UDK: 004.9:811.16’374
811.16'322

ЈЕЗИЧКИ РЕСУРСИ ЦЕНТРАЛНЕ И ЈУЖНЕ ЕВРОПЕ У ОКВИРУ ПЛАТФОРМЕ МЕТА-РАЗМЕНА

Мaћej Огродничук (maciej.ogrodniczuk@ipipan.waw.pl),
Институт за рачунарство, Пољска академија наука
Радован Гарабík (garabik@kassiopeia.juls.savba.sk),
Институт за лингвистику „Људовит Штур“, Словачка академија наука
Светла Коева (svetla@dcl.bas.bg),
Институт за бугарски језик, Бугарска академија наука
Цветана Крстев (cvetana@matf.bg.ac.rs),
Универзитет у Београду, Филолошки факултет
Пјотр Пезик (piotr.pezik@gmail.com),
Универзитет у Лођу
Тибор Пинтер (pinter.tibor@nytud.mta.hu),
Институт за лингвистичка истраживања, Мађарска академија наука
Адам Пжепјорковски (adam.przepiorkowski@ipipan.waw.pl),
Институт за рачунарство, Пољска академија наука
Ђерђ Сасак (szaszak@tmit.bme.hu),
Катедра за телекомуникације и информатику медија,
Будимпештански универзитет за технологију и економију
Марко Тадић (marko.tadic@ffzg.hr),
Универзитет у Загребу, Филозофски факултет
Тамаш Варади (varadi@nytud.hu),
Институт за лингвистичка истраживања, Мађарска академија наука
Душко Витас (vitas@matf.bg.ac.rs),
Универзитет у Београду, Математички факултет

превод с енглеског Јелена Бајић

Апстракт

Намера овог рада је да пружи кратак преглед једног од најновијих подухвата усмерених на изградњу паневропске инфраструктуре у области језичких технологија, МЕТА-НЕТА – мреже изврсности која окупља 54 истраживачка центра из 33 земље, а посебно пројекта CESAR земаља учесница из Централне и Јужне Европе. Једна од главних активности у оквиру пројекта је одабир ресурса и алата који ће се сакупљати, валидирати, стандардизовати, унапређивати, проширивати и поравнавати међу језицима и чувати у отвореном систему за размену ресурса МЕТА-РАЗМЕНА (META-SHARE).

У овом прилогу се највећа пажња посвећује представљању репозиторијума у коме

се складиште метаподаци из одабраних извора, методологије и критеријума за њихову селекцију и даје се детаљан приказ ресурса и алата обезбеђених у оквиру пројекта у 2011. години. У фокусу су прво појмови модела метаподатака, МЕТА-РАЗМЕНА и синхронизоване мреже сервера са метаподацима, након чега се у чланку представљају методологија и критеријуми за селекцију ресурса кроз одређивање број бодова појединачних ставки на основу поузданих критеријума за прављење процене, као што су: доступност ресурса, квалитет и квантитет сличних доступних ресурса, покривеност, зрелост, одрживост и прилагодљивост.

Такође су представљене беле књиге о језицима настале у оквиру мреже МЕТА-НЕТ – низ извештаја о стању сваког европског језика понаособ, по питању језичких технологија, као и упутства за лиценцирање која је предложила заједница учесника у МЕТА-РАЗМЕНИ. Тим упутствима се промовише отвореност и бесплатно коришћење података и алата путем стандардизованих и добро дефинисаних начина заштите ауторских права.

Кључне речи

Језички ресурси и алати, метаподаци, репозиторијум ресурса, отворена лингвистичка инфраструктура, беле књиге о језицима, словенски језици, бугарски, хрватски, мађарски, пољски, српски, словачки

Увод

Како циљеви информационог друштва све више постају животна стварност, изазови са којима се суочава језик, као средство за комуникацију својствено људској врсти, без обзира на технологију којом се преноси, постају једно од фундаменталних питања. Отуда и значај језичких технологија, као кључних технологија које отварају нове могућности, а потребне су у бројним областима, од уклањања језичких баријера и очувања културног наслеђа, до екстракције знања путем анализе текста. То су само неке од области које могу занимати читаоце овог часописа.

Језичке технологије у пресудној мери зависе од језичких ресурса , велике количине пажљиво анализираних и анотираних података, као и одговарајућих алата и стандардизованих метода обраде тих ресурса. Брига да се обезбеди да подаци и алати буду широко доступни на стандардан и добро документован начин, укључујући и јасан статус по питању интелектуалне својине уродила је формулисањем идеје инфраструктуре у области језичких технологија. Један од најновијих паневропских подухвата је МЕТА-НЕТ, мрежа изврсности заснована на четири пројекта који у њој учествују. Међу њима је и пројекат CESAR о коме је објављен извештај у једном од претходних бројева овог часописа (Varadi, 2011).

Намера овог чланка, чији аутори су партнери на пројекту CESAR је да представи преглед инфраструктуре МЕТА-РАЗМЕНЕ, као и опис прве групе језичких ресурса и алата које је припремио конзорцијум CESAR.

1. МЕТА-НЕТ и МЕТА- РАЗМЕНА

МЕТА-НЕТ је мрежа изврсности посвећена развоју технолошких темеља вишејезичног европског информационог друштва. Језичке технологије ће:

- омогућити комуникацију и сарадњу на више језика,
- обезбедити говорницима свих језика укључених у пројекат једнак приступ информацијама и знању,
- доприносити даљем развоју и унапређивању функционалности умрежене информационе технологије.

Потребно је учинити знатан заједнички

напор на целом континенту на плану истраживања и инжењеринга у области језичких технологија са циљем развоја апликација које омогућују аутоматско превођење, вишејезичко управљање информацијама и знањем, као и продукцију садржаја на свим европским језицима. Уложени напор ће довести до стварања интуитивних језичких сумећа за технологију применљиву у низу области, од кућних апарата, машина и возила, до рачунара и робота.

Ради остварења тих циљева, META-NET ствара Технолошки савез вишејезичне Европе - META (енг. Multilingual Europe Technology Alliance). С обзиром да окупља истраживаче, оне који пружају приступ технологији на комерцијалним основама, приватне и пословне кориснике језичких технологија, стручњаке за језик и друге стране заинтересоване за информационо друштво, META ће припремити потребни, амбициозни, заједнички подухват усмерен на развијање језичких технологија, као средства за остварење визије Европе уједињене у јединствено дигитално тржиште и информациони простор.

1.1. Модел метаподатака

Једна од најдрагоценијих карактеристика META-РАЗМЕНЕ је глобално сагледавање ресурса, чиме значајно превазилази идеју репозиторијума. Осим што пружа каталог језичких ресурса и алата (података, алата, технологија), намера је такође да чува податке који могу да се употребе за унапређење њиховог коришћења, као што су референтни документи (радови у којима се описује ресурс и са њима повезани извештаји, упутства о скуповима етикета, смернице за продукцију језичких ресурса, итд.), информације о особама и организацијама укљученим у њихово стварање и коришћење (на пример, о стваоцима ресурса, финансијерима, дистрибутерима, итд.), линкови ка сродним покретима и активностима (на пример, ка пројектима којима је финансирано стварање језичких ресурса, у оквиру којих је експлоатисан неки од језичких ресурса, итд.) или лиценце (за дистрибуцију језичких ресурса).

Модел метаподатака META-РАЗМЕНА се базира на три појма из компонентне инфраструктуре метаподатака (Broeder at al. 2008):

- компонентама – у њима су садржани елементи који окупљају семантички кохерентна својства у слабије гранулиран скупове
- елементима – они кодирају специфичне дескриптивне карактеристике језичких ресурса,
- релацијама – оне повезују међусобно повезане ресурсе из репозиторијума META-РАЗМЕНА (на пример, изворне и изведене, сирове и анотиране ресурсе, одређени језички ресурс и алат употребљен приликом његовог стварања, итд.)

Заједнички модел метаподатака META-РАЗМЕНА је, у складу са овим принципом, припремљен и кодиран уз помоћ нотације XML схема. Она нуди два основна нивоа описа ресурса: минималну схему која пружа основни опис ресурса, а задржава се ради обезбеђивања компатибилности и максималну схему са вишим нивоом грануларности која даје детаљније информације о сваком ресурсу. Слично томе, класе елемената су подељене у групе према важности и зависности: обавезне, обавезне у зависности од услова (морају да буду попуњене када су испуњени специфични услови), препоручене (стваоцима језичких ресурса се саветује да укључе те информације) и необавезне.

У овом моделу постоје два нивоа таксономије језичких ресурса: први ниво се односи на тип ресурса (корпус, лексички/концептуални ресурс – терминолошки ресурси, листе речи, семантички лексикони, онтологије, језички опис и технологија/алат. Други ниво

омогућава поткласификацију у зависности од типа (језик, медиј, домен, формат, анотацијске карактеристике, итд.).

1.2. Репозиторијум META-РАЗМЕНА

META-РАЗМЕНА је отворена дистрибуирана платформа за дељење и размену ресурса коју обезбеђује META-NET. Сервери META-РАЗМЕНЕ формирају ланац међусобно повезаних чворова који корисницима омогућавају да лако и без препрека приступају ресурсима, а власницима ресурса обезбеђују поуздану сумеђу за уређивање и обављање административних послова.

Репозиторијуми META-РАЗМЕНЕ садрже описе ресурса у облику метаподатака који одговарају управо описаном моделу метаподатака, а попуњавају се и ажурирају прикупљањем података из постојећих иницијатива и пројеката на којима се сарађује, као и постављањем нових описа ресурса путем програма за уређивање метаподатака и одржавањем дозвола за приступ. Чим метаподаци новог ресурса постану доступни неком од чворова META-РАЗМЕНЕ, мрежа аутоматски умножава и синхронизује описе између осталих регистрованих чворова (видети слику 1).

Слика 1. Резултат претраге репозиторијума META-РАЗМЕНА [1], после филтрирања језика (пољски) и врсте вишејезичног ресурса (паралелни). Пронађено је четири таква ресурса а све их је обезбедио CESAR.

1 http://www.meta-share.eu/

2. Попуњавање META-РАЗМЕНЕ

Концепти META-НЕТА и META-РАЗМЕНЕ су први пут тестирани на паневропском нивоу почетком децембра 2011. године, када су сва три братска ICT-PSP пројекта учинила своју прву траншу ресурса доступним. У овом раду се фокусирамо на један од тих пројеката, CESAR (Central and South-east europeAn Resources (Ресурси Централне и Југоисточне Европе).

2.1. Пројекат CESAR

Технологије за обраду природних језика у пресудној мери зависе од језичких ресурса и алата који су применљиви, корисни и доступни. Намера пројекта CESAR (у који је укључено 9 партнера из 6 земаља), је да се, кроз блиску сарадњу и усклађивање циљева са савезом META-NET, позабави овим питањима и то путем даљег развоја, унапређивања стандардизовања и међусобног повезивања широког спектра језичких ресурса и алата и омогућавања да они буду доступни, што представља допринос отвореној лингвистичкој инфраструктури.

Главни циљ пројекта је да се учини доступним свеобухватни скуп језичких ресурса и алата којима су обухваћени бугарски, хрватски, мађарски пољски, српски и словачки језик. Из чињенице да су пројектом обухваћени ови језици произилази додатна корист од пројекта – антиципирање и испуњавање захтева који се могу предвидети, а тичу се ресурса ових језика. Ослањајући се на широк спектар већ постојећих ресурса и националних и интернационалних активности, пројекат ствара и попуњава свеобухватну платформу језичких ресурса којом и управља. Та платформа омогућава и подржава вишејезичне и међујезичке производе и услуге великих размера. Међу ресурсима већ укљученим у пројекат (чији број стално расте) су: међусобно разменљиве једнојезичне и вишејезичне говорне базе

података, једнојезични и двојезични корпуси, речници, worднети и алати за обраду података релевантни у области језичких технологија, као што су: токенизатори, лематизатори, тагери и парсери.

Главни циљеви пројекта CESAR су:

- описивање: ситуације у земљама учесницама по питању употребе језика; производа и услуга који се заснивају на познавању језика, језичких технологија и ресурса; главних актера (истраживачи, индустрија, државна управа и друштво); државне политике и програма, преовлађујућих стандарда и пракси; тренутног нивоа развоја, главних покретача и препрека;
- давање доприноса паневропском систему за размену дигиталних ресурса кроз прикупљање, документовање, повезивање и унапређивање ресурса да би се достигли договорени стандарди и смернице;
- сарадња са другим партнерским пројектима, нарочито са пилот пројектом ICT-PCP 6.1 који се одвија у исто време када и CESAR и мрежом изврсности, META-HET, а тамо где то може бити од користи, и са другим релевантним мултинационалним форумима и активностима, као што су FlaReNET[2] и CLARIN[3] – да би се обезбедила доследност приступа, пракси и стандарда, чиме се лакше постиже шира доступност и лакши приступ квалитетним језичким ресурсима и алатима, односно, њихово поновно коришћење;
- помоћ при прављењу опсежних, некомерцијалних репозиторијума које стварају чланови заједнице и који су међусобно повезани, а могу их користити истраживачи у области језика, појединци који се баве развојем софтвера и стручњаци;
- мобилизација заинтересованих страна, државних органа и агенција које обезбеђују финансијска средства, у земљама учесницама и у региону, а то се постиже развијањем свести, организовањем састанака и других догађаја усмерених на остварење тог циља;
- јачање сарадње између главних партнера на пољу технологије у региону, настављајући ранију сарадњу на пројектима TELRI[4], MULTEXT-East[5] и другим;
- превазилажење технолошког јаза између овог региона и других делова Европе, елиминисањем очигледних и важних недостатака у инфраструктури језичких ресурса и алата.

Главни стубови активности у оквиру пројекта CESAR се огледају у унапређењу ресурса и алата (у смислу величине, покривености, прецизности, одзива, тачности), прилагођавању ресурса и алата да би се ускладили са договореним стандардима међусобног уклапања, даљем развоју ресурса и алата тако што се комбинују са другим ресурсима и алатима да би се постигао предвиђени ниво међусобног уклапања и у прилагођавању корисничких сумећа да би се испунили захтеви корисника. Посебан напор је уложен да би се успоставио заједнички стандард ресурса и алата укључених у пројекат, да би се побољшало и олакшало

2 Fostering Language Resources Network (www.flarenet.eu/) – Мрежа за развој језичких ресурса

3 Common Language Resources and Technology Infrastructure (www.clarin.eu/external/) – Заједничка инфраструктура језичких ресурса и технологија

4 Trans-European Language Resources Infrastructure (//telri.nytud.hu/) - Трансевропска инфраструктура језичких ресурса

5 Multilingual Text Tools and Corpora for Central and Eastern European Languages (nl.ijs.si/) –Вишејезични алати за обраду текста и корпуси за језике Централне и Источне Европе

постизање предвиђене међусобног уклапања између њих, као и да би се извршила процена њихових начина лиценцирања и питања која се тичу права интелектуалне својине.

Циљ пројекта CESAR је да се подстакне међујезичка комуникација, сарадња и учешће који се остварују уз помоћ информационо-комуникационих технологија и тиме допринесе стварању јединственог паневропског дигиталног тржишта. Најважнији ресурси обухваћени пројектом CESAR ће бити повезани и учињени међусобно уклопивим уз коришћење капацитета репозиторијума МЕТА-РАЗМЕНА. Циљна заједница корисника практично обухвата све заинтересоване стране на савременом дигиталном тржишту: обичне крајње кориснике, професионалне крајње кориснике (пословне кориснике, администрацију, медије, образовне установе, библиотеке итд.), као и стручњаке за различите области (истраживаче, представнике индустрије, креаторе политике и др.). Пројекат настоји да пажљиво истражи потребе различитих врста корисника – од појединаца до великих мултинационалних организација – из перспективе тренутног статуса, као и скорије будућности.

Овај пројекат, такође, својим драгоценим ресурсима даје допринос МЕТА-РАЗМЕНИ која ће једног дана извесно постати значајан елемент тржишта језичких технологија за истраживаче технологија обраде природних језика и оне који раде на њиховом развоју, људе који се професионално баве језиком (преводиоце који се баве писменим и усменим превођењем, стручњаке за локализацију садржаја и софтвера, итд.), као и за индустрију, нарочито мала и средња предузећа која учествују у целокупном процесу развоја технологија у области обраде природних језика, од истраживања, до иновативних производа и услуга.

Сарадња на обухваћеним ресурсима и алатима ће се одвијати (а материјал подносити) у три фазе. Резултат активности (као што су горепоменута надоградња, стандардизација, хармонизација, повезивање, решавање проблема везаних за права интелектуалне својине) је већ постигнут у такозваној „првој транши“ ресурса, која је објављена почетком децембра 2011. године (објављивање „друге транше“ – описа активности и метаподатака је заказано за крај августа 2012, а треће за крај фебруара 2013. године). Прва транша обухвата 33 ресурса за шест језика и садржи 20 корпуса (19 писаних и један говорни), 2 речника, 4 wordнета, 1 лексикон и 6 говорних база података. У овој транши се налазе подаци које су приложили партнери на пројекту, али се планира да се у преостале две транше у оквиру пројекта CESAR укључе и други релевантни центри у земљама учесницама, да би се повећао број унапређених ресурса.

2.2. Методологија и критеријуми за одабир ресурса

Први корак је подразумевао развој методологије на основу које би била рађена процена идентификованих језичких ресурса. Партнерима је подељен упитник у намери да се добију сугестије какав приступ применити што се тиче процедуре процене. Потврђено је да се ни једна од тренутно постојећих методологија не може самостално прихватити као стандард. Уместо тога, конзорцијум је направио листу четири општа индикатора који се сматрају репрезентативним и индикативним за одабир језичких ресурса. Индикатори одређују опште захтеве који треба да буду испуњени приликом селекције. За сваки индикатор су дефинисане различите групе прецизних критеријума. Критеријуми су описани у одељцима који следе.

2.2.1. Општа процена ресурса

Процес унапређивања ресурса и алата се, у оквиру овог индикатора, одвија у три тока: надоградња ресурса, проширивање и поравна-

вање међу језицима. Међу тим индикаторима се прави даља класификација према следећим критеријумима:

За надограђене ресурсе:

- Сви одабрани ресурси су **најсавременији** представници типа коме припадају за одређени језик (да, не)
- **Представници исте вредности** су сви **укључени** у одабир (да, не)
- Тренутни статус ресурса је такав да су они **супериорног квалитета**, макар на регионалном нивоу, те нема потребе за претераним даљим развојем (да, не)
- Питања лиценцирања су решена тако да је могућ слободан приступ обради и ресурсима, као и материјалима везаним за ресурсе или конзорцијум успева да постигне договор са одговарајућим носиоцима ауторских права (да, не)

За проширене/повезане ресурсе:

- **Проширење** ресурса је **од велике вредности** за заједницу, макар на регионалном нивоу (да, не)
- Нагласак је дат на **обезбеђивање додатних блокова за постојеће алате, пре него на значајно реструктурирање** (да, не)
- **Додатни ресурси су интегрисани** са постојећим **само ако доводе до битног побољшања квалитета** ресурса који произилазе из тог процеса (да, не)
- **Ако је одабрано више представника једног типа ресурса** за неки језик, врло је вероватно да ће они бити **међусобно повезани** да би биле искоришћене предности добрих страна оба решења (да, не)
- Ако м**ање развијеним, мада још увек веома популарним ресурсима** могу бити од користи побољшања, са обзиром да имају добро развијене еквиваленте, разматра се и њихово унапређивање (да, не)
- **Искуство других чланова** конзорцијума/других конзорцијума се у великој мери **користи** у процесу проширивања националних ресурса да би се обезбедило чврсто утемељење за рад на више језика (да, не)
- **Даје се предност алатима који су применљиви на све језике** или који се могу користити за рад на више језика (да, не)

За ресурсе поравнате на више језика:

- За сваки појединачни језик се не користи **више од једног алата одређеног типа** (да, не)
- Када год је могуће, бира се највећи сет језика (да, не)
- Алати за обраду језика у **NooJ** (да, не)
- Жели се у великој мери постићи **независност од језика** (да, не)
- **Квалитет резултата** је од огромне важности (да, не)

О исправности спецификације се не може судити без познавања ширег контекста употребе, адекватности, итд., језичког ресурса. Ради процене квалитета, квантитета и значаја, сваки случај ће се детаљно анализирати, узимајући у обзир регионалне одлучујуће факторе, популарност формата изван институције из које потиче, итд. Овај индикатор захтева комплексну процену језичких ресурса у контексту читавог сета утврђених критеријума. Партнери не само што оцењују да ли одабрани ресурси испуњавају утврђене критеријуме, већ и дају конкретне примере и детаљна објашњења на основу темељне анализе.

2.2.2. Укупан број бодова појединачних ставки

Примењујући приступ пројекта Европске уније, NEMLAR (Network for Euro-Mediterranean Language Resources (Мрежа евромедитеранских језичких ресурса)) (у вези BLARK-а за арапски – за више информација

о BLARK-у, погледати тачку 2.2.5), детаљније се прецизирају појмови доступности, квалитета, квантитета и стандарда и узимају се у обзир у процесу одабира језичких ресурса. Развијена је техника која допуњује приступ NEMLAR-а, а притом дефинише егзактне мере за аспекте квалитета и квантитета и инкорпорира стандардизацију у домен квалитета. Процес евалуације се састоји од следећих корака: спецификације броја бодова појединачних ставки у оквиру сваке мере за сваки од ресурса; агрегације ставки у јединствени збир (укупан број бодова појединачних ставки); показивања корисности језичких ресурса у даљој обради; одабира ресурса који испуњавају предефинисане услове. Одређене су следеће ставке које се бодују:

i. Доступност: Доступно коме?, По којој цени?, Колико је једноставно поновно коришћење (степен прилагодљивости)?;
ii. Квалитет: Усклађеност са стандардима (да ли је ресурс заснован на заједничком стандарду?), Поузданост (Унутрашња доследност, т.ј., да ли је ресурс заснован на добро дефинисаним спецификацијама?), Релевантност у односу на задатак (да ли ресурс одговара одређеном задатку?), Релевантност у односу на окружење (да ли је се ресурс може уклопити са другим ресурсима?)
iii. Квантитет (само ресурси).

Најмањи могући укупан број бодова је 8, а највећи 25 бодова. Утврђени критеријуми за одабир језичких ресурса налажу да укупан број бодова буде мањи од минималног броја од 16 или једнак том броју бодова. Укупан број бодова за ресурсе који се одабирају за пројекат се може израчунати пре него што се предузму било какви радови на плану надоградње. Тај процес је у директној вези са снижавањем укупног број бодова, што се може користити као конкретни индикатор успешности пројекта.

2.2.3. Беле књиге о језицима

Серија белих књига META-NETA, „Језици у европском информационом друштву“ представља извештаје о стању сваког европског језика понаособ, што се тиче језичких технологија и објашњава најважније ризике и могућности.

У белим књигама о језицима се даје преглед тренутне ситуације у области подршке језичким технологијама. Рејтинг постојећих ресурса и алата се заснива на стручним проценама неколико водећих експерата, при чему се примењују следећи критеријуми (распон бодова за сваки од критеријума је од 0 до 6).

i. Квантитет: Да ли алат/ресурс постоји за језик о коме је реч? Што више алата/ресурса постоји, рејтинг је виши;
ii. Доступност: Да ли су алати/ресурси доступни, т.ј. отвореног кода, односно, да ли се могу слободно користити на било којој платформи или су доступни по високој цени или под веома рестриктивним условима?;
iii. Квалитет: У којој мери најбољи постојећи алати, апликације или ресурси задовољавају одговарајуће критеријуме којима се одређује ефикасност алата, односно, индикаторе квалитета ресурса?;
iv. Покривеност: До ког степена најбољи алати испуњавају одговарајуће критеријуме покривености? До ког степена ресурси репрезентују циљни језик или подјезике?;
v. Зрелост: Да ли се алат/ресурс може сматрати зрелим, стабилним и спремним за тржиште? Могу ли се најбољи доступни алати/ресурси користити у форми у којој се испоручују или се морају адаптирати?;
vi. Одрживост: Колико добро се алат/ресурс може одржавати/интегрисати у постојеће информатичке системе?;
vii. Прилагодљивост: Колико добро се нај-

бољи алати/ресурси могу прилагодити/проширити на нове задатке/домене/жанрове/типове текстова/ситуације у којима се користе, итд.?

Корист корист коју доносе језичке технологије варира од језика до језика, у зависности од фактора као што су комплексност одређеног језика, величина заједнице његових говорника и постојање активних истраживачких центара у области у којој се тим језиком говори. Најважнији фактори за надоградњу ресурса су квалитет и зрелост података, у случајевима када није планирана никаква додатна обрада, осим аутоматске конверзије. Да би ресурси били вишејезички поравнати, сви фактори су од кључне важности, а ресурс треба да представља најбољи доступни резултат. Са друге стране, ресурси који имају најмање бодова су подједнако прикладни за даљу обраду, пошто мањи број бодова указује на потребу да се побољша стање у области одређеног језика.

Ситуацију у језицима обухваћеним пројектом су описали и податке прикупили партнери који сарађују на CESAR-у – за бугарски (Blagoeva et al., 2011), хрватски (Tadić et al., 2011), мађарски (Simon & Lendvai, 2011), пољски (Miłkowski, 2011), српски (Vitas et al., 2011), и словачки (Šimková et al., 2011).

2.2.4. Пропорција између одабраних ресурса развијених унутар и изван конзорцијума

Ресурси могу бити класификовани према томе да ли се развијају унутар конзорцијума, изван њега или на оба поменута начина. Ова информација пружа додатне доказе који се могу поредити са уоченим помањкањима, тако да се извесни напори могу усмерити на даљу идентификацију језичких ресурса изван конзорцијума.

2.2.5. BLARK и закључак

BLARK (Basic Language Resources Kit - Основни прибор за процену језичких ресурса) је концепт дефинисан захваљујући заједничкој иницијативи ELSNET-а[6] и ELRA-е[7]. BLARK се дефинише као минимални сет ресурса неопходних за било какво истраживање које води ка комерцијалном производу или за образовање за све. BLARK укључује многе разнородне ресурсе, попут (једнојезичних и вишејезичних) корпуса писаног и говорног језика, једнојезичних и двојезичних речника, збирки термина и граматика, тагера, морфолошких анализатора, парсера, анализатора и препознавача говора, итд. ELDA[8] (Evaluations and Language Resources Distribution Agency (Агенција за процене и дистрибуцију језичких ресурса)) је израдила подробан извештај у коме се дефинише (минимални) сет језичких ресурса који треба учинити доступним за што већи број језика.

BLARK може да пружи смернице како да се одреди приоритет за почетни одабир и да се постигне да он не зависи од локалних преференци. Постоје информације о покривености језичких ресурса и алата за 13 језика који се говоре у ЕУ (како за велике језике, тако и за оне са мањим бројем говорника), на основу колективног искуства и знања које стичу припадници језичке заједнице. До сада, ни за један језик заступљен у пројекту CESAR не постоје званичне информације добијене применом BLARK-а. Са друге стране, класификација из белих књига о језицима и резултати израчунавања укупног број бодова појединачних ставки за алате и ресурсе за појединачне језике обухваћене пројектом пружају обиље доказа за и против одабира одређеног ресурса. Стога,

6 European Network of Excellence in Language and Speech (Европска мрежа изврсности у језику и говору) (www.elsnet.org)

7 European Language Resources Association (Evropsko udruženje za jezičke resurse) (www.elra.org)

8 Evaluations and Language Resources Distribution Agency (Агенција за процене и дистрибуцију језичких ресурса) (www.elda.org)

супротно нашим почетним намерама, BLARK није уврштен у индикаторе на основу којих се врши процена за пројекат CESAR.

Да закључимо, комбинација четири индикатора (сваки је прецизиран према различитим групама критеријума) се користи у процесу одабира језичких ресурса у пројекту CESAR. Први индикатор је општи и стога се процена тог индикатора врши у складу са општим критеријумима типа да/не. Сви ресурси који се оцењују за одређени језик се набрајају у оквиру критеријума: „сви одабрани ресурси су најсавременији представници типа коме припадају". Следећа два индикатора, укупан број бодова појединачних ставки и беле књиге о језицима се заснивају на нумеричкој процени ресурса према претходно установљеним квалитативним и квантитативним критеријумима и конвенцијама за њихово мерење. Извор података коме се даје предност у нашој анализи су табеле за сваки језик појединачно, које се праве на основу оцена свих категорија унапред дефинисаних у белим књигама о језицима. Четврти индикатор је допунски - он није од највеће важности за сам одабир, али наговештава где треба усмерити напоре да би се елиминисали недостаци у одабиру.

2.3. Проблеми лиценцирања

Четири конзорцијума - укључујући и CESAR – који чине заједницу META-РАЗМЕНЕ су већ дефинисали принципе лиценцирања и факторе које треба узети у обзир у вези с тим. Основни циљеви су промовисање слободног и бесплатног коришћења података или алата када год је то могуће, да би се обезбедио неометан приступ ресурсима о којима је реч. Водећи се принципима Creative Commons и Open Data Commons, развијен је низ модела лиценцирања, који је назван „META-SHARE Commons". Осим ових шаблона, за неке ресурсе се користе изворна Creative Commons (CC) и GPL лиценца. Све лиценце којима се омогућава слободно коришћење, дозвољавају даљу дистрибуцију ресурса и алата. Лиценце могу имати следеће стандардизоване и добро дефинисане одредбе о законским правима коришћења (акроним наведен у загради популарише организација Creative Commons):

- клаузулу којом се налаже навођење имена изворног даваоца лиценце (BY)
- клаузулу којом се налаже некомерцијална употреба ресурса/алата (NC)
- клаузулу којом се налаже поновно депоновање свих прерада или производа под истим условима лиценцирања (зове се „ауторство - делити под истим условима" у организацији Creative Commons, SA)
- клаузулу којом се забрањује прерада дела (ND)

Ове додатне клаузуле се могу комбиновати, осим SA и ND, које се међусобно искључују. Наравно, поновно лиценцирање постојећег ресурса или алата често није изводљиво, нарочито ако су власници ауторских права исувише бројни или ако су на снази друга ограничења која се односе на податке. Тачније речено, OpenSource/OpenContent лиценце и лиценце које омогућавају слободан приступ често нису применљиве из већ постојећих разлога правне (углавном због заштите ауторских права), стратешке или друге природе. Стога је развијена нова група лиценци којом се обезбеђују стандардизовани шаблони за лиценцирање којима се забрањује било какво редистрибуирање без изричите дозволе изворног даваоца лиценце. Такве рестриктивне лиценце могу бити комерцијалне или некомерцијалне, издаване уз накнаду или без накнаде, а могу чак садржати и додатно ограничење да се „не дозвољава прерада дела".

Конзорцијум CESAR из сета до сада представљених шаблона, за сваки ресурс, бира лиценцу која је најпримеренија за дати ресурс. Конзорцијум је упоредио моделе лиценци са

прописима и праксом на националном нивоу, имајући у виду могућност постојања различитих режима, с обзиром на врсте власништва (приватни сектор са једне и јавни са друге стране), врсту садржаја (подаци са једне и софтвер са друге стране), или од раније актуелне аранжмане са власницима оригиналног садржаја који је прерађен при изради ресурса. Ресурси произашли из пројекта треба да буду усклађени са правним принципима и одредбама META-РАЗМЕНЕ које је комплетирао или изменио конзорцијум, а прихватили власници права на дати ресурс или алат.

Могућа је, мада не и обавезна варијанта која подразумева да су неки ресурси или алати вишеструко лиценцирани. То је углавном случај са OpenSource/OpenContent ресурсима, а обично је у питању група лиценци GPL, CC-BY-SA и GFDL.

3. Активности и циљеви

Прва транша ресурса је доспевала за објављивање 1. децембра 2011. године. Партнери у конзорцијуму CESAR су тог датума објавили 52 језичка ресурса или алата и учинили их доступним путем репозиторијума CESAR у оквиру META-РАЗМЕНЕ (као и у централним репозиторијумима META-РАЗМЕНЕ). Касније се планира синхронизација репозиторијума, пошто, за сада, то још увек није урађено. Због бројности ресурса и чињенице да су у наставку представљени заједно са ресурсима и алатима предвиђеним за наредне транше – другу и трећу – овде неће бити дата детаљна статистика из разлога мањка простора.[9]

3.1. Бугарска

Језички ресурси се развијају у многим истраживачким центрима у Бугарској – на Универзитету у Софији (на пример, говорни корпуси), Универзитету у Пловдиву (на пример, електронски речници), Новом бугарском универзитету (ресурси из домена преводилачких меморија), Југозападном универзитету (паралелни корпуси) и другим. Међу најистакнутијим су истраживачки институти Бугарске академије наука - Институт за математику и информатику, Институт за комуникационе и информационе технологије и Институт за бугарски језик. У Институту за бугарски језик се тренутно развија неколико важних ресурса и алата и о њима се извештава у оквиру пројекта CESAR. Овде ћемо поменути неке од најистакнутијих језичких ресурса, а комплетан списак се може погледати на веб локацијама CESAR-а и META-НЕТА.

Бугарски национални корпус (BulNC) је корпус фокусиран на бугарски језик који је доступан јавности и који се непрестано проширује (тренутно садржи 469,5 милиона токена). Осмишљен је као јединствени оквир за текстове настале у различитим медијима (писани – изговорени текстови), периодима (синхрони – дијахрони) и на различитом броју језика (једнојезични – паралелни, при чему је један од језика бугарски).

Тако било који језик X из бугарско-X паралелног корпуса има једнак третман (пошто је део јединственог оквира BulNC-а) по питању разноврсности и избалансираности типова текста, начина описа метаподатака, фаза које претходе обради и анотације, формата у коме се подаци чувају и упита у претраживачима. Тренутно се у паралелном корпусу налазе текстови на 33 језика. Међутим, они нису једнако заступљени: највећи паралелни корпус је бугарско-енглески (81,4 милиона токена за бугарски), 21 паралелни корпус је величине 30–52 милиона токена, следе корпуси величине од 1–10 милиона токена, а остали су мањи од једног милиона.

Два ручно анотирана корпуса су такође део

9 Додатне информације се могу наћи на адреси META-РАЗМЕНЕ, репозиторијума у чијем саставу се налази CESAR http://nlp.ipipan.waw.pl/metashare и у наставку овог документа.

Бугарског националног корпуса – BulPoSCop – тагиран врстама речи, који укључује 150 хиљада речи и BulSemCor – тагиран значењем речи, са приближно 100 хиљада речи. Свакој лексеми (простој или сложеној речи), која се јавља у одређеном контексту у BulSemCor-у, бива ручно додељено јединствено семантичко или граматичко значење из бугарског ворднета. Систем за претраживање BulNC-а[10] је осмишљен тако да подржава једнојезични бугарски корпус и паралелне корпусе на исти начин. Систем проналази елементе који одговарају упиту у свим документима, без обзира на ком су језику. Упитни језик подржава термине: „реч, граматичко својство и релација“ (дозвољени су упити за облике речи, синониме, хипериониме и сличне придеве). На вебу је доступан и сервис који омогућава увид у статистичке податке о колокацијама.

Бугарски wordnet (BulNet) је један од најкомплетнијих и најконзистентнијих лексичких ресурса за бугарски језик (поређења ради, број литерала у бугарском ворднету је много већи од листе речи у стандардном правописном речнику). Сетови синонима из BulNet-а су повезани са принстонским WordNet-ом 3.0 путем релација међујезичке еквиваленције, и стога је он део вишејезичне лексичко-семантичке мреже, такозваног глобалног ворднет-а. Бугарски ворднет је по величини једнак једној четвртини ворднет-а енглеског језика и један је од највећих у Европи. BulNet дистрибуира ELDA.

Обимни морфолошки речници бугарског које развија више центара постоје већ дуже време. Они омогућавају аутоматску анализу и синтетизовање облика речи, а тиме и прављење парадигме (састављене од свих могућих облика) дате речи, препознавање датог облика као дела парадигме и приписивање граматичких својстава. Морфолошки речник бугарског је одабран зато што је на располагању конзорцијуму, заснован је на најновијем издању правописног речника бугарског језика и користи се за развој софтвера за проверу правописа у текстовима на бугарском, а један од циљева је да се он дистрибуира у оквиру CESAR-а.

3.2. Хрватска

Једини партнер из Хрватске који учествује у пројекту CESAR је Филозофски факултет Универзитета у Загребу. Његов Институт и Катедра за лингвистику се баве израдом рачунарских корпуса од 1967. године, а факултет се сматра водећом националном институцијом у области корпусне и рачунарске лингвистике у Хрватској. Ресурси и алати у првој транши ресурса и алата за хрватски језик, у оквиру CESAR-а, укључују следеће ресурсе:

- *Хрватски национални корпус* (HNK) је репрезентативни корпус писаних текстова, на савременом хрватском стандардном језику, објављиваних од 1990. године. Корпус је аутоматски лематизован, а морфосинтаксички опис је тагиран уз помоћ хибридног система за лематизовање и тагирање, CroTag. Документи су анотирани према жанру, типу и другим информацијама. Читав корпус се састоји од публицистике, белетристике и текстова који су мешавина та два жанра. То је псеудокорпус, наиме, само је сумења за постављање упита, користећи клијент Bonito, доступна без икаквих ограничења, док се изворни текстови не могу дистрибуирати јер су ауторска права заштићена. Клијент Bonito омогућава постављање комплексних упита, захваљујући развијеном упитном језику, а као резултате претраживања даје не само конкорданце, већ и листе речи, колокације и друге врсте података о дистрибуцији, итд., токена, лема и морфосинтаксичких описа. Ти подаци се могу бесплатно преузети са Интернета и користити

10 http://search.dcl.bas.bg

за све врсте даље обраде.

- *Хрватски морфолошки лексикон* (HML) је флективни лексикон који аутоматски генерише Хрватски флективни генератор из око 110.000 лема, чиме се добија преко 4.000.000 облика речи. Он је резултат рада заснованог на теоријским предлошку објављеном 1992. године (видети Tadić 1994). Почетна група лема је сакупљена из неколико постојећих једнојезичних и двојезичних речника хрватског језика, док су додатне одреднице сакупљене путем корпуса или аутоматским проширивањем почетне листе лема (видети Bekavac, Šojat 2005, и Oliver, Tadić 2004). Исправљене су познате системске грешке у аутоматски генерисаном излазном резултату који је затим кодиран по систему кодирања UTF-8, а похрањен у формату MULTEXT-East Lexica, т.j. *lemma[TAB]word-form[TAB] MSD*. Скуп морфорсинтаксичких етикета је усклађен са препорукама MULTEXT-East v4.0 за хрватски језик. Постоје, међутим, неколики додаци: у презименима није наведен род (-), уведена је додатна поткласификација адвербијала, итд. Хрватски морфолошки лексикон је у овом тренутку псеудолексикон, доступан искључиво путем сумеће за постављање упита хрватском серверу за лематизацију на вебу или позивањем PHP скрипта.
- *Паралелни хрватско-енглески корпус* (Hr-En p-corp) је паралелни једносмерни (хрватско-енглески) корпус савременог хрватског стандардног језика, састављен од чланака објављиваних у новинама *Croatia Weeклу* од 1998. до 2000. године. Узорци корпуса су добијени у потпуности у дигиталном облику, конвертовани у XML, поравнати уз помоћ програма Vanilla Aligner, ручно проверени и похрањени у формату TMX. Корпус је могуће преузети путем платформе за дистрибуцију МЕТА-РАЗМЕНА.
- *Валенцијски лексикон хрватских глагола*, верзија 2.0008 (CROVALLEX 2.0008) представља покушај формалног описа валенцијских оквира хрватских глагола. CROVALLEX 2.0008 је развијен у оквиру докторске дисертације (Mikelić Preradović, 2008). Функционални генеративни опис који развија чешки лингвиста Петр Згал са сарадницима од шездесетих година XX века се користи као теорија на којој се заснива CROVALLEX 2.0008 за опис валенцијских оквира одабраних глагола. CROVALLEX 2.0008 садржи 1740 глагола. Они су одабрани из фреквенцијског речника хрватског језика, а критеријум је био да њихова апсолутна фреквенција буде изнад 10.

Алати:

- *Хрватски лематизацијски сервер* (HLS) је услуга доступна на вебу и омогућава лематизацију, етикетирање текстова на хрватском врстама речи и морфосинтаксичким описом. Подржава уношење података на два начина. Ако се подаци уносе путем формулара на вебу, сервер подржава директне упите и дозвољава да се уносе леме или облици речи, а резултат упита су сви облици унете леме или све леме којима облик речи може припадати. У оба случаја, уз резултате су дате и морфосинтаксичке етикете. Ако се подаци преносе на сервер, HLS очекује вертикализован текст, на савременом хрватском стандардном језику, на који је примењен систем кодирања UTF-8 и враћа датотеку у зип формату која садржи резултате обраде датотеке

постављене на сервер. У овом тренутку, постоји ограничење да величина датотеке не може прећи 50.000 токена. Обрадом се добија потпуна анализа сваког токена, т.ј. сваки редак у вертикализованом корпусу која се односи на лему, врста речи и морфосинтаксички опис. Сумеђа на вебу дозвољава корисницима да изаберу потребни ниво обраде: само лематизацију, лематизацију са етикетирањем врстама речи или лематизацију са етикетирањем морфосинтаксичким описом. Те две врсте етикетирања су у складу са спецификацијама MULTEXT-East v4.0 за хрватски језик. Након регистрације као корисник из академске заједнице или комерцијални корисник, може се обезбедити позивање PHP скрипт прилагођено захтевима корисника. Такође, постојећи хрватски лематизацијски сервер ће прерасти у услугу доступну на вебу, која ће омогућавати лематизацију и морфосинтаксичко етикетирање вертикализованих текстова на хрватском на које је

3.3. Мађарска

Мађарску у пројекту представљају два партнера, Институт за лингвистичка истраживања Мађарске академије наука (RILHAS) – који је такође координатор пројекта CESAR – и Катедра за телекомуникације и информатику медија Будимпештанског универзитета за технологију и економију (BME-TMIT). Профили те две институције се разликују, стога су ресурси и алати које су оне унеле у пројекат прилично комплементарни.

Институт за лингвистичка истраживања Мађарске академије наука је водећи институт специјализован за мађарску лингвистику (општу, теоријску и примењену лингвистику, лингвистику уралске групе језика и фонетику, као и за израду енциклопедијског речника мађарског језика). То је један од првих центара у Мађарској који је у свој програм укључио рад великих размера у области рачунарске лингвистике.

Одељење за језичке технологије основао је 1997. године (првобитно под именом Одељење за корпусну лингвистику) његов садашњи шеф Тамаш Варади. Одељење редовно учествује у пројектима из области језичких технологија, односно, из домена корпусне лингвистике, плитког парсинга, развоја онтологија, машинског превођења и изградње језичких ресурса. Стечено је значајно искуство у истраживачком раду и постигнути изузетни резултати, нарочито на плану развоја лингвистичких ресурса. Одељење је учествовало у неколико успешних међународних пројеката који су за циљ имали, са једне стране, прихватање извесних процеса развијених за западноевропске језике, који се данас сматрају делом стандарда за анализу мађарског (MULTEXT-East, Gramlex[11]) и, са друге стране, развој нових стандарда за стварање лингвистичких ресурса (база података електронских речника, CONCEDE[12], CLARIN), као и машинског превођења[13]. Истраживачи из овог одељења су стекли значајна знања о рачунарски заснованим системима за обраду језика и о технологијама развијеним или примењеним у овим пројектима. Такође имају активну улогу у њиховом прилагођавању потребама мађарског језика.

Седам ресурса и три алата понуђених у првој транши су углавном доступни, али не за широку публику, већ су више намењени за некомерцијално коришћење у академској заједници.

---

11 Lexiques grammaticaux et morphologiques (Граматички и морфолошки речници) (www-igm.univ-mlv.fr/~laporte/Copernicus/)

12 Consortium for Central European Dictionary Encoding (Конзорцијум за кодирање речника централноевропских језика (www.itri.brighton.ac.uk/projects/concede)

13 Internet Translators for all European Languages (Машински преводиоци за све европске језике доступни на Интернету (iTranslate4.eu)

Ресурси:

- *Сегедински корпус* – Морфосинтаксички анотиран корпус од 1,2 милиона речи (са базом података подељеном у шест различитих тематских целина, од којих свака има приближно 200 хиљада речи) у коме је ручно отклоњена вишезначност. 1,2 милиона речи обухвата 155.500 различитих облика речи и још 250 хиљада знакова интерпункције. Датотеке које сачињавају корпус су доступне у XML формату, а њихова унутрашња структура је описана TEIxLite DTD схемом (дефиниција типа документа).
- *Сегединска банка стабала* (treebank)– Ручно проверена банка стабала од 1,2 милиона речи. Одређивање означених синтагми и њихових односа помаже даљу језичку обраду, између осталог и семантичку анализу текстова. Урађено је интензивно маркирање синтаксичких структура у 82.000 реченица (1,2 милиона речи + 250 хиљада знакова интерпункције) садржаних у верзији 2.0 Сегединског корпуса. Датотеке банке стабала се чувају у XML формату, а њихова унутрашња структура је описана TEIxLite DTD схемом.
- *Сегедински корпус за препознавање именованих ентитета* – Ручно анотиран део Сегединске банке стабала који се састоји од кратких пословних вести. Користе се следеће категорије именованих ентитета (засноване на систему CoNLL): ЛИЦЕ, ОРГАНИЗАЦИЈА, ЛОКАЦИЈА и ОСТАЛО.
- *Мађарски WordNet* – Мађарски WordNet је вишејезична онтологија, што значи да је већина његових синсетова (скупова синонима) мапирана на еквивалентне појмове у енглеском (принстонском) WordNet-у в.2.0. Онтологија је такође повезана са одредницама у Једнојезичном речнику мађарског језика са објашњењима, као и са одредницама садржаним у лексикону валенцијских оквира мађарских глагола.
- *Мађарски корпус на вебу* – Овај корпус садржи преко 1,48 милијарди речи (од чега је 589 милиона морфолошки филтрирано), што га чини убедљиво највећим корпусом мађарског језика, а доступан је у целини захваљујући либералној лиценци Open Content. Мађарски корпус на вебу је настао у оквиру пројекта WordSword у Центру за медијска истраживања и едукацију и састоји се од 18 милиона страница преузетих са .xy домена.
- *Корпус Hunglish* – Паралелни корпус Hunglish је слободно доступни паралелни мађарско-енглески корпус, поравнат до нивоа реченице, који садржи око 2 милиона реченица. Корпус се може претраживати путем услуге претраживања којој се приступа на вебу. Страница на којој се налази ова услуга има преко 200.000 посета месечно.
- *morphdb.hu* – Мађарска лексичка база података и морфолошка граматика. morphdb.hu је представљена у облику формалног описа који захтева hunlex, офлајн компилатор ресурса који нуди лингвистички мотивисан језик за морфолошки опис и омогућава принципијелно и флексибилно одржавање и проширивање ресурса. Стога, morphdb.hu, уз помоћ h3unlex-a и hunmorph-a, обезбеђује главне језичке ресурсе за проверу правописа, стеминг, морфолошку анализу и извршавање бројних других задатака у области анотације.

Алати:

- *hunalign* – Моћан бесплатни алат за

поравнавање до нивоа реченице који се користи за изградњу паралелних корпуса. Овај алат је развијен у оквиру пројекта Hunglish са циљем изградње (мађарско-енглеског) корпуса Hunglish. hunalign поравнава двојезичне текстове на нивоу реченице. Он ради са текстовима на оба језика који су већ токенизовани и сегментирани на реченице. У најједноставнијем случају, резултат је низ реченичних парова на два језика (двојезичне реченице). Ако постоји речник, овај алат га користи и комбинује те информације са информацијама из Гејл-Черчовог алгоритма који се заснива на дужинама реченица. Ако не постоји речник, прво се ослања на информације о дужини реченица и гради аутоматски речник на основу тог поравнавања. Затим, поравнава текст по други пут користећи аутоматски речник. Као и већина алата за поравнавање, hunalign не узима у обзир промене реда реченица, и није у могућности да врши унакрсно поравнавање, тј. да поравнава сегменте А и Б у једном језику са одговарајућим сегментима Б' и А' у другом језику.

- *hunmorph* – Алат и програмска библиотека отвореног кода за проверу правописа, стеминг и морфолошку анализу, развијен углавном за аглутинативне језике. hunmorph се заснива на проширењу програмског кода MySpell-a, поновној примени познатог програма за проверу правописа, Ispell, чиме се добија генеричка библиотека за анализу речи. У овом тренутку, развој ове библиотеке је кренуо у два правца. Сада је проширена верзија MySpell-a, под називом HunSpell, део вишејезичног програмског пакета LibreOffice (а претходно и OpenOffice.org). hunmorph је програм прилагођен потребама морфолошке анализе. Систем hunmorph је састављен од три компоненте: а) ocamorph анализатор уклања афиксе, без обзира о ком језику је реч, б) morphdb.hu је лексичка база података и морфолошка граматика коју може користити ocamorph (детаљније о овоме на адреси: http://mokk.bme.hu/resources/morphdb.hu), ц) hunlex је офлајн компонента за управљање ресурсима која употпуњује учинак наше извршне компоненте (runtime layer) описним језиком високог нивоа и подесивим прекомпилатором.
- *huntoken* – Токенизатор заснован на правилима и детектор границе међу реченицама за текстове на мађарском (и енглеском) језику. Он ради са обичним текстуалним датотекама у којима су карактери кодирани коришћењем карактерских скупова ISO Latin-1 или Latin-2 а резултат је токенизована датотека у XML формату. Прецизност одређивања тачних граница речи и реченица износи 98%. Може се користити са оперативним системима Unix, Linux, Mac OSX као и Windows.

BME-TMIT је веома важан актер у области истраживања говорних технологија и апликација у том домену у Мађарској. Њихову групу за говорну технологију је 1969. године основао проф. Геза Гордош. Главна делатност групе је истраживање говорних технологија у областима препознавања говора, синтезе говора, као и говорних база података. У TMIT је развијено и осмишљено неколико корпуса и алата, укључујући опште и специјализоване корпусе за синтезу говора и препознавање говора или истраживање говорних технологија и неколико алата за обраду звука и текста (алати за синтезу и препознавање, алати за фонетско транскрибовање, принудно поравнавање,

прозодијски сегментатор, вежбе слуха, итд.)

Један сет ових ресурса и алата је понуђен за размену путем МЕТА-РАЗМЕНЕ. Наша општа упутства за лиценцирање налажу да се обезбеђује отворен и ако је могуће слободан приступ ресурсима и алатима у некомерцијалне и образовне сврхе. Ипак, неколико ресурса је већ доступно или ће бити доступно и у комерцијалне сврхе. Наш допринос укључује:

- **Говорне корпусе за потребе истраживања у области говорних технологија и говорне комуникације (укључујући и невербалну комуникацију):**

BABEL – Мађарски корпус разговетног говора је база података компатибилна са базом података EUROM1 и садржи снимке добијене од 60 говорника подељених у три групе.

База података о емоцијама – Говорни корпус који садржи исказе пуне емоција, у коме су обележене емоције (обележено је 8 основних емоција).

База података говорних гестова – лексикон звуковних гестова који садржи 770 токена.

Медицинска база података– Медицинска база података је говорни корпус сачињен од исказа људи који болују од различитих поремећаја говора органског порекла.

База података форманата из изговорених речи – говорни корпус мађарског језика, ручно анотиран. База података форманата је састављена од 3000 ставки, изговорених речи мађарског језика, користи се у истраживачке и образовне сврхе. То је референтна база података за информације о мађарским формантима.

- **Опште говорне корпусе за препознавање говора или синтезу говора:**

MRBA – Мађарски референтни говорни корпус садржи снимке говора – континуираног читања текста. У бази података се налазе искази 332 различита говорника који читају. Искази су снимљени на акустички различитим локацијама.

MTBA – Мађарски телефонски говорни корпус је база података о гласовној комуникацији говорника мађарског језика путем јавне комутационе мреже (PSTN) и мобилног телефона. База података садржи податке засноване на дефиницији из базе SpeechDatE о равнотежи између дијалеката, годишта и пола говорника и вокабулару. Оно што је важно и што ову базу података разликује од базе SpeechDatE је чињеница да су

MTÜBA – Мађарски говорни корпус телефонских клијената обухвата телефонске позиве снимљене у позивном центру једне компаније за пружање услуга. Садржи дијалоге између оператера и клијената. Дата је ортографска транскрипција говорних исказа, клаузе су сегментиране (аутоматски, након чега су уследиле ручне корекције).

Мађарски MALACH – Мађарски говорни корпус састављен од сведочанстава преживелих жртава холокауста, тј. животне приче старијих људи о Другом светском рату, углавном оних који су преживели холокауст.

Говори из мађарског парламента – корпус доступан јавности у коме се налазе приближно транскрибовани говори из парламента.

База двогласа за конверзију текста у говор (TTC) – ручно анотиран говорни корпус садржи 5500 записа говора на мађарском, немачком и шпанском. Звучни записи двогласа засновани на логатомима се користи за конверзију текста у говор. Гласови за мађарски: мушки и женски, за немачки: мушки, за шпански; мушки.

Говорна база на нивоу речи – ручно анотиран говорни корпус за мађарски језик, садржи два сата говора. То је списак речи који се чита ради представљања структуре говора на мађарском језику на сегменталном нивоу (групе CV, VC, VV, VVV, CC, CCC, CCCC). Гласови: мушки и женски.

Говорна база прочитаног текста за конверзију

текста у говор (TTC) – говорни корпус за мађарски језик, полуаутоматски анотиран, садржи 20 сати говора. Састоји се од реченица прочитаних ради конверзије текста у говор одабиром јединица и за потребе система за конверзију текста у говор који се заснивају на скривеном Марковљевом моделу (SMM). Гласови: мушки и женски.

Лексичка база именованих ентитета – лексикон транскрибован за потребе конверзије текста у говор, садржи збирку имена (особа, правних лица, места) и њихову транскрипцију у говор. Користи се за читач мађарских имена и адреса који претвара појмове у говор.

База изговорених бројева за конверзију текста у говор (TTC) – говорни лексикон за мађарски, немачки и енглески (3 сата). Садржи елементе бројева за висококвалитетне системе за читање бројева, времена и датума у (на пример, финансијским) информационим системима. Гласови на мађарском: мушки и женски, на немачком: мушки, на енглеском: мушки.

Мађарски речник изговора – вокабулар изговора за облике речи мађарског језика. То је једина референтна база података о изговору мађарског.

- **Мултимедијални корпуси:**

База вести емитованих на радију и телевизији – Овај корпус је направљен за 10 европских језика у оквиру Интересне групе за вести емитоване у електронским медијима, односно, активности COST278 (Žibert et al., 2005). Материјал на мађарском је састављен од снимака у трајању од 3 сата и 30 минута који су транскрибовани и анотирани у складу са конвенцијама NIST-a (National Institute of Standards and Technology - Национални институт за стандарде и технологију, САД).

База емитованих предавања – База емитованих предавања садржи видео снимке емитованих предавања на разне научне теме намењених широј публици.

3.4. Пољска

Пољске ресурсе и алате у пројекту CESAR дају два партнера из те земље: Институт за рачунарство, Пољске академије наука и Универзитет у Лођу.

Институт за рачунарство Пољске академије наука (IPIPAN – http://www.ipipan.eu/) је водећи национални центар за фундаментална истраживања у области рачунарских наука, као и за примењена истраживања у домену вештачке интелигенције и информационих система. Међу алатима и решењима које је развила Група за језичко инжењерство[14] IPIPAN-а су тагери, плитки и дубоки парсери, као и различити алати за екстракцију информација засновани на машинском учењу и на правилима. IPIPAN је такође развио велике, лингвистички анотиране корпусе, који се примењују на бројне начине у области екстракције информација и ископавања из текстова (Text Mining) и у општим и у специјализованим доменима.

Универзитет у Лођу је још један велики истраживачки и образовни центар у Пољској. Истраживачка група PELCRA[15] Катедре за лингвистику овог универзитета је већ дуго активна у домену пољских језичких ресурса. Последњих година, њене активности су усмерене на прикупљање података за корпусе, укључујући и мултимедијалне говорне податке и развој језичких алата и услуга који се примењују у истраживањима и технологији. Чланови групе су доказано искусни у области развоја система за обраду језика, за текстове како општег, тако и специјализованог типа на пољском и енглеском језику.

Прва транша је садржала 10 пољских ресурса и алата:

14 http://zil.ipipan.waw.pl/

15 http://pelcra.ia.uni.lodz.pl/

- *PoliMorf речник флексије* (прелиминарна верзија 0.5) – нови морфолошки речник пољског језика који је плод стандардизације, спајања и аутоматизованог исправљања грешака у два најважнија морфолошка речника пољског – Morfeusz-y SGJP и Morfologik-y који садржи 6,8 милиона облика речи са флективним наставцима и њихове морфолошке описе. У следеће верзије овог ресурса ће бити унети ручно исправљени подаци. Први припојени ресурс, Morfeusz SGJP се заснива на SGJP -у, граматичком речнику пољског који је резултат дугогодишњег рада неформалне групе на чијем челу је проф. Зигмунд Салони. Рад је започет осамдесетих година, дигитализацијом листе одредница у једанаестотомном речнику пољског језика Дорошевског (1958–1969). Граматички опис у SGJP -у се базира на новим концептима предложеним у другој половини XX века, а многа детаљна решења су предложили чланови тима (Токарски, Грушчињски, Салони). За PoliMorf су коришћени подаци из другог издања SGJP-а. Лексемама којих има 244.341 одговара 4.223.981 облик речи (при чему се синкретички облици исте лексеме броје као једна јединица). Флексија је у SGJP -у представљена флективним парадигмама, којима су облици описани уз помоћ основе заједничке свим облицима и наставака на бази којих се облици међусобно разликују. Други припојени ресурс, Morfologik, је још један морфолошки речник пољског отвореног кода. Садржи 216.992 лексема и 3.475.809 облика речи. Овај речник је настао обогаћивањем ispell/hunspell речника пољског морфолошким информацијама. Нажалост, структура изворног речника није била довољна да би било могуће поуздано утврђивање појединих информација, попут прецизне поткласификације мушког рода именица. Те информације су додате ручно уз коришћење хеуристичких метода. Као инспирација за скуп етикета овог речника је послужио скуп етикета IPI PAN. Морфологик се, међутим, разликује од тог скупа и Morfeusz-a по томе што никада не дели ортографске речи („ниске од белина до белине“) на краће речи из речника, т.ј. не узима се у обзир такозвана аглутинација. Осим тога, из разлога непостојања информација у речнику испелл, неки облици речи нису у потпуности анотирани и означени су као неправилни. Међутим, додата су још нека обележја за повратне глаголе која не постоје у изворном скупу етикета IPI PAN. То је уведено због алата за проверу граматичке исправности, LanguageTool који је веома користио поменути речник.
- *Поткорпус Националног корпуса пољског од милион речи* – ручно анотиран узорак Националног корпуса пољског од 1,5 милијарде речи (Narodowy Korpus Języka Polskiego, NKJP)[16] је плод заједничке иницијативе четири институције: Института за рачунарство Пољске академије наука (координатор), Института за пољски језик Пољске академије наука, Пољских издавача научних публикација PWN и Катедре за рачунарску и корпусну лингвистику Универзитета у Лођу. Регистрован је као истраживачко-развојни пројекат Министарства науке и високог образовања. На списку извора из којих потичу подаци у корпусу су класична књижевна дела, дневне

16 http://nkjp.pl/

новине, специјализоване периодичне публикације и часописи, транскрипти разговора и разни ефемерни текстови и текстови са Интернета. Ресурсе карактерише велика разноврсност тема и жанрова. У говорном делу корпуса су заступљени говорници оба пола, различитих годишта из различитих делова Пољске.

- *Корпус пољског Сејма* (верзија 1.0) – збирка анотираних исказа посланика пољског Сејма из мандата 1-6 (од 1991 до 2011. године). Подаци у корпусу су доступни у једној варијанти формата TEI P5[17] која је примењена у NKJP-у, а садрже информације о сегментирању текста (на пасусе, реченице, токене), морфосинтаксичким описима у којима је отклоњена вишезначност (лемама, етикетама врстама речи, етикетама морфосинтаксичким описом), синтаксичком опису (синтаксичким речима и групама) и именованим ентитетима (личним именима, локацијама организацијама). Ти подаци су драгоцен извор лингвистичких информација, с обзиром да је у питању велика збирка квази-говорног материјала (114 милиона сегмената) која чини основу аудио и видео снимака седница, што је започето 2011. године. Планира се да се ти снимци поступно додају корпусу. Следеће верзије корпуса ће бити проширене и у њих ће бити укључени сродни садржаји – посланичка питања и транскрипти састанака одбора Сејма.
- *Пољски WordNet* (Słowosieć, верзија 1.5) – мрежа лексичко-семантичких релација, електронски тезаурус развијен на Технолошком универзитету у Вроцлаву, структурисан по моделу принстонског WordNet-а, као и оних који су изграђени у оквиру пројекта EuroWordNet. Пољски WordNet описује значење лексичке јединице постављајући је у мрежу семантичких релација попут хиперонимије, меронимије, антонимије, итд. Да би се смањили трошкови пројекта, пољски WordNet је полуаутоматски. Лексичке релације су аутоматски препознаване у великим корпусима пољског, да би потом биле понуђене лексикографима преко графичке сумеће. Данас је пољски WordNet један од највећих на свету. Обухвата 103.000 лексичких јединица у 74.000 синсетова.
- *Пољски алат за препознавање именованих ентитета* (NERF, verzija 0.2) – статистички алат за препознавање именованих ентитета, заснован на методу моделирања уз помоћ условних случајних поља (CRF - Conditional Random Fields). Овај алат је настао у оквиру пројекта Национални корпус пољског језика. Модификован је тако да може да препознаје дрволике структуре именованих ентитета (т.ј. структуре са рекурзивно уграђеним именованим ентитетима) применом метода етикетирања здруженим обележјима. (JLT - Joined Label Tagging). Ово је једноставан метод кодирања структура именованих ентитета у облику низа етикета. Овај метод омогућава кодирање разних додатних информација категоријалне природе о именованим ентитетима - тип, подтип, врста деривације – на нивоу етикета, а затим и њихово препознавање захваљујући методу CRF који из тога произилази. Овај алат се може конфигурисати тако да користи различите врсте запажања током процеса обуке и препознавања, на

17 http://nlp.ipipan.waw.pl/TEI4NKJP/

пример: лексичке информације са текстуалног нивоа, или граматичке информације са морфосинтаксичког нивоа.

- *Пољски ресурси именованих ентитета* (прелиминарно издање) – група ресурса који се односе на именоване ентитете, допуњени додатним ресурсима специфичним за пољски језик, добијеним са веба. Флективни облици су генерисани, уз помоћ генератора Morfeusz SGJP, када год је то било примерено. Подаци из Пољских ресурса именованих ентитета су коришћени у поступку анотације именованих ентитета у Националном корпусу пољског језика (NKJP). Подаци из ресурса су описани уз примену комплексне хијерархије имена: личних имена и презимена, имена градова, држава, планина, области, река, институција, односних придева и назива за становништво, изведених из имена држава, окидача именованих ентитета (месеца, дана, положаја, итд.). Тренутно садржи 153 хиљаде одредница.
- *Корпус LUNA.PL* садржи 500 изговорених дијалога који се одвијају између људи на пољском језику (13.000 исказа). Корпус је анотиран на неколико нивоа, од транскрипције дијалога и њихове морфосинтаксичке анализе, до семантичке анотације концепата, предиката и анафоре. Анотација на морфосинтаксичком и семантичком нивоу је урађена аутоматски, а затим ручно исправљена. На концептуалном нивоу, анотацијом је обухваћено око 200 појмова из онтологије направљене специјално за потребе пројекта. Скуп оквира за анотацију на нивоу предиката је дефинисан као ресурс типа FrameNet.
- *Корпус LUNA-WOZ.PL* садржи 69 изговорених дијалога који се одвијају између људи и рачунара на пољском језику (5,5 хиљада исказа). Корпус је анотиран на неколико нивоа, од транскрипције дијалога и њихове морфосинтаксичке анализе, до семантичке анотације концепата.
- Највећу збирку транскрибованих спонтаних разговора на пољском прикупља тим PELCRA на Универзитету у Лођу од 2000. године, прво у оквиру референтног корпуса PELCRA-е, а затим Националног корпуса пољског језика. Укупно, корпус има скоро 2 милиона речи из транскрипција конверзација снимљених у неформалној атмосфери, а често се дешавало да неки од говорника нису знали да су снимани (мада су били обавештени и сагласили су се са могућношћу да буду снимани, а потом дозволили да се снимци транскрибују). До сада су ови подаци били доступни само путем онлајн сумеђе, али је у оквиру пројекта CESAR, један део тих података, укупно 1,8 милиона речи, постао доступан у формату TEI P5, пошто су се поставила нека питање из домена приватности. Осим тога, део транскрипција (најмање 200.000 речи) је одабран и поравнат према времену са изворним снимцима на нивоу исказа и постао је доступан, у складу са лиценцом типа CC, путем репозиторијума META-РАЗМЕНА као мултимедијални временски анотиран корпус конверзација изговорених на пољском језику. Такви ресурси су без сумње веома драгоцен извор података о језику који се могу применити за моделирање конверзација на пољском и за препознавање говора у посебним ситуацијама.
- Иако за пољски језик постоји више паралелних ресурса којима јавност има слободан приступ, односно, ресурса доступних у складу са отво-

реном лиценцом, у вези са њима се јављају проблеми који се у значајној мери негативно одражавају на њихову употребљивост и међусобну укопивост. Једна од дужности Универзитета у Лођу, као дела пољског огранка пројекта CESAR је да се компилирају сви пољски паралелни корпуси у неколико стандардних формата, што ће их учинити лако доступним за обраду природних језика и рачунарски потпомогнуто превођење. До сада су четири групе паралелних корпуса објављене у форматима TEI 5 i XliFF под лиценцом Creative Commons. Неки од тих корпуса су и пре били доступни, али нису имали библиографску анотацију (Acquis Communautaire), други представљају нове ресурсе састављене од ресурса доступних без накнаде (CORDIS, RAPID), али постоје и паралелни текстови добијени директно од издавача и ручно поравнати на нивоу реченице (часопис *Academia*). За тај ресурс је било потребно постићи посебне договоре са издавачима око ауторских права, а то је добар пример утицаја пројекта CESAR у смислу разјашњавања статуса језичких ресурса и алата, по питању права интелектуалне својине, да би били доступни за коришћење.

Осим обезбеђивања језичких ресурса и алата, партнери из Пољске који раде на пројекту CESAR су успели да додатно развију свест о постојећим језичким ресурсима и алатима као и да учврсте везе између главних актера обраде природних језика у земљи кроз оснивање новог веб портала „Рачунарска лингвистика у Пољској“ (CLIP, http://clip.ipipan.waw.pl) у априлу 2011. године. На овој веб локацији се налазе исцрпне информације о језичким ресурсима и алатима, истраживачким центрима, пројектима и курсевима из области језичког инжењерства који се односе на пољски. Поред тога, веб локација је настала са намером да повеже иницијативе које се баве језиком, истраживачке институције, истраживаче, представнике власти и индустрије, нудећи им свеобухватне информације о доступним језичким технологијама. Један од основних принципа примењених при осмишљавању веб локације је функционисање уз помоћ wики-ја, чиме је представницима свих група језичких ресурса и алата у Пољској, који за то имају дозволу, омогућено да директно уређују садржаје. Такав приступ се показао врло плодним, а спољашњи уредници су већ унели неколико модификација и додатака постојећом садржајима. Колико је нама познато, ова веб локација је тренутно највећи репозиторијум референци на пољске језичке ресурсе и алате доступне јавности.

3.5. Србија

Група за језичке технологије на Катедри за рачунарство и информатику Математичког факултета, Универзитета у Београду је основана пре више од 30 година, 1978. године, са главним циљем да се развије формални опис српског и стварају и користе ресурси и алати за тај језик. Језгро Групе за језичке технологије сада чине истраживачи са неколико факултета Универзитета у Београду, а група је блиско повезана са већином институција у Србији које се баве језичким технологијама.

Током своје дуге историје, Група за језичке технологије је развила значајан број језичких ресурса и алата, међу којима су најважнији:

- Електронски речници, као српски WordNet и морфолошки речници општих лексикона и личних имена, простих и вишечланих облика;
- Корпуси који садрже једнојезичне, двојезичне и вишејезичне текстове који су општег типа или усмерени на одређене области;

- Различити алати за развој или коришћење тих ресурса, од којих је најразноврснији *LeXimir*, намењен развоју лексичких и текстуалних ресурса и његов пандан на вебу *VebRanka* за проширивање упита приликом претраге.

Неки од ресурса од важности за српски језик се такође развијају у другим деловима српске академске заједнице. Значајна истраживања у области технологије говорне интеракције се врше на Технолошком факултету у Новом Саду и у компанији AlfaNum која наставља те активности. Они су осим говорних база података развили и лексичку базу података са више од 4.000.000 акцентованих облика речи српског језика. На основу тих ресурса су развијене разне комерцијалне апликације за конверзију текста у говор и аутоматско препознавање говора. Неки врло значајни мултимедијални ресурси се развијају на Институту за балканске студије Српске академије наука и уметности у Београду. У тој бази података се налази дигитализован аудио, видео, фото и текстуални етнографски материјал сакупљен у Србији у оквиру истраживања на терену.

Улога Групе за језичке технологије у пројекту CESAR је да открива све ресурсе и алате који се праве за српски језик и да их учини видљивим, не само у истраживачкој заједници, већ и за све оне који развијају и користе језичке технологије. За прву траншу ресурса који се испоручују МЕТА-РАЗМЕНИ, група је одлучила да припреми сопствене ресурсе, који су не само најзрелији, већ су и највише коришћени или за њих постоји највеће интересовање разних корисника. Одабрани су следећи ресурси:

- *Корпус савременог српског језика (SrpKor)* – састоји се од 4.523 текста, укупне дужине преко 113 милиона речи[18]. Лематизација и етикетирање врстом речи су извршени уз помоћ алата TreeTagger. Текстови који су ушли у SrpKor укључују белетристику чији аутори су српски писци XX и XXI века, научне чланке из различитих области (како друштвених, тако и природних наука), правна акта и текстове опште природе (новине, часописе, магазине, фељтоне). Доступан је путем сумеђе на вебу од 2003. године и редовно га користи преко 300 корисника, углавном слависта из целог света.
- *Српски WordNet (SrpWN)* – представља хијерархијску лексичко-семантичку мрежу која садржи синсетове са глосама и различитим семантичким релацијама, као што су антонимија, меронимија, каузација, доменске категорије, итд. Првобитна верзија српског WordNet-a је настала у оквиру пројекта Balkanet који је финансирала ЕУ.
- *Српски лематизовани корпус анотиран врстама речи (SrpLemKor)* – садржи узорак разних текстова из SrpKor-a, лематизованих и етикетираних врстама речи уз помоћ TreeTagger-a. Састоји се од вести објављених у дневним новинама „Политика”, извесног броја фељтона објављених у новинама „Политика” и „Данас”, белетристике чији аутори су српски писци из XX века, као и разних научних текстова из различитих области и текстова из домена права. За разлику од SrpKor-a, где је приступање корпусу и његово претраживање могуће само путем веб сумеђе, овај корпус је могуће преузети са интернета.
- *Француско-српски поравнати корпус (SrpFranKor)* – обухвата књижевне и текстове из новина из француских или српских извора и њихове преводе. Поравнати су на подреченичном нивоу. Корпус садржи 25 књижевних текстова

18 korpus.matf.bg.ac.rs

и 15 новинских чланака. Приступ корпусу и његово претраживање су омогућени путем сумеђе на вебу. Детаљније о овој теми пише (Утвић, 2011).

- *Вишејезично издање Вернового романа „Пут око света за 80 дана“ (Verne80days)* – садржи 17 издања романа Жила Верна „Пут око света за 80 дана“, изворни текст на француском и 16 превода. Преводи су поравнати са француском, енглеском или српском верзијом, пребачени у формат TMX, а ту верзију је могуће преузети са интернета.
- *Организовање дигитализованог материјала (InfoBeaver)* – апликација за прикупљање и представљање мултимедијалних информација. Подржава мултимедијалне документе и омогућава претраживање базе података на основу различитих критеријума. Демо верзија илуструје могућности апликације уз помоћ неколико података о пројекту CESAR и његовим учесницима.

Сви ресурси су доступни под CC BY-NC лиценцом док се на алат примењује лиценца GPL. Драго нам што можемо да кажемо да су неке ресурсе, који су учињени доступним у првој транши, већ преузимали, односно, да су њима приступали заинтересовани ван Србије.

3.6. Словачка

Обрада природних језика у Словачкој, нарочито што се тиче ресурса, се углавном одвија на Одељењу за Словачки национални корпус Института за лингвистику Људовит Штур. У институту је сакупљен највећи и најзначајнији извор података о савременом писаном словачком језику – база података, Словачки национални корпус. Политика одељења је од његовог оснивања 2003. године увек била да се ресурси и алати чине доступним стручној и најширој јавности путем либералних лиценци Open Source и Open Content, ако је могуће у оквирима постојећих ограничења (која се углавном тичу ауторских права) и да се промовише значај обраде природних језика и дигиталних ресурса у општој и примењеној лингвистици и у другим областима повезаним са лингвистиком. Та политика се наставља у пројекту CESAR, а нарочито се подношење ресурса у мрежу МЕТА-РАЗМЕНА смишљено користи да се разјасне постојећи уговори којима су регулисана ауторска права, као и да се власници ауторских права

Највидљивији ресурс у домену обраде природних језика је Словачки национални корпус, свеобухватни, репрезентативни корпус савременог писаног словачког језика (Garabík 2010). Корпус је пројекат који је у току, а у тренутку када ово пишемо садржи преко 770 милиона токена, са аутоматском морфосинтаксичком анализом и лематизацијом. Корпус је у ствари први велики фокусирани истраживачки пројекат у области обраде природних језика у Словачкој и означио је нову еру у словачкој рачунарској лингвистици. Из тог разлога су многи ресурси донекле фокусирани на корпус - претходних година, у многим пословима се морало кренути од нуле, пошто је мали број ресурса и алата за обраду природних језика за словачки био употребљив.

За МЕТА-РАЗМЕНУ смо одабрали ресурсе који су добро развијени, репрезентативни и корисни (или чак неопходни) за истраживања и апликације у домену обраде језика. Први и најважнији је верзија 5.0 Словачког националног корпуса. Нажалост, због ограничења наметнутих због заштите ауторских права (текстови у корпусу су обухваћени постојећим уговорима о лиценцирању постигнутим са власницима ауторских права, али се ти уговори односе само на њихово укључивање у корпус, а приступ је омогућен само путем сумеђе за претраживање), корпус не може бити даље дистрибуиран. Зато је доступан само као псеудокорпус, путем специјализованог

клијента Tcl/Tk (Bonito) или путем сумеђе на вебу (Bonito2).

Овај корпус употпуњава и Говорни корпус словачког језика (Šimkova et al. 2008), верзија 3.0, који је састављен од 300 сати ручно транскрибованих снимака стандардног словачког говорног језика, односно 1,6 милиона токена. За разлику од корпуса писаног језика, овај корпус садржи податке које је институт углавном наменски снимао. Пошто обична нарација не подлеже обавези заштите ауторских права (осим ако не садржи уметничке елементе, на пример, рецитовање поезије), а сви говорници су се сагласили да њихови снимци буду унети у базу података, а корпус је тростуко лиценциран. У питању је комбинација лиценци Affero GPL, Creative Commons „ауторство - делити под истим условима“, и GNU Free Documentation Licence, лиценце која омогућава слободан приступ документацији.

База облика речи, одговарајућих лема и морфосинтаксичких етикета (Garabík 2005) је фундаментални ресурс неопходан за било какву обраду словачког (језика богате морфологије). База података садржи 77 хиљада лема, што даје 2,5 милиона облика речи.

Паралелни корпуси су представљени двема најразвијенијим базама података. Прва је Словачко-чешки паралелни корпус који садржи 1,5 милиона поравнатих реченичних парова. Корпус се састоји од три врсте превода (углавном белетристике, али и текстова друге врсте): превода са чешког на словачки, са словачког на чешки и са трећег језика и на чешки и на словачки. Други велики паралелни корпус је Словачко-енглески корпус од 700 хиљада поравнатих реченичних парова. У том корпусу се готово искључиво налази белетристика написана на енглеском језику и преведена на словачки, али садржи и известан број текстова који спадају у публицистику.

Плановима за будућност је обухваћено неколико других добро развијених ресурса (који су развијени у Института за лингвистику Људовит Штур или другде). Ти ресурси су банка стабала за словачки језик, словачки WordNet, други паралелни корпуси, речник словачких колокација и други. CESAR и META-РАЗМЕНА представљају добру прилику да се промовишу идеје отворених лиценци за заштиту ауторских права, да се разјасни правни статус старијих ресурса и да се они ускладе са савременим стандардима, као и да се нагласи значај језичких технологија у Словачкој.

4. Закључци

Један од кључних циљева пројекта CESAR је изградња отворене инфраструктуре језичких технологија, што ће омогућити да достигнућа земаља учесница у области језичких технологија буду глобално доступна путем META-РАЗМЕНЕ, отворене платформе за размену ресурса. Језички ресурси и алати су, у земљама које учествују у пројекту CESAR, развијани засебно и притом се није водило рачуна о било каквој заједничкој инфраструктури или међусобној усклађености. Пројекат CESAR је путем репозиторијума META-РАЗМЕНА учинио да најрелевантнији и најразвијенији подаци за све главне европске језике буду доступни и представља покушај да се они представе кроз јединствену сумеђу и инфраструктуру. Стога, пројекат CESAR представља значајан корак ка отварању језичких ресурса и омогућавању да они буду доступни на стандардизован и потпуно документован начин, чиме би се оснажио Европски истраживачки простор.

Литература

Božo Bekavac and Krešimir Šojat. 2005. Lexical acquisition through particular adjectival endings for Croatian. Proceedings of the Workshop on Computational Modeling of Lexical Acquisition, University of Split, Split, 2005.

Broeder, Daan, Thierry Declerck, Erhard W. Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari and Peter Wittenburg. 2008. *Foundation of a Component-based Flexible Registry for Language Resources and Technology*, In Calzolari, N. (Ed.) *Proceedings of the 6th International Conference of Language Resources and Evaluation*, Ed. Nicoletta Calzolari, 1433-1436. European Language Resources Association (ELRA).

Blagoeva, Diana, Svetla Koeva and Vladko Murdarov. 2011. *Languages in the European Information Society – Bulgarian*, META-NET White Paper Series, Berlin.

Garabík, Radovan, Svetla Koeva, Maciej Ogrodniczuk, Marko Tadić, Tamas Váradi and Duško Vitas. 2011. Detecting Gaps in Language Resources and Tools in the Project CESAR, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Ed. Zygmunt Vetulani, 37-41, , Poznań: Fundacja Uniwersytetu im. A. Mickiewicza.

Garabík, Radovan. 2005. Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. *Proceedings of the conference Slovko 2005*, Ed. Radovan Garabík. Bratislava: Veda.

Garabík, Radovan. 2010. Slovak National Corpus tools and resources. *In: Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2010)*. Eds. Laclavík, M., Hluchý, L., 2 – 7, Bratislava.

Mikelić Preradović, Nives. 2010. *Approaches to the Development of the Machine Lexicon for Croatian Language*, PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences,.

Miłkowski, M. 2011. Languages in the European Information Society – Polish, META-NET White Paper Series, Berlin.

Oliver, Antonije and Marko Tadić. 2004. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. IV, 1259-1262. Genoa-Paris: European Language Resources Association (ELRA).

Simon, Eszter and Lendvai, P. 2011. *Languages in the European Information Society – Hungarian*, META-NET White Paper Series, Berlin.

Šimková, Mária, Radovan Garabík, Agáta Karčová, Katarína Gajdošová, Michal Laclavík, Jozef Juhár, Karol Furdík, Peter Ďurčo, Helena Ivoríková, Jozef Ivanecký and Július Zimmermann. 2011. *Languages in the European Information Society – Slovak*, META-NET White Paper Series, Berlin.

Šimková, Mária, Radovan Garabík, Agáta Karčová and Katarína Gajdošová: Hovorený korpus slovenčiny. 2008. In: *Čeština v mluveném korpusu*. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu, 227–233.

Tadić, Marko, Dunja Brozović-Rončević and Amir Kapetanović. 2011. *Languages in the European Information Society* – Croatian, META-NET White Paper Series, Berlin.

Tadić, Marko. 1994. *Računalna obradba morfologije hrvatskoga književnoga jezika*. PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences.

Утвић, Милош. «Анотација корпуса савременог српског језика.» *INFOteka* 12, 2: (2011) 39-51.

Varadi, Tamas. 2011. Introducing the CESAR project. *Infotheca* 12(1), 71-74.

Vitas, Duško., Ljuba Popović, Cvetana Krstev, Mladen Stanojević and Ivan Obradović. 2011. *Languages in the European Information Society – Serbian*, META-NET White Paper Series, Berlin.

Žibert, J., et al. 2005. The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results. In: *Eurospeech 2005: 9th European Conference on Speech Communication and Technology*. Lisboa, Portugal, 629-632.