Петта национална конференција со меѓународно учество ЕТАИ'2000
Fifth National Conference With International Participation ETAI'2000
Охрид, Република МАКЕДОНИЈА – Ohrid, Republic of MACEDONIA
21–23. IX 2000

I1-2

КОРЕКТОР М – СОФТВЕРСКИ ПАКЕТ ЗА КОРЕКТУРА И РАЗДЕЛУВАЊЕ НА СЛОГОВИ НА ТЕКСТОВИ НА МАКЕДОНСКИ ЈАЗИК

Дејан Ѓорѓевиќ, Драган Михајлов, Никола Грчевски

Електротехнички факултет, Универзитет Св. Кирил и Методиј,
П. Фах 574, МК-1001 Скопје, Република Македонија
dejan@cerera.etf.ukim.edu.mk

Извадок **- Коректор М е софтверски пакет за автоматска коректура и разделување на слогови за македонски јазик која се вградува во процесорот на текстови Microsoft Word 97 или Microsoft Word 2000. Програмата овозможува коректура на текстови во електронска форма напишани на македонски јазик, при што се откриваат грешките при внесување на текстот и најчесто се нуди замена за погрешно внесениот збор. Покрај ова пакетот содржи и модул кој овозможува пренесување на дел од последниот збор од крајот на редот во наредниот (хифенација) со правилна поделба на зборовите на слогови според правилата на македонскиот правопис.**

Клучни зборови **– коректура, разделување на слогови, македонски јазик**

1. ВОВЕД

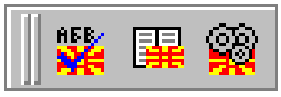
Во последните години со напредокот на информатичката технологија сé поголем дел од текстовите наменети за печатење и секојдневна кореспонденција настануваат на компјутер. При внесувањето на текстови најчесто се користи тастатурата, а при тоа доаѓа и до грешки во внесувањето. Отстранувањето на овие грешки е макотрпна работа која им се доверува на коректорите. За таа цел за повеќето јазици направени се програми за автоматска коректура на текстовите. Во овој труд ќе биде презентирана една таква програма наменета за македонскиот јазик развиена на Електротехничкиот факултет во Скопје наречена Коректор М.

Коректор М е програма за автоматска коректура и разделување на слогови за македонски јазик која се вградува во процесорот на текстови Microsoft Word 97 или Microsoft Word 2000. Програмата вклучува база на македонски зборови (вградени речници) и овозможува проверка на правописот (spell checking) и разделување на слогови (hyphenation) на македонските текстови внесени во Microsoft Word 97 или Word 2000. При тоа се откриваат грешките при внесување на текстот и најчесто се нуди замена за погрешно внесениот збор. Овозможено е и користење на кориснички речник во кој корисникот може да додава сопствени зборови. Со нагодување на некои од опциите, оставена е можноста функционирањето на Коректор М максимално да се прилагоди кон потребите на корисникот.

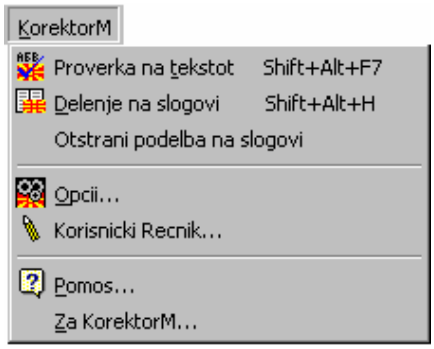
2. ИНСТАЛАЦИЈА И КОРИСНИЧКИ ИНТЕРФЕЈС

Инсталацијата е автоматска по што Коректор М се интегрира во околината на Microsoft Word 97/2000 со посебно мени и алатник, со што се обезбедува максимална функционалност.

По инсталацијата програмата се вградува во околината на Microsoft Word и додава алатник (toolbar) (Сл. 1) и ново мени "KorektorM" со неколку опции (Сл. 2).



Сл. 1



Сл. 2

Алатникот се состои од 3 икони од кои првата е за повикување на проверка на текстот, втората за делење на слогови, а третата за поставување на опциите на програмата.

Проверката на текстот може да се повика и преку менито KorektorM со избор на ставката "Proverka na tekstot" или од тастатура со притисок на комбинацијата тастери Shift+Alt+F7. Слично и делењето на слогови може да се повика и со избор на ставката "Delenje na slogovi" или од тастатура со притисок на комбинацијата тастери Shift+Alt+H.

Поставување на опциите на Коректор М освен преку алатникот, може да се повика и со избор на ставката "Opcii…" од KorektorM менито.

Со избор на ставката "Za KorektorM…" се добиваат основните информации за верзијата на програмата, додека со избор на ставката "Pomos..." се повикува вграденото упатство за работа со програмата.

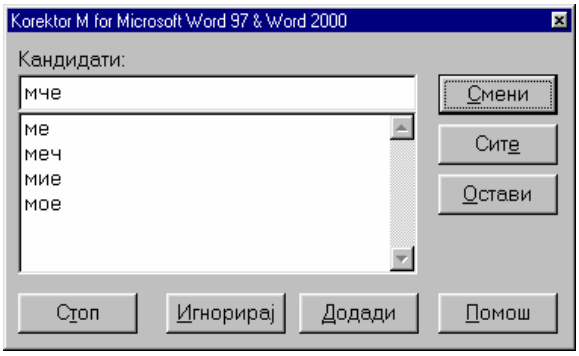
Опцијата за разделување на слогови во текстот на документот додава дополнителни знаци помеѓу слоговите, кои се невидливи за корисникот. Овие дополнителни знаци ја зголемуваат физичката големина на документот. Заради ова дадена е можноста да се отстрани претходно направеното разделување на слогови доколку тоа не е потребно во документот или се менува начинот на разделување, на пример се применува разделување само на последните зборови во редот на документот. Отстранувањето на претходно направената поделба на слогови се прави со избор на опцијата "Otstrani podelba na slogovi" од менито на KorektorM.

3. ПРОВЕРКА НА ТЕКСТ

Проверката на текстот вклучува пронаоѓање и исправање на грешките при внесување на текстот. Проверката на зборовите започнува од местото каде се наоѓа курсорот (дури и ако тој е на средина од некој збор) и продолжува до крајот на текстот. Во случај во моментот на активирањето на проверката, да постои обележан (селектиран) дел од документот, тогаш проверката се врши само на обележаниот дел од текстот.

Коректор М успешно пронаоѓа и помага при исправањето на грешки од неколку типови кои најчесто се среќаваат во типографијата. При проверката за секој збор од текстот се проверува неговото присуство во интерните речници. Ако зборот е пронајден во овие речници, се третира како правилно напишан и се продолжува со проверка на следниот збор од текстот. Кога ќе се наиде на збор кој не е присутен во ниту еден од речниците, се јавува грешка и се прави обид да се помогне при исправањето на грешката. Во овој случај Коректор М ќе се обиде да изгенерира зборови кандидати за замена на непостоечкиот (најверојатно погрешно напишан) збор, и тоа земајќи ги предвид следниве можни грешки:

- Некоја буква од зборот е испуштена
- Некоја буква од зборот е вишок
- Некоја буква од зборот е заменета со друга
- На некои две соседни букви им се заменети местата



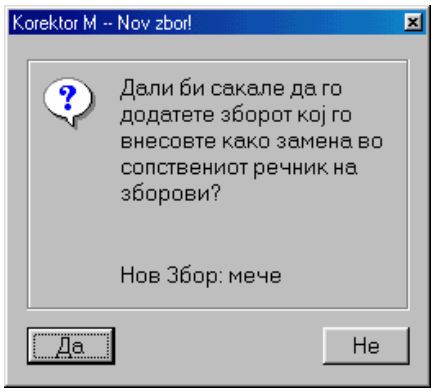
Сл. 3

За секој збор кој нема да биде пронајден во интерната база на зборови, на екранот се појавува форма како на (Сл. 3) која го прикажува погрешниот збор и листата на зборови кандидати за замена. По ова корисникот може да избере една од неколкуте предложени акции. Со избор на некој од понудените кандидати од листата и избор на копчето "Смени" се врши замена на погрешно напишаниот збор во текстот со избраниот од листата. Доколку зборот е точен, но е пријавен како грешка бидејќи не бил пронајден во речниците (на пр. лични имиња или помалку користени и стручни зборови) тој може да се остави неизменет во текстот доколку се избере копчето "Остави".

Коректор М овозможува и проширување на неговиот речник со додавање на нови зборови. Ако сакаме зборот пронајден како погрешен (бидејќи го немало во интерниот речник) да биде додаден во корисничкиот речник, тоа можеме да го сториме со избор на копчето “Додади”. По ова додадениот збор станува рамноправен со зборовите во вградениот речник и во иднина ќе биде препознаван како валиден збор.

Ако пак, за некој непронајден збор кој не сакаме да го додадеме во корисничкиот речник, сакаме во натамошниот тек на проверката на тековниот текст сите негови појави да не се пријавуваат како грешки и да бидат игнорирани, можеме да избереме “Игнорирај”.

Доколку за некој погрешен збор, правилниот збор кој е валидна замена за грешката не е понуден во листата на кандидати, можно е тој рачно да се поправи во полето на формата каде што е прикажан погрешниот збор. По рачната поправка, со избор на копчето “Смени” се врши замена на погрешниот збор од текстот со рачно поправениот. При тоа програмата ќе праша дали сакаме нововнесениот збор да го додадеме во корисничкиот речник со форма како на Сл. 4 на која можеме да избереме нововнесениот збор да биде или да не биде додаден во корисничкиот речник.



Сл. 4

Доколку пак сакаме сите понатамошни грешки за даден збор автоматски да се заменуваат со некој од листата избран или рачно корегиран збор, можеме да го избереме копчето “Сите”.

По ова сите понатамошни појави на дадениот збор автоматски (без прашање за потврда) ќе бидат заменувани со избраниот (или рачно внесениот) збор за замена.

Со избор на копчето “Стоп” може да се прекине процесот на проверка на текстот, додека со копчето “Помош“ се повикува упатството за користење на програмата.

Коректор М овозможува и проверка на уште една честа грешка при внесување на текстови. Ова е грешката на повторени зборови, односно дуплицирани соседни зборови. Ова е честа грешка при спонтаното пишување, но во некои случаи дуплицирани соседни зборови можат и регуларно да се појават во текстот, па ваквите случаи Коректор М ги пријавува со форма која е само предупредување на што корисникот може да избере дали едниот дупликат да се отстрани или да остане.

4. РАЗДЕЛУВАЊЕ НА СЛОГОВИ

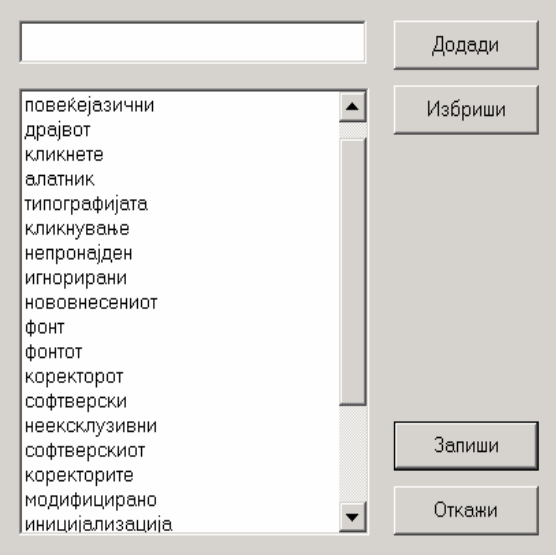
Разделувањето на слогови врши делење на зборовите на слогови за нивно правилно пренесување во наредниот ред. При ова во сите зборови кои ќе се најдат на крајот на редот се вметнуваат знаците за “optional hyphen” кои ќе обезбедат правилно пренесување на зборот во следниот ред според правилата на македонскиот правопис. Делењето на слогови на зборовите започнува од редот во кој се наоѓа курсорот и продолжува до крајот на текстот. Доколку во моментот на активирањето на делењето на слогови, постои обележан (селектиран) дел од документот, тогаш поделбата на слогови се изведува само над обележаниот дел од текстот.

Можно е Коректор М да се нагоди при делењето на слогови да врши поделба на секој збор во документот или само на зборовите кои навистина се пренесуваат во следниот ред. Доколку за обработка на текстот се користи исклучиво MS Word, тогаш нема потреба од разделување на слогови на секој збор во документот со што се добива во брзина, а исто така физичкото зголемување на документот е минимално. Доколку текстовите (документите) кои се подготвуваат во MS Word се доработуваат во некој друг текст процесор, како на пример во Adobe PageMaker или QuarkXress, тогаш неопходно е вклучување на опцијата за делење на слогови на сите зборови во документот. На овој начин делењето на слогови во дорабoтката нема да зависи од поставувањето на маргините на страната, ниту пак од поставувањето на текстот во колони и сл.

5. КОРИСНИЧКИ РЕЧНИК

При работата со Коректор М корисникот има можност за проширување на речникот со додавање на сопствени зборови, притоа формирајќи сопствен кориснички речник. За секој збор пријавен како погрешен затоа што не бил пронајден во интерниот речник, корисникот може со избор на копчето “Додади” да го додаде збо-

рот во корисничкиот речник, по што тој станува рамноправен со зборовите од интерниот речник. При ваквото додавање на зборови во корисничкиот речник може несакајќи да се додаде погрешен збор или да се додаде збор кој нема често да се користи, па затоа постои можноста за измена на овој речник, што се прави со изборот на опцијата “Korisnicki Recnik...” од менито KorektorM со која се активира форма за менување на речникот (Сл. 5).



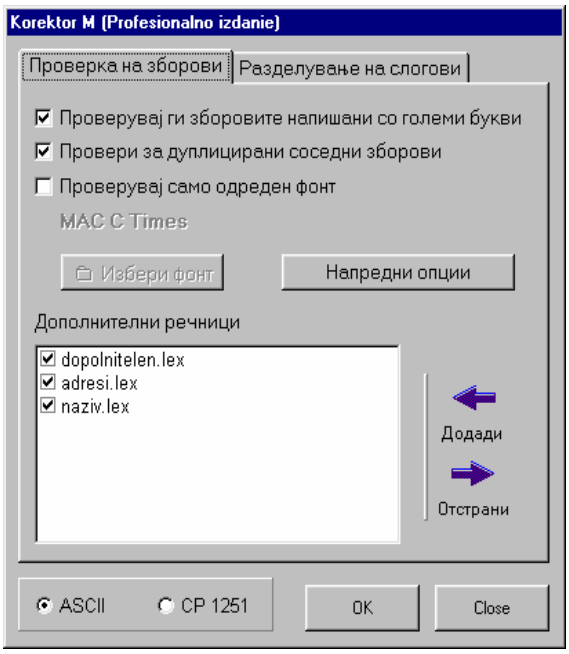
Сл. 5

Оваа форма овозможува додавање и бришење на одредени зборови од речникот. Бришењето се изведува со избор на зборот кој сакаме да го избришеме од корисничкиот речник и со избор на копчето “Избриши”. Преку оваа форма овозможено е и директно додавање на зборови во корисничкиот речник со едноставно внесување на зборот кој сакаме да го додадеме во речникот во полето лево од копчето “Додади” и избор на копчето “Додади”. По извршените измени во корисничкиот речник, се користи опцијата “Запиши” со која се потврдуваат промените и се напушта формата. Со избор на опцијата “Откажи”, се напушта формата без запишување на измените.

6. НАГОДУВАЊЕ НА ОПЦИИТЕ

Нагодување на опциите се избира преку третата иконата од алатникот или со избирање на ставката “Opcii...” од менито KorektorM. Со ова се отвора формата преку која може да се нагодат опциите за начинот на работа на програмата. Во оваа форма преку јазичињата (tabs) на врвот: “Проверка на зборови” и “Разделување на слогови” се избираат формите за нагодување на параметрите за работа на програмата за проверка на текстот и делењето на слогови, соодветно.

На дното на формата може да се избере дали Коректор М ќе работи со текстови напишани во кодниот распоред ASCII со користење на македонски кирилични фонтови (како MAC C Times, MAC C Swiss, Macedonian Tms, Matka, Treska, Pulshelvetika7, …) или со текстови напишани во кодниот распоред CP-1251 со користење на стандардните Windows повеќејазични фонтови (како Arial, Times New Roman, Curier New, Vedrana, …). Овој избор важи и за проверката на текстот и за делењето на слогови.



Сл. 6

Во горниот дел на формата постојат неколку опции кои можат да бидат вклучени или исклучени, а се однесуваат на третирањето на делови од текстот напишани со големи букви, ограничување на доменот на проверка односно делење на слогови само за делови од текстот напишани со одреден фонт и пријавување на исти соседни зборови.

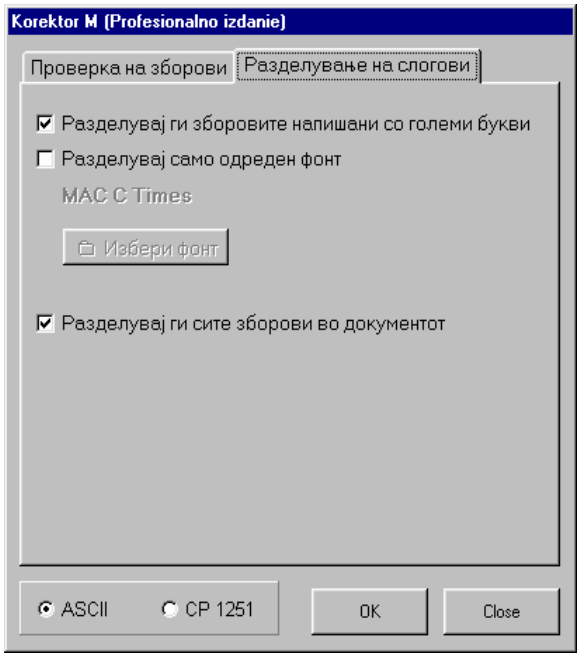
Со избор на јазичето “Проверка на зборови” се активира формата за нагодување на опциите за работа на програмата при проверката на текстовите (Сл. 6).

Во рамките на оваа форма можно е да се вклучат или исклучат неколку опции:

- Проверувај ги зборовите напишани со големи букви
- Провери за дуплицирани соседни зборови

- Проверувај само одреден фонт
- Дополнителни речници
- Напредни опции

Со избор на јазичето “Разделување на слогови” се активира формата за нагодување на опциите за работа на програмата при разделување на слогови (Сл. 7).



Сл. 7

Во рамките на оваа форма можно е да се вклучат или исклучат неколку опции:

- Разделувај ги зборовите напишани со големи букви
- Разделувај само одреден фонт
- Разделувај ги сите зборови во документот.

7. ИЗВЕДБА

При изработката на програмата Коректор М користени се Microsoft Visual C++ 6.0, Borland Delphi 3.0 и Microsoft Visual Basic for Applications. Проверката на текстот се врши на тој начин што, со помош на вградената поддршка на Microsoft Office пакетот за Visual Basic for Applications (VBA), текстот кој се пишува или кој веќе постои, се испитува збор по збор со помош на надворешен модул (динамички поврзана библиотека - DLL) која го врши пребарувањето низ базата на зборови и генерирањето на кандидатите.

За оваа намена е направен речник од околу 300.000 зборови од македонскиот јазик, кои се сместени во база организирана како модифицирано reTRIEval стебло [5], [6]. За потребите на развојот на овој речник оформен е корпус од 10.000.000 зборови кој секојдневно се надградува. Корпусот е статистички обработен за да се оцени фреквенцијата на најчесто употребуваните зборови во македонскиот јазик. Најчесто употребуваните зборови потоа се поделени во зборовни групи (именки, глаголи, придавки, ...) и за секој од нив е извршено автоматско или полуавтоматско генерирање на сите зборовни форми.

За именките на пример е извршено рачно генерирање на множината и во некои случаи на избројната и збирната множина (затоа што не постои правило кое би овозможило негово автоматизирање) и автоматско генерирање на вокативната и членуваните форми на именката. За некои именки полуавтоматски се генерирани и нивните членувани деминутивни и пежоративни форми. За глаголите е изведено автоматско менување по лица и времиња, како и полуавтоматско генерирање на глаголски прилог и глаголска именка (каде тоа е можно). За придавките е извршено полуавтоматско генерирање на компаратив и суперлатив и нивно автоматско членување.

За потребите на модулот за делење на слогови не се користи базата на македонски зборови, туку се применуваат директно правилата на македонскиот правопис [1], [2], [3]. Секој збор кој треба да се подели на слогови се пренесува на функција од DLL библиотеката која во зборот вметнува “optional hyphen” знаци помеѓу слоговите. Овие знаци понатаму MS Word ги интерпретира како места на кои треба да се подели зборот при неговото пренесување во наредниот ред.

8. ЗАКЛУЧОК

Претставена е програма која овозможува автоматска коректура и разделување на слогови за текстови на македонски јазик. Програмата работи во рамките на процесорот на текстови Microsoft Word и овозможува пронаоѓање и помага при корекцијата на погрешно напишаните зборови во текстот. Вклучена е и можноста за креирање на кориснички речник во кој корисникот според потребите може да додава сопствени зборови. Дополнителна флексибилност е постигната со овозможување на корисникот да промени многу од параметрите на работа на програмата со што се постигнува максимално прилагодување на програмата кон потребите на конкретниот корисник.

За потребите на изградба на речник на македонски зборови во дигитално читлива форма, оформен е корпус од македонски текстови кој е користен како основа за градба на речникот. Истиот може во иднина да се користи и за компјутерска анализа на синтаксата на македонскиот јазик, со цел за развој на компјутерска проверка на граматичката коректност на македонски текстови и автоматска лектура.

9. ЛИТЕРАТУРА

[1] Блаже Коневски, „Граматика на македонскиот литературен јазик", Култура, Скопје 1976.

[2] Крум Тошев, Божо Видоески, Тодор Димитровски, Кирил Конески, Рада Угринова-Соколовска, „Правопис на македонскиот литературен јазик со правописен речник", Графички завод „Гоце Делчев", Скопје, 1970.

[3] Лилјана Минова-Ѓуркова, „Синтакса на македонскиот стандарден јазик", Радинг, Скопје, 1994.

[4] Дејан Ѓорѓевиќ, „Систем за оптичко препознавање на македонски кириличен текст за помош на лица со оштетен вид", магистерски труд, Скопје, Октомври, 1997.

[5] Душан Чакмаков, „Компјутерски алгоритми", книга во подготовка.

[6] E. M. Remgold, J. Nievergelt, N. Deo, "Combinatorial Algorithms: Theory and Practice", Prentice-Hall, New Jersey, 1977.

- -

Summary

KOREKTOR M – A SOFWARE PACKAGE FOR SPELL CHECKING AND HYPHENATION OF MACEDONIAN TEXTS

Dejan Ǵorǵeviќ, Dragan Mihajlov, Nikola Grčevski

Faculty of Electrical Engineering, Ss. Cyril and Methodius University
P.O.Box 574, MK-1001 Skopje, Republic of Macedonia
dejan@cerera.etf.ukim.edu.mk

Key words – spell checking, hyphenation, Macedonian language.

Korektor M is a software package for automatic spell checking and hyphenation for the Macedonian language, that works within the word processor Microsoft Word 97 or Microsoft Word 2000. It enables checking for spelling mistakes for Macedonian texts and helps in their correction offering a substitution for the mistyped word. Korektor M is also equipped with a module that enables proper hyphenation of the words when breaking a word between syllables at the end of a line, according to the rules of the Macedonian grammar.