

УНИВЕРЗИТЕТ „СВ. КИРИЛ И МЕТОДИЈ“
МЕЃУНАРОДЕН СЕМИНАР ЗА МАКЕДОНСКИ ЈАЗИК, ЛИТЕРАТУРА И КУЛТУРА

**XXXVII НАУЧНА КОНФЕРЕНЦИЈА
НА МЕЃУНАРОДЕН СЕМИНАР
ЗА МАКЕДОНСКИ ЈАЗИК, ЛИТЕРАТУРА
И КУЛТУРА**

(Охрид, 15-16 јули 2010)

ЛИНГВИСТИКА



Скопје, 2011

Бранко Тошовиќ

КОРПУСОТ НА МАКЕДОНСКИ ЈАЗИК ВО РАМКИТЕ НА КОРПУСОТ ГРАЛИС

Во овој текст е претставена основната концепција и структура на корпусот на македонски јазик, кој се развива во рамките на соработката меѓу Филолошкиот факултет „Блаже Конески“ во Скопје и Институтот за славистика на Универзитетот „Карл Франц“ во Грац.

Клучни зборови во реферативот: македонскиот јазик, корпус, корпус Гралис

0. Корпусот Гралис претставува информациски и аналитички комплекс за собирање, обработка и евалуација на пишувани текстови, на говорени текстови и на визуелни информации за систематско истражување на словенските јазици, кој е достапен на Интернет. Корпусот Гралис се состои од два големи дела: од текстуален и од говорен корпус. Текстуалниот корпус претставува збирка од паралелни текстови за различни словенски јазици, кој е достапен на Интернет.

de | en | hks | ru

GRALIS
korpus

Impressum | Hilfe | Linkliste | Logout

Institut für Slavistik | GAMS | GRALIS

UNI
GRAZ

Gruppe: Alle

└ Korpus: Gralis BKS-Korpus

└ Primärsprache: Bosnisch/Bosniakisch

└ Autor: -

└ Genre: -

└ Funktionaler Stil: -

Verbundene Sprachen: Kroatisch, Serbisch

Kontextgröße: Ein Satz

Anzeigeumfang: 20 Wörter als HTML

☒ Morphosyntaktische Annotation

Suchoptionen +

Übermitteln

1. Целта на проектот „Корпус Гралис“ е да се создаде комплексна платформа за испитување на врските меѓу сите словенски литературни јазици, која е достапна на Интернет. Притоа разликувам четири словенски корелативни системи: 1. интракорелативен систем, кој ги содржи односите во рамките на еден словенски јазик, на пример корелациите во македонскиот јазик, 2. интеркорелативен систем, кој ги

опфаќа врските меѓу многу блиските словенски јазици, какви што се: а) македонскиот и бугарскиот или б) српскиот, хрватскиот и бошњачкиот јазик, 3) супракорелациски систем, кој содржи релации меѓу територијално групирани јазици како јужнословенските јазици: македонскиот, бугарскиот, српскиот, хрватскиот, бошњачкиот и словенечкиот, 4) суперкорелациски систем, кој означува интеракции меѓу територијално оддалечени јазици, на пример македонскиот и рускиот јазик, 5) екстракорелациски систем, кој го претставува системот на односи меѓу (а) еден словенски јазик или повеќе словенски јазици и (б) еден несловенски јазик или повеќе несловенски јазици, на пример меѓу македонскиот и германскиот.

Тие корелациски системи се инкорпирирани во повеќе поткорпуси. Првата група ги опфаќа поткорпусите на сите словенски јазици. Втората група служи за истражување на меѓусловенските односи на интеркорелациско, супракорелациско и суперкорелациско ниво. Постојат хомогени и хетерогени поткорпуси Гралис. Хомогени се тие што ги покриваат трите групи словенски јазици – јужнословенската, источнословенската и западнословенската – и служат за испитување интеркорелациски и на супракорелациски план. Хетерогени се тие што опфаќаат јазици од две територијално оддалечени групи, на пример јужнословенските и источнословенските јазици. Споменатите поткорпуси служат за анализа на словенските јазици во нивните суперкорелации.

Освен тоа, има еден сосем изолиран поткорпус што не е поврзан со ниеден јазик, јазична група или корелациски систем. Тој може метафорички да се нарече островски корпус бидејќи служи за индивидуални, посебни потреби на студенти, докторанти и други истражувачи. Што значи тоа? Тоа значи дека ако сакаме да анализираме некој свој текст со една, две или повеќе јазични варијанти, односно да бараме форми, зборови, изрази и конструкции само во тој текст и во неговите јазични верзии, можеме сето тоа да го одделиме во еден поткорпус. Таа можност веќе ја користат моите студенти за изготвување дипломски работи. На пример, минатата година една студентка на овој начин собра материјал за дипломската работа „Разлики меѓу бошњачката, хрватската и српската верзија на Дејтонскиот договор“. Таа шанса можат да ја искористат и македонските студенти. За таа цел, текстот даден во Word треба да се претвори во табела и секоја реченица да се исече и да се внесе во посебен ред. Ако постојат две или повеќе јазични верзии, тогаш секој текст се пишува во посебна вертикална колона во табелата. Притоа е многу важно речениците да бидат паралелизирани по принципот: 1 – 1 (еден – еден), односно реченицата на јазикот **А** да има како еквивалент реченица на јазикот **Б**.

Во овој индивидуален, изолиран поткорпус се наоѓаат и многу-јазичните корпуси на познати писатели. Засега постојат два такви поголеми корпуса: 1. корпус текстови од добитникот на Нобеловата

награда Иво Андриќ, 2. корпус текстови од најпреведуваниот писател од поранешна Југославија, Зоран Живковиќ. Веќе две години подготвувате уште еден поткорпус, и тоа од Бранко Ќопиќ.

Би било многу убаво во рамките на македонскиот корпус Гралис да внесеме и дела од истакнати македонски писатели. Во моментот се работи на монолингвалниот и на мултилингвалниот корпус на македонскиот јазик. Во првата фаза од изработката на овие корпуси се предвидени македонски дела од уметничколитературниот стил што се преведени на германски јазик (Блаже Конески, Славко Јаневски и др.). Дигитализацијата ја врши Димитар Попоски, а за изборот на текстовите и за нивната електронска обработка се грижи Емилија Бојковска. При тоа напоменувам дека при изработката на корпусите на писателите строго ги почитуваме авторските права и дека при пребарувањето во корпусот Гралис не е можно да се добие резултат што е поголем од една реченица.

Ако зборуваме само за македонскиот јазик, тоа значи дека кога ќе го завршиме целиот проект, ќе биде можно да се анализираат неговите интракорелациски, интеркорелациски, супракорелациски и екстракорелациски односи со другите словенски јазици. Но во сегашнава ситуација, тој јазик можеме да го споредиме само со српскиот, со бугарскиот и со германскиот јазик.

Комунистичка меморија

Mazedonisch	Ако не ме терал тој мене , јас него сум го терала .
Deutsch	Und wenn er mich nicht dazu anhielt , dann hielt ich ihn dazu an .

Македонскиот поткорпус се подготвува во соработка меѓу Институтот за славистика во Грац, Центарот за моделирање на информации при Универзитетот во Грац и Филолошкиот факултет „Блаже Конески“ во Скопје. Тој ќе биде реализиран во периодот од 2009 година до 2012 година. Договорот за австриско-македонскиот научноистражувачки проект „Монолингвален и мултилингвален електронски корпус на македонскиот јазик“ беше потпишан на 18 декември 2008 година, а неговата цел е истражување, едукација и популаризирање на македонскиот јазик.

Grupa:	Sve
└ Korpusi:	Gralis Mak-Korpus Andričev Gralis-Korpus Čopičev Gralis-Korpus Gralis BKS-Korpus Gralis BKS-Korpus (N) Gralis Ist-Korpus Gralis Ist-Zap-Korpus Gralis Južn-Korpus Gralis Južn-Zap-Korpus Gralis Mak-Korpus Gralis Ru-N(j)em-Korpus Gralis Slav-Korpus Gralis Zapad-Korpus
└ Primarni jezik:	
└ Autor:	
└ Žanr:	
└ Funkcionalni stil:	

Корпусот на македонски јазик ќе биде составен од два дела: 1) текстови за проучување, презентирање и популаризирање на македонскиот јазик (Монолингвален корпус на македонскиот јазик), 2) текстови за контрастивно истражување на македонскиот јазик во однос на (а) другите словенски јазици, (б) германскиот јазик (Мултилингвален корпус на македонскиот јазик).

Австриската страна а) ќе обезбеди сервер и серверска инфраструктура, б) ќе го стави на располагање мултилингвалниот корпус Гралис како платформа за развој и презентирање на македонскиот корпус, в) ќе побара од австриските соодветни институции финансиска поддршка со која би се покриле патните трошоци до Грац за стручни лица од Македонија по одделни дисциплини (фонетика, фонологија, зборообразување, лексикологија, фразеологија, граматика итн.) за одржување на годишните колоквиуми на кои ќе се анализира развојот на корпусот и ќе се врши негово истражување, г) ќе се обиде да обезбеди средства за печатење зборник со материјалите од годишните колоквиуми, д) еднаш годишно ќе покани на гостување еден или двајца македонисти вклучени во проектот на два-три дена.

The screenshot shows the GRALIS Korpus web interface. The top navigation bar includes links for Impressum, Hilfe, Linkliste, Logout, Institut für Slavistik, GAMS, and GRALIS. The left sidebar contains a search and filter menu with the following options: Gruppe (Alle), Korpus (Gralis Mak-Korpus), Primärsprache, Autor, Genre, and Funktionaler Stil. The main content area displays search options: Verbundene Sprachen, Kontextgröße (Ein Satz), and Anzeigeumfang (20). A map of Europe is visible on the right side of the interface.

Македонската страна а) ќе го врши изборот на македонските текстови, скенирањето и корекцијата до создавањето на нивната верзија за Интернет и за серверот, којашто ќе ја подготвува австриската страна,

б) ќе ги решава прашањата на авторски права за текстовите објавени на македонски јазик, в) ќе вклучи студенти кои според европски програми, како што е ЦЕЕПУС, ќе имаат можност да добијат стипендија за престој во Грац, каде што би можеле да работат на корпусот во консултации со австриски соработници на проектот, д) ќе помогне во изготвувањето на зборникот на македонски јазик со материјалите од годишните колоквиуми.

Австриската и македонската страна заеднички ќе работат врз усовршување на паралелизацијата на два или на повеќе јазици.

Проектот опфаќа две фази. Во првата се подготвуваат македонските текстови, се внесуваат во корпусот и се дава основната метаинформација, семантичка и граматичка информација (2009–2012). Во втората фаза се истражува собраниот материјал и резултатите се презентираат на годишните колоквиуми.

Договорните страни се обврзуваат дека нема да го даваат собраниот материјал и заеднички направените програми за цели кои ги надминуваат рамките на овој проект без претходна писмена согласност од другата страна.

Корпусот Гралис има засега интерфејс на неколку јазици – на англиски, германски, руски, српски, хрватски и бошњачки јазик.

При пребарувањето има две можности: а) да се бара одреден збор во контекст од една реченица или б) да се добие список со сите форми од некој збор. Резултатот на пребарувањето се дава во два формата: 1. за речениците во html-формат, за списокот – во html- и Excell-формат.

Мостот на Дрина	
Mazedonisch	На тоа место каде што Дрина избликува со целата тежина на својата водна маса, зелена и запенета, од привидно затворениот склоп на црните и стрмни планини, стои еден голем, складно изделкан МОСТ од камен, со еднаесет широки арки.
Bulgarisch	На това място, гдето Дрина с цялата тежест на своите зелени и пенливи водни маси избликва от затворения наглед масив на черните и стръмни планини, има голем, стројно изграден от камък мост с единаесет широки свода.
Serbisch	Na tom mestu gde Drina izbija celom težinom svoje vodene mase, zelene i zapenjene, iz prividno zatvorenog sklopa crnih i strmih planina, stoji veliki, skladno srezani most od kamena, sa jedanaest lukova širokog raspona.

Корпусот Гралис спаѓа во корпусите со морфосинтактичка анотација. Тоа обележување се добива со помош на специјалната програма Gralis MorphoGenerator.

mesto		
Kasus	Singular	
	n	
Nominativ	mesto	N-nsn----6SnO01
Genitiv	mesta	N-nsg----6SnO01
Dativ	mestu	N-nsd----6SnO01
Akkusativ	mesto	N-nsa----6SnO01
Vokativ	mesto	N-nsv----6SnO01
Instrumental	mestom	N-nsi----6SnO01
Lokativ	mestu	N-nsl----6SnO01

mesta		
Kasus	Plural	
	n	
Nominativ	mesta	N-npn----6SnO01
Genitiv	mesta	N-npg----6SnO01
Dativ	mestima	N-npd----6SnO01
Akkusativ	mesta	N-npa----6SnO01
Vokativ	mesta	N-npv----6SnO01
Instrumental	mestima	N-npi----6SnO01
Lokativ	mestima	N-npl----6SnO01

Во поткорпусот за српскиот, за хрватскиот и за бошњачкиот јазик се обележени сите зборови како зборовни групи. Зборовите (1) што се деклинираат или (2) што имаат парадигма (именки, заменки, придавки, броеви, прилози) имаат комплетна морфосинтаксичка ано- тација, а засега се анотирани само 200 најчести глаголи. Со 71 правило успеав да генерирам околу 28 илјади придавки, но за генерирање на околу 40.000 именки ми беа потребни дури 306 правила. Глаголите се особено сложени (само глаголот *biti/jesam* има повеќе од 200 позиции).

Мостот на Дрина

Mazedonisch	Тоа е оној дел од мостот што се вика порта .
Bulgarisch	Това е онаа част от него , която се нарича " капия " .
Serbisch	<u>To je onaj deo mosta koji se zove</u> kapija .

Кога ќе се добие резултат од барањето на формите од една реченица, за сите подвлечени зборови се нудат соодветни парадигми. Нужно е само да се кликне на таа форма. Ако сакаме да ги погледнеме сите побарани и добиени форми во една реченица, можеме да притис- неме на дадената форма и ќе го добиеме следниов резултат:

Wortliste

Korpus: Gralis Andrić-Korpus; **Primärsprache:** Mazedonis

Suchbegriff(e): .*МОСТ.* (Groß- und Kleinschreibung: Nein; Such

347	■	МОСТОТ
93	■	МОСТ
4	■	МОСТОВИ
2	■	ПРЕМОСТИ
2	■	САМОСТОЈНИ
2	■	САМОСТОЈНОСТ
1	■	МОСТАРСКО
1	■	МОСТОВИОТ
1	■	МОСТОВИТЕ
1	■	МОСТОГ
1	■	ПРЕМОСТЕНО
1	■	ПРЕМОСТИЛЕ
1	■	ПРЕМОСТИМЕ
1	■	САМОСТОЈНО

На самиот крај од парадигмата се наоѓа информацијата за фреквенцијата на сите корпусни форми во избраните јазици.

Покрај морфосинтаксичката и лексикографската информација, сите добиени резултати при барањето ја следат метаинформацијата. Таа ги нуди основните податоци за изворот: 1. кој е авторот, 2. кога и каде е издадена публикацијата, 3. колку страници има, 4. на кој функционален стил и жанр ѝ припаѓа итн.

Wortliste

Korpus: Gralis Andrić-Korpus; **Primärsprache:** Mazedonisch

Suchbegriff(e): .*и.* (Groß- und Kleinschreibung: Nein; Suche nach Le

7795	и
890	ги
726	ја
488	тој
390	си
358	ни
355	тие
350	кој
345	или
336	сите
297	им
268	која
242	кои
238	нив
205	својата
201	има
189	години
165	секој
159	оние
154	покрај
147	ниту
144	своите
138	кај
137	својот
135	ти
133	работи

Има можност и да влеземе во програмата Lexikarium, која овозможува да ја добиеме лексикографската информација за бараните зборови. Овде ќе најдеме и информација за разликите меѓу зборовите на избраните јазици (засага само за српскиот, хрватскиот и бошњачкиот).

Постои уште една програма што е поврзана со корпусот Гралис. Таа се вика Гралис-Акцентариум. Веројатно се знае дека српскиот, хрватскиот и бошњачкиот јазик имаат многу сложен акцентски систем. Постојат четири акценти: долг нагорен (*rúka*) и краток нагорен (*nòga*), долг надолен (*mājka*) и краток надолен (*saŋ*) и таканаречена задакцентска (постакцентска) должина (*Nema mnogo ženā*). Освен тоа, овие јазици, за разлика од македонскиот, немаат фиксиран, туку слободен т. е. неопределен акцент. Затоа е доста тешко да се учи и да се иследува овој прозодиски систем. Од таа причина, со моите соработници во Институтот за славистика во Грац ја развив програмата Акцентариум, која нуди можност да добиете информација за акцентот на сите форми од корпусот Гралис. За македонскиот корпус тоа нема примарно значење.

Текстовите што се внесуваат во паралелниот корпус мора да имаат најмалку еден еквивалентен текст на другиот јазик. Тоа значи

дека треба да се бара соодветен превод, на пример, од македонски на руски или обратно.

Корпусната инфраструктура на Гралис веќе е разработена и нормално функционира. Таа **не** чинише малку (туку многу) – околу 100.000 [сто илјади] евра. Јас имам голема желба да ја искористам за изработка на корпусот за македонскиот јазик. Се надевам дека ќе имам соодветна поддршка за тоа и од македонската страна.

Овој проект е важен за македонистиката затоа што, според моите сознанија, не постои таков комплексен паралелен и повеќејазичен корпус за македонскиот јазик. Со помош на корпусот Гралис ќе биде можно тој јазик да се учи и да се иследува во неговите (а) внатрешни односи, (б) во словенските интеркорелации, супраколерации, суперкорелации и екстракорелации и (в) во македонско-несловенските релации. Тој корпус ќе биде особено релевантен во рамките на интеркорелационските односи (меѓу македонскиот и бугарскиот јазик) и на супракорелационските односи (меѓу македонскиот и српскиот, хрватскиот и бошњачкиот јазик). На суперкорелационското рамниште доминира врската меѓу македонскиот и најголемиот словенски јазик – рускиот. На екстракорелационски план за Македонија е веројатно интересен односот меѓу македонскиот и албанскиот јазик.

Засега корпусот Гралис содржи околу пет милиони, кажано на англиски *token*, на руски *словоформ*, на македонски *збороформи (одредници)*, на српски *pojavnice*.

Постојат два типа корпус – пишуван и говорен. Вториот е многу скап и сложен, тако што е доста редок. Дури и многу големите национални корпуси немаат поткорпус за устен, разговорен јазик. За среќа корпусот Гралис има таканаречен *Speech-Corpus*, вистина само за српскиот, хрватскиот и бошњачкиот. Неговата инфраструктура може да се искористи и за евентуална изработка на повеќејазичен македонски говорен корпус.

Das Korpus der mazedonischen Sprache im Rahmen des Gralis-Korpus

Für das Studium, den Unterricht und die Popularisierung der mazedonischen Sprache haben die Philologische Fakultät „Blaže Koneski“ der Universität Skopje und das Institut für Slawistik der Karl-Franzens-Universität Graz eine Vereinbarung zur Gründung des österreichisch-mazedonischen Projektes zur Schaffung eines mono- und multilingualen Korpus der mazedonischen Sprache getroffen, das im Zeitraum zwischen 1. Jänner 2009 und 1. Jänner 2012 durchgeführt wird.

Das Korpus der mazedonischen Sprache wird aus zwei Teilen bestehen: 1. Texten für das Studium, das Erlernen, die Vorstellung und die Popularisierung der mazedonischen Sprache (monolinguales Korpus der mazedonischen Sprache), 2. Texten für eine kontrastive Analyse der mazedonischen Sprache im Bezug auf (a) andere slawische Sprachen und (b) die deutsche Sprache (multilinguales Korpus der mazedonischen Sprache).

Das Projekt besteht aus zwei Phasen. In der ersten werden die mazedonischen Texte aufbereitet, in das Korpus integriert und mit den grundlegenden Meta-, semantischen und grammatikalischen Informationen versehen (2009-2012). In der zweiten Phase kommt es zu einer Analyse des bearbeiteten Materials, wobei die Ergebnisse auf den jährlichen Kolloquien 2010 und 2010 präsentiert werden.

Die Erstellung der monolingualen und multilingualen Korpus der mazedonischen Sprache erfolgt in Zusammenarbeit mit dem Zentrum für Informationsmodellierung in den Geisteswissenschaften.