: Gate Developper

Jovan Kostov

Јован КОСТОВ
Истражувачки центар E.A. 4514 – PLIDAM
ИНАЛКО, Париз

КОРПУС И ОТВОРЕН СОФТВЕР ЗА АВТОМАТСКА ОБРАБОТКА НА МАКЕДОНСКИОТ ЈАЗИК: ПРИМЕРИ ОД GATE DEVELOPPER

Апстракт: Во последните дваесетина години, интересот за автоматската обработка на природните јазици[1] во македонската научна фела постојано се зголемува. Потребата од користење софтвер за третман и обработка на пишани текстови, како и на говорни секвенции од природниот јазик е голема, не само во областа на македонската лингвистика туку и во други домени, како на пример, интернет-пребарувањето, кое се чини, станува составен дел од секојдневјето на сите генерации. Од таа гледна точка, автоматската обработка на јазикот не претставува само апстрактна област во која лингвистите си поигруваат со компјутерите, туку составен дел од нашето секојдневје. Сепак, изработката на еден ваков вид софтвер бара многу време и ресурси (не само финансиски туку и човечки и научни), исто толку колку и обуката на корисниците што ќе се служат со него.

Во тесна врска со компјутерската лингвистика е и корпус-лингвистиката чија цел е да ги проучува јазиците врз основа на корпуси од еден јазик. Спротивно на постојните размислувања, македонскиот јазик не е единствениот јазик што не располага со „национален корпус“. Впрочем, и еден „голем“ јазик како францускиот нема еден единствен корпус што би можеле да го оквалификуваме како таков, бидејќи француската школа претполага дека секој пишан или кажан дискурс е дел од еден глобален корпус, кој може да се истражува со цел да се најдат константите и варијантите на јазикот. Во оваа статија ќе се осврнеме на неколку корисни функционалности на софтверот Gate Developper[2] со кој може да се врши обработка на секаков вид пишани корпуси во UTF-8 поддршка.

Клучни зборови: Gate Developper, компјутерска лингвистика, автоматска обработка на македонскиот јазик, корпус, етика, моделизација, морфосинтакса, ексцерпција, глагол, регуларни изрази.

1 Вовед

Од самото осамостојување на Република Македонија, па сè до денес, идејата за создавање македонски национален корпус е речиси секојдневно присутна во научните дискусии од областа на македонистиката. Трнливите прашања за обработка и структурирање на корпусот се оние на кои македонистите најмногу им се осврнуваат, барајќи, притоа, соодветни параметри кои ќе бидат стожер за да се изберат „добри“ текстови, запазувајќи ги основните јазични критериуми за еден таков проект. Потребата од корпус е голема ако се знае дека тој не претставува само алатка за понатамошен опис на македонскиот јазичен стандард, туку и за истражувања од

[1] Терминот „компјутерска лингвистика“ е широко распространет термин во најголемиот дел од светските јазици. Во македонскиот јазик, често се користи и терминот „сметачка лингвистика“, но „пресметувањето“ е само еден од многуте процеси на оваа дисциплина чии методи не се единствено математички, туку ги земаат предвид и автохтоните јазични правила. Од таа гледна точка, претпочитаме да ги користиме термините „компјутерска лингвистика“ или „автоматска обработка на јазиците“.
[2] Заради практични причини, во понатамошниот текст ќе го користиме единствено името Gate.

областа на многу други дисциплини на лингвистиката, како, на пример, дијахронијата, стилистиката и социолингвистиката. Параметрите, пак, не се предмет единствено на лингвистиката, туку потегнуваат и други прашања, пред сè од областа на правото (на репродукција и на редистрибуција на авторски дела). Во нашата статија ќе се обидеме да дадеме еден поинаков поглед врз поимот „корпус“ од гледна точка на информатичката лингвистика, на начин којшто можеби не е својствен за македонската лингвистичка описна традиција.

Оваа статија содржи два дела: првиот се осврнува на дефинициите на поимот корпус, како и на основните потешкотии во процесот на негово создавање од лингвистичка, правна и етичка природа. Во него ќе направиме и краток преглед врз досегашните обиди за создавање македонски корпус. Во вториот дел, ќе дадеме еден краток опис на главните функционалности на помагалото Gate со кое можеме да вршиме јазични истражувања врз корпуси создадени за определена задача. Пред сè, ќе го разгледаме процесот на моделизација, анотација и ексцерпција, кои претставуваат основа за понатамошно изучување на контекстот на определените зборови или зборовни групи. Во него ќе дадеме и кратка анализа на ексцерпираните форми и неколку идеи што би можеле да се искористат во понатамошните истражувања врз македонскиот јазик со ова помагало.

2 Корпус и корпусна лингвистика: дефиниции, проблеми и правни аспекти

2.1 Општи дефиниции и значење на поимот „корпус“ во лингвистиката

Во голем број европски и светски култури, поимот „корпус“ опишува голема маса текстуални или говорни продукции на еден јазик. Најеминентен пример за национален корпус е Британскиот национален корпус (British National Corpus), кој ги има следниве карактеристики[3]:

- Претставува збир од говорни и пишани документи од последните стотина години;
- Содржи 100 милиони зборови;
- Неговата структура е направена според TEI-стандардот[4] за организација на текстови во дигитална форма;
- Содржи различни жанрови кои овозможуваат изучување на јазичната варијација.

Водичот за анотација: http://www.natcorp.ox.ac.uk/docs/URG/ (последна консултација на 28 септември 2016 година) ги содржи сите информации за методологијата според која е составен, како и за начинот на кој може да се користи овој корпус. Од почетокот на неговото основање па до денес, Британскиот национален корпус постојано се збогатува. Во 2007 година, тој е поделен на два подкорпуса од кои едниот (BNC-Sampler) е збир на два вида текстови од по еден милион зборови во пишана и во говорна форма, додека другиот (BNC-Baby) е корпус од четири милиони примероци на транскрибирани зборовни форми. Важно е да се напомене дека овој корпус е составен од текстови извадени од најразличен вид контексти со цел да бидат застапени сите жанрови и говорни ситуации, што е од големо значење за изучување на

[3] Карактеристиките на Британскиот национален корпус претставуваат и критериуми за негово креирање. Деталните информации можат да бидат консултирани на веб-страницата http://www.natcorp.ox.ac.uk/corpus/index.xml (последна консултација: 28 септември 2016 година).
[4] Text-encoding initiative – иницијатива за структурирање на јазични документи.

јазичната варијација. Со други зборови, Британскиот национален корпус го има предзнакот „национален“, но не и „стандарден“, бидејќи содржи и јазични творби од нестандарден карактер.

Во други јазични култури, пак, поимот „корпус“ не означува збир од голем број текстови, туку претставува примерок од репрезентативни текстуални мостри врз основа на определени критериуми. Според оваа дефиниција се води и француската јазично-описна традиција во која ќе се најдат, на пример, следниве формулации: „корпус на јазични продукции на изучувачи на францускиот како втор јазик“, „корпус на новинарски текстови“, „корпус од криминални приказни“, итн. Во француската лингвистичка школа, корпусите се најчесто предмет на истражувања и тестови на информатичката лингвистика, и како резултат на тоа, бележиме повеќе иницијативи за нивна експлоатација од страна на лингвистите кои работат врз францускиот, но и врз многу други јазици. За изучување на францускиот јазик, академската фела се повикува пред сè на платформата на Националниот центар за научни истражувања (CNRS) и на истражувачката лабораторија ATLIF позната под името Centre national de ressources textuelles et lexicales (Национален центар за текстуални и лексички ресурси)[5] во кој спаѓа и корпусот Frantext. Овој корпус содржи литературни дела на француски јазик од десетина векови. На крај, би сакале да посочиме дека многу лингвисти ја сметаат целокупната текстуална и медиска продукција на Интернет како еден голем дигитален корпус кој може да се користи за јазични истражувања за чија намена се развиени голем број техники (Tanguy & Hathout, 2007).

Според сите овие примери, би го дефинирале корпусот како збир од повеќе текстови напишани на еден јазик врз основа на најразлични критериуми кои можат да бидат од лингвистичка (жанр, контекст, итн.) и нелингвистичка природа (најчесто демографски критериуми). Во одредени случаи, како на пример, во рамките на истражувањата од областа на дијалектологијата, овие два вида критериуми се тесно врзани еден за друг и неопходно е да се земат предвид при конституција на еден корпус на говорители (демографски критериум) од одреден дијалект (демографско-лингвистички критериум) со говорна традиција (лингвистички критериум: текстовите се достапни само во нивната говорна форма). Во оваа статија ќе се осврнеме врз еден корпус од три документи[6] напишани на современ македонски јазик во временскиот интервал помеѓу 1990 и 2015 чии текстови се кодирани во UTF-8[7] формат.

2.2 Правни и етички проблеми при конституција на корпусите

Во секоја научно-истражувачка работа, од несомнено голема важност е и етиката на истражувачот. Централно прашање на секој научник кога работи врз јазична творба која не е негова е дали може да ја искористи и да ја репродуцира неа во рамките на своите истражувања. Од таа гледна точка, јазичарите се наоѓаат пред голема дилема дали некому нешто би му должеле ако искористат нечие авторско дело за потребите на нивните научни истражувања. Со цел да најдеме соодветно решение за овој проблем, се обидовме да ја проучиме соодветната законска регулатива за авторските права на пишаните текстови при нивно користење во научно-образовни процеси од информатички и технолошки карактер.

[5] http://www.cnrtl.fr/ (последна консултација: 28 септември 2016 година).

[6] Насловите и референциите на овие дела се дадени во посебен дел во заглавјето „Литература“ на крајот на оваа статија.

[7] Unicode UTF-8 е стандарна текстуална поддршка со чија помош се кодираат азбуките на повеќе светски јазици. Unicode-стандардот е дел од нормата ISO/CEI 10646.

Законот за авторското право и сродните права[8] предвидува непречено користење без паричен надоместок за авторските дела, согласно со членот 52 став 1 алинеа 4 кој вели дека: „Користењето авторско дело без плаќање надоместок се однесува на [...] користење дела заради илустрирање во образовни или научни истражувања до степен оправдан со некомерцијалната цел што треба да се постигне, под услов да се наведе името на авторот и изворот, освен доколку тоа не е возможно“. Од таа гледна точка, секое авторско дело може да се користи во рамките на научни истражувања врз македонскиот јазик и да стане дел од еден истражувачки корпус кој ќе ги задоволи потребите на истражувачите во областа на корпус-лингвистиката. Авторското дело напишано на македонски јазик е дел од македонското културно наследство и претставува сведок за еволуцијата на нашиот јазик на планот на дијахронијата, независно од шпекулацијата за неговиот „стандарден карактер“.

2.3 Поимот „корпус“ во македонистиката: иницијативи и постојни проекти

Како што веќе напоменавме во воведниот дел на оваа статија, поимот „корпус“ е предмет на повеќе дискусии во круговите на македонистиката. Во досегашните лингвистички истражувања, можеме да издвоиме неколку иницијативи за корпусно истражување започнати во различни контексти, како во Република Македонија, така и надвор од нејзините граници. Руска Ивановска-Наскова (Ivanovska-Naskova, 2006) бележи повеќе иницијативи и истражувања од кои произлегуваат Македонскиот пишан корпус (Macedonian writen corpora) (Mitrevski, 2006) и иницијативата Multext-EAST (Vojnovski V. & al., 2005; Ivanovska A. & al., 2005) со која се врши „тренирање“ на процесот на морфосинтаксичко етикетирање на зборовните групи (анг. part of speech tagging). Како една од поуспешните иницијативи, и покрај нејзиниот затворен карактер е и корпусот Gralis[9] што претставува резултат на работата на повеќе истражувачи од Универзитетот во Грац на иницијатива на Бранко Тошовиќ.

На крај, би го спомнале корпусот на SAM 97[10] кој претставува корпус од стотина македонски прозни текстови (есеи, раскази и романи) во кој се врши пребарување на зборови во контекст. Корпусот е поврзан со платформата Makedonski.info - речник составен со компилација на информациите од сите издадени македонски речници досега. За оваа приватна иницијатива допрва се изготвува специфицчна документација која ќе се осврне на методологијата и на функционалностите на ова помагало. Во отсуство на други истражувања од ваков вид и како резултат на сè помасовното користење на Интернет како канал за пребарување на информациите, ова помагало е единственото од таков карактер што се користи за потребите на говорителите, но и како преведувачка и лекторска алатка, токму поради неговата достапност и отворениот пристап.

Со оглед на сите овие иницијативи, можеме да заклучиме дека прашањето за создавање македонски корпус не е само апстрактна тематика на истражувањата на македонистите, туку и практично прашање кое си го поставуваат многу говорители на македонскиот јазик, водени од потребата за правилно изразување и пишување во согласност со постојните правописни и правоговорни норми. За жал, сите овие иницијативи наидуваат на одреден анимозитет од научната фела поради многу различни, а понекогаш и нерационални причини. Една од нив е недоволната

[8] За потребите на оваа статија ги проучивме различните верзии на овој закон преземени од официјалната веб-страница на Министерството за култура достапни на адресата http://www.kultura.gov.mk/index.php/legislativa/2011-03-04-10-39-07/281-zakon-za-avtorskoto-pravo-i-srodnite-prava (последна консултација: 28 септември 2016 година).

[9] http://www-gewi.uni-graz.at/gralis/ (последна консултација: 28 септември 2016 година).

[10] http://www.makedonski.info/literature/show/ (последна консултација: 28 септември 2016 година).

застапеност на информатичките технологии во процесот на изучување на јазикот: голем дел од лингвистите немаат доволни познавања од програмирањето и програмските јазици со кои се вршат истражувањата од областа на информатичката лингвистика и корпус-лингвистиката, за разлика од други научно-образовни контексти во кои овие дисциплини се дел од наставните програми. Во секој случај, нивното постоење не е само случајно, туку произлегува од една голема потреба на која, сега за сега, македонските образовни институции не можат да одговорат.

3 Помагала за автоматска обработка на јазиците и структурирање на корпуси

Автоматската обработка на јазици, дисциплина која им припаѓа наизменично и на информатиката и на лингвистиката, е еден од домените во кои корпусот има големо значење. Нејзина главна задача е создавање софтвер за лингвистичка обработка и анализа на природните јазици во нивната говорна и пишана форма. Со оглед на тоа што оваа статија се осврнува на пишаната форма, ќе се обидеме да дадеме еден општ преглед на неколку функции на помагалото Gate[11] што служи за обработка на пишаниот јазик и да ја објасниме поврзаноста на информатичката лингвистика и на корпус-лингвистиката. Ќе видиме, пред сè, како се врши анотација на зборовните групи и на дел-реченицитe, како и нивна ексцерпција од литературни корпуси за понатамошно изучување.

3.1 Што е Gate?

Gate (скратеница од *General architecture for text engineering*) е отворен софтвер за автоматска обработка на текстови напишани на разни поддршки. Ова помагало служи, пред сè, за семантичка и морфосинтаксичка анотација на текстови од најразлични јазици, чии азбуки (меѓу кои и македонската) можат да бидат кодирани со Unicode UTF-8 стандардот. Според постоечката документација, употребата на ова помагало е широка: методите и принципите на Gate се користат за пребарување на информации од областа на медицината и на фармацевтската индустрија, изучување на мислењето на потрошувачите, анализа на говорот на социјалните мрежи, анализа на побарувачката на работна рака во агенциите за вработување, итн. На корисниците на ова помагало им се овозможува истражување врз нивни лични корпуси со посебна методологија на јазична моделизација. Според тоа, Gate не се користи исклучиво за јазичните истражувања, туку наоѓа практична примена во сите сфери каде што е потребна јазичната анализа на морфолошко, синтаксичко и семантичко рамниште.

Во оваа статија ќе ги илустрираме можностите на ова помагало за пребарување на информации што можат да бидат моделизирани на синтаксичкото рамниште. Ќе прикажеме неколку функционалности што би можеле да бидат корисни за анотација, пребарување и екцерпција на дел-реченици од текстуални корпуси и на тој начин ќе илустрираме како овој софтвер би можел да придонесе во понатамошните истражувања во областа на македонистиката. Ќе се осврнеме пред сè на операциите за креирање корпус, сегментација (Gate Unicode Tokenizer[12]), зборовна анотација со помош на специфични речници (Gazetteers[13]) и реченична анотација со помош на граматики

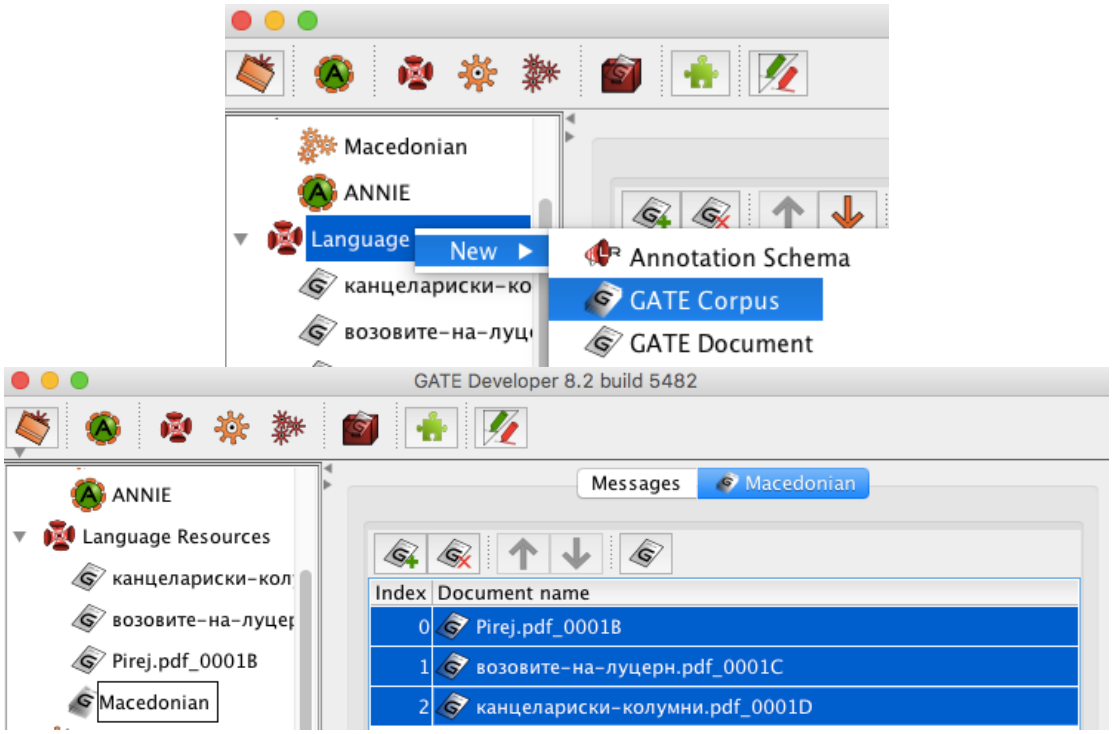
[11] https://gate.ac.uk/ (последна консултација: 29 септември 2016).
[12] Токенизатор (анг. Tokenizer) е помагало што се користи за сегментација на еден текст (што може да се претстави како низа симболи - анг. string) на повеќе зборовни единици (анг. tokens).
[13] Речници.

(JAPE[14]). Овие неколку операции се неопходни во процесот на ексцерпција како една од најважните и најмакотрпните задачи во лингвистичките истражувања.

3.2 Работен корпус

За потребите на оваа статија, работевме врз 3 романи од македонската современа проза: *Пиреј* од Петре М. Андреевски и *Канцелариски колумни* и *Возовите на Луцерн* од Лилјана Пандева. Текстовите во формат PDF се кодирани во UTF-8 поддршка и обработливи со помош на корпус-генераторот на Gate. Се работи за документи што се наоѓаат на компјутерот на којшто е инсталиран Gate, но би сакале да напоменеме дека помагалото нуди можност и за далечинско обработување на корпуси со спецификација на УРЛ-референцата на серверот на којшто се наоѓа еден документ што сакаме да го анализираме.

Првата етапа на нашата анализа ја засега организацијата на корпусот со помош на операцијата *Language resources => New => GATE Corpus:*



Илустрација 1: Процесирање на работен корпус со Gate

Gate го пречистува текстот и креира работна верзија на нашиот корпус, ослободена од сите други метазаписи што ги содржи документот во неговиот првобитен формат (pdf, doc, docx, итн.). Со оваа операција, корпусот е спремен за обработка и врз него може да се врши анотација и пребарување според соодветната методологија на Gate што ќе ја објасниме во продолжение на ова поглавје.

3.3 Моделизација

Оваа етапа е клучна во процесот на обработка на еден корпус. Целта на оваа студија е обработка на природниот јазик, земајќи ги предвид сите негови особености и правила на различни лингвистички нивоа (морфологија, синтакса, итн.). Според тоа,

[14] Анг. Java Annotation Patterns Engine.

најбитна етапа за да ги ексцерпираме сите контексти каде што се појавуваат нашите дел-реченици што сакаме да ги анализираме е нивната моделизација. Примерите врз кои ќе можеме да илустрираме неколку од бројните можности на Gate се примери од глаголската група и можат да бидат анализирани на морфосинтаксичкото рамниште и опишани како низи од симболи или од зборови. Да претпоставиме дека сакаме да ги анализираме сите дел-реченици што содржат честичка негација или сврзник (ако, да, не, ќе) проследена со заменка за индиректен и/или директен предмет и со глагол. Овие конструкции можеме да ги опишеме на следниов начин:

(честичка){1,*} + (директен|индиректен предмет){1,*} + глагол

Оваа моделизација е направена со помош на т.н. „регуларни изрази“ (анг. regular expressions) и во неа имаме три вида зборови: неменливи зборови (во кои ги вклучивме честичките, негацијата и сврзниците), заменки за индиректен и за директен предмет, и глаголи. Изразот значи дека неменливите зборови и заменките за директен и индиректен предмет се задолжителни барем еднаш, можат да стојат и повеќепати еднододруго (пр.: „не ќе му ја земев“, „да ми го рече“, итн.), додека глаголот е задолжителен. Треба да се напомене дека примерите не го опфаќаат заповедниот начин, за кој е потребна поинаква линеарна моделизација и проширување на правилата. Овој начин на прикажување на синтаксичките конструкции што сакаме да ги ексцерпираме се нарекува „мотив“. Мотивот е средство за опис на една низа симболи или зборови и се користи при пребарување со регуларните изрази кои претставуваат моќно помагало за опис на пишаниот јазик. Овие изрази се користат, пред сè, за автоматско препознавање на низи на симболи или зборови.

3.4 Множества (*Gazetteers*)

Некои низи претставуваат константи, т.е. редовно се среќаваат во дадени реченични контексти. Така, на пример, неменливите зборови, заменките за директен и заменките за индиректен предмет можеме да ги претставиме како множества на неменливи низи симболи (ниво 1), кои можат да се сретнат во низата зборови и симболи што ја сочинуваат глаголската конструкција од нашиот пример (ниво 2). За таа цел, ќе ги складираме во посебни речници (анг. Gazetteers) за да можеме да ги пребаруваме одеднаш и да ја упростиме моделизацијата со што ќе создадеме еден глобален мотив способен да ги ексцерпира сите низи што функционираат според овие примери:

- ќе одам;
- ќе му го дадам;
- да ми ја вратиш;
- не ти го пишувам итн.

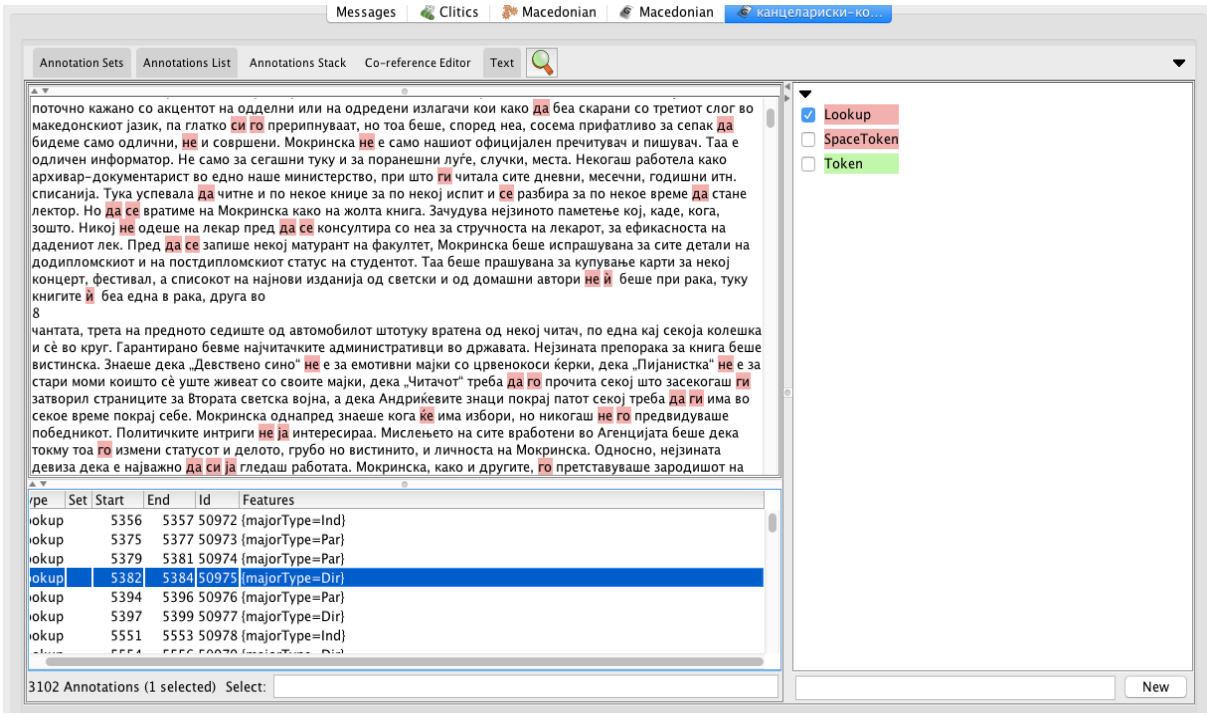
Речниците се листи на зборови со барем еден или повеќе заеднички именители. Во примерите што сакаме да ги ексцерпираме, можеме да ги групираме зборовите според нивната морфолошка категорија: честички, заменки за директен предмет и заменки за индиректен предмет. Речниците се текстуални документи во кои секој елемент има еден единствен збор и можат да бидат креирани со помош на обичен уредувач на текст од типот на Notepad ++[15]. Секој збор содржан во нив стои на посебна линија и може да има една или две спецификации (Major и Minor Type) што служат

[15] https://notepad-plus-plus.org/fr/ (последна консултација: 24 септември 2016).

како етикети на елементите на едно множество. Множествата содржат една единствена референтна листа (*.def) и повеќе лексички листи (*.lst). Еве како изгледа структурата на референтната листа на множествата што ќе ги искористиме при анотацијата на честичките и на заменките за директен и индиректен предмет:

direct.lst:Dir
indirect.lst:Ind
particles.lst:Par

Речниците ни овозможуваат да направиме „дискриминатори“ што ќе ни помогнат за полесна синтаксичка моделизација на зборовните групи, пред да се пристапи кон нивна ексцерпција. Во рамките на Gate, овие множества можат да се компилираат со помош на функцијата Hash Gazetteer[16] која е дел од апликацијата ANNIE[17] и претставува интегрален дел од помагалото. Резултатот од анотацијата со помош на множествата е следниот:



Илустрација 2: Резултати добиени по анотација на корпусот со помош на речници

Од оваа илустрација можеме да забележиме дека неменливите зборови се обележани како „Par“, заменките за директен предмет како „Dir“, а заменките за индиректен предмет како „Ind“. Сите обележани зборови се наоѓаат во сетот „Lookup“, со што корисникот, но и помагалото, располагаат со доволен број дискриминатори за понатамошна синтаксичка моделизација на овие конструкции и за нивна ексцерпција. За таа цел, ќе видиме како функцираат т.н. JAPE граматики, кои овозможуваат идентификација на зборовни низи, како и за ексцерпција и проучување на нивниот поширок контекст со помош на дискриминаторите опишани погоре.

[16] Целосната процедура за анотација на зборови со помош на множествата е опишана на веб-страницата https://gate.ac.uk/releases/gate-5.2.1-build3581-ALL/doc/tao/splitch13.html (последна консултација: 25 септември 2016).
[17] A Nearly-New Information Extraction System.

3.5 JAPE граматики

Овие граматики ни овозможуваат детален опис и ексцерпција на еден реченичен контекст и анотација на зборовните групи. JAPE граматиките се всушност алгоритми кои ги користат, меѓу другото, постојните анотации содржани во сетот Lookup, вклучувајќи ги и елементите што им припаѓаат на множествата, но и необележани зборови (содржани во сетовите Token и SpaceToken). Нивната функција може да се определи според реченичниот контекст. Структурата[18] на JAPE граматиките се состои од два типа правила: правила за опис и правила за анотација. Правилата за опис (услови) овозможуваат анализа на контекстот врз основа на определени дискриминатори (честичките и заменките), додека правилата за анотација (условни операции) се применуваат доколку мотивот што го аплицираме одговара на описот на една моделизирана конструкција. За ексцерпција на конструкциите што ги моделизиравме, граматиката што ги опфаќа сите случаи ја има следната архитектура:

```
Phase: firstpass
Input: Lookup Token
Options: control = appelt

Rule: inddir
Priority: 100
(
{Lookup.majorType == "Par"}
{Lookup.majorType == "Ind"}
{Lookup.majorType == "Dir"}
{Token.kind == "word"}
):mkGl
-->
:mkGl.inddir = {kind="indirektendirekten", rule="inddir"}

Rule: dir
Priority: 100
(
{Lookup.majorType == "Par"}
{Lookup.majorType == "Dir"}
{Token.kind == "word"}
):mkGl
-->
:mkGl.dir = {kind="direkten", rule="dir"}

Rule: ind
Priority: 100
(
{Lookup.majorType == "Par"}
{Lookup.majorType == "Ind"}
{Token.kind == "word"}
):mkGl
-->
```

[18] Функциите на JAPE се детално опишани на веб-страницата https://gate.ac.uk/sale/tao/splitch8.html (последна консултација: 25 септември 2016).

:mkGl.ind = {kind="indirekten", rule="ind"}

Сложеноста на правилата на една ваква граматика за секој лингвист кој нема познавања од областа на информатиката и програмските јазици е значителна. Сепак, ако се знае дека правилата на овие граматики функционираат како силогизми, нивната синтакса може многу лесно да биде совладана. Искуството и обуката овозможуваат конструкција на алгоритми способни да пребаруваат многу посложени синтаксички конструкции од оние што ги изучуваме во нашата статија. Правилата на овие алгоритми можат да бидат опишани на следниов начин[19]:

Командна линија		**Интерпретација**
Правила	Rule: inddir Priority: 100	Назив на правилото и приоритет на негово извршување во операцијата на анотација.
Опис	({Lookup.majorType == "Par"} {Lookup.majorType == "Ind"} {Lookup.majorType == "Dir"} {Token.kind == "word"})	= почеток на мотивот = Мотивот започнува со честичка... ...па со заменка за индиректен предмет... ...па со заменка за директен предмет... ...па продолжува/завршува со збор. = крај на мотивот =
Анотација	:mkGl --> :mkGl.inddir = {kind="indirektendirekten", rule="inddir"}	Анотација на низата зборови која ги задоволува критериумите на мотивот.

Табела 1: Структура на алгоритмите на граматиките JAPE

Оваа граматика е составена според анотациите на зборовите содржани во множествата кои ги објаснивме претходно и затоа се повикува на етикетите што ги избравме за да ги обележиме неменливите зборови, заменките за индиректен и заменките за директен предмет (MajorType). Множествата и JAPE граматиката се вклучени во една глобална апликација[20] што содржи, исто така, и модул за пречистување на корпусот (Document Reset PR), како и модул за сегментирање на зборовите (GATE Unicode Tokenizer). Редоследот на извршување на сите задачи во рамките на оваа апликација е следниов:

- Пречистување на корпусот (Document Reset PR)
- Сегментација на текстовите на зборови (Gate Unicode Tokenizer)
- Анотација на зборовите од множествата (Hash Gazetteer)
- Анотација и ексцерпција на моделизираните зборовни групи (JAPE Transducer)

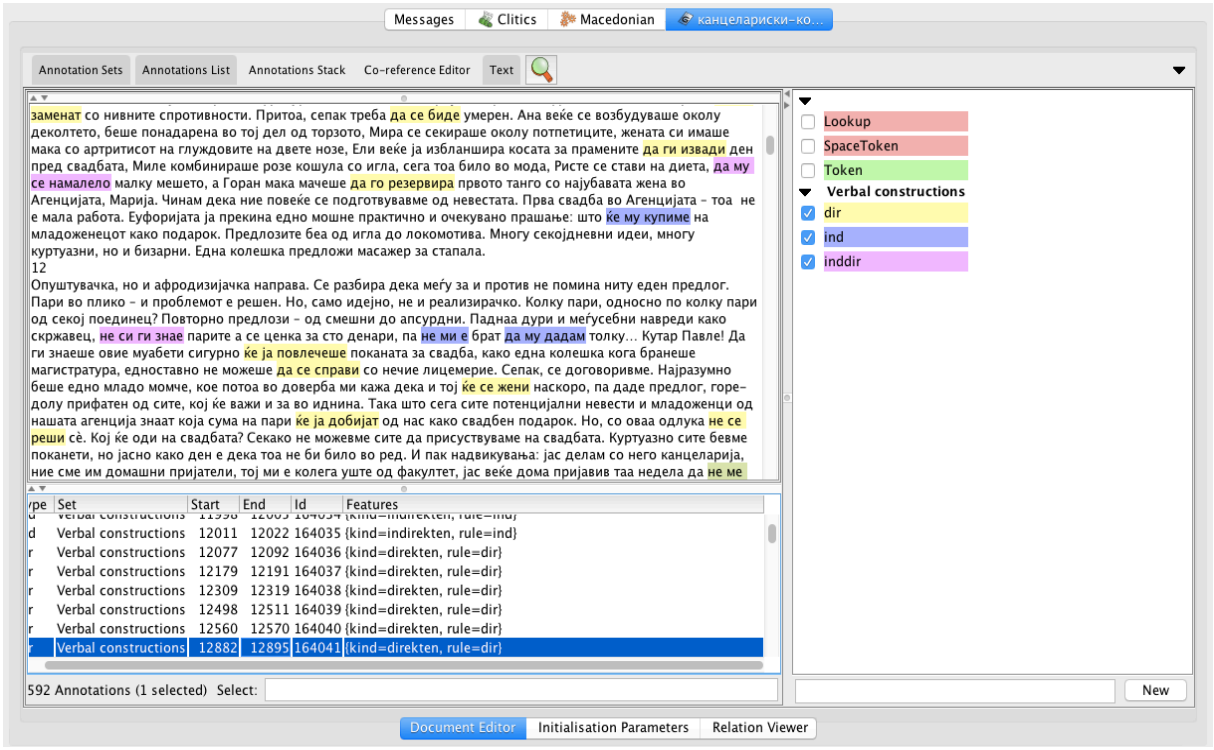
Со извршување на овие апликации врз корпусот, во сетот етикети Lookup се обележани елементите од трите константни множества (неменливите зборови,

[19] Во оваа табела објасnуваме само еден од мотивите на алгоритамот, кој овозможува да се ексцерпираат сите конструкции што содржат неменлив/и збор/ови, индиректен предмет и глагол.
[20] Пајплајн или нишка (анг. pipeline).

заменките за директен и заменките за индиректен предмет), додека во сетот Verbal constructions ги имаме трите глаголски конструкции што сакаме да ги ексцерпираме:

– Неменлив/и збор/ови + заменка за индиректен предмет + глагол
– Неменлив/и збор/ови + заменка за директен предмет + глагол
– Неменлив/и збор/ови + заменка за директен и/или индиректен предмет + глагол

Резултатите се прикажуваат во соодветни листи при самата консултација на еден од текстовите во нашиот корпус, со кликање на командите Annotation Sets и Annotation List на горниот дел од прозорецот. Командата Text овозможува консултација на реченичниот контекст на ексцерпираната форма при нејзина консултација во листата на долниот дел на екранот. Преку десното мени, корисникот избира (штиклира) кои анотации сака да ги прикаже на екранот. Во нашиот пример, сетот Lookup ги содржи дискриминаторите (неменливите зборови и заменките за директен и индиректен предмет), додека сетовите содржани во листата Verbal constructions ги содржат конструкциите што сакаме да ги ексцерпираме:



Илустрација 3: Резултати добиени по анотација на корпусот со JAPE граматики

Од илустрацијата 3 се забележува дека сите форми од типот (неменлив збор){1,*} + (директен|индиректен предмет){1,*} + глагол што сакавме да ги ексцерпираме се коректно анотирани во нашиот корпус. Корисникот може да направи увид и да ги проучува сите случаи поединечно со кликнување на листата што се појавува на долниот дел на прозорецот, при што во главниот прозорец ќе се појави поширокиот реченичен контекст за подетално проучување на еден пример од ексцерпираните форми. На долниот лев дел се појавува и бројот на означени конструкции од овој тип: вкупно 592 низи зборови во документот *Канцелариски колумни* се идентификувани како конструкции што се идентификувани од нашиот мотив за пребарување и ексцерпција.

4 Заклучоци и перспективи за понатамошни истражувања

Во оваа статија, најпрвин го дефинираме поимот „корпус“. Видовме дека неговото значење е различно според усвоеното гледиште и лингвистичката традиција. Потоа, проучивме неколку елементи од законската рамка според која можеме да се водиме при создавањето на еден корпус во македонскиот научен контекст. Во нашиот случај, законот предвидува слободно користење на сите пишани дела без претходна согласност од авторите во рамките на научните истражувања и на технолошките процеси коишто произлегуваат од нив. Во продолжение на нашата статија, дадовме еден мал приказ на неколку функционалности на помагалото Gate Developper што можат да бидат полезни при автоматското ексцерпирање, преку примери од глаголската група.

Несомнено е дека корпусот претставува една важна компонента во рамките на дескриптивната лингвистика и прашањата за неговото постоење би требало да бидат заменети со конкретни иницијативи за собирање на пишани и говорни документи, со цел македонскиот јазик да добие уште подетален и поусовршен опис од постојниот. Впрочем, корпусот е гаранција за квалитетот на описот на македонскиот јазик на морфосинтаксичко и на семантичко ниво и сите досегашни иницијативи што ги спомнавме можат да бидат искористени за понатамошни истражувања од областа на морфологијата и на синтаксата кои се, според нас, сè уште недоволно опишани. Токму поради тоа не можеме да очекуваме логични одговори и конечни решенија за многу прашања поврзани со зборовите и нивните функции во исказите.

Од друга страна, фактот што македонската кирилица е дел од големото Unicode UTF-8 семејство ни отвора еден голем спектар можности за користење на многу различни помагала за автоматска обработка на јазиците и за третман на пишани текстови, и тоа не само Gate Developper туку и Unitex[21], UIMA[22], NooJ[23], иако многубројните обиди за креирање корпуси, за жал, не наидоа секогаш на потребната поддршка од македонистичката научна фела. Сите тие изолирани иницијативи претставуваат поволна потпора за понатамошни истражувања врз македонскиот јазик кој, во доба на дигитализација, не може да биде поштеден од современите тенденции на науката. А дигитализацијата не ги засега единствено потребите на научниците, туку и оние на обичните зборувачи кои, сега за сега, не изобилуваат со методи за полесно и поедноставно изучување на јазичната норма.

Би ја заклучиле оваа статија повикувајќи се на една мисла на Крсте Мисирков која ни служеше како премиса за сите досегашни и ќе ни служи за сите понатамошни истражувања: секоја научна задача ни налага да видиме што сме направиле и што треба да направиме за понатаму. А она што ни претстои е, пред сè, еден поотворен поглед кон иницијативите за автоматска обработка на македонскиот јазик кои можат да му послужат на секој лингвист што се фатил во костец со нашиот богат јазик, но со млад јазичен стандард.

[21] http://www-igm.univ-mlv.fr/~unitex/ (последна консултација: 26 септември 2016).
[22] https://uima.apache.org/ (последна консултација: 26 септември 2016).
[23] http://www.nooj-association.org/ (последна консултација: 26 септември 2016).

ЛИТЕРАТУРА

Андреевски, Петре М. (1983), Пиреј, Наша Книга, Скопје.

Конески, Блаже (1952), Граматика на македонскиот литературен јазик, Култура, Скопје.

Минова-Ѓуркова, Лилјана (2011), Синтакса на македонскиот стандарден јазик, 2-ри Август С, Штип.

Митревски, Џорџ (2006), *Македонски електронски корпус: дизајн, имплементација, пристап*, Предавања на 38 Семинар за македонски јазик, литература и култура, 149‡257 (електронска верзија достапна на https://www.academia.edu/9969265/ - последна консултација: 15 септември 2016), УКиМ, Семинар за македонски јазик, литература и култура, Скопје.

Пандева, Лилјана (2012), Возовите на Луцерн, ПНВ Публикации, Скопје.

Пандева, Лилјана (2014), Канцелариски колумни, ПНВ Публикации, Скопје.

Aepli, Noëmi, Samardžić, Tanja & von Waldenfels, Ruprecht (2014), *Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote*, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) - Coling 2014 (електронско издание достапно на http://www.aclweb.org/anthology/W14-5309.pdf - последна консултација: 15 септември 2016), Dublin.

Ivanovska-Naskova, Ruska (2006), *Developent of the First LRs for Macedonian: CurrentProjects*, Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation, 1837‡1841 (електронско издание достапно на http://www.lrec-conf.org/proceedings/lrec2006/pdf/68_pdf.pdf - последна консултација: 15 септември 2016), Genoa.

Rafajlovska, Aneta & Zdravkova, Katerina (2015), *Représentation des expressions composées en macédonien en tant qu'entrées lexicales en Unitex*, 22ème Traitement Automatique des Langues Naturelles - atelier TASLA (електронско издание достапно на http://www.atala.org/taln_archives/ateliers/2015/TASLA/tasla-2015-court-001.pdf - последна консултација: 15 септември 2016), Caen.

Tanguy, Ludovic & Hathout, Nabil (2007), Perl pour les linguistes, Hermes science publication, Cachan.

Vojnovski, Viktor, Dzeroski, Sašo & Erjavec, Thomaž (2005), *Learning POS Tagging from a Tagged Macedonian Text Corpus*, Proceedings of the 8th International Multiconference on in Information Society, 11-17 October 2005, Ljubljana Conference on Data Mining and Data Warehouses (SKKID 2005), 199‡202 (електронско издание достапно на https://www.researchgate.net/publication/240087842_LEARNING_POS_TAGGING_FROM_A_TAGGED_MACEDONIAN_TEXT_CORPUS - последна консултација: 15 септември 2005), Ljubljana.

Ivanovska, Aneta, Zdravkova, Katerina, Džeroski, Sašo & Erjavec, Thomaž (2005), *Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns,* Proceedings of the

8th International Multiconference on in Information Society, 11-17 October 2005, Ljubljana Conference on Data Mining and Data Warehouses (SKKID 2005), 195‡198 (електронско издание достапно на https://www.researchgate.net/profile/Katerina_Zdravkova/publication/246670126_LEARNING_RULES_FOR_MORPHOLOGICAL_ANALYSIS_AND_SYNTHESIS_OF_MACEDONIAN_NOUNS/links/02e7e52a022ade30fa000000.pdf - последна консултација: 15 септември 2005), Ljubljana.

British National Corpus: http://www.natcorp.ox.ac.uk/corpus/index.xml (последна консултација: 28 септември 2016 година).

Техничка документација за British National Corpus: http://www.natcorp.ox.ac.uk/docs/URG/ (последна консултација на 28 септември 2016 година).

Centre National de Ressources Textuelles et Lexicales: http://www.cnrtl.fr/ (последна консултација: 28 септември 2016 година).

Закон за авторското право и сродните права: http://www.kultura.gov.mk/index.php/legislativa/2011-03-04-10-39-07/281-zakon-za-avtorskoto-pravo-i-srodnite-prava (последна консултација: 28 септември 2016 година).

Корпус Gralis Mak: http://www-gewi.uni-graz.at/gralis/ (последна консултација: 28 септември 2016 година).

Корпус Makedonski.info: http://www.makedonski.info/literature/show/ (последна консултација: 28 септември 2016 година).

Gate Developper: https://gate.ac.uk/ (последна консултација: 29 септември 2016). Notepad ++: https://notepad-plus-plus.org/fr/ (последна консултација: 24 септември 2016).

Процедура за анотација на зборови со Gate: https://gate.ac.uk/releases/gate-5.2.1-build3581-ALL/doc/tao/splitch13.html (последна консултација: 25 септември 2016).

JAPE граматики: https://gate.ac.uk/sale/tao/splitch8.html (последна консултација: 25 септември 2016).

Unitex: http://www-igm.univ-mlv.fr/~unitex/ (последна консултација: 26 септември 2016).

UIMA: https://uima.apache.org/ (последна консултација: 26 септември 2016).

NooJ: http://www.nooj-association.org/ (последна консултација: 26 септември 2016).