Георге Гоце МИТРЕВСКИ
Снежана ВЕНОВСКА-АНТЕВСКА

МАКЕДОНСКИ ЈАЗИЧЕН КОРПУС
(идеја, можности, реализација)

Во денешни услови речиси е невозможно да се вршат јазични истражувања без помош на современата техника. Кон крајот на 80-тите години на 20-от век компјутерската техника, сосема скромно почна да се употребува и во македонската лингвистика. Така, ИМЈ во желбата да не заостане зад другите славистички центри започна со употреба на новите технички помагала, со еден компјутер 2,86 и во текот на последниве десетина години речиси целата работа се одвива преку компјутери. Факт е дека не успеавме преку мрежа да ги поврземе барем деловите од петте одделенија, а исто така факт е дека и во поглед на компјутеризацијата многу не можеме да се пофалиме иако се знае колкава е потребата од добра техничка опременост. Имајќи ги пред себе искуствата на други славистички и лингвистички центри и информациите преку директна размена на податоци и преку интернет пред нас се поставува потребата од навремено претставување на македонскиот јазик во електронска форма, а тоа го става во преден план формирањето, поточно изработката на македонскиот јазичен корпус или на националниот корпус. Идејата за разработка на ова прашање е поттикната од предавањето на Радмила Хоракова од Братислава, Словачка, која имаше предавање за хомонимијата во ИМЈ, во септември 2003. Меѓу другото, таа не поттикна да размислуваме за македонски јазичен корпус, претставувајќи го Словачкиот национален корпус кој го изработува Александар Хорак. Со оглед на поставеноста на словачкиот јазик во славистиката, степенот на истражувањата врз него, најдовме многу допирни точки и се родија идеите како да се направи македонскиот јазичен корпус. Ова предавање може да го сметаме како почеток

```
          l.s.  full_art         molivyt     f
- -------------- -------------- ------------- -
6 Clitic                            -        -
- -------------- -------------- ------------- -
7 Animate                           -        -
- -------------- -------------- ------------- -
8 Owner_Number                      -        -
- -------------- -------------- ------------- -
9 Owner_Person                      -        -
---------------- -------------- ------------- -
10Owned_Number                      -        -
================================ ============= =
```

Во првата редица е назначен редовниот број на атрибутот. Во втората редица е назначен називот на атрибутот. Во третата редица се назначени можните вредности на атрибутот. Во четвртата редица се дава пример на именка со одредената вредност, а во последната редица е назначен кодот соодветен за одредената вредност. Да видиме неколку примери.

Комбинации:

POS	Type	Gend	Numb	Case	Def	Clit	Anim	OwnN	OwnP	OwdN	Example
N	[cp]	[mfn]	[sp]	-	n	-	-	-	-	-	1.
N	[cp]	m	s	-	[sf]	-	-	-	-	-	2.
N	[cp]	[mf]	-	v	-	-	-	-	-	-	3.
N	[cp]	[mfn]	p	-	y	-	-	-	-	-	4.
N	[cp]	[fn]	s	-	y	-	-	-	-	-	5.
N	c	m	t	-	-	-	-	-	-	-	6.

Примери:

```
1. narod,          Ncms-n
   zhena,          Ncfs-n
   selo, Ncns-n
   Ivan, Npms-n
   Penka,          Npfs-n
2. naroda,         Ncms-s
   narodyt         Ncms-f
3. narode,         Ncm-v
   zheno,          Ncf-v
4. narodite,       Ncmp-y
   zhenite,        Ncfp-y
   selata,         Ncnp-y
5. zhenata         Ncfs-y
   seloto,         Ncns-y
6. naroda,         Ncmt
```

Морфосинтактичките особености на зборовите во корпусот ќе бидат додадени, или припоени, кон секој збор со помош на компјутерска програма наречена tagger. Програмата ќе треба да биде создадена од страна на специјалисти на компјутерска и корпусна лингвистика. Програмата ќе треба да биде способна да го анализира одредениот текст на ниво на збор, да ги определи морфосинтактичките особености на зборот, и да го прилепи кон зборот соодветниот морфосинтактички опис во MULTEXT-East format. Сигурно во некои текстови во корпусот ќе има помалку или повеќе грешки во компјутерската анотација. Во такви случаи текстовите ќе треба да се прегледат и исправат рачно. Рачно проверување на еден повеќе милионски корпус сигурно нема да биде практично. Затоа меѓународните стандарди препорачуваат рачно прегледување на одредена количина од корпусот, обично не помалку од околу десет проценти од корпусот.

На најниско ниво еден корпус треба да овозможи пребарување во корпусот на одредени зборови и правење на обичен конкорданс во KWIC (Key Word In Context), или „главен збор во контекст". Современите јазични корпуси се разликуваат по начинот на нивниот пристап, обемност и сложност. Да видиме неколку примери.

Кога се пребарува Босанскиот корпус, резултатите може да се побараат во форма на KWIC конкорданс, или може да се побара распределување на резултатите според лингвистичката форма на зборовите или према изворите на текстовите. Еве неколку примери:

- **sębi"** Svi primjeri riječi **sebi**.
- **"kak.*"** Sve riječi koje počinju slovima **kak**.
- **".*ovati"** Sve riječi koje završavaju nizom **ovati** (=infinitivi, na pr. kritikovati).
- **".*t" "ć.*"** Svi nizovi koji se sastoje od dvije susjedne riječi kod kojih prva završava na **t** a druga počinje sa **ć** (=puni oblik budućeg vremena, na pr. vidjet ćeš).
- **"da" []{0,7} "se"** Riječ **da** iza koje slijedi **se** koje može biti odvojeno od **da** sa najviše sedam riječi.
- **"u" []* "u" []* "u" within p** Paragrafi koji u sebi sadrže bar tri primjera riječi **u**.

- **kak** u novinama ili časopisima (Kodovi koji počinju sa PU).
- **[word=".*t" $ ori=".*94"] []* "ć.*"** Nizovi riječi (koje ne moraju biti susjedne) gdje se prva riječ završava na **t** a druga počinje sa **ć** u djelima koja su objavljena 1994. godine.

Системот за пребарување во Хрватскиот национален корпус е сличен на Босанскиот. Еве како изгледа страницата за пребарување во тој корпус:

Upit nad 30-milijunskim korpusom:

(Zamjensko pisme: *, npr. sve riječi koje počinju sa *imenic...*: imenic* Napomena: Ako se u upitu nalazi *, pretraživanje korpusa znatno će se usporiti.) Kodnu stranicu treba podesiti na Central European Alphabet (Windows) (CP 1250) Oni koji preko tipkovnice ne mogu dobiti: č ć đ š ž neka ih kopiraju s ovog mjesta.

Unesite riječ:

Odaberite (pot)korpus:

Cijeli korpus | Pošalji | Briši

Еден дел, јавниот, од Чешкиот национален корпус може да се пребарува од секого преку интернет страница, а комплетниот корпус може да се пребарува само со посебна дозвола. Јавниот дел се состои од 20 милиони зборови избрани од големиот сто милионски корпус. Пребарување во јавниот корпус дозволува пребарување само на посебни зборови, а не фрази, и контекстот на зборот не може да биде поголем од 60 букви. Големиот корпус овозможува безграничен контекст, пребарување на фрази, пребарување базирано на морфолошки карактеристики (именки, заменки, итн.). Истиот овозможува запишување на резултатите (конкордансот) на локална дискета. Еве како изгледа страницата за пребарување во јавниот дел на корпусот

Czech Corpus
Public access

Look for the word:

with a width of context 25 (max. 60) characters.

Selection of concordances: linear random

Search Clear

Еве како изгледа страната за пребарување во јавниот дел на Британскиот национален корпус

Simple Search of BNC-World

Please enter your query: Solve it! Cancel

You can search for a single word or a phrase, for example **Dogged** or **brown bread**. *Use the _ character to match any single word, for example* **bread _ butter** finds bread and butter, bread or butter *etc. Use the = character to restrict searches by part of speech, for example* **house=VVB** *finds only verbal uses of house. Use braces { and } to enclose a regular expression, for example* **{s|iau|ng}** *finds* sing, sang or sung

Рускиот национален корпус е пример за најдобар и најефикасен систем за пребарување во јазичен корпус. Целиот корпус е достапен на секого, а морфосинтактичките карактеристики на пребаруваните зборови може да бидат најопределени од сите други јавни корпуси. Еве како изгледа страницата за пребарување во овој корпус:

Национальный корпус русского языка

Поиск в корпусе: основной корпус
задать свой корпус

Поиск точных форм

искать

Лексико-грамматический поиск

Слово 1 грамм. признаки выбрать

Расстояние: от 1 до 1

Слово 2 грамм. признаки выбрать

искать очистить

Определување на граматичките особености на пребаруваниот збор е тоа што го одликува квалитетот на овој корпус од сите други јавни корпуси:

Часть речи
- существительное
- прилагательное
- глагол
- наречие
- мест-сущ
- мест-прил
- мест-наречие
- числительное
- числ-прил
- предлог
- союз
- частица
- междометие
- предикатив
- вводное слово

Падеж
- именительный
- звательный
- родительный
- родительный 2
- дательный
- винительный
- винительный 2*
- творительный
- предложный
- предложный 2

Степень / Краткость
- сравнительная
- сравнительная 2*
- превосходная
- полная форма
- краткая форма

Род
- мужской
- женский
- средний

Антропонимы
- фамилия
- имя
- отчество

Лицо
- первое
- второе
- третье

Время
- настоящее
- будущее
- прошедшее

Прочее
- словарная форма
- аномальная форма*
- искаженная форма*
- несловарная форма**

Наклонение / Форма
- изъявительное
- повелительное
- инфинитив
- причастие
- деепричастие

Залог
- действительный
- страдательный
- медиальный

Число
- единственное
- множественное

Одушевленность
- одушевленное
- неодушевленное

Вид
- совершенный
- несовершенный

Переходность
- переходный*
- непереходный*

Отмена

Стандардизиран, проверен корпус на македонскиот јазик кој претставува балансиран пресек на жанрови ќе биде непроценлива алатка за истражувања на македонскиот јазик. За да биде корисен, корпусот треба да биде достапен на македонски и на странски лингвисти. Тој претставува непроценливо средство не само за истражувачи, туку и за педагози, студенти и за јавноста. Кога ќе биде готов, корпусот би требало да се состои од 100 милиони зборови, составен од широк избор на современи пишувани и говорни текстови. Корпусот ќе претставува една точна слика на тоа како македонскиот јазик се употребува во денешна Македонија.

Пред сè корпусот треба да биде национален, како што е случајот и со корпусите на другите јазици.

Литература

1. BNC - British National Corpus. http://info.ox.ac.uk/bnc/
2. Tomaž ERJAVEC, Nancy IDE, 1998: The MULTEXT-East Corpus. First International Conference on Language Resources and Evaluation, LREC'98. Ur. Antonio Rubio, Natividad Gallardo, Rosa Castro, Antonio Tejada. Granada. 971-974. http://nl.ijs.si/ME/
3. TEI94 - Guidelines for Electronic Text Encoding and Interchange, 1994. Ur. C. M. Sperberg-McQueen, Lou Burnard. Chicago, Oxford. http://www-tei.uic.edu/orgs/tei/
4. Lou BURNARD, C.M. SPERBERG-MCQUEEN, 1995: TEI Lite: An Introduction to Text Encoding for Interchange. http://www.uic.edu/orgs/tei/lite/
5. ICNC - The Institute of the Czech National Corpus. http://ucnk.ff.cuni.cz
6. Fillmore, C., Ide, N., Jurafsky, D., and Macleod, C. (1998). An American National Corpus: A Proposal. *Proceedings of the First International Language Resources and Evaluation Conference*,Granada, Spain, 965-70.
7. Ide, N., Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. *Proceedings of Corpus Linguistics 2001*, Lancaster UK.
8. Korpus bosanskih tekstova na Univerzitetu u Oslu http://www.tekstlab.uio.no/Bosnian/Korpus2.html
9. Czech National Corpus http://ucnk.ff.cuni.cz/english/index.html
10. Corpus of Estonian Written texts 1983-87 http://psych.ut.ee/gling/en/corpusb/index.html
11. Hrvatski nacionalni korpus http:// www.hnk.ffzg.hr/index.html
12. The Oslo Corpus of Tagged Norwegian Texts http:// www.tekstlab.uio.no/norsk/bokmaal/english.html
13. Национальный корпус русского языка http:// www.ruscorpora.ru\index.html
14. Multext-East Resources, Version 3

EAGLES (Expert Advisory Group on Language Engineering Standards) http:// www.ilc.cnr.it/EAGLES96/home.html

бидејќи неговата цел е запознавањето со можноста за оформување на еден таков проект, но наедно и желба да се разменат искуства и да се поттикнат сите заинтересирани да се вклучат во изградбата на корпусот. Она што може да се најде како информација за корпусите кај другите словенски јазици и начинот како е поставено прашањето околу функционирањето на другите корпуси потврдува дека постоењето на корпус во електронска форма и неговата достапност и претставеност на интернет е еден вид фундаментално истражување во хуманистичко-општествените науки. Овде ќе се обидеме да ви претставиме само дел од она што може да го отвори прашањето за македонскиот корпус како основно за соодветно претставување на македонскиот јазик не само во рамките на словенските туку и во рамките на европските и на светските јазици. Иако е доволно, кога ќе се видат другите корпуси, да се каже дека без такво претставување на еден јазик, во денешни услови, е исто како да не постои таков јазик. Пред сè е потребно да се почне од основните прашање, а тоа се:

Што е корпус?

Што значи постоењето на корпусот за еден научник што се занимава со јазични истражувања?

Корпусот на еден јазик претставува збирка од различни видови текстови и јазични материјали дадени во електронска форма. Лингвистите со помош на информатичарите врз автентични јазични материјали ја создаваат таа збирка од текстови. Богатството на корпусот се огледа во бројноста на материјалите од различни функционални стилови и застапеноста на сите функционални стилови. Факт е дека картотеките на ИМЈ од петте клучни области (историја, дијалектологија, ономастика, современ јазик и лексикологија) претставуваат корпуси на македонскиот јазик на дијахрониски и на синхрониски план. Но, тие картотеки се речиси од затворен тип и се користат само од определен број научници и со определена цел. Нивното претставување во електронска форма во рамките на еден Македонски јазичен корпус би имало двојно значење. Првото е заштита на картотеките, односно заштита на собраниот материјал и второ, отвореност за јазични истражувања. Дел од овие картотеки веќе се во електронска форма, така што нивно-

то вградување во МЈК најлесно би можело да се реализира. Во практиката тоа ќе значи дека секој што е заинтересиран за определено јазично прашање поврзано со македонскиот јазик, од кој било дел од светот, со помош на комјутер и интернет врска ќе може да добие информација преку македонскиот јазичен корпус.

Денес, корпусот служи како појдовна точка на секое лингвистичко истражување. Факт е дека македонскиот јазик нема ваков корпус во кој би биле вградени репрезентативни текстови на историски и на современ план во електронска форма. Нивното вградување во еден јазичен корпус ќе значи можност за повеќенаменски истражувања.

Значи, првата фаза за изработката на Корпус е електронска обработка на текстови од различни функционални стилови, тука може да се има предвид дека определени текстови веќе се во електронска форма, така што лесно би се вградиле во Корпусот. Тој материјал би се претставил на одделна веб-страница со сите податоци за постоењето на Корпусот. Едноставно кажано, во комјутерот се внесуваат текстови, се класифицираат, се групираат зборовите и со помош на одделна методологија може да се види кои зборови, колку пати, каде и во која форма се појавуваат. Така се добива статистички опис на употребата на еден определен збор во сите негови форми и во сите позиции и во различни текстови,, тоа овозможува повеќенаменска употреба на дадените материјали и целосна претстава за дадениот збор во дадениот јазик.

Македонскиот јазик нема таков корпус, додека пак другите словенски јазици започнаа со изработка на сопствените национални корпуси, и овде не треба да се најдеме во позиција да не биде претставен нашиот јазик, имајќи ги предвид можностите што ги нуди Корпусот. Меѓутоа искуството од другите може да ни овозможи полесно да започнеме со создавањето на македонски јазичен корпус. Ако се види сето она што го направиле Словаците кои неколку години се подготвувале за оформување на Корпусот, а потоа во 2002 го претставиле на интернет, тогаш нивниот пример може да ни служи за да го оформиме и ние Корпусот. Тие имаат текстови во електронска форма од различни функционални стилови, и работат со помош на програмата конкорданс за пребарување на бара-

ните зборови. Многу подалеку во оваа проблематика се во хрватскиот јазик кои го имаат Хрватскиот национален корпус, каде што внесуваат дела од познати хрватски автори и други материјали. Чесите и Полјаците се многу повеќе навлезени во оваа проблематика, работејќи на неколку различни корпуси, сите потоа во координација. Исто така и Русите имаат направено Корпус со обемна литература и со неверојатно голем број информации. ФИДА или Корпусот на словенечкиот јазик веќе дава информации за над 100 милиони зборови. Разгледувајќи ги корпусите на другите словенски јазици мора да се запрашаме каде сме ние, кога ќе почнеме да размислуваме за вакво претставување на македонскиот јазик, дали може да направиме нешто врз основа на постоечките материјали во електронска форма и врз основа на искуствата од другите.

Целта поставена на почетокот може да опфати три подрачја: составување и обработка на македонскиот јазичен корпус како темел на основните јазични материјали кои може да послужат за различни видови истражувања во лексикологијата и во лексикографијата (изработка на разни видови речници, еднојазични и повеќејазични, терминолошки итн.); дополнување на правописниот речник; дијахрониско и синхрониско истражување на македонскиот јазик по определени проблеми и по определени автори. Целосноста во информациите овозможува побргу да се дојде до податоци потврдени преку примери.

Реализацијата на ова прашање нè враќа кон основното прашање што може да биде оправдување за постоење на еден вака голем проект, проект кој кога ќе започне треба да биде во континуитет, долгорочен, со постојано надоградување и со следење на најновите движења. Кон прашањето зошто ни е потребен корпус, се навраќаме кон потребата за соодветно претставување на нашата држава и кон нашата желба за европска интеграција, иако создавањето на Корпусот не е само со таа намена, поточно соодветно да бидеме претставени во Европа, бидејќи како што споменав на почетокот заштитата на македонскиот јазик во денешни услови е можна само ако биде претставен пред другите европски и светски јазици, токму преку Корпус.

Ова објаснување за потребата за постоење на Корпус може да биде многу долго елаборирано и сликовито претставено преку другите корпуси, нивните почетоци итн., итн., но на крајот би дошле до ист заклучок дека започнувањето со планови за формирање на Корпус треба да биде приоритетна задача.

Пред сè, треба да се овозможи изработка на проект со такво име и да се обезбедат средства за неговата реализација. Во другите словенски јазици зад изработката на Корпусите застанува државата, Владата, Претседателот на државата, моќни фирми, Министерства и донатори. За оформувањето и финансирањето на еден таков проект е потребна пошироко елаборирање со конкретен план што и како би се работело во определена фаза.

За нас би било добро носител на еден таков проект да биде ИМЈ, со тоа што би се вклучиле сите заинтересирани страни (Филолошкиот факултет, МАНУ, Електротехнички, информатичари, ПМФ и др.)

Корпусот би можел да се шири на поткорпуси, слично како во другите словенски корпуси, така на пример, во Хрватскиот национален корпус (ХНК) во 1996 год. имаше обработено пет поткорпуси, и тоа, печат (8 примероци со по 25 000 пројави, со 195 052 зборoформи), драма (20 примероци со по 10 000 пројави, 203 208 зборoформи), проза (20 примероци со по 10 000 пројави, 205 816 зборoформи), поезија (20 примероци со по 10 000 пројави, 201 667), учебници (58 примероци со по 3 450 пројави, 202 005 зборoформи). Овде е опфатен периодот 1935-1978, а во самиот Корпус биле вклучени неколку организации (Заводот за лингвистика, Хрватскиот електронски текстовен архив). Денес нивните материјали може да се најдат на интернет заедно со сето она што е поврзано со ХНК, но исто така во нивните планови е доработка на 30 милионскиот корпус и претставување на цд-ром.

Корпусот не би значел комплетно одбегнување на печатењето книги, особено речници, туку би овозможил поголема достапност и информираност, а од друга страна целосна и брза изработка на определени лингвистички прашања во врска со македонскиот јазик, а со помош на другите корпуси и компаративни истражувања, значи треба да се предвиди една голема репрезентативност. Ако нам ни тре-

баше 100 години да минат па повторно да се запрашаме што направивме и што треба да направиме за однапред, не треба да чекаме уште 100, туку треба да ги здружиме сите можни капацитети, секако од различни области и да започнеме со Корпусот. Никој не би требало да се почувствува непоканет, туку сосема спротивното, сите што може да помогнат со нешто (идеи, насоки, информации) треба да се вклучат. Не треба да губиме во време барајќи виновници за нешто што не сме сториле досега, а наедно губејќи го времето во тие трагања. Лично, постоењето на Македонскиот јазичен корпус треба да биде цел и задача на секој што може нешто да придонесе.

Овде во најкуси црти ви го претставуваме прашањето за Корпусот од аспект на некој, што се интересира за оваа проблематика како лингвист, во рамките и со можностите што би побудиле интерес кај некој што се занимава со јазични проучувања, а сака да си помогне преку современа технологија што му е достапна, со едно жалење што кај нас нема специјализирани студии по информатичка лингвистика кои би овозможиле многу полесно решавање на вакви и слични прашања.

Имајќи го предвид сето тоа потребни се стратегии за дизајнирање и за конструирање на Македонски национален корпус сличен и споредлив со други национални корпуси, како на пример, Британскиот национален корпус, Американскиот национален корпус, Чешкиот национален корпус и други.

Повеќето таканаречени големи светски јазици веќе имаат свои национални корпуси. Британскиот национален корпус се употребува како мерило за квалитетот на повеќето корпуси на другите јазици обработени во последните десет години. Бидејќи македонскиот јазик нема свој национален корпус, се предлага изработка на еден таков корпус кој би се базирал на искуствата на другите светски корпуси и на меѓународни стандарди за кодирање на корпуси.

Македонскиот национален корпус (МНК) ќе се состои од балансиран збир на текстови кои ја карактеризираат состојбата на современиот македонски пишуван и говорен јазик. МНК ќе се употребува за собирање на квалитативни и

квантитативни јазични податоци и докази за градење на помали специјализирани корпуси, за конкорданси, за создавање на низа списоци на лингвистички елементи, и за изучувања на фреквентноста на граматички елементи. Така, МНК ќе биде наменет, пред сѐ, за употреба при изучување на граматичката структура на македонскиот јазик и за пронаоѓање на некои постојани промени во јазикот во еден определен временски период. Во иднина, описите на граматичката структура на македонскиот јазик, како и стручните речници ќе може да бидат градени врз база на овој корпус.

Во својата почетна форма корпусот треба да биде синхроничен и да се состои од текстови составени од 1990 година. Текстови пишувани во претходни години може да бидат додавани постепено. Текстовите кои ќе влезат во корпусот ќе треба да бидат само оние составени од македонски автори кои живеат или живееле во Македонија. Што значи дека се исклучуваат текстови кои се преводи од странски јазици, текстови пишувани од автори на кои македонскиот јазик не им е мајчин јазик, и текстови пишувани од Македонци кои живеат во други јазични средини. За да биде македонскиот јазик претставен целосно потребно е корпусот да вклучува примери од сите синтактички и семантички феномени на јазикот и да се состои од текстови на различни теми и од различни жанрови, текстови објавени во книги, учебници, весници, неделни и месечни списанија, проспекти, писма, итн, како и текстови пишувани од машки и женски личности од различна возраст. Тука, исто така, ќе треба да се има предвид и возраста на авторите. Хрватскиот национален корпус, на пример, исклучува текстови пишувани од личности под 16 години.

МНК ќе треба да има и динамичен составен дел во кој ќе се додаваат текстови во определени интервали. Американскиот национален корпус, на пример, предвидува додавање 10% нови текстови секој пет години. Ова ќе овозможува изучување на периодичните промени во јазикот.

МНК ќе биде корисен за лингвистите само ако е нешто повеќе од збирка на зборови. Корпусот може да биде најкорисен ако биде кодиран. или анотиран, според некој познат стандард со кодирање на текстови. Во светот на

компјутерската лингвистика има неколку такви стандарди. Надворешните својства на текстови, како на пример наслов, автор, жанр, итн, обично се кодираат во согласност со стандардот TEI (Text Encoding Initiative – Иницијатива за кодирање на текстови). За кодирање на содржината на текстовите би се употребил европскиот стандард EAGLES, кој е веќе прифатен како стандард од многу други современи системи за компјутерско изучување на електронски текстови. Прифаќањето на овој стандард ќе овозможи полесно спроведување на компаративни изучувања на македонскиот корпус споредно со други корпуси.

Токенизацијата е следниот степен во обработката на текстот пред да влезе во корпусот. Токенизација подразбира делење на текстот на посебни зборови. Интерпункциските знаци ќе треба да се кодираат посебно (исто како и зборовите), за да може, на пример, корисникот да пребарува зборови кои се проследени со запирка, или пак реченици кои завршуваат со извичник.

За кодирање на текстовите со серија на граматички и семантички карактеристики МНК ќе употребува проверени методи од компјутерската и корпусната лингвистика. Колку побогата е анотацијата на посебните зборови во текстот, толку поголем е и истражувачкиот квалитет на корпусот. Затоа, секој текст во МНК ќе треба да биде кодиран на начин кој би ги покажал морфолошките, семантичките и синтаксичките својства на зборовите во секој посебен текст.

За да биде корпусот максимален текстовите ќе бидат кодирани според светскиот стандард XCES (XML Corpus Encoding Standard) за анотирање на електронски текстови, кој беше подготвен за да ги овозможи потребите за работење со корпуси. За анотирање на ниво на посебни зборови ќе се употреби стандардот MULTEXT-East, кој се состои од јазични средства базирани на морфосинтаксичкиот систем EAGLES, и опфаќа повеќе централни и источноевропски јазици, како на пример српскиот, словенечкиот и бугарскиот. Multext-East пропишува хармонизирани лексички спецификации и ги формулира релевантните нотации кои се употребуваат за градење на лексикони и анотирани корпуси направени во овие јазични групи.

Главните заеднички граматички категории во MULTEXT-East се следните:

Part-of-Speech	Code	Atts
Noun	N	10
Verb	V	15
Adjective	A	12
Pronoun	P	17
Determiner	D	10
Article	T	6
Adverb	R	6
Adposition	S	4
Conjunction	C	7
Numeral	M	12
Interjection	I	2
Residual	X	0
Abbreviation	Y	5
Particle	Q	3

Во првата редица е назначен називот на видот на зборот. Во втората редица е кодот за видот на зборот, назначен со една голема буква. Во третата редица е бројот на атрибути за секој вид збор. Така, на пример, кодот за именките е латинската буква N, а именката може да има најмногу 10 атрибути. Да разгледаме еден пример на MULTEXT-East системот за морфосинтактички опис на именки во бугарскиот јазик.

Noun (N)

P	ATT	VAL	Example	C
1	Type	common	kniga	c
		proper	Ivan	p
2	Gender	masculine	stol	m
		feminine	masa	f
		neuter	vreteno	n
3	Number	singular	momtche	s
		plural	stolove	p
	l.s.	count	(dva) stola	t
4	Case	nominative	narod	n
		vocative	narode	v
5	Definiteness	no	utchitel	n
		yes	zhenata	y
	l.s.	short_art	moliva	s