Procesiranje slovenskega jezika v razvojnem okolju NooJ

Kaja Dobrovoljc

Trojina, zavod za uporabno slovenistiko
Dunajska cesta 116, 1000 Ljubljana
kaja.dobrovoljc@trojina.si

Povzetek

Z razvojem področja gradnje in označevanja obsežnih računalniških zbirk avtentičnih besedil se tudi v slovenskem jezikoslovju povečuje število raziskav, ki temeljijo na njihovem preučevanju. Kljub vse večjemu uveljavljanju korpusnih metod pa trenutno v slovenističnem prostoru obstaja precejšen razkorak med mnogimi raziskovalnimi priložnostmi, ki jih obstoječi korpusi ponujajo, ter njihovo dejansko izrabo. To stanje je v določeni meri tudi posledica pomanjkanja enostavnih orodij za kompleksnejšo obdelavo korpusnih besedil. Kot primer računalniško zmogljivega, a jezikoslovnemu uporabniku prijaznega orodja v prispevku predstavljamo NooJ, jezikoslovno razvojno okolje za izdelavo obsežnih formaliziranih opisov naravnih jezikov in njihovo uporabo v besedilnih korpusih. Na primeru izbranih jezikovnih virov in pravil iz pilotnega modula za slovenščino predstavimo najpomembnejše funkcionalnosti tega razvojnega okolja in prednosti njegovega povezovanja z že obstoječimi viri in orodji za slovenščino.

Slovene Language Processing with NooJ

With continued development in the field of corpus compilation and annotation, there is also an increase in the quantity of linguistic research dealing with their analysis. However, despite the growing recognition of corpus-based methods in Slovene language studies, there exists a considerable gap between many research opportunities arising from the available corpora and their actual realization. To some extent, this discrepancy can be attributed to the lack of user-friendly tools for complex corpus processing. As an example of such computationally powerful and yet linguist-friendly tool, we present NooJ, a linguistic development environment for construction of large-coverage formalized descriptions of natural languages and their application to corpora. By drawing on the examples from the initial module for Slovene, we present some of its most important features, and the prospects of its integration with other existing resources and tools for Slovene language processing.

1. Uvod

Z razvojem področja gradnje in obdelave obsežnih računalniških besedilnih zbirk se povečuje tudi število jezikoslovnih raziskav, ki temeljijo na njihovem preučevanju. Tudi v slovenskem jezikoslovju korpusni pristop postaja vse bolj uveljavljena metoda raziskovanja, toda zdi se, da so raziskovalci pogosto ujeti med načelno zavezo k raziskovanju jezikovne rabe v obsežnih zbirkah avtentičnih besedil na eni strani in nezmožnostjo kompleksne računalniške obdelave, ki jo taka analiza zahteva, na drugi. Posledično se raziskovalci pri načrtovanju metodologije pogosto omejujejo na preučevanje manjših (pod)korpusov, ki omogočajo obvladljivejšo, ročno analizo, ali pa predmet svojega raziskovanja prilagajajo omejenim funkcionalnostim spletnih korpusnih vmesnikov.

Naprednejši prostodostopni spletni konkordančniki, kot sta NoSketchEngine[1] in CUWI[2] (Erjavec, 2013), sicer omogočajo oblikovanje zapletenejših korpusnih poizvedb v obliki regularnih izrazov, a ti pri opisovanju kompleksnejših struktur hitro postanejo težko obvladljivi, prednastavljeni pristopi k analizi izluščenih konkordanc pa se v omenjenih orodjih osredotočajo predvsem na distribucijsko analizo in merjenje besedne povezovalnosti. V slovenskem prostoru tako obstaja realna potreba po orodju, ki bi tudi raziskovalcem brez znanja programiranja omogočilo enostavno opisovanje, luščenje, označevanje in urejanje kompleksnih jezikovnih pojavov na različnih ravneh.

Kot primer takega računalniško zmogljivega, a jezikoslovnemu uporabniku prijaznega orodja v prispevku predstavljamo razvojno okolje NooJ, za katerega je bil pred kratkim razvit tudi pilotni modul slovenski jezik (Dobrovoljc, 2014a).

2. Programska oprema

NooJ je jezikoslovno razvojno okolje za izdelavo obsežnih formaliziranih opisov naravnih jezikov in njihovo uporabo v besedilnih korpusih. Opisi naravnih jezikov so formalizirani v obliki elektronskih slovarjev in na grafu temelječih slovnic (pravil), s katerimi lahko na razmeroma preprost način opisujemo jezikovne pojave na različnih ravneh površinske zgradbe besedil, od besednih oblik do zapletenejših skladenjskih in besedilnih enot.

Za opis jezikovnih prvin in razčlenjevanje besedil NooJ med drugim uporablja končne pretvornike (*finite-state transducers*, FST) za prepoznavanje in označevanje nizov črk in/ali besed, končne avtomate (*finite-state automata*, FSA) za korpusna poizvedovanja, rekurzivne mreže prehodov (*recursive transition networks*, RTN) za izdelavo slovnic z več povezanimi grafi končnih stanj ter napredne rekurzivne mreže prehodov (*enhanced recursive transition networks*, ERTN), ki z uporabo spremenljivk in omejitev omogočajo raznorazne besedilne pretvorbe.

NooJ je bil kot nadgradnja okolja INTEX[3] (Silberztein, 1993) pod avtorstvom istega razvijalca izhodiščno izdelan v ogrodju .NET, v okviru projekta CESAR pa je bila izdelana tudi njegova odprtokodna različica v ogrodju Java (Silberztein, Váradi, Tadić, 2012). Obe različici z grafičnim vmesnikom sta za prenos na voljo na uradni spletni strani orodja.[4] Priročnik (Silberztein, 2003: 206-

[1] http://nl.ijs.si/noske/.
[2] http://nl.ijs.si/cuwi/.

[3] Kot odgovor na izhodiščno zaprtost sistema INTEX je bil leta 2002 izdelan zelo podoben odprtokodni sistem Unitex: http://www-igm.univ-mlv.fr/~unitex/.
[4] http://www.nooj4nlp.net/.

211) opisuje tudi različico za uporabo v ukazni vrstici in programskih vmesnikih (noojapply), ki pa ni javno objavljena, saj se po objavi odprtokodne različice program ne posodablja več.

Na spletni strani so objavljeni tudi jezikovni moduli, tj. zbirke jezikovnih virov (korpusov, leksikonov in slovnic) za procesiranje posameznih jezikov. V nasprotju s programsko opremo je odločitev o dostopnosti in rednem posodabljanju vsebin modulov prepuščena njihovim avtorjem, zato se obseg, kakovost in dostopnost virov za posamezne jezike pogosto precej razlikujejo. Trenutni nabor modulov vključuje 23 jezikov raznolikih oblikoslovnih tipov, pisav in jezikovnih družin, med njimi tudi južnoslovanske (bolgarščina, hrvaščina, srbščina, slovenščina).

V nadaljevanju na izbranih primerih iz obstoječe odprtokodne različice modula za slovenščino[5] predstavimo najpomembnejše značilnosti tega razvojnega okolja, in sicer možnost uvoza označenih in neoznačenih korpusov (3), leksikalne podatkovne zbirke (4), grafični vmesnik za izdelavo oblikoslovnih in skladenjskih pravil (5), možnosti označevanja besedil (6) in konkordančnik za njihovo analizo (7). Te in druge funkcionalnosti orodja NooJ so podrobneje in z več slikovnimi ponazoritvami predstavljene v uradnem priročniku za uporabnike (Silberztein, 2003), primeri konkretne uporabe orodja za različne namene obravnave naravnega jezika in nekatere novejše funkcionalnosti, s katerimi bi veljalo posodobiti obstoječi priročnik, pa so predstavljeni v zbornikih vsakoletnih konferenc (NooJ International Conference).

3. Korpusi

NooJ lahko procesira eno (.not) ali več besedilnih datotek (.noc) v več kot 150 različnih vhodnih formatih (npr. html, MS-Word, pdf, rtf itd.) in različnih načinih kodiranja. Uporabniki lahko besedila ustvarijo v integriranem urejevalniku ali pa jih vanj uvozijo kot zunanje datoteke, pri čemer je treba v obeh primerih določiti tudi izbrani razmejevalnik besedil (npr. odstavek, XML element ali določen izraz v datoteki).

NooJ poleg golih, neoznačenih besedil omogoča tudi uvoz že označenih besedil (v formatu XML), kar pomeni, da lahko uporabniki skupaj z besedilom v program uvozijo tudi podatke o njegovi strukturi, prevodu, slovničnih oznakah ipd. Pri uvozu datoteke v formatu XML v NooJ se imena elementov znotraj izbranega razmejevalnika besedila avtomatsko pretvorijo v skladenjske oz. semantične oznake (ime elementa postane ime skladenjske oz. semantične kategorije za vsebino elementa). Kot bomo videli v nadaljevanju, je izjema zgolj posebni element <LU>, ki napoveduje osnovne jezikovne enote (besedne pojavnice).

Z namenom preizkusa in prikaza uvoza besedil z različnimi tipi metapodatkov smo v slovenski modul vključili tri obstoječe korpuse (tabela 1): neoznačeno besedilo romana Telesni čuvaj (Mazzini, 2000), korpus ccKres (Logar Berginc et al., 2012), označen s statističnim označevalnikom Obeliks (Grčar et al., 2012), in ročno označeni korpus ssj500k (ibid.).

Korpus	Št. bes. pojavnic	Označenost
Telesni čuvaj.not	78.367	neoznačen
ssj500k.not	500.295	oblike, skladnja, NER
ccKres.noc	10.000.532	oblike

Tabela 1: Velikost in označenost korpusov v slovenskem modulu za NooJ.

Korpusa ssj500k in ccKres sta v svoji izhodiščni različici zapisana v formatu XML TEI P5[6], kakršnega predvidevajo krovne specifikacije za zapis korpusov, razvitih v okviru projekta Sporazumevanje v slovenskem jeziku[7]. Kot prikazuje slika 1, smo morali pred uvozom v NooJ izhodiščni format omenjenih korpusov prilagoditi tako, da smo besedne pojavnice (v procesu tokenizacije označene z <w>) pretvorili v format, na podlagi katerega NooJ vsebino elementa <LU> prepozna kot besedno obliko, vrednost neobveznega atributa LEMMA kot lemo in vrednost obveznega atributa CAT kot slovnično kategorijo (besedno vrsto). Morebitni drugi metapodatki o pojavnici lahko tema dvema sledijo v obliki poljubno poimenovanih atributov in njihovih vrednosti; v našem primeru je to informacija o celotni oblikoskladenjski oznaki (atribut MSD), pri skladenjsko razčlenjenem korpusu pa tudi o identifikatorju pojavnice v stavku (atribut ID), skladenjski odnosnici (HEAD) in skladenjskem razmerju (DEPREL).

```
<s id="ssj1.1.5">
   <LU LEMMA="ta" CAT="P" MSD="Pd-nsg" ID="t1" HEAD="t5" DEPREL="Obj">Tega</LU>
   <LU LEMMA="se" CAT="P" MSD="Px------y" ID="t2" HEAD="t5" DEPREL="PPart">se</LU>
   <LU LEMMA="sploh" CAT="Q" MSD="Q" ID="t3" HEAD="t0" DEPREL="Root">sploh</LU>
   <LU LEMMA="biti" CAT="V" MSD="Va-r1s-y" ID="t4" HEAD="t5" DEPREL="PPart">nisem</LU>
   <LU LEMMA="zavesti" CAT="V" MSD="Vmep-sm" ID="t5" HEAD="t0" DEPREL="Root">zavedel</LU>
   <c ID="t6" HEAD="t0" DEPREL="Root">.</c>
</s>
```

Slika 1: Stavek korpusa ssj500k v formatu NooJ XML.

4. Slovarji

Leksikoni oz. slovarji (dictionaries, .nod) v orodju NooJ opravljajo vlogo podatkovnih zbirk za prepoznavanje eno- in večbesedne leksike ter opisovanje njihovih oblikoslovnih, skladenjskih, pomenskih in drugih lastnosti. Tipična leksikonska iztočnica je opisana kot niz leme, besednovrstne oznake ter drugih lastnosti (slika 2). Njihov nabor in poimenovanja poljubno določajo uporabniki, v slovarju pa so lahko leksikalni enoti pripisane kot značilke (npr. +splošni) ali kot niz lastnosti in njene vrednosti (+vrsta=splošni). Vrednosti atributov lahko vsebujejo metajezikovne informacije (npr. +vrsta=splošni), besedilo v izhodiščnem jeziku (npr. +sinonim=zlahka oz. +sinonim="brez napora") ali besedilo v tujem jeziku (npr. +ANG=effortlessly).

```
lahko,R+Type=general+FLX=LAHKO
nizko,R+Type=general+FLX=LAHKO
ozko,R+Type=general+FLX=LAHKO
```

Slika 2: Tipični zapis iztočnic v leksikonih NooJ na primeru prislovov iz leksikona Sloleks.

[5] http://www.nooj4nlp.net/pages/slovene.html.

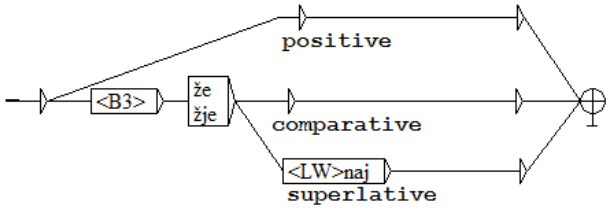
[6] www.tei-c.org/P5/Guidelines/.

[7] http://www.slovenscina.eu/.

4.1. Vzorci pregibanja

Med leksikalnimi atributi s posebnim privzetim pomenom sta najpomembnejša +FLX in +DRV, ki leksikonsko iztočnico (lemo) povezujeta s pripadajočim pregibnim oz. besedotvornim pravilom. Pravila za pregibanje so formalizirana v ločenih datotekah (.nof), in sicer bodisi v obliki besedila bodisi v obliki grafa. Vrednost atributa +FLX (vedno zapisana z velikimi črkami) tako iztočnico v leksikonu poveže z enako poimenovanim pravilom, ki v procesu transformacije osnovne oblike z dodajanjem črk in/ali uporabo posebnih operatorjev (npr. za brisanje, premikanje naprej/nazaj, podvajanje znakov ipd.) iz osnovne ustvari ustrezne pregibne oblike in jim pripiše morebitne dodatne lastnosti.

Slika 3 tako prikazuje opis pregibnega vzorca LAHKO za prislove z variantnim stopnjevanjem (npr. *lahko-laže/lažje-najlaže/najlažje*). Vzorec določa, da je oblika za nedoločno stopnjo enaka lemi, obliki za primerniško in presežniško obliko pa se tvorita tako, da se osnovni obliki brez zadnjih treh črk (<B3>) dodata končnici *-že* in *-žje*, pri čemer se na začetek obeh presežniških variant doda še predpona *naj-* (<LW>naj).




```
LAHKO = <E>/=positive | <B3>že/=comparative |
<B3>žje/=comparative | <B3>že<LW>naj/=superlative
| <B3>žje<LW>naj/=superlative;
```

Slika 3: Primer pregibnega vzorca LAHKO za obrazilno stopnjevanje prislovov v grafični (zgoraj) in besedilni obliki (spodaj).

Pri poskusu prenosa referenčnega oblikoslovnega leksikona Sloleks[8], ki v formatu XML LMF opisuje več kot 100.000 lem s pripadajočimi pregibnimi oblikami, se je kmalu izkazalo, da izluščeni oblikoslovni vzorci vsebujejo precej nedoslednosti, napak in neskladnosti z jezikovno rabo, zato je pred njihovo formalizacijo in povezovanjem z iztočnicami (lemami) nujna še dodatna faza ročne evalvacije. Predlog cevovodnega procesa za polavtomatsko validacijo oblikoslovnih vzorcev za slovenščino smo že preizkusili na primeru obrazilnega pregibanja prislovov (Dobrovoljc, 2014b), rezultati pa so v obliki formaliziranih vzorcev ter posodobljenega in razširjenega leksikona prislovov vključeni tudi v trenutni modul.

Do zaključka procesa sistematične evalvacije vzorcev drugih besednih vrst je celoten leksikon Sloleks v trenutno različico modula zato vključen v obliki, ki ne predvideva povezave s konkretnim formaliziranim vzorcem, temveč posamične oblike z vsemi oblikoskladenjskimi lastnostmi[9] navaja kot ločene leksikonske enote (slika 4). To vpliva na hitrost, ne pa tudi na rezultat procesiranja, saj pripisane oznake ne glede na format vsebujejo enak nabor informacij o slovničnih lastnosti oblik.

```
garažisti,garažist,N+Type=common+Gender=m
asculine+Number=plural+Case=instrumental
```

Slika 4: Primer zapisa iztočnice leksikona Sloleks, pretvorjenega v format za uporabo v orodju NooJ.

5. Slovnice

Drugi temeljni vir jezikovnih opisov v orodju NooJ so slovnice (grammars), ki so v nasprotju z opisovanjem končnega nabora leksike v slovarjih namenjene opisovanju pravil za produktivnejše slovnične pojave na vseh jezikovnih ravneh, pri tem pa se na različne načine povezujejo tudi z leksikalnimi informacijami. Poleg že omenjenih slovnic za pregibanje in besedotvorje, ki opisujejo pregibne lastnosti leksikonskih iztočnic, NooJ vključuje tudi vmesnika za oblikovanje oblikoslovnih pravil (morphological grammars), ki opisujejo morfemske lastnosti besednih oblik, in skladenjskih pravil (syntactic grammars), ki opisujejo skladenjske in semantične lastnosti eno- ali večbesednih izrazov.

Uporabniki oba tipa pravil opisujejo v obliki eno- ali večnivojskih grafov, pri katerih niz vozlišč med začetnim in končnim vozliščem grafa označuje bodisi zaporedje črk ali morfemov znotraj pojavnice (oblikoslovne slovnice) bodisi zaporedje ene ali več pojavnic (skladenjske slovnice). Oba tipa pravil v svojih vozliščih in njihovih oznakah omogočata uporabo spremenljivk in nekaterih privzetih operatorjev, npr. za male in velike začetnice, naglašene in nenaglašene črke, dolžino besede, začetek ali konec stavka ipd.

Vmesnik za izdelavo slovničnih grafov vsebuje tudi nekaj koristnih funkcij, ki uporabnikom na vizualno razumljiv način omogočajo sprotno validacijo pravil, npr. preverjanje seznama pozitivnih in negativnih primerov, generiranje vseh možnih poti grafa (oblik, skladenjskih vzorcev ipd.) in razhroščevanje.

5.1.1. Oblikoslovne slovnice

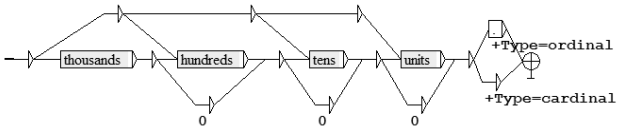
Kot smo že omenili v poglavju o leksikalnih zbirkah, se besedne oblike privzeto analizirajo z vpogledom v leksikon, ki obliko povezuje s pripadajočimi slovničnimi in drugimi lastnostmi. Vendarle pa je nekatere besedne oblike namesto eksplicitnega zapisovanja v leksikalnih zbirkah smiselneje opisovati s posebnimi generičnimi pravili, ki hkrati opisujejo več različnih oblik z enakimi slovničnimi lastnostmi, kot so npr. števniki. Oblikoslovne slovnice (.nom) so tako namenjene predvsem opisovanju tovrstnih produktivnih oblikotvornih pravil za besedne oblike s podobnimi lastnostmi. Najpogosteje se uporabljajo za poenotenje oblikovnih variant, prepoznavanje in označevanje neologizmov, ustvarjanje povezav med besedotvorno povezanimi besedami ipd.

Med oblikoslovnimi slovnicami v slovenskem modulu v prispevku izpostavljamo dve, ki prikazujeta dva možna namena njihove uporabe. Prva slovnica (slika 5) prepoznava rimske številke, določa njihove slovnične

[8] http://www.slovenscina.eu/sloleks/opis.

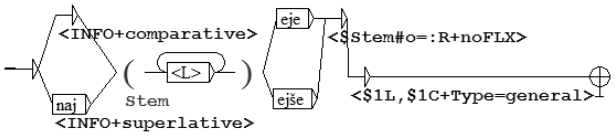
[9] Oblikoslovni metapodatki vseh omenjenih virov temeljijo na naboru oznak, razvitem v okviru projektov MULTEXT-East (Erjavec, 2010) in JOS (Erjavec in Krek, 2008). S tega vidika so slovenski viri kompatibilni tudi z viri drugih jezikov, vključenih v projekt MULTEXT-East, ki imajo obenem razvite tudi svoje module znotraj okolja NooJ (Stanković et al., 2012).

lastnosti (zapis, vrsta) in jih pretvarja v ekvivalentni arabski zapis (npr. *MMLXIV* v *2064*). To je tipični primer pravila, s katerim na hiter in preprost način opišemo večje skupine besednih oblik (npr. vse možne rimske številke), med katerimi so v leksikone običajno vključene le najpogostejše.



Slika 5: Oblikoslovna slovnica za prepoznavanje in pretvarjanje rimskih številnikov.

Druga oblikoslovna slovnica ponazarja primer pravila za procesiranje neznanih besednih oblik, ki smo ga uporabili pri evalvaciji leksikona Sloleks. Slovnica namreč preverja, ali se za prislove, ki so v izhodiščnem leksikonu označeni kot nestopnjevani (npr. *zavzeto*), v rabi pojavljajo tudi obrazilno stopnjevane (primerniške in presežniške) oblike (npr. *zavzeteje, najzavzeteje*). Slovnični graf (slika 6) tako poišče neznane besedne oblike, ki se končajo z enim od obeh možnih obrazil (*-eje* ali *-ejše*), in preveri, ali je izluščeni koren (spremenljivka $Stem) z dodano končnico -o ($Stem#o) v izhodiščnem slovarju označen kot nestopnjevani prislov (:R+noFLX). Če je pogoj izpolnjen, besedna oblika podeduje lemo ($1L) in besedno vrsto ($1C) prislova v leksikonu, doda pa se ji oznaka za primernik oz. presežnik.



Slika 6: Slovnica za prepoznavanje obrazilno stopnjevanih prislovov.

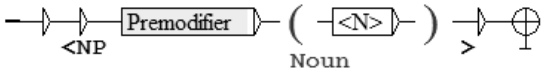
5.1.2. Skladenjske slovnice

Skladenjske ali lokalne slovnice (.nog) so namenjene opisovanju in označevanju eno- ali večbesednih izrazov. Tipično se uporabljajo za skladenjsko in pomensko razčlenjevanje, luščenje in označevanje imenskih entitet ter drugih prekinjenih ali neprekinjenih stalnih besednih zvez, pa tudi za avtomatsko razdvoumljanje besednih oblik v kontekstu ter besedilne pretvorbe, kot sta npr. parafraziranje in prevajanje.

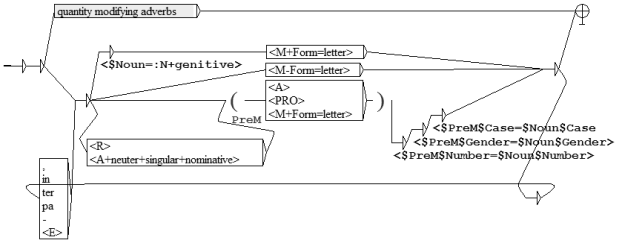
V primerjavi z drugimi na pravilih temelječimi aplikacijami za računalniško obdelavo besedil je ena izmed glavnih prednosti okolja NooJ dejstvo, da sta skladenjska in oblikoslovna raven medsebojno povezljivi. To pomeni, da lahko uporabniki v skladenjska pravila z različnimi operatorji za spremenljivke vključujejo tudi raznorazne pretvorbe njihovih vrednosti (npr. lematizacijo, pregibanje, priklic določene leksikalne lastnosti), določajo omejitve (npr. pogoj ujemanja v izbranih slovničnih lastnostih) in spreminjajo njihovo zaporedje.

Slovenski modul vsebuje nekaj osnovnih slovnic za prepoznavanje in označevanje različnih vrst imenskih entitet (npr. lastnih imen, besednih številnikov, datumov in časovnih izrazov) ter nekaterih površinskih skladenjskih struktur. Primer druge skupine je denimo slovnica, ki v oblikoskladenjsko označenih besedilih prepoznava podredno zložene samostalniške besedne zveze, tj. zveze samostalnika in njegovih (nestavčnih) določil oz. prilastkov.

Na najvišji ravni pravila (slika 8) je besedna zveza opredeljena kot samostalnik, pred katerim stoji eno ali več določil. Vdelani podgraf (slika 9) pa nato podrobneje določa možne strukture določil, tj. končen nabor količinskih prislovov (*nekaj, malo, veliko* ipd.) ter neomejen niz številnikov in/ali prilastkov (spremenljivka $PreM), ki se morajo z jedrnim samostalnikom ujemati v spolu, sklonu in številu. Ujemalni prilastki imajo lahko tudi sami določila v obliki prislovov ali pridevnikov v imenovalniku srednjega spola ednine (npr. *slovensko* v zvezi *slovensko-francoski odnosi*).



Slika 7: Slovnica za prepoznavanje podredno zloženih samostalniških besednih zvez (prvi nivo).



Slika 8: Slovnica za opis levih prilastkov v podredno zloženih samostalniških besednih zvezah (drugi nivo).

6. Označevanje

V prispevku smo že večkrat nakazali, da oblikoslovne in skladenjske slovnice niso namenjene zgolj prepoznavanju jezikovnih pojavov, temveč tudi njihovemu označevanju.

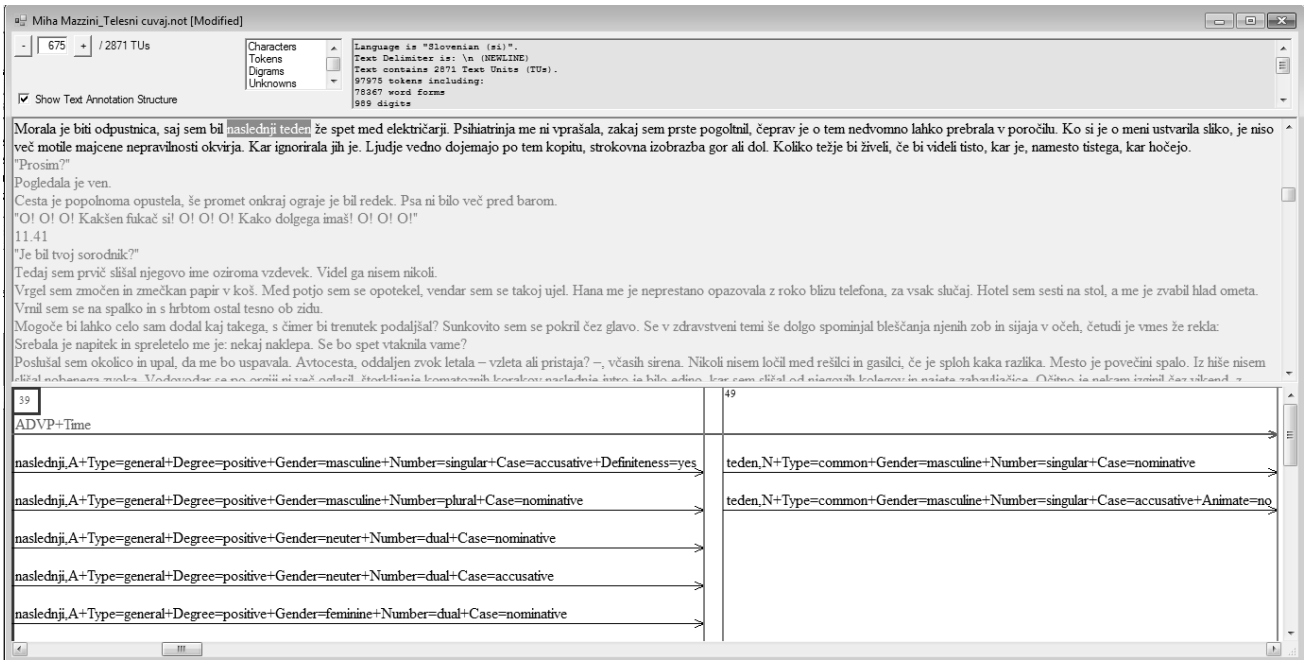
Pri označevanju besedil NooJ ustvarja pare (*pozicija, informacija*), ki označujejo, da ima določen niz v besedilu določene lastnosti. Te binarne oznake se v sistemu shranijo v t. i. strukturo označenega besedila (text annotation structure, TAS), pri čemer se ta nenehno usklajuje z izhodiščno besedilno datoteko, ki se sama nikoli ne spreminja. Binarne oznake se lahko nanašajo tako na posamezne besede (npr. označevanje besede *miza* s kategorijo samostalnika), na del besede (npr. razčlenjevanje pojavnice *karkoli*) ter na neprekinjene in prekinjene večbesedne enote (npr. *okrogla miza* ali *potegnti* nekaj *iz klobuka*). Poleg dodajanja oznak v strukturo označenega besedila NooJ omogoča tudi izvoz označenih besedil in uvoz že označenih besedil (v datotekah XML).

NooJ pri označevanju vedno vrne vse možne oznake, a obenem omogoča tudi njihovo nadaljnje odstranjevanje (razdvoumljanje) na avtomatski (s pravili v obliki

skladenjskih slovnic), polavtomatski (s filtriranjem po seznamu konkordanc) ali ročni način.

Uporabnik pred kakršnimkoli označevanjem v nastavitvah sam izbire relevantne vire (slovarje in slovnice) za jezikoslovno analizo. Določa lahko tudi njihovo zaporedje (stopnjo pomembnosti), pri čemer se nižje uvrščeni viri upoštevajo zgolj pri analizi pojavov, ki jih višje uvrščeni viri niso obravnavali. Ta mehanizem se tako tipično uporablja predvsem za procesiranje neznanih besed (z viri, uvrščenimi za leksikone) oz. za popravke v tokenizaciji večbesednih enot (z viri, uvrščenimi pred leksikone).

Vse informacije v strukturi označenega besedila so v korpusnem vmesniku vizualizirane pod besedilom (slika 9), pri čemer so leksikalne in skladenjske oznake barvno ločene, uporabnik pa jih lahko v primeru ročnega razdvoumljanja tudi ureja.

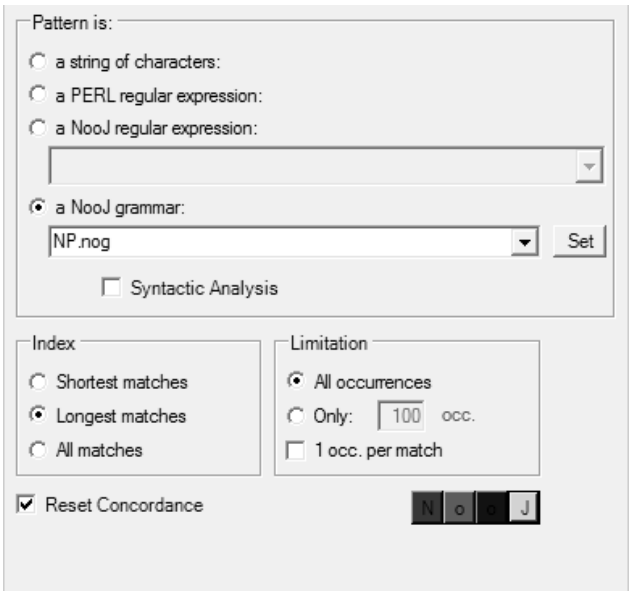


Slika 9: Prikaz strukture označenega besedila z nerazdvoumljenimi oblikoskladenjskimi oznakami po aplikaciji skladenjske slovnice za prepoznavanje časovnih izrazov.

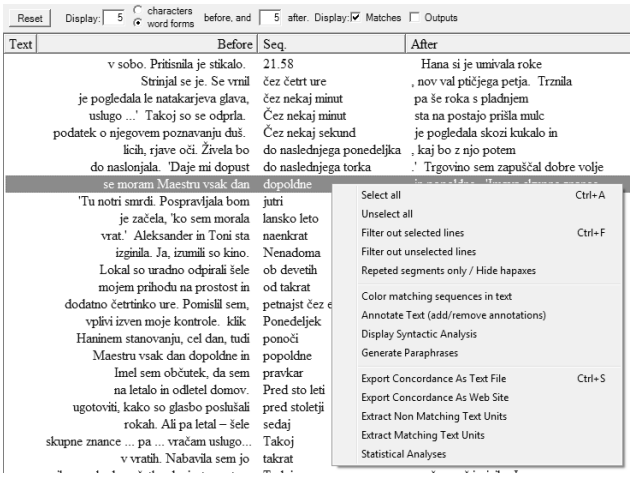
7. Konkordančnik

Poleg zmogljivega opisovanja in procesiranja besedil NooJ odlikuje tudi vmesnik za luščenje korpusnih konkordanc. Kot prikazuje slika 10, lahko uporabniki po besedilu in njegovih raznolikih oznakah poizvedujejo s črkovnimi nizi, različnimi tipi regularnih izrazov in tudi z neposrednim vnosom (skladenjskih) slovnic, ki se tako uporabljajo tudi oz. predvsem za izdelavo kompleksnih korpusnih poizvedb, ki ne spreminjajo označevalne strukture.

Uporabniki lahko priklicane konkordance (slika 11) poljubno urejajo, filtrirajo in izvažajo, jim dodajajo ali odstranjujejo oznake (npr. pri polavtomatskem razdvoumljanju), ogledajo pa si lahko tudi nekatere statistične izračune (pogostost, standardna vrednost, relevantnost za posamezno besedilno enoto, podobnost besedišča).



Slika 10: Vmesnik za iskanje po korpusu, v katerem je kot iskalni pogoj vneseno pravilo za prepoznavanje samostalniških besednih zvez.



Slika 11: Konkordančni niz za skladenjsko slovnico, ki prepoznava časovne izraze.

8. Zaključek

V prispevku smo na omejenem naboru primerov jezikovnih virov in pravil iz pilotnega modula za slovenščino predstavili nekaj temeljnih značilnosti razvojnega okolja NooJ. V primerjavi s splošnimi konkordančniki in drugimi vmesniki za analizo korpusnih besedil NooJ jezikoslovnim in drugim raziskovalcem ponuja možnost naprednejše obdelave korpusnih besedil, ne da bi ti za to potrebovali napredno računalniško predznanje. Odlikujeta ga predvsem vmesnik za razmeroma preprost opis raznolikih jezikovnih pojavov v obliki grafov ter možnost njihove takojšnje uporabe na korpusnih besedilih.

Čeprav je NooJ prvenstveno namenjen razvoju samostojnih, na pravilih temelječih orodij za strojno označevanje jezika, menimo, da se znotraj slovenskega prostora njegov največji potencial skriva v povezovanju z drugimi, že obstoječimi jezikovnimi viri in orodji za strojno procesiranje slovenščine.

V prvi vrsti imamo v mislih možnost oblikovanja kompleksnih korpusnih poizvedb po površinski in označeni strukturi besedila, denimo za luščenje podatkov iz površinskoskladenjsko razčlenjenih korpusov, govornih korpusov ali drugih korpusov, ki poleg slovničnih lastnosti besednih oblik vsebujejo tudi druge vrste in ravni jezikoslovnih oznak.

Druga obetavna možnost uporabe orodja NooJ je v približevanju obstoječih korpusnih virov jezikoslovnim raziskovalcem, ki so te doslej zaradi označevalnih napak ali nestrinjanja z označevalnim sistemom pogosto zavračali kot nezanesljive. Funkcionalnosti orodja NooJ omogočajo preprosto izdelavo hibridnih orodij za nadgradnjo ali dopolnitev oznak v izhodiščnih virih, npr. s pravili za usmerjeno odpravljanje napak, prilagajanje specifičnim raziskovalnim potrebam oz. teoretskim nazorom ter druge oblike hevrističnega usmerjanja statističnih jezikovnih modelov.

Nenazadnje NooJ kot odprtokodna programska oprema predstavlja tudi priročno komunikacijsko stičišče jezikoslovne in računalniške skupnosti, saj jezikoslovcem omogoča preprosto formalizacijo opazovanih jezikovnih pojavov, informatikom pa njihovo brezšivno implementacijo v širše računalniške sisteme.

Literatura

Dobrovoljc, K., 2014a. Introduction to Slovene Language Resources for NooJ. V S. Koeva, S. Mesfar in M. Silberztein (ur.), *Formalising Natural Languages with NooJ 2013: Selected Papers from the NooJ 2013 International Conference*. Newcastle: Cambridge Scholars Publishing. 27-40.

Dobrovoljc, K., 2014b. Re-evaluating morphological dictionaries: the case of adverbs in Slovene. *NooJ 2014 International Conference*. [v objavi]

Erjavec, T. in S. Krek, 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. V: T. Erjavec in J. Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 49-53.

Erjavec, T., 2010. MULTEXT-East version 4: multilingual morphosyntactic specifications, lexicons and corpora. V: N. Calzolari (ur.): *Proceedings of the 7th International Conference on Language Resources and Evaluations, 19-21 May 2010, Valletta, Malta*. 2544–2547.

Erjavec, T., 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1:24-49.

Grčar, M, S. Krek in K. Dobrovoljc, 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 89-94.

Logar Berginc, N., M. Grčar, M. Brakus, T. Erjavec, Š. Arhar Holdt in S. Krek, 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Mazzini, M., 2000. *Telesni čuvaj: verzija 1.72.* Ljubljana: Študentska založba.

Silberztein, M., 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Pariz: Elsevier Masson.

Silberztein, M., 2003. *NooJ Manual*. Dostopno na: http://www.nooj4nlp.net/NooJManual.pdf.

Silberztein, M., T. Váradi in M. Tadić, 2012. Open source multi-platform NooJ for NLP. *Proceedings of COLING 2012: Demonstration Papers.* 401-408.

Stanković, R., M. Utvić, D. Vitas, C. Krstev in I. Obradović, 2012. On the Compatibility of Lexical Resources for NooJ. V: K. Vučković, B. Bekavac, & M. Silberztein (ur.): *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the 2011 International Nooj Conference*. Cambridge Scholars Publishing. 96–109.