УДК 811.163.3:001.102-028.27

Руска Ивановска-Наскова

ПРОЕКТ *ЕЛЕКТРОНСКИ РЕСУРСИ ЗА МАКЕДОНСКИОТ ЈАЗИК:* СОСТОЈБА И ПЕРСПЕКТИВИ (NIP.UKIM.21-22.20)

Проектот *Електронски ресурси за македонскиот јазик: состојба и перспективи (ЕРМак)* на Филолошкиот факулетет „Блаже Конески“ – Скопје се реализираше во периодот од 31.1.2022 до 31.1.2023 година во рамките на програмата на Универзитетот „Св. Кирил и Методиј“ за поддршка на научноистражувачката дејност на единиците.

Поширокаата цел на проектот ЕРМак беше да се направи преглед на електронските ресурси за македонски јазик, да се разгледаат во поширок контекст што се однесува и на други јазици и да се одредат правците во кои треба да се насочат идните истражувања за македонскиот јазик на ова поле. Потесната и главна цел на проектот беше да се размисли подетално за идејата за создавање референтен корпус на македонскиот јазик, да се испита како би требало да изгледа еден ваков ресурс и да се даде предлог како би можел да се реализира. ЕРМак беше, исто така, можност да се испитаат капацитетите на УКИМ за реализација на проект од вакви размери.

Во реализација на проектните активности учествуваа:

– д-р Руска Ивановска-Наскова, редовен професор, Филолошки факултет „Блаже Конески“, главен истражувач

– д-р Анета Дучевска, редовен професор, Филолошки факултет „Блаже Конески“, истражувач

– д-р Бранислав Геразов, вонреден професор, Факултет за електротехника и информациски технологии, истражувач и

- д-р Николче Мицкоски, асистент-истражувач, Лексикографски центар „Георги Старделов“ во МАНУ, и стручњак од практиката, Филолошки факултет „Блаже Конески“, истражувач.

Целите на проектот беа остварени преку самостојни и заеднички истражувачки активности што се одвиваа во две фази. Во првата фаза беше направена анализа на постојните електронски ресурси за македонски јазик и на јазични корпуси на други јазици, додека, во втората фаза беа дефинирани параметрите на иден референтен корпус на македонскиот јазик. Истражувачките активности подразбираа анализа на електронски ресурси и на стручна литература, пишување научни трудови и изготвување елаборат. Во текот на проектот беа одржани вкупно пет работни средба на тимот, а дел од резултатите беа претставени на XLIX меѓународна научна конференција на LV летна школа на Меѓународниот семинар за македонски јазик, литература и култура, одржана на 4.9.2022 година во Охрид. Резултатите на проектот се претставени во три статии во печат и во текст-елаборат за иден референтен корпус на македонскиот јазик. Од истражувањето произлезе уште еден дополнителен резултат, а тоа е преглед на библиографски единици за електронски ресурси за македонскиот јазик, поднесен како прилог за објавување во стручно списание.

Истражувањата спроведени во рамките на проектот покажаа дека терминот *електронски* или *јазични ресурси* (англ. *language resources*) опфаќа различни видови јазични податоци и описи во електронски формат што се користат во автоматската обработка на јазиците. Во проектот беа земени предвид повеќе видови вакви ресурси за македонскиот јазик создадени во последните две децении: различни видови корпуси (еднојазични, паралелни, споредливи), програми за обработка на корпуси (тагери, парсери, лематизатори), електронски речници и терминолошки бази на податоци, алатки од доменот на говорните технологии, конјугатори итн. Во прегледот се опфатени околу 30 електронски ресурси и 60 библиографски единици што го опишуваат нивниот развојот и примена. Анализата покажа дека во изминативе години има напредок на ова поле во однос на алатките за македонски јазик и дека тие најчесто преставуваат резултат на индивидуални иницијативи или, пак, се дел од поголеми проекти во кои учествуваат, пред сè, истражувачи од странство. Прегледот на ресурсите укажа дека е потребно создавање нови алатки, а кога станува збор за покомплексни и пообемни електронски ресурси, дека би било добро во нивната реализација да се вклучат повеќе истражувачи и институции. Исто така, истражувањето ја истакна и неопходноста ресурсите да бидат направени според воспоставените светски стандарди, што ќе го олесни процесот на создавање, а ресурсите ќе ги направи споредливи и компатибилни со ресурси за други јазици.

Анализата на корпусите на други јазици беше фокусирана на референтни јазични корпуси, односно на корпуси чија цел е што поверодостојно да ги претставаат карактеристиките на стандардната варијанта на јазикот. Овие корпуси се составени од текстови од различни функционални стилови, имаат големи диманзии и дозволуваат различни видови пребарувања. Прегледот на состојбата со ваков вид ресурси укажа на тоа дека голем број јазици имаат свој референтен корпус. Во словенски и балкански јазичен контекст македонскиот јазик е еден од ретките без ваков јазичен ресурс. Анализата на референтните корпуси на повеќе јазици беше направена врз основа на следните параметри: обем, внатрешна структура (број и видови поткорпуси), длабочина на означување, можности за пребарувања и обработка на резултатите, пристапност. Исто така, за остварување на главната цел на проектот беше важно и да се дојде до податоци за тоа кои институции биле вклучени во нивната изработка, со колкави човечки и материјални ресурси учествувале и дали биле создадени како самостоен проект или, пак, биле дел од проект со друга главна цел. Истражувањето покажа дека квалитетните референтните корпуси честопати се плод на соработка на повеќе инстиитуции и на големи тимови на истражувачи и стручњаци од доменот на лингвистиката и информациските технологии. Изработката на референтен корпус вообичаено се смета за стратешки долгорочен проект од национален интерес и затоа поголемиот дел од овие проекти во значителен дел се поддржани, пред сè, од државата.

Главната цел на проектот ЕРМак беше остварена преку изготвување *Елаборат за поставување основни параметри на иден корпус на македонскиот јазик*. Текстот на Елаборатот е артикулиран во девет поглавја, претставени на 30 страници текст. Автори на Елаборатот се сите учесници во проектот. Во првата глава насловена *Вовед* објаснета е мотивацијата за подготовка на ваков документ, претставени се истражувачите-автори на Елаборатот и целната група за која е наменет овој документ. Во втората глава, *Состојбата со електронски ресурси за македонскиот јазик и потребата од нивно развивање*, се дава кус осврт на развојот на интересот на оваа проблематика во научната и академската средина кај нас и пошироко, даден е преглед на ресурсите и идентификуван е токму недостатокот на референтен корпус на македонскиот јазик, како најгорливо прашање поврзано со оваа тема. Третата глава дава одговор на прашањето: што претставува јазичен корпус? Првиот дел од третата глава се однесува на примената на овие електронски ресурси, додека, вториот дава преглед на различни видови корпус. Во последниот дел детално е претставен процесот на консултирање корпуси преку конкретни примери за различни

видови пребарувања (основни и напредни пребарувања, пребарувања во означен корпус, лексичка честота, колокации, пребарување поткорпуси). Четвртата глава се однесува на лингвистичките параметри на референтните корпуси. Во неа се разгледува прашањето за видот на текстовите (пишани и говорни), за времето на создавање на текстовите, како и за нивниот домен и големина. Оваа глава се осврнува, исто така, и на степенот на обработка на текстовите, односно на процесот на означување на текстовите, на длабочината на означувањето и на неопходноста за рачно коригирање на автоматски означениот корпус. Тука се разгледува, исто така, и прашањето за обемот на корпусот, а на крајот е даден и конкретен предлог за можен профил на референтен корпус на македонскиот јазик. Во петтата глава, *Технички параметри на јазичен корпус*, се претставени главните технички аспекти на референтните јазични корпуси, особено, како ќе биде обработуван, чуван и користен материјалот. Поконкретно, тука се опишани практиките на именување датотеки, обележување заглавие на текст, прашања поврзани со форматот на текстот, на звучните записи и на транскрибираните текстови. Во овој дел се разгледуваат, исто така, и хостирањето, користењето, изработката, управувањето и одржувањето на корпусот. И овој дел, како и претходниот, дава можен профил на иден референтен корпус на македонскиот јазик во однос на техничките параметри. Во шестата глава се наведени потребните чекори за обединување на досегашните потфати во еден голем национален проект за македонски јазичен корпус. Во седмата глава, која се однесува на рамковната финансиска и временска конструкција за остварување на проектот, се констатира дека првата голема етапа на овој долгорочен проект би траела три години и дека е потребна дополнителна студија за изводливост на проектот, која излегува од рамките на овој проект. Во осмата глава се претставени неколку главни заклучоци и препораки. Последната глава содржи библиографски единици на кои се упатува во Елаборатот.

Активностите во проектот го актуализираат прашањето за изработка на референтен корпус на македонскиот јазик, како неопходен чекор во осовременувањето на методологијата со која се истражува македонскиот јазик и во заштита на македонскиот јазик. Сознанијата од проектот ќе послужат како основа за подготовка на подетално разработен предлог-проект за создавање електронски корпус на македонскиот јазик, како и за поттикнување на интересот за оваа проблематика кај пошироката научна јавност и кај институциите што се грижат за македонскиот јазик.