vzpomínám si na to, jak *Aeneas vykřikl* na svou matku, když se mu zjevila v lese v podobě lovkyně.

Ve větě (115) je mezi slovy *Aeneas* a *vykřikl* víc mezer než jedna a Kontrola na ně na vyžádání uživatele upozorní.

4.3.6 PŘÍPONY U ČÍSEL

Tento typ formální chyby, kterou Kontrola odhalí, spočívá v nesprávném psaní spojovníku u slov typu *23-letý*, *5-stranný* (správně bez spojovníku: *23letý*, *5stranný*), dále v chybném psaní *ti* po číslovce: *ve 45ti letech* (správně *ve 45 letech*). Kontrola navíc chápe spojovník — ve spojeních typu *Praha 2-Vinohrady* jako nesprávný (činí tak ovšem v rozporu s normou!) a navrhuje opravit toto spojení na *Praha 2 — Vinohrady*, tedy s pomlčkou oddělenou z obou stran mezerami.

4.3.7 CHYBA V PSANÍ ZKRATEK

Pravopisné chyby se dělají i v psaní zkratek, Kontrola např. upozorní na chybu v psaní zkratky *tj.*:

(116) Má zvláštní pohled na svou práci: podstata, *t.j.* chemické složení farmak ho nezajímá.

4.3.8 SPOJOVNÍK A ROZDĚLOVNÍK

Kontrola upozorní na (často se vyskytující) chyby v psaní spojovníku mezi slovy (zvláště čísly) tam, kde by se správně měla psát pomlčka:

(117) *Za normalizace v letech *1970-1989* zažívala naše země hlubokou stagnaci.[7]

4.3.9 PŘÍLIŠ DLOUHÁ VĚTA

Ve verzi, jež je součástí systému Microsoft Office 2003, Kontrola odmítá prověřovat věty delší než tři sta slov. Toto omezení ve vyšších verzích (Microsoft Office 2007, 2010, 2013) již není.

7 Hvězdička je větě předřazena jenom kvůli chybnému užití spojovníku.

4.4 NESLEDOVANÉ JEVY

Jelikož gramatika češtiny (a ostatně každého jazyka) představuje velice složitý systém mnoha vzájemně spolupracujících langueových pravidel, nebylo možné v poměrně krátkém čase (cca patnácti měsíců) vyhrazeném zadavatelem postihnout úplně všechny jevy v náležitém rozsahu a hloubce zpracování. Zejména byly jen velmi povrchně zpracovány následující typy chyb nebo nebyly zpracovány vůbec:

— Chyby ve valenci, tj. v identifikaci chyb v obligatorních a fakultativních doplněních hlavně sloves, adjektiv a substantiv: Bylo by třeba podrobně zachytit takové chyby, jako je například absence obligatorního doplnění nějakého slova, nadbytečnost doplnění, nesprávná forma doplnění apod. Je ovšem jasné, že nesmírně složitou problematiku valence lze v jakékoli kontrole gramatiky zachytit jen zčásti.
— Chyby v záporové shodě: zhruba řečeno, je-li v klauzi záporný zájmenný lexém (např. *žádný*, *nikdo*) či záporné příslovce (např. *nikde*, *nikam*), musí být finitní sloveso záporné (srov. např. Petkevič, 2004).
— Chyby v syntaktických vztazích na (syntakticky a/nebo slovosledně) velkou vzdálenost: kontrola například neodhalí chybnou neshodu (různého typu) dvou slov v téže klauzi, která však netvoří souvislý úsek, ale je tvořena nejméně dvěma částmi, mezi nimiž se nachází jiná, vnořená klauze, přičemž každé ze slov vstupujících do gramatického vztahu kongruence se nachází v jiné části.
— Chyby tkvící v sémantice, konkrétněji v sémantické nekompatibilitě slovních spojení (**Stručně* se u mě zastavil na kus řeči.).

V eventuální další verzi *Kontroly české gramatiky* bude nutno uvedené typy chyb odhalovat v co největší míře. Navíc se dá využít i korpusu chyb brněnské provenience nazvaného výstižně *Chyby* a obsahujícího nejrůznější typy klasifikovaných a anotovaných chyb (blíže viz Pala et al., 2003).

5. ALGORITMUS ZPRACOVÁNÍ VSTUPNÍHO TEXTU

Celé zpracování gramatické, pravopisné, stylistické a formální kontroly vstupního textu v Kontrole se skládá z následující posloupnosti kroků (ty jsou vzápětí popsány podrobněji):

1. Aktivuje se modul větné segmentace (segmentátor).
2. Aktivuje se modul tokenizace (tokenizér).
3. Aktivuje se modul „pražské“ morfologické analýzy.
4. Aktivuje se skupina interpunkčních pravidel GCfaze0 pro formální prověrku textu.
5. Aktivuje se skupina pravidel GCfaze1; tato pravidla jsou nezávislá na morfologické disambiguaci.
6. Poprvé se použije skupina pravidel GCfaze2, jejichž činnost je obecně sice závislá na výsledku disambiguace, ale před tímto použitím se dosud žádná disambiguační pravidla neuplatnila.

7. Aktivace disambiguačních pravidel; pokud některé disambiguační pravidlo odstraní všechny morfologické značky (tagy) u nějakého slova, aktivuje se skupina pravidel GCfaze2.

Pokud Kontrola nalezne ve větě chybu, oznámí ji uživateli. Ten chybu buď ignoruje a aplikuje Kontrolu na další větu, nebo chybu opraví a Kontrola začne prověřovat nejprve opravenou větu a pak věty další.

Popíšeme nyní jednotlivé kroky podrobněji.

1. krok: V tomto počátečním kroku je vstupní text zpracován nestochastickým větným segmentátorem založeným na pravidlech.[8] Segmentátor rozdělí vstupní text rozčleněný do odstavců na věty.

2. krok: Při tokenizaci je vstupní text rozdělen programem zvaným tokenizér[9] na tokeny, tj. ortografická slova, což jsou řetězce písmen a čísel nacházející se, zhruba řečeno, mezi mezerami a/nebo interpunkčními znaménky. Interpunkční znaménka jsou oddělena od slov, s nimiž jsou ve vstupním textu spojena. Výsledné tokeny dále vstupují do procesu morfologické analýzy.

3. krok: Aktivace morfologické analýzy pražské provenience (srov. Hajič, 2000; 2004),[10] která každému tokenu přiřadí na základě morfologického slovníku jeho lemma/ta a všechny jeho možné morfologické interpretace (včetně slovnědruhových) bez ohledu na kontext: je-li token homonymní, obdrží více morfologických interpretací a/nebo lemmat. Probíhá tedy standardní zpracování textu jako například při značkování korpusových textů.

4. krok: Použije se skupina interpunkčních pravidel GCfaze0 pro formální kontrolu textu. Tato skupina obsahuje skupiny formálních pravopisných pravidel popsané výše v oddílu 4.3 a navíc test (pouze ve verzi Microsoft Office 2003!) zjišťující, zda věta není příliš dlouhá (max. tři sta slov). Pokud je delší než tři sta slov, Kontrola její gramatickou správnost prověřuje až počínaje verzí Microsoft Office 2007.

5. krok: Aktivace skupiny pravidel GCfaze1; tato pravidla jsou nezávislá na morfologické disambiguaci. Skupina GCfaze1 obsahuje skupiny pravidel pro odhalování těchto chyb:

8 Jeho autorem je Pavel Květoň.

9 Jeho autorem je rovněž Pavel Květoň.

10 Původním autorem je Jan Hajič; v průběhu času se na dlouholetých úpravách, opravách a doplňování morfologického slovníku podílely zejména Milena Hnátková, Jaroslava Hlaváčová a Hana Skoumalová. Uvádíme tu ještě odkaz na morfologický slovník: <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9> a také odkaz na ryze stochastický tagger Jana Hajiče: <http://hdl.handle.net/11858/00-097C-0000-0001-4904-2> (v Kontrole však jeho tagger nebyl použit). Je možné si jej vyzkoušet pomocí aplikace: <http://lindat.mff.cuni.cz/services/morphodita/>.

— chybějící nebo přebývající čárka ve větě (viz oddíl 4.1.11),
— stylistická pravidla pro některá slova, jež jsou ve větě použita nadbytečně (oddíl 4.2),
— skupina pravidel pro zpracování některých jednoduchých slovesných jevů (oddíl 4.1.2),
— pravidla zpracovávající některé chybné slovní tvary (4.1.12).

6. krok: První aktivace skupiny pravidel GCfaze2, což jsou nedisambiguační pravidla specifická pro Kontrolu (dále budeme nedisambiguační pravidla obsažená ve skupinách GCfaze1 a GCfaze2 označovat jako GC pravidla podle anglického označení *g*rammar *c*hecker). Jejich činnost je obecně sice závislá na výsledku disambiguace, ale před touto jejich prvou aktivací při zpracování vstupní věty se dosud žádná disambiguační pravidla nepoužila, tj. zpracovává se věta v podobě výstupu z morfologické analýzy. Skupina GCfaze2 obsahuje skupiny (morfo)syntaktických, foneticko-fonologických (vokalizace předložek) a pravopisných pravidel popsané výše v oddílu 4.1.

7. krok: Aktivuje se skupina disambiguačních pravidel: jednotlivá pravidla se na větu aplikují v nekonečné smyčce, dokud se ve větě u jednotlivých slov disambiguací stále odstraňují morfologické značky, jež jsou v daném kontextu nesprávné. V této fázi se vždy před každou aplikací disambiguačního pravidla „uchová" aktuální stav, a pokud nějaké disambiguační pravidlo odstraní u nějakého slova všechny zbývající značky, má disambiguační systém za to — za předpokladu, že disambiguační pravidla opravdu vystihují gramatiku češtiny a neobsahují chybu —, že ve vstupní větě je chyba. Aktivuje se proto skupina pravidel GCfaze2 (viz 6. krok výše; při této aktivaci jsou tato pravidla nyní závislá na výsledku předchozí disambiguace).

Disambiguační pravidla jsou obecně pravidla užívaná k nestochastické morfologické disambiguaci (korpusových) textů (srov. Petkevič, 2006; Jelínek et al., 2011), a to zejména textů obsažených v synchronních korpusech současné češtiny řady SYN (SYN2005, SYN2006PUB, SYN2009PUB, SYN2010, SYN2013PUB, SYN, viz např. Český národní korpus — SYN2010, 2010). Pravidla mají přísně formální podobu: jsou to počítačové programy psané ve speciálním programovacím jazyce LanGr (srov. Květoň, 2006).[11] Disambiguační systém se skládá ze tří hlavních podsystémů:

— lemmatizace: stanovení náležitého lemmatu daného slovního tvaru v daném kontextu,
— slovnědruhová disambiguace, která zjednoznačňuje slovnědruhovou homonymii slovních tvarů jejich slovnědruhově jednoznačnou interpretací,
— disambiguace morfologických vlastností morfologicky víceznačných slov.

11 Autorem tohoto jazyka je Pavel Květoň.

Disambiguačními pravidly užitými v Kontrole pro Microsoft Office 2003, resp. 2007, se značkovaly tehdy vyvíjené korpusy (tj. v období 2004–2007). V květnu 2014 je pravidel okolo 2500 (v době, kdy byla vyvinuta Kontrola pro verzi Microsoft Office 2003, jich bylo cca 1950; při vývoji Kontroly pro verzi Microsoft Office 2007 jich bylo cca 2200). Pravidla jsou soustředěna ve skupinách pravidel, které mají hierarchickou stromovou strukturu: nadřazená skupina obsahuje skupiny podřazené, jež obsahují zase své podřazené skupiny atd., až hierarchicky nejnižší skupiny obsahují pravidla, tedy koncové listy stromového systému. Skupiny se aktivují v pořadí daném prvním průchodem stromu (tj. rekurzívně podle pořadí v každé skupině).

Níže uvádíme příklad jednoduchého disambiguačního pravidla: je to pravidlo využívající samozřejmého faktu, že v češtině nemohou stát ve větě tři předložky těsně za sebou (mezi znaky /* a */ jsou vloženy komentáře):

```
Rule TriPredlozky {
/* tři předložky za sebou */

prep1 = ITEM IsSafe Preposition; /* kontext */
possnotprep2 = ITEM Possible not Preposition; /* disambiguační místo */
prep3 = ITEM IsSafe Preposition; /* kontext */

LEAVE ONLY not Preposition IN possnotprep2;

}; // konec pravidla
```

Pravidlo se nazývá TriPredlozky a stanoví, že nachází-li se mezi dvěma slovy, která jsou předložkami a nemají žádnou jinou slovnědruhovou interpretaci (v našem pravidle jsou to slova označená jako *prep1* a *prep3*), slovo, jež má nepředložkovou interpretaci (v našem příkladovém pravidle je to slovo označené jako *possnotprep2*), pak toto slovo není předložka a disambiguační systém ponechá (příkaz LEAVE ONLY) u tohoto slova jen všechny možné nepředložkové morfologické interpretace. Pravidlo se s úspěchem uplatní například na větu:

(118) Sedl si *na místo v* tramvaji.

Předpokládáme, že po fázi morfologické analýzy nebo po provedené částečné disambiguaci věty (118) jsou slova *na* a *v* slovnědruhově jednoznačné předložky a tvoří takzvaný kontext, který se pravidlem nezmění, jen je v pravidle využíván. Slovo, jehož interpretace podléhá změně, nazýváme disambiguačním místem, v našem příkladu je to — ludi gratia — slovo *místo*. Pravidlo u něj ponechá pouze všechny nepředložkové interpretace, tj. interpretaci substantivní, případně spojkovou (a odstraní tak interpretaci předložkovou).

Na rozdíl od disambiguačních pravidel jsou GC pravidla specifická pro Kontrolu české gramatiky. I tato pravidla jsou ovšem psána v programovacím jazyce LanGr. Každé pravidlo patří k nějaké skupině pravidel jistého typu (např. *Chyba ve shodě pří-*

sudku s podmětem). Obsahuje podobně jako pravidla disambiguační také popis zkoumané chybné slovosledné konfigurace, ale na rozdíl od nich neprovádí disambiguaci. Neobsahuje tudíž disambiguační příkazy, nýbrž zahrnuje:

— ohlášení (vysvětlení) konkrétní chyby v rámci obecného typu,
— návrh opravy,
— prioritu chyby.

Ohlášení (vysvětlení) konkrétní chyby je text, který v rámci obecného typu chyby (např. *Chyba ve shodě přísudku s podmětem*) podrobněji popisuje konkrétní chybu, např. *Určitý slovesný tvar se neshoduje s podmětem věty ve jmenném rodě nebo čísle*. Návrh opravy jednak vymezí místo v textu, jehož se oprava týká, jednak obsahuje pokyn, jak text změnit, např. *Zaměňte skupinu_slov_1 za skupinu_slov_2*. Pravidlo ještě obsahuje tzv. prioritu chyby, což je celé kladné číslo, které odráží závažnost chyby: čím vyšší číslo, tím je chyba závažnější nebo specifičtěji určená. Může se totiž stát, že vstupní věta je chybná z více důvodů, a jejich odlišnou závažnost/specifičnost vystihují právě odlišné priority, přičemž při konkurenci více chyb, jejichž kontexty se překrývají, Kontrola hlásí chybu s nejvyšší prioritou; pokud se kontexty nepřekrývají, hlásí se více chyb (pro každý disjunktní kontext chyba s nejvyšší prioritou, viz podrobněji dále v tomto oddílu).

Pro ilustraci uveďme dvě GC pravidla. První z nich je pravidlo, které uživatele upozorňuje na nesprávné kladení čárky ve specifické slovní konfiguraci. Aplikaci pravidla předvedeme na větě:

(119) *Udělal *to, proto že* měl strach z nadřízeného.

Kontrola nejprve podtrhne úsek vyznačený kurzívou a potom po kliknutí na pravé tlačítko myši vyzve uživatele

Zaměňte *to, proto že* za *to, protože*

Klikne-li nyní uživatel na tlačítko *Gramatika*, ohlásí systém typ chyby: „Chybějící nebo přebývající čárka ve větě“. Příslušné GC pravidlo je totiž typu

Typ: Chybějící nebo přebývající čárka ve větě
[„Gramatika“,“Gramatika+Styl“]

a patří, jak vidno, k modům *Gramatika* a *Gramatika+Styl* a do skupiny GC pravidel zabývajících se nesprávným kladením čárek. Jelikož na chybu „přišlo“ pravidlo z této skupiny, vydala Kontrola výše uvedené hlášení: „Chybějící nebo přebývající čárka ve větě“. Součástí pravidla je i podrobnější vysvětlení chyby (nejen její typ): „Spojky **proto** a **že** je v této větě třeba psát dohromady (**protože**) nebo je oddělit čárkou.“

Pravidlo vypadá takto:

```
Rule Protoze {
/* Komentář: GC pravidlo pro chybějící čárky (a špatně — odděleně psané — formy) u spojení proto že, proto aby, pro to že… */

ErrorPriority 74;

ITEM SentenceStart;

SEQUENCE OF lower form!=„že“;
/* Poznámka: „Že“ je vyloučeno kvůli typu věty: „Říkal, že ho to mrzí, proto že za námi také přišel.“ */

ITEM Comma;
ItemProto = ITEM lower form ==„proto“;
ItemZe = ITEM lower form ==„že“;

Error Report „Spojky proto a že je v této větě třeba psát dohromady (protože) nebo je oddělit čárkou.“;

SUGGEST DELETION OF ItemProto;
SUGGEST WORDFORM „protože“ ON ItemZe;

}; // konec pravidla Protoze
```

Mimo vysvětlující komentáře a poznámky pro autora pravidel (nikoli pro uživatele Kontroly) obsahuje pravidlo:

(i) prioritu (zde 74) a

(ii) popis posloupnosti slov zkoumaného úseku:

— formální počátek věty,
— posloupnost slovních tvarů, v níž se nenachází slovo *že*,
— čárku,
— slovo *proto* následované slovem *že*,
— ohlášení chyby (Error Report), které se uživateli objeví, bude-li chtít chybu podrobněji vysvětlit (tlačítko *Vysvětlit*),
— návrh opravy: odstranění slova *proto*,
— nahrazení slova *že* slovem *protože*.

Obecně systém umožňuje nahradit (případně prázdný) řetězec slov jiným (případně prázdným) řetězcem slov. V uvedeném příkladu dochází k náhradě

proto že → protože

což je náhrada velmi jednoduchá. Je-li ve větě například chybná předložková rekce

(120) *Udělal to *pro naší* potěchu.

navrhne příslušné GC pravidlo opravu:

naší → naši.

Systém tedy využívá nejen morfologické analýzy zkoumaného textu: z analýzy se dozví lemma tvaru *naší*, tj. *náš*, a jelikož navrhuje akuzativní tvar po předložce *pro*, musí tento tvar také zkonstruovat. Ze znalosti lemmatu a nutnosti nabídnout akuzativní tvar tohoto lemmatu syntetizuje tvar *naši*. V tomto kroku své činnosti používá tedy Kontrola ve svých pravidlech morfologickou syntézu.

Uveďme nyní ještě jedno ne úplně triviální pravidlo, jež upozorňuje uživatele na chybnou atrakci typu:

(121) Čirou náhodou pracujete v obchodním *zastoupením* firmy IBM.

Pravidlo se nazývá AtrLocInsSgGC, je typu

```
Typ: Chyba ve jmenné skupině
[„Gramatika“,“Gramatika+Styl“]
```

a vypadá takto (mezi znaky /* a */ jsou opět komentáře):

```
rule AtrLocInsSgGC {
 /* lokálová předložka není uspokojena */

ErrorPriority 90;

predl = ITEM (IsSafe Preposition and Not Instrumental) and (Possible Locative);

vycpavka = SEQUENCE OF (((IsSafe AdjectiveSynt) and (Possible ((Masculine or Neuter) and
Locative and Singular))) or (IsSafe Adverb or Particle));

synadj = ITEM (IsSafe AdjectiveSynt) and (Possible (Masculine or Neuter) and
Locative and Singular);

SEQUENCE OF IsSafe Adverb;
```

```
subs = ITEM IsSafe Noun and ((Instrumental and Singular) or (Dative and Plural));

POST-SENTENCE ITEM MustNotBe AdjectiveVerbal;
/* vylučuje případy jako „na prvním místem pohrdajícího závodníka“ */

Error Report „Po předložce vyžadující šestý pád musí být podstatné jméno v šestém pádě.“;

Suggest Tag {case = Loc, number = Sg} ON subs;

}; // konec pravidla
```

Toto pravidlo se zabývá slovem označeným jako *subs*: je to potenciálně homonymní substantivní tvar v Isg nebo Dpl. Tento tvar je tedy slovnědruhově nehomonymní substantivum, které je navíc v Isg nebo Dpl, přičemž „nebo“ není vylučovací, tvar tedy může mít obě interpretace (uvedený tvar má tyto vlastnosti buď již po fázi morfologické analýzy, nebo se k nim dospělo procesem disambiguace). Po stanovení vysoké priority (90) následuje popis celkového kontextu, přičemž pořadí jednotlivých elementů odráží slovosled (skupin) slov:

— *predl* je slovnědruhově jednoznačná (tj. nehomonymní) předložka, která nemá instrumentálovou valenci a může mít valenci lokálovou,
— *vycpavka* představuje (případně prázdnou) posloupnost syntaktických adjektiv, která mohou být v maskulinu či v neutru a v Lsg, a/nebo adverbií nebo částic; do této posloupnosti mohou patřit slova patřící ke všem těmto třem slovním druhům,
— *synadj* je nehomonymní syntaktické adjektivum, které může být v maskulinu či neutru a zároveň v Lsg,
— poté následuje (případně prázdná) posloupnost adverbií (SEQUENCE OF IsSafe Adverb),
— *subs* je zmíněný chybný tvar, který je třeba opravit
— a konečně následuje buď konec věty, anebo tvar, který určitě není slovesným adjektivem (komentář uvádí důvod takovéto specifikace).
— Poté následuje chybové hlášení (Error Report), které obdrží uživatel: *Po předložce vyžadující šestý pád musí být podstatné jméno v šestém pádě.*

Je tedy vymezen kontext, v němž je tvaru *subs* užito chybně, a zbývá navrhnout uživateli opravu

```
Suggest Tag {case = Loc, number = Sg} ON subs
```

což je příkaz znamenající: udělej z tvaru identifikovaného identifikátorem *subs*, který je v Isg nebo Dpl, tvar Lsg. Kontrola zde z chybného tvaru nejprve vnitřně vytvoří jeho lemma a poté morfologickou syntézou vytvoří jeho tvar v Lsg.

Jakmile Kontrola ohlásí uživateli chybu, čeká na zásah uživatele. Uživatel zasáhne do celého procesu tím, že opraví příslušné místo ve větě (může ovšem také chybu

ignorovat a tím celé zpracování dané věty končí a Kontrola přechází ke zpracování další věty). Po uživatelově opravě Kontrola prověřuje opravený text. Pokud Kontrola nezjistí v prověřované větě gramatickou, pravopisnou ani stylistickou chybu, sama mlčky přechází ke zpracování další věty.

GC pravidla obsahují celočíselnou prioritu (v rozsahu 1–100), která vystihuje závažnost a specifičnost chyby a příslušného pravidla, které ji odhaluje. Priorita přichází ke slovu, když je ve vstupní větě víc chyb, a Kontrola hlásí chyby podle závažnosti a specifičnosti: na nejzávažnější či nejspecifičtější chybu upozorní uživatele nejdříve. Je-li v prověřované větě chyb víc, mohou nastat dva případy: (a) v případě, že kontexty, v nichž došlo k chybám, jsou disjunktní, tj. textově se nepřekrývají, hlásí se více chyb najednou (pro každý takový kontext jen chyba s nejvyšší prioritou); (b) pokud se kontexty překrývají, hlásí se pro takové kontexty pouze jediná chyba — ta s nejvyšší prioritou. V obou případech se tedy nehlásí chyby s nižšími prioritami pro daný chybový kontext nebo pro skupinu překrývajících se chybových kontextů. Po uživatelově opravě je věta kompletně prověřena znovu. V důsledku opravy chyby se mohou zároveň opravit i dosud nehlášené chyby s nižší prioritou (systém o nich věděl, ale neohlásil je). Kontrola je tedy ctižádostivá: chce nalézt všechny chyby ve větě, a tak se opakovaně aktivuje, dokud věta není správně opravena, nebo dokud uživatel nepřejde na větu další.

Upozorněme nyní na rozdíl mezi morfologickou disambiguací a Kontrolou české gramatiky; ten je dán jejich odlišnými cíli a úkoly:

— Disambiguace chová apriorní důvěru ke správnosti textu: bezelstně předpokládá, že věta je správná, a odstraňuje (v daném kontextu) nesprávné značky. Pokud je věta gramaticky či pravopisně nesprávná, snaží se zpracovat co nejvíce jejích správných úseků a nehlásí žádné chyby. Spouští se dávkově, tj. nekomunikuje interaktivně s uživatelem.
— Kontrola je naopak ke správnosti textu spíše nedůvěřivá: vyhledává chyby, označuje chybná místa, hlásí chyby a navrhuje jejich opravy, ale teprve po důkladné prověrce kontextu celé věty (aby nehlásila chyby ve správných větách). Při hlášení chyby předpokládá interakci s uživatelem a jeho intervence.

Gramatika, která řídí Kontrolu, je vyjádřena, jak už bylo řečeno, jako velmi složitý systém kontextových pravidel implementovaných jako počítačové programy psané ve speciálním jazyce LanGr: v tomto programovacím jazyce jsou psána jak disambiguační pravidla, tak GC pravidla. Disambiguačních pravidel bylo v době vytváření druhé verze Kontroly pro Microsoft Office 2007 celkem cca 2200, GC pravidel 820. Softwarová implementace obou podsystémů pravidel zaručuje, že Kontrola zpracuje vstupní text na obrazovce velmi rychle, neboť uživatel nemůže čekat na prověření textu déle než pár vteřin. Uvážíme-li celý složitý proces zpracování od segmentace a tokenizace vstupního textu, přes morfologickou analýzu až po disambiguační pravidla a GC pravidla (přičemž převážná většina pravidel se na danou větu uplatňuje mnohokrát v cyklu), je vysoká rychlost systému velmi náročný požadavek, který však Kontrola splňuje. Poznamenejme ještě, že při zpracování textu se neuplatňují statistické metody.

Jak bylo řečeno, ústřední součástí Kontroly jsou především GC pravidla a také pravidla disambiguační. Neprovádí se tedy syntaktická analýza (*parsing*), nevytvářejí se tak žádné syntaktické struktury ani na výstupu, ani vnitřně. Veškerá gramatika, především syntax, se vyvozuje na úrovni slovních druhů a morfologie a vyvozená syntaktická pravidla se pak formálně implementují. Parser spolu s formální gramatikou obsahující pravidla je úkol mnohem složitější, nicméně někteří badatelé používají plnohodnotné parsery i na prověrku gramatické správnosti textu. Tak si počínají např. Jakubíček et al. (2010), kteří se mimo jiné snaží stanovit správné a chybné postavení interpunkčních znamének na základě činnosti chart-parseru *synt*.

6. ÚSPĚŠNOST *KONTROLY ČESKÉ GRAMATIKY*, SROVNÁNÍ SE SYSTÉMEM *GRAMMATICON*

Na základě důkladného testování *Kontroly české gramatiky* bylo zjištěno, že Kontrola odhalí cca 30–40 % chyb v českých textech. Velmi užitečná je hlavně v odhalování absence čárek, neboť tato chyba je mezi gramotnými pisateli snad nejrozšířenější. Kontrola se „ztrácí" při zpracování velmi dlouhých vět a také vět interpunkčně velmi členitých, tj. vět obsahujících závorky, pomlčky, mnoho čárek, zkratky končící tečkou, neznámá slova.

Srovnání s *Grammaticonem* vyznívá pro Kontrolu lépe, a to zejména z těchto důvodů:

— Kontrola dokáže zpracovat všechny hlavní typy gramatických chyb v českých textech, aniž by hlásila nepřijatelně vysoký počet chyb domnělých (*false flags*): tj. jen zřídkakdy neoprávněně hlásí chybu, a je tedy úspěšná zejména při zpracování gramaticky a pravopisně správného textu. Naopak *Grammaticon* poměrně často neoprávněně hlásí chybu (zvláště u slovesných skupin, u parazitických slov, ve jmenných skupinách, ve shodě přísudku s podmětem i ve slovosledu), viz tabulku 1 níže.
— Kontrola si vede velmi dobře v odhalování chyb v kladení čárek a v tomto směru *Grammaticon* jasně předčí (má přibližně desetkrát vyšší úspěšnost).
— Kontrola hlásí vždy jen jednu (nejzávažnější) chybu pro daný kontext (nebo pro překrývající se kontexty), *Grammaticon* někdy hlásí touž chybu i vysvětlení dvakrát.
— *Grammaticon* častěji nerozpozná frapantní chybu, což se Kontrole přihodí spíše výjimečně.
— Kontrola pokrývá patrně větší množství jevů (autor však toto tvrzení již nedokáže exaktně doložit).
— Kontrola se mnohem méně věnuje stylu (což je z hlediska počítačového zpracování nejasná kategorie) než *Grammaticon*.
— Hlášení Kontroly jsou výstižnější a jasnější.

			Kontrola č. gramatiky		Grammaticon	
	#slov	#všech chyb	rozpoznané chyby	domnělé chyby	rozpoznané chyby	domnělé chyby
test1.rtf	831	108	34	4	38	78
Gctest.rtf	416	55	50	3	22	33
grchtext.rtf	1179	90	37	9	33	27
Celkem	2226	253	121	16	92	134

TABULKA 1: Srovnání výsledků testování obou systémů na třech souborech: test1.rtf, Gctest.rtf a grchtext.rtf.

V tabulce 1 jsou pro srovnání uvedeny výsledky testování obou systémů na třech souborech: soubory test1.rtf a Gctest.rtf obsahují vzorové věty, soubor grchtext.rtf obsahuje běžný český text. Tabulka zachycuje počet všech slov, počet všech chyb a z nich pak chyby rozpoznané oběma systémy a chyby domnělé, tj. neoprávněně hlášené (*false flags*). Je vidět, že z hlediska obou typů chyb je na tom Kontrola výrazně lépe (s výjimkou kategorie rozpoznaných chyb v souboru test1.rtf).

Z provedených srovnávacích testů se bez rozsáhlého testování bohužel nedá hodnověrně odvodit pokrytí (*recall*) a především přesnost (*precision*) výkonu obou systémů. Hodnověrné testy by se musely provést na několika zkoumaných vzorcích vět různých typů a byly by velice pracné, zejména pak výpočet přesnosti.

7. SHRNUTÍ POZITIVNÍCH A NEGATIVNÍCH RYSŮ *KONTROLY ČESKÉ GRAMATIKY*

V bodech nyní shrňme klady a zápory Kontroly. Kontrola má tyto kladné stránky:

— Je opatrná, zřídkakdy neoprávněně hlásí chybu, je tedy úspěšná při zpracování gramaticky a pravopisně správného textu.
— Jejím východiskem je jemná lingvistická analýza.
— Velmi dobře dokáže odhalit jasné chyby.
— Velmi dobře si vede v odhalování chyb v kladení čárek.
— Pokrývá poměrně široký okruh jevů.
— Její hlášení jsou srozumitelná, systém je uživatelsky přívětivý.
— Rychlost zpracování prověřovaného textu je naprosto uspokojivá.
— Zpracovává i ryze formální chyby (například víc mezer či interpunkčních znamének za sebou).

Kontrola má ovšem i zápory:

— Až na pár jevů nezachycuje valenci.
— Nezpracovává záporovou shodu.
— Je zmatena při prověrce strukturně i formálně (mnoho interpunkce) složitých a dlouhých vět.

— Občas nehlásí i vyloženou chybu.
— Má rezervy ve zpracování konkrétních slovních tvarů, v nichž se často dělají chyby (např. zájmenných tvarů *mě*, *mně*).

8. MOŽNOSTI DALŠÍHO ROZVOJE *KONTROLY ČESKÉ GRAMATIKY*

Současná podoba Kontroly české gramatiky má samozřejmě rezervy a leccos se na ní dá vylepšovat, zdokonalovat a rozvíjet. Toto vylepšování má ovšem smysl jen v případě, že společnost Microsoft projeví o další verzi Kontroly zájem. Níže naznačujeme, jakým směrem by se další rozvoj Kontroly měl ubírat především. Obecně by se Kontrola měla zlepšovat v obou parametrech: v pokrytí a přesnosti (jakkoli se tyto ukazatele zjišťují velmi namáhavě). V této souvislosti uveďme zvláště tyto požadavky:

— Další vylepšování Kontroly se nyní s výhodou může opřít o dnes už mnohem kvalitnější morfologickou disambiguaci (ve srovnání s obdobím vývoje Kontroly, tj. před cca deseti lety), o větší morfologický slovník, o rozsáhlejší a lépe značkované korpusy a vůbec o pokrok v oblasti počítačového zpracování češtiny. Je však zapotřebí vyvinout ještě mnohem více disambiguačních pravidel i GC pravidel, zejména v oblasti valence všech klíčových slovních druhů: sloves, substantiv, adjektiv a adverbií. Velmi dobře zde mohou posloužit jako vhodná datová základna syntaktické stromové struktury (*treebanks* či *parsebanks*) a stále větší a lépe značkované korpusy češtiny (zvláště řady SYN).
— Je vhodné zjistit, jakých chyb se uživatelé v současnosti dopouštějí nejčastěji, a přihlédnout v tomto směru ke snižující se písemné gramotnosti mluvčích češtiny, příp. k dysortografii.
— Je nutné získávat od uživatelů zpětnou vazbu: co jim dnes na Kontrole nejvíc vadí, co by od ní obzvláště očekávali.

9. ZÁVĚR

V tomto příspěvku byl podrobně popsán automatický počítačový systém Kontroly české gramatiky, který automaticky prověřuje vstupní český text, hledá v něm gramatické, některé pravopisné a formální chyby a v omezené míře i chyby stylistické a hlásí je uživateli, většinou spolu s iniciativním návrhem, jak chybu opravit. Systém je založen na gramatických vztazích mezi slovy — vychází z preskriptivní gramatiky a pravopisu spisovné češtiny — a sleduje tak prohřešky proti této gramatice a pravopisu. Nezabývá se obyčejnými chybami v pravopisu slov bez ohledu na kontext (obyčejnými překlepy), neplní tedy funkce spell checkeru. Gramatické, pravopisné i stylistické vztahy mezi slovy jsou implementovány v podobě kontextových syntaktických, morfologických, foneticko-fonologických a též některých stylistických pravidel. Tato pravidla jsou dvojí povahy: pravidla disambiguační a specifická pravidla pro vlastní kontrolu gramatiky. Kontrola české gramatiky je implementována v systému Microsoft Office verze 2003, 2007, 2010 a 2013 a je v nich k dispozici od poloviny roku 2005 jako součást editoru Microsoft Word.

LITERATURA:

Akademická pravidla českého pravopisu (1993). Praha: Academia.

Český národní korpus — SYN2010 (2010). Praha: Ústav Českého národního korpusu FF UK v Praze. Dostupné z WWW: <http://www.korpus.cz>.

Hajič, Jan (2000): Popis morfologických značek — poziční systém [online]. In: Marie Kopřivová — Jan Kocek (eds.), *Manuál korpusového manažeru Bonito*. Cit. 10. 11. 2014. Dostupné z WWW: <http://ucnk.ff.cuni.cz/bonito/znacky.php>.

Hajič, Jan (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum.

Havránek, Bohuslav — Jedlička, Alois (1981): *Česká mluvnice*. Praha: Státní pedagogické nakladatelství.

Jakubíček, Miloš — Horák, Aleš (2010): Punctuation detection with full syntactic parsing. *Research in Computing Science, Special Issue: Natural Language Processing and Its Applications*, 46, s. 335–343.

Jelínek, Tomáš — Petkevič, Vladimír (2011): Systém jazykového značkování současné psané češtiny. In: Vladimír Petkevič — Alexandr Rosen (eds.), *Korpusová lingvistika Praha 2011: 3, Gramatika a značkování korpusů* [Studie z korpusové lingvistiky, 16]. Praha: Nakladatelství Lidové noviny — Ústav českého národního korpusu FF UK v Praze, s. 154–170.

Květoň, Pavel (2006): *Rule-Based Morphological Disambiguation* [disertační práce]. Praha: MFF UK v Praze.

MČ2: Komárek, Miroslav — Kořenský, Jan — Petr, Jan — Veselková, Jarmila (eds.) (1986): *Mluvnice češtiny: 2, Tvarosloví*. Praha: Academia.

MČ3: Daneš, František — Grepl, Miroslav — Hlavsa, Zdeněk (eds.) (1987): *Mluvnice češtiny: 3, Skladba*. Praha: Academia.

Microsoft Office™ (2003, 2007, 2010, 2013).

Pala, Karel — Rychlý, Pavel — Smrž, Pavel (2003): Text corpus with errors. In: Václav Matoušek — Pavel Mautner (eds.), *Text, Speech and Dialogue: 6th International Conference, TSD 2003, České Budějovice, Czech Republic, September 8–12, 2003: Proceedings*. Berlin: Springer, s. 90–97.

Petkevič, Vladimír (2004): Využití pravidel pro negaci v automatickém značkování českých korpusů. In: Zdeňka Hladká — Petr Karlík (eds.), *Čeština — univerzália a specifika 5: Sborník 5. mezinárodního setkání bohemistů v Brně*. Praha: Nakladatelství Lidové noviny, s. 143–150.

Petkevič, Vladimír (2006): Reliable morphological disambiguation of Czech: rule-based approach is necessary. In: Mária Šimková (ed.), *Insight into the Slovak and Czech Corpus Linguistics*. Bratislava: Veda, s. 26–44.

PMČ: Karlík, Petr — Nekula, Marek — Rusínová, Zdenka (eds.) (1995): *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny.

Pravidla českého pravopisu (1993). Praha: Academia.

Richter, Michal — Straňák, Pavel — Rosen, Alexandr (2012): Korektor — a system for contextual spell-checking and diacritics completion. In: Martin Kay — Christian Boitet (eds.), *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai: The COLING 2012 Organizing Committee, s. 1–12.

Šmilauer, Vladimír (1966): *Novočeská skladba*. Praha: Státní pedagogické nakladatelství.

Šmilauer, Vladimír (1972): *Nauka o českém jazyku*. Praha: Státní pedagogické nakladatelství.

Vladimír Petkevič | Ústav teoretické a komputační lingvistiky FF UK v Praze
<Vladimir.Petkevic@ff.cuni.cz>