https://doi.org/10.18485/climb.2018.6.1.ch23
004.9:811.111'374
Оригинални научни рад
Примљен: 10.04.2018.
Прихваћен: 04.05.2018.

dr Miloš D. Đurić*
Univerzitet u Beogradu
Elektrotehnički fakultet

NEKI ASPEKTI SLOŽENICA I SISTEMA ZA ANALIZU SLOŽENICA PREMA ISTRAŽIVAČKIM REZULTATIMA IZ RAČUNARSKE LINGVISTIKE

Sažetak

Engleske leksičke jedinice, poput *electric circuit*, *ion path*, *interrupting capacity*, i francuske leksičke jedinice, poput *circuit électrique*, *trajectoire ionique*, *intensité maximum de rupture*, proučavane su u relevantnoj lingvističkoj literaturi pod nazivima *složenice*, *kompleksni nominali*, *kompleksne imeničke fraze*, da navedemo samo nekolicinu. U literaturi su predložene raznovrsne klasifikacije i taksonomije da bi se identifikovale semantičke relacije u okviru mnoštva interpretacija koje bi ovi termini trebalo da prikazuju. Pored toga, ovi pokušaji su rezultirali velikim porastom taksonomija koje su imale za cilj da dokuče sintaksu i semantiku složenica, a koje su u najboljem slučaju bile arbitrarne, što je u krajnjoj liniji zamračilo pre nego osvetlilo fenomen složenica. Dolaskom okvira generativne semantike Levijeve za kompleksne nominale i potonjih istraživanja računarske lingvistike ove relacije su preciznije opisane.

Prvi deo mog rada je deskriptivno istraživanje termina 'složenica'. Drugi deo pruža kratak prikaz semantičkih klasifikacija složenica na osnovu istraživačkih podataka iz računarske lingvistike. Treći deo je deskriptivni prikaz sistema TANKA, HAIKU, SENS, SESEMI, MINDNET, SEMEVAL-2010 i G_hoSt-NN. Četvrti deo ukratko prikazuje primenu sistema sens na analizu složenica u diskursu elektrotehnike. U petom delu se iznose relevantne zaključne napomene.

Ključne reči: složenice, semantičke relacije, sistemi za analizu složenica, MINDNET, SENS, SESEMI, TANKA, HAIKU, SEMEVAL-2010, G_hoSt-NN, računarska lingvistika.

* Elektrotehnički fakultet, Bulevar kralja Aleksandra 73, Beograd
MilosDDjuric@hotmail.com

1. Deskriptivno istraživanje termina „složenica“

U okviru odrednice *Oksfordskog rečnika lingvistike*, termin „složenica“ se određuje kao reč formirana od dve ili više jedinica, koje su i same reči (Matthews 2005: 66). Navodi se primer iz engleskog (složenica *blackboard*, nastala od *black* i *board*) i iz nemačkog jezika (složenica *Schreibmaschine*, nastala od *schreiben* i *Maschine*). Na dalje, ovaj autor izjednačava proces stvaranja složenica i kompoziciju, i kaže da su to alternativni termini koji se odnose na proces stvaranja složenica (Matthews 2005: 66). U literaturi se kaže da su složene reči ili složenice sastavljene od dve ili više slobodnih morfema[1] (Kristal 1988: 240). U relevantnoj literaturi se navodi da se termin *složen* odnosi na jezičku jedinicu sastavljenu od elemenata koji mogu da funkcionišu samostalno ukoliko se nađu na drugom mestu (Kristal 1996: 430).

U višejezičnim rečnicima nalazimo sledeće ekvivalente za termin *složena reč*/*složenica*: *mot composé* (u francuskom), *compound word* (u engleskom) i *Kompositum* i *Zusammensetzung* (u nemačkom jeziku) (Đokić 2001: 195).U monolingvalnom engleskom rečniku *The Penguin English Dictionary* navode se čak tri odrednice za leksemu *compound* od kojih svaka sadrži još po nekoliko pododrednica sa različitim značenjima, ili varijacijama značenja. Za složenicu se kaže da je to reč koja se sastoji od komponenti, koje su reči, oblici koji se kombinuju i afiksi, a primeri su *houseboat* i *anthropology*. U monolingvalnom francuskom rečniku *Le Micro Robert* određuje leksemu *mot composé* kao jedinicu koja se sastoji od nekoliko reči ili koja ima prefiks, npr. *antigel*, *chemin de fer*, *chou fleur*. Dakle, i ovde nalazimo prilično široko određenje, pošto u odrednici figurira *plusieurs éléments*, dakle, ne samo dva već mogućnost više elemenata složenice.

Najopštije definicije određenja složenica nalazimo u relevantnoj literaturi, u kojoj se složenice određuju kao gramatički i semantički osamostaljene leksički oblici sa jedinstvenim značenjem i gramatičkom funkcijom (Fišer-Popović, 1981: 160). U pogledu broja konstituenata složenice, takođe uzimamo prilično široka shvatanja, prema kojima je složenica reč koja se

1 Kako bi približili tadašnjoj domaćoj (jugoslovenskoj) lingvističkoj javnosti Kristalov rečnik, prevodioci sa engleskog na srpskohrvatski, profesor Ivan Klajn i Boris Hlebec, potrudili su se da navedu skoro uvek srpskohrvatske primere, pa tako i za složenice navode primere: *parobrod*, *gulikoža* i *klinčorba* (Kristal 1988: 240).

sastoji od dve ili više reči (Fabb, 1998: 66). U relevantnoj opštelingvističkoj literaturi, imeničke složenice se određuju kao jedinice čija interna struktura može da sadrži sled imenica, prideva i imenice (npr. *telephone bell*, *paper knife*), ali čije sintaksičko ponašanje odgovara ponašanju proste imenice (Smith & Wilson 1979: 272). Iako navodi kanoničke/binarne složenice, ni Ranko Bugarski ne uvodi restrikciju kada je u pitanju broj elemenata, pošto konstatuje da je kompozicija kombinovanje slobodnih morfema, čiji su proizvod složenice, a navodi primere: *zlodelo*, *mimohod*, *starmali* (Bugarski 1996: 169). Prema reprezentativnoj literaturi, „kompozicijom se spajaju u jednu riječ dva leksema koja inače mogu biti samostalni korijeni u tvorbi riječi: na taj se način dobivaju složenice." (Škiljan 1980: 107). U pogledu interne strukture složenica, usvojili smo jednu našu raniju podelu složenica prema broju konstituenata na kanoničke (binarne, bazične, dvokonstituentske) i nekanoničke (višečlane) složenice (Đurić 2016: 22-39). Mi smo usvojili prilično široka određenja složenice ubrajajući tu složenice (npr. *Christmas party*, *hedge-hop*, *chemistry laboratory*) koje se pominju u reprezentativnoj lingvističkoj literaturi (Chomsky & Halle 1995: 91). Takođe, u literaturi se pominju i *višeimenske složenice* koje se mogu parafrazirati na više načina (Bartolić 1979: 47). U citiranom radu navodi se složenica *beam intensity modulation terminal*, koji se može interpretirati kao: „a terminal for the modulation of the intensity of a beam", „a terminal for modulating the intensity of a beam" ili pak „a terminal which/that modulates the intensity of a beam" (Bartolić 1979: 47).

Imajući navedene definicije u vidu, neki primeri iz našeg korpusa za engleske složenice i francuske složenice su[2]: *electric circuit - circuit électrique*, *ion path - trajectoire ionique*, *interrupting capacity - intensité maximum de rupture*. U morfologiji se složenica određuje kao derivirani oblik koji proizilazi iz kombinacije dve ili više leksema (npr. *space* + *ship* > *spaceship*) (Aronoff & Fudeman 2005: 236). Slično tome, složenica se definiše i kao jedinica koja se sastoji od dve ili više osnova (Quirk et al. 1972: 1019). Isto tako, složenicu posmatraju i kao semantičku jedinicu koja se sastoji od dva ili više elemenata (Aitchinson 2001: 56-60). Neki autori insistiraju na tome da su elementi/konstituenti složenice postojeće reči u prirodnom jeziku, pa određuju složenicu kao konstrukt dobijen kombino-

2 Uvek navodimo prvo englesku složenicu, a onda francusku složenicu (koja je ujedno i prevodni ekvivalent engleske složenice).

vanjem dve ili više *postojećih* reči (emfaza M. D. Đurić), čije značenje nije uvek predvidljivo na osnovu značenja delova složenice (Trask 2000: 30). S druge strane, izvesni autori konstatuju da ne postoji ograničenje u pogledu vrsta kombinacija elemenata, makar kada je u pitanju engleski jezik (Fromkin & Rodman 1983: 121). U literaturi se kaže i da složenice imaju više od jedne leksičke morfeme (Napoli 1996: 229). Neki autori određuju kompoziciju kao vezivanje dve ili više leksema kako bi se obrazovala nova leksema (Lardiere 2006: 77). Možda najšire određenje nalazimo možda kod Pinkera, koji za složenicu navodi da je to reč formirana pridruživanjem drugih reči (Pinker 1995: 475). Generativna semantičarka Judith N. Levi kaže da se sve složenice deriviraju isključivo putem dva sintaksička procesa, odnosno, brisanjem predikata i nominalizacijom predikata (Levi 1978: 6).

Kada određuju složenice, francuski autori ističu sintaktičke osobenosti složenica i konstatuju da ove jedinice korespondiraju različitim vrstama sintaksičkih konstrukcija (Dubois & Lagane 2004: 227). Takođe, izvesni francuski autori insistiraju na tome da u kombinacijama učestvuju postojeće reči, pa kažu da se složenice obrazuju spajanjem nekoliko postojećih reči (Grevisse & Goosse 1995: 63). Francuski autori uvode i niz testova kojima se potvrđuje identitet složenica (Marcellesi 1977: 195). Francuski autori daju isto široka shvatanja složenice, uključujući tu i složenice sa strukturom imenica+predlog+imenica, poput *pomme de terre* i *moulin à vent* (Mitterand 1972: 53). Citirani autor posvećuje i poseban odeljak koji se odnosi na ortografiju složenica.

Prilikom razmatranja složenica u specifičnim registrima, lingvisti ističu da složenice obezbeđuju morfološko-semantičko i leksičko-semantičko jedinstvo termina, čime se istovremeno dobija kako na ekonomičnosti, tako i na preciznosti jezika (Vujić 2004: 156).

U okviru računarske lingvistike, izvesni autori konstatuju da se o definisanju složenica kontroverzno raspravlja u lingvističkoj literaturi, kao i činjenicu da jedva da postoje neki opšteprihvaćeni kriterijumi za određivanje složenica, naročito u pogledu toga koja su svojstva neophodna i dovoljna da bi se jezička jedinica smatrala složenicom (Ziering 2018: 61). Jedan od izazova u računarskoj lingvistici jeste i dekodiranje složenica. U okviru metodike nastave engleskog jezika problem dekodiranja se razrešava uspostavljanjem paralela između kategorija definicija i kategorija slo-

ženica (Master 2003: 3). Naravno, logičan problem koji se nameće jeste i broj konstituenata kao i redosled kojim se ove komponente interpretiraju. Dodatni problem nastaje ako se doda kontrastivna dimenzija, pošto kod engleskih i francuskih složenica sintaksički centar se različito locira (Di Sciullo 2005).

Odluka da pored jezičkih opisa i jezičkih tumačenja složenica, odgovore potražimo i u istraživačkim naporima računarske lingvistike, motivisani su, između ostalog, i konstatacijom iz literature, da su izazovi, koje su predstavile složenice i njihova semantika, generisale značajana tekuća interesovanja za interpretaciju složenica u zajednici proučavalaca prirodne obrade jezika (Butnariu et al. 2010: 100). Stoga smo posvetili istraživačku pažnju i ovoj oblasti u okviru koje se, između ostalog, tretiraju i složenice.

2. Semantičke klasifikacije složenica na osnovu istraživačkih podataka iz računarske lingvistike

Složenice su se našle i u fokusu računarske lingvistike, a naročito zanimljive su semantičke relacije koje vladaju između konstituenata ovih semantičkih jedinica. Pogledaćemo koncizan deskritpivan prikaz semantičkih relacija prema istraživačkim rezultatima iz računarske lingvistike.

Pojedini autori posmatraju složenice kao svojevrsni semantički okvir sa izvesnim brojem slotova koji se ispunjavaju zahtevanim ispunjivačem ili preferiranim ispunjivačem (npr. Finin 1980). Citirani autor čak i tvrdi da su složenice najvećim delom nominalizacije, čak i u slučajevima gde centar (imenica) nije morfološki deriviran iz glagola. Finin je u svom sistemu utvrdio ukupno šesnaest semantičkih relacija.

Primenili smo njegove relacije na diskurs elektrotehnike, pa navodim po dva primera za svaku relaciju, od kojih prvi pripada opštem diskursu, a drugi diskursu elektrotehnike. Relacije su: 1. agens + glagol (*cat scratch*, *program run*), 2. objekat + glagol (*engine repair*, *propagation delay*), 3. instrument + glagol (*knife wound*, *loop gain*), 4. vozilo + glagol (*boat ride*, *carrier beat*), 5. recipijent + glagol (*Unicef donations*, *library program*), 6. mesto + glagol (*ocean fishing*, *lighthouse valve*), 7. vreme + glagol (*summer rains*, *daylight*), 8. izvor + glagol (*Chicago flights*, *Winchester disk*), 9. odredište + glagol (*target shooting*, *target capacitance*), 10.

uzrok + glagol (*drug killings*, *coincidence gate*), 11. glagol + agens (*repair man*, *software analyst*), 12. glagol + objekat (*throwing knife, timesharing system*), 13. glagol + instrument (*cooking fork*, *programming language*), 14. glagol + lokacija (*meeting room*, *storage location*), 15. glagol + vreme (*election year*, nema odgovarajućeg primera u diskursu elektrotehnike i računarstva), 16. glagol + izvor (*shipping depot*, *storage tube*).

Istraživački rezultati u računarskoj lingvistici fokusiraju se, između ostalog i na različite stepene leksikalizacije koje složenice mogu da ispoljavaju u diskursu (Isabelle 1984: 509). Citirani autor svoju pažnju posvećuje složenicama koje sadrže argument i predlaže pet semantičkih relacija koje obuhvataju vrstu relacije i podatak o tome na šta se ova relacija preslikava. Autor klasifikuje sledeće složenice: 1. složenice koje sadrže semantičku relaciju mera (*measure*), preslikavaju predmete na količinu (npr. *oil temperature*), 2. složenice koje sadrže semantičku relaciju oblast (*area*), preslikavaju predmete na njihove sastavne delove (npr. *box top*), 3. zbirne složenice koje preslikavaju individualne članove na skup (npr. *tank group*), 4. reprezentacione složenice, koje preslikavaju predmete na reprezentaciju predmeta (npr. *circuit diagram*), i 5. složenice koje sadrže relaciju „ostalo“ (npr. *component location*). Klasifikacija ovog autora nije sveobuhvatna i čini nam se nedovoljno precizno definisanom, naročito peta grupa složenica koje sadrže relaciju „ostalo“ jer dozvoljava klasifikovanje velikog broja složenica, u koju bismo mogli da svrstamo brojne primere, ali bi oni bili nedovoljno diskriminisani u klasifikatornom smislu.

Prva istraživačica koja se temeljno bavila relacijama u računarskoj lingvistici jeste Rosemary Leonard (1984). Svoju tipologiju imeničkih složenica ona predstavlja kao skup od osam tipova (to su: Sentence, Locative Sentence, Locative, Annex, Equative, Material, Additive, Reduplicative). Računarska implementacija ove autorke sastoji se od leksikona i skupa pravila za semantičku interpretaciju. Leksikon ima četiri skupa semantičkih obeležja koja podržavaju klasifikaciju isključivo binarnih imeničkih složenica. To su: 1. primarna obeležja, koja obuhvataju: (1) ‘has an associated verb’, (2) locative, (3) material, (4) human; 2. sekundarna obeležja koja markiraju glagolske relacije; 3. tercijarna obeležja koja ukazuju na semantičko polje; 4. kvartijarna obeležja koja ukazuju na relativnu veličinu u okviru semantičkog polja. Ova autorka primenjuje svoj program da testira korpus koji obuhvata 445 imeničkih složenica, a koji je podskup njenog

originalnog korpusa koji sadrži 1944 imeničke složenice. U pogledu uspešnosti svog sistema, može se reći da je sistem prilično pouzdan pošto ima 76% korektnih interpretacija, a jedini nedostatak ovog sistema jeste analiza složenica u nasumično odabranom tekstu, odnosno, tekstu koji nije prethodno ručno obrađen u smislu obezbeđivanja neophodnih semantičkih informacija u leksikonu.

Jedan od izazova koji se postaljva pred istraživače u oblasti računarske lingvistike jeste i kompozitnost složenica. Konstatuje se da složenice ispoljavaju kontinuum kompozitnosti. Pošto nam je naročito zanimljiva i relevantna kontrastivna dimenzija ovih istraživanja, onda smo pregledali radove koji se upravo bave predviđanjem kompozitnosti složenica iz kontrastivnog ugla. Konsultovani autori predlažu okvir za analizu predviđanja kompozitnosti složenica upotrebom distribucionih semantičkih modela na materijalu engleskog, francuskog i portugalskog (Cordeiro et al. 2019: 1-2).

Ovo naravno nije iscrpan pregled svih analiza, ali zbog prostorne ograničenosti, zadržali smo se samo na opisanim sistemima. Sada ćemo pogledati druge sisteme prema istraživačkim rezultatima iz računarske lingvistike. U nastavku rada, prikazaćemo sisteme TANKA, HAIKU, SENS, SESEMI, MINDNET, SEMEVAL-2010 i GhoSt-NN.

3. Deskriptivni prikaz sistema TANKA, HAIKU, SENS, SESEMI, MINDNET, SEMEVAL-2010 i G_hoSt-NN

Sistem TANKA (*Text ANalysis for Knowledge Acquisition*) jeste poluautomatski sistem koji može samostalno uvežbavati. Analiza pomoću sistema TANKA prepoznaje semantičke odnose koji su signalizovani putem površinskih jezičkih pojava (Barker, Delisle & Szpakowicz 1998: 60).

Ostali parametri koji su relevantni za našu analizu, a koji su pokriveni sistemom TANKA jesu sledeći. Pomoću ovog sistema postiže se interaktivna semantička reprezentacija tehničkih tekstova. Na dalje, sistem pokriva semantičke relacije centra i modifikatora. Uglavnom su pokrivene kanoničke složenice. Modifikatori u ovom sistemu obuhvataju prideve i predloške fraze. Takođe, sistem prepoznaje semantičke relacije preko površinske strukture. Sistem TANKA vrši detaljnu sintaksičku analizu koristeći javno dostupne liste vrsta reči i leksikone kako bi proizveo tentativnu se-

mantičku analizu. Nedostatak sistema nalazi se u tome što se analiza vrši pretežno na materijalu složenica sa strukturom N+N, odnosno kanoničkih/binarnih/dvokonstituentskih složenica.

Sistem HAIKU izvlači semantičke infomacije iz engleskih tehničkih tekstova. Na dalje, sistem HAIKU vrši poluautomatsku semantičku analizu na tri nivoa: 1. Između klauza, 2. u okviru klauza i 3. u okviru imeničkih fraza. U nedostatku prethodno kodirane semantike, ovaj sistem angažuje i pomoć kooperativnog korisnika koji nadgleda semantičke odluke (Barker 1998: 196).

Sistem HAIKU ima tri podmodula, a to su: 1. sintaksička analiza DIPETT (*Domain Independent Parser of English Technical Texts*), 2. Modul CLR (*Clause Level Relationships*) i 3. Modul NMR (*Noun Modifier Relationships*). Kako bi smanjio opterećenje korisnika, pošto je u pitanju poulautomatski sistem, HAIKU sistem prvo pokušava da izvrši automatsku analizu poređenjem ulaznih struktura koje su slične strukturama u tekstu koji je već semantički analiziran. Pošto sistem HAIKU nema pristup velikoj količini prethodno analiziranih tekstova, započinje akviziciju potrebnih podataka inkrementalno.

Sada ćemo pogledati kako se dalje usavršilo tretiranje relacija u okviru računarske lingvistike kod lingvistkinje Lucy Vanderwende. Ova lingvistkinja proučava relacije sa ciljem da razvije računarski sistem koji bi uspešno tretirao kompleksne sledove, a posebnu pažnju poklanja *imeničkim sledovima* (*noun sequences*) koje ona izjednačava sa složenicama (Vanderwende, 1995: 1). Ona je razvila sistem SENS (System for Evaluating Noun Sequences), odnosno mali skup opštih pravila za interpretiranje imeničkih složenica. Algoritam koji je razvila sastoji se od primene svih pravila na imeničke nizove i evaluacije rezultata na osnovu ponderisanja pripisanih pravilima.

Ova pravila pristupaju semantičkim informacijama za svaku komponentu u okviru složenice, a ove semantičke informacije se sastoje od semantičkih obeležja i semantičkih atributa, koji imaju kompleksne vrednosti. Ona uspostavlja i novu klasifikacionu shemu koja je okosnica analize složenica ove autorke u okviru računarske lingvistike. Klasifikaciona shema se sastoji od četrnaest bazičnih klasa, od kojih je svaka formulisana „wh“ pitanjem.

Nakon uspostavljanja SENS-a i klasifikacione sheme sa četrnaest bazičnih klasa, ova autorka kreira sistem SESEMI (System for Extracting SEMantic Information). Sistem SESEMI zapravo je implementacija pristupa koji automatski stvara veoma detaljne semantičke resurse iz onlajn rečnika. Ovaj pristup obuhvata potpunu sintaksičku analizu teksta definici-

je složenice i obrasce koji identifikuju semantičke informacije tako što ih sravnjuju u odnosu na strukturu koju obezbeđuje analiza. Ovi obrasci identifikuju kako semantička obeležja, koja imaju binarne vrednosti i semantičke atribute.

Sistem SENS vrši interpretaciju imeničkih složenica u neograničenom tekstu. Obrada ovakve vrste teksta ima najmanje dve posledice: 1. značenje reči ne može da se utvrdi pre analize, 2. raspon leksikona zabranjuje ručno kodiranje leksičkih unosaka. SENS algoritam se fokusira na primenu svojih generalnih pravila na značenje centra i modifikatora. Pravila se mogu grupisati prema tome da li testiraju semantičke atribute na centru, modifikator ili deverbalni centar. Autorka konstatuje da iako sistem SENS ne može da pristupi istoj vrsti informacija koje ljudi koriste da interpretiraju složenice, može ipak da definiše skup semantičkih obeležja dovoljnih da se utvrde koja su moguća tumačenja najverovatnija (Vanderwende 1993: 173).

Sistem MINDNET funkcioniše pomoću nekoliko automatski usvojenih baza podataka za leksičko-semantičke relacije. Tvorci sistema MINDNET su članovi Grupe za obradu prirodnog jezika (*Natural Language Processing Group*) u okviru odeljenja Microsoft Research. Sistem koristi MEG (*Microsoft English Grammar*), a korpus je elektronska enciklopedija *Microsoft Encarta*. Stopa uspešnosti je visoka, mada se radi o ograničenom korpusu u okviru enciklopedije *Encarta*. Autori konstatuju da MindNet predstavlja leksički resurs koji se automatski generiše putem obrade teksta. Takođe, za sistem MINDNET nigde se eksplicitno ne navodi da posebno tretira složenice, ali pošto analizira semantičke relacije, svakako se može primeniti i na složenice, između ostalog.

Naročito relevantne su strukture semantičkih relacija koje se u okviru ovog sistema posebno tretiraju. Naime, sprovodi se veoma temeljna analiza, pošto se posmatraju hijerarhijski zbir semantičkih relacija (*semrel*) koji se automatski ekstrahuje iz izvorne rečenice kao kao *semrel* struktura. Svaka *semrel* struktura sadrži sve semantičke relacije ekstrahovane iz jedne izvorne rečenice (Vanderwende, Kacmarcik, Suzuki & Menezes 2005: 8).

Sistem SEMEVAL-2010 ima više zadataka u okviru kojih deluje. Zadatak br. 9 bavi se imeničkim složenicama oslanjajući se na parafraziranje glagola i predloga (Butnariu et al. 2010: 100). Predlaže se zadatak u kome učestvujući sistemi moraju da procene kvalitet parafraze za skup imenič-

kih složenica koji se testira (Butnariu et al. 2010: 103). Pažnja je posvećena, uglavnom, kanoničkim složenicama, tipa *apple pie* i *malaria mosquito*, ali ne zanemaruju ni nekanoničke složenice, poput *caffeine withdrawal headache*. Analiziran je korpus od 250 složenica sa strukturom N+N. Zanimljivo je da je obuhvaćen Levi-250 dataset koji obuhvata primere iz studije Levijeve (Levi 1978), a koji je relevantan za naše primere, pošto diskurs elektrotehnike obiluje i složenicama sa strukturom Adj+N kojima Levijeva daje poseban status, za razliku svih istraživača pre nje, a i brojnih posle nje. Autori sistema SemEval-2010 navode stopu uspešnosti ovog sistema, koja iznosi 68.3%.

Sistem G_hoSt-NN je reprezentativni zlatni standard za nemačke složenice sa strukturom N+N, međutim ima primenu i na engleske složenice tipa *fireworks* i *fruit cake*. Ovaj sistem analizira modifikatori iz leksikona, a centar iz onlajn baze podataka. Sistem analizira ukupno 868 složenica koje su anotirane sa korpusnim frekvencijama. Takođe, pravi distinkciju između centara koji imaju jedan, dva ili više od dva značenja. Autori sistema izveštavaju o stopi uspešnosti od 72%, ali treba napomenuti da je samo manji skup primera iz engleskog, sistem je uglavnom primenjen na nemački jezik. U pogledu kognitivne obrade i reprezentacije složenica, autori eklektički spajaju različite klasifikacije iz reprezentativne i relevantne literature, ali se ipak fokusiraju na sledeće faktore: 1. Faktore zasnovane na frekvenciji, 2. Produktivnost tj. morfološka veličina porodice, 3. Semantičke varijable sagledane kroz odnos modifikatora složenice i njenog centra, 4. Efekat ambigviteta modifikatora i centra (Schulte im Walde et al. 2016: 2285).

Iako su opisani automatski sistemi, posvetili bismo malo pažnje i jednom poluautomatskom prepoznavanju odnosa imeničkih modifikatora. Ovaj sistem za analizu NMR (noun modifier relationship) odnosa pripisuje semantičke odnose u kompleksnim imeničkim frazama. Semantički odnosi između reči i fraza često se markiraju eksplicitnim sintaksičkim ili leksičkim pokazateljima koji pomažu da se ovakvi odnosi prepoznaju u tekstu. Sistemi koji analiziraju takve nominale moraju da kompenzuju nedostatak površinskih indikatora drugom vrstom informacije. Stoga se autori sistema za NMR analizu, s pravom, pitaju: Kako da definišemo leksičku semantiku i izgradimo velike semantičke leksikone? (Barker & Szpakowicz 1998: 96). U pogledu zagrađivanja, sistem citiranih autora zagrađuje imenicu-centar i sled premodifikatora u parove modifikator-centar pre dode-

ljivanja NMR-a. Ono što je bitno istaći, naročito s obzirom da u našem korpusu ima dosta primera složenica sa strukturom „pridev+imenica" i višečlanih složenica, ovaj sistem smatramo relevantnim pošto dozvoljava bilo koji broj prideva ili imenica u ulozi premodifikatora (Barker & Szpakowicz 1998: 98).

Kada poredimo performanse računarskih sistema u veoma grubim crtama, možemo da zaključimo sledeće. Najpre, svaki od pomenutih sistema uglavnom je modelovan tako da analizira samo kanoničke ili samo nekanoničke složenice. Čak i kada su šire koncipirani i obuhvataju kanoničke i nekanoničke složenice, u datom trenutku ovi sistemi ne mogu da solidno vrše analizu obe kategorije. U narednom delu pogledaćemo konkretnu primenu sistema SENS na analizu složenica u diskursu elektrotehnike.

4. Primena sistema SENS na analizu složenica u diskursu elektrotehnike

Pogledali smo sve klasifikacije iz analiziranih radova i konstatovali da sistem SENS može da se primeni u zadovoljavajućoj meri na složenice iz diskursa elektrotehnike. Razlozi su sledeći. Algoritam sistema SENS primenjuje opšta pravila na značenje centra i modifikatora. Ova pravila se mogu grupisati prema tome da li testiraju semantičke atribute na centru, modifikator ili pak deverbalni centar. U ovom sistemu, relacije su iskazane putem „wh" pitanja[3]. Videćemo u sledećoj tabeli, koju smo preuzeli iz literature, ali smo modifikovali i proširili našim primerima iz diskursa elektrotehnike.

„Wh" pitanja sistema SENS za semantičke relacije	Konvencionalni naziv relacije	Primer složenice u opštem diskursu	Primer složenice u diskursu elektrotehnike
Who/what?	Subject	*press report*	*cathode drop*
Whom/what?	Object	*accident report*	*commutation switch*
Where?	Locative	*garden party*	*bus topology*

3

When?	Time	*night attack*	*daylight*
Whose?	Possessive	*family estate*	*collision frequency*
What is it part of?	Whole-Part	*duck foot*	*picture element*
What are its parts?	Part-Whole	*daisy chain*	*network software*
What kind of?	Equative	*flounder fish*	*emitter follower*
What with?	Instrument	*paraffin cooker*	*plasma display*
What for?	Purpose	*bird sanctuary*	*field coil*
Made of what?	Material	*alligator shoe*	*crystal filter*
What about?	Topic	*budget vote*	*network topology*
What does it cause?	Causes	*disease germ*	*noise diode*
What causes it?	Caused-by	*drug death*	*induction balance*

Tabela 4.1. „Wh" pitanja za semantičke relacije sistema SENS, konvencionalni nazivi relacija sa primerima složenica u opštem diskursu i diskursu elektrotehnike.

Kao što se to može videti u tabeli 4.1., složenice iz diskursa elektrotehnike mogu da se uklope u semantičke relacije ovog sistema. Međutim, u okviru ovog sistema, donošenje odluke o pripadnosti određenoj klasi prilično je otežano, čak i za ljudske anontatore i parsere. Na primer, kako da odlučimo da li konstituent pripada relaciji „Lokativ" ili „Celina-Deo" (*Whole-Part*), pošto u prethodnoj literaturi niko nije ponudio nikakve kriterijume za donošenje ovakvih odluka (Vanderwende, 1995: 39). Upravo zbog toga, citirana autorka predlaže niz testova koji mogu da posluže kao heuristika prilikom odlučivanja da li je SENS klasifikacija odgovarajuća, mada napominje da u odsustvu konteksta, ljudsko rasuđivanje može da ostane neubedljivo. Test sa „*wh*" pitanjima i odgovorima koristi se kao kriterijum za donošenje odluke da li je složenica primereno klasifikovana, mada se ne prenebregavaju i drugi testovi, poput testa parafraze i testa koordinacije.

Skup pravila sistema SENS obezbeđuje najverovatniju interpretaciju složenice tako što testira konfiguraciju semantičke informacije i na modifikatoru i na centru složenice. Relevantnu ulogu ima skup semantičkih obeležja i atributa koji su dovoljni za interpretaciju imeničkih složenica. Utvr-

đeno je ukupno 27 semantičkih obeležja i atributa, koji su neophodni za intrepretaciju putem SENS-a, čiju primenu možemo videti u sledećoj tabeli.

Br. obeležja	Semantičko obeležje i atribut
1	±
2	caused-by
3	has-object
4	±
5	instrument-for
6	location-of
7	±
8	purpose
9	subject-of
10	by-means-of
11	causes
12	has-part
13	hypernym
14	located-at
15	made-into
16	object-of
17	role
18	±
19	classifier
20	±
21	has-subject
22	±
23	±
24	made-of
25	part-of
26	±
27	time-of

Tabela 4.2. Obeležja i atributi prema Vanderwende (1995: 257).

Citirana autorka je i u jednom ranijem radu utvrdila 17 relacija (Ab-

stract, Caused–By, Event, Food, Group–Noun, Has–Object, Has–Subject, Human, Is–For, Location–Noun, Location–Of, Made–Of, Material, Means, Object–Of, Subject–Of, Time) i utvrdila je 11 semantičkih relacija između imenica u složenicama (1. Who/What? [Subject of a deverbal head], 2. Whom/What? [Object of a deverbal head], 3. Where [Locative], 4. When? [Time], 5. Whose? [Possessive], 6. How? [Instrument], 7. What for? [Purpose], 8. What kind of? [Equi/General], 9. Made of what? [Material], 10. What does it cause? [Causes], 11. What causes it? [Caused-by]) (Vanderwende 1994).

Kada smo provukli primere iz korpusa, uglavnom su se uklapali u klasifikacije sistema SENS i ostale klasifikacije ove autorke, što ne znači nužno da i drugi sistemi nisu u istoj meri solidni. Međutim, moramo da napomenemo da je ovaj sistem primenjiv i na francuske složenice. Pa tako za engleske primere navedene u tabeli 4.1. pronalazimo francuske ekvivalente koji se uklapaju u klasifikacije citirane autorke: *goutte de cathode*, *commutateur de commutation*, *topologie de bus*, *lumière du jour*, *fréquence de collision*, *élément d'image*, *logiciel de réseau*, *émetteur suiveur*, *écran plasma*, *bobine de champ*, *filtre à cristaux*, *topologie de réseau*, *diode de bruit*, *balance d'induction*.

Dakle, vidimo da sistem SENS može uspešno da klasifikuje i tretira i izvesne francuske složenice. Te smo stoga skloni da zaključimo da je pogodan za engleske i francuske složenice u diskursu elektrotehnike. U narednom delu rada iznosimo tentativne zaključne napomene.

5. Zaključne napomene

Pre same analize pogledali smo različite sisteme, prikazane u relevantnoj literaturi, no učinila nam se možda najprikladnijom analiza koju vrši sistem SENS. Nije zanemarljiva razlika između sistema SENS i drugih sistema, pošto SENS ima višestruke klase tamo gde ostali sistemi klasifikacije imaju jednu ili dve. Naravno, ovaj naš nalaz ograničen je na primere složenica u diskursu elektrotehnike, i ne pravimo šire generalizacije, barem ne na sadašnjem stupnju istraživanja. Sledeća prednost sistema SENS u odnosu na druge sisteme je to što on ima jedinstvenu klasu tamo gde drugi sistemi prave više distinkcija. Treće, postoje neke složenice koje ne

mogu da se tretiraju sistemom klasifikacije koji predlaže SENS.

Ukoliko posmatramo razvoj klasifikacija u istraživanjima prema rezultatima iz računarske lingvistike, može se uočiti da sistemi koji se zasnivaju na tradicionalnim klasifikacijama uglavnom značenje konstituenata analiziraju na vrlo opštem nivou sa očiglednim preplitanjem tradicionalnih gramatičkih kategorija (kao što su subjekat i objekat) i semantičkih (kao što je npr. *material*). Kasnije klasifikacije koji se oslanjaju na generativnu gramatiku (npr. okvir predložen u Levi 1978), bivaju preciznije određene i usložene.

Takođe, ako se uporede pomenuti sistemi, kod semantičke interpretacije neki od njih insistiraju na parafrazi, a drugi predlažu nešto drugačija rešenja, pošto se parafraze smatraju nepotpunim mehanizmom. Onda se neki sistemi fokusiraju se na funkcionalne uloge. U nekim od radova iz oblasti računarske lingvistike, složenice se posmatraju kao semantički okviri koji se ispunjavaju ili zahtevanim ispunjivačem ili preferiranim ispunjivačem, dok drugi sistemi podrazumevaju računarsku obradu kojom je obuhvaćen leksikon, ali i skup pravila za njegovu interpretaciju.

Ipak je možda najprimenljiviji sistem za složenice u diskursu elektrotehnike SENS. Ovaj sistem obuhvata skup opštih pravila kojima se mogu interpretirati složenice, ali u najvećoj meri se odnosi na imeničke složenice.

U našem radu smo se oslonili na računarsko-lingvističke i izvesne generativne pristupe složenicama, pošto smo najviše podataka o semantičkoj klasifikaciji našli upravo u okvirima ovih pristupa. Iako svaki od tih pristupa na neki način upućuje na određeni tip semantičkih relacija u složenici, nijedan od njih zapravo ne razrešava pitanje na koji način možemo sa sigurnošću interpretirati semantičke odnose između konstituenata jedne složenice.

U sistemu SENS i ostalim pomenutim sistemima primećeni su različiti pristupi analizi semantičkih relacija konstituenata unutar složenica. Naravno, dodatni izazov je kontrastivna analiza engleskih i francuskih složenica, pošto se javlja asimetrija u pogledu centra. Iako su autori primenili različite tehnike, eksperimentalni rezultati su i dalje daleko od potpuno zadovoljavajućih, naročito ako se imaju u vidu nekanoničke složenice koje postavljaju zadatak zagrađivanja. Čini se da je potrebno uključiti još teorijskih pristupa, poput onoga koji predlaže Levijeva, ali i vršiti dalje empi-

rijske provere, kako bi se utvrdile uspešnosti datih sistema. Naravno, ove zaključne napomene nisu definitivne, već tentativne i zahtevaju dalje elaboracije i rafiniranije klasifikacije semantičkih odnosa konstituenara kod kanoničkih i nekanoničkih složenica u engleskom i francuskom jeziku u specifičnim diskursima.

Literatura:

Aitchison, James. *Cassell's Dictionary of English Grammar*. London: Cassell & Co, 2001. Print.

Aronoff, Mark & Kirsten Fudeman. *What is Morphology?*. Malden, MA: Blackwell, 2005. Print.

Barker, Ken. "A trainable bracketer for noun modifiers." *Advances in Artificial Intelligence: The 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, Eds. Robert E. Mercer & Eric Neufeld. Vancouver: Springer, 1998. 196–210. Print.

Barker, Ken & Stan Szpakowicz. "Semi-automatic recognition of noun modifier relationships." *Proceedings of The Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics* (*COLING-ACL'98*), Eds. Christian Boitet & PeteWhitelock. San Francisco, CA: Morgan Kaufmann Publishers, 1998. 96–102. Print.

Barker, Ken, Sylvain Delisle & Stan Szpakowicz. "Test-driving TANKA: Evaluating a Semi-Automatic System of Text Analysis for Knowledge Acquisition." *Advances in Artificial Intelligence: The 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, Eds. Robert E. Mercer & Eric Neufeld. Vancouver: Springer, 1998. 60-71. Print.

Bartolić, Ljerka. „Imenske složenice u tehničkom engleskom jeziku." *Strani jezici*, VII, 1-2, (1979): 47-58. Print.

Bugarski, Ranko. *Uvod u opštu lingvistiku. Sabrana dela – Knjiga 6*. Beograd: Čigoja & XX vek, 1996. Print.

Butnariu, Cristina, Su Nam Kim, Preslav Nakov, Dairmuid Ó Séaghdha, Stan Szpakowicz & Tony Veale. "Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions." *Proceedings of the 5th International Workshop on Semantic Evaluation – SemEval-2010*, Eds. Katrin Erk & Carlo Strapparava. Uppsala: Association for Computational Linguistics, 2010. 39–44. Print.

Chomsky, Noam & Morris Halle. *The Sound Pattern of English*. The First MIT Press Paperback Edition, Third Printing. Cambridge, Massachusetts: The MIT Press, 1995. Print.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart & Carlos Ramisch. "Unsupervised compositionality prediction of nominal compounds." *Computational Linguistics*, Volume 45, Issue 1, (2019): 1-57. Print.

Di Sciullo, Anna-Maria. "Decomposing compounds." *Skase Journal of Theoretical Linguis-*

tics, Vol. 2, No. 3, (2005): 14-33. Print.
Dubois, Jean & René Lagane. *La nouvelle grammaire du français*. Paris: Larousse et SEJER, 2004. Print.
Đokić, Nada. *Rečnik lingvističke terminologije: srpsko-francusko-englesko-nemački*. Beograd: Mrlješ d.o.o., 2001. Print.
Đurić, Miloš D. „Kanoničke i nekanoničke složenice u engleskom i francuskom jeziku." *Prevodilac*, XXXIII, 76, 3-4, (2016): 22-39. Print.
Fabb, Nigel. "Compounding." *Handbook of Morphology*. Eds. Andrew Spencer & Arnold M. Zwicky. Oxford: Blackwell, 1998. 66-83. Print.
Finin, Timothy Wilking. "The semantic interpretation of nominal compounds." *Proceedings of the First National Conference on Artificial Intelligence*. Stanford and Menlo Park, California: AAAI Press, 1980. 310-315. Print.
Fišer-Popović, Ana. „O nekim semantičkim osobinama imeničkih složenica u engleskom jeziku tehničkih nauka." *Strani jezik struke u teoriji i praksi*. Ed. Nadežda Vinaver. Beograd: Udruženje univerzitetskih nastavnika i drugih naučnih radnika Srbije i Osnovna organizacija nastavnika stranih jezika na nefiloloških fakultetima i višim školama, 1981. 160-167. Print.
Fromkin, Victroria & Robert Rodman. *An Introduction to Language*. Third Edition. New York: Holt, Rinehart and Winston, 1983. Print.
Grevisse, Maurice & André Goosse. *Nouvelle grammaire français*. 3e édition. Louvain-la-Neuve et Paris: DeBoeck et Duculot, 1995. Print.
Isabelle, Pierre. "Another look at nominal compounds." *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics*. Ed. Yorick Wilks. Stanford, California: The Association for Computational Linguistics, 1984. 509-516. Print.
Kristal, Dejvid. *Enciklopedijski rečnik moderne lingvistike*. Beograd: NOLIT, 1988. Print.
Kristal, Dejvid. *Kembrička enciklopedija jezika*. Beograd: NOLIT, 1996. Print.
Lardiere, Donna. "Words and their parts." *An Introduction to Language and Linguistics*. Ed. Ralph W. Fasold and Jeff Connor-Linton. Cambridge: Cambridge University Press, 2006. 55-96. Print.
Levi, Judith N. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press, 1978. Print.
Marcellesi, Jean-Baptiste. "Le lexique." *La linguistique*. Ed. Pierre Caussat et al. Paris: Larousse, 1977. 187-197. Print.
Master, Peter. "Noun compounds and compressed definitions." *English Teaching Forum*, Vol. 41, No. 3, (2003): 2-9. Print.
Matthews, P. H. *The Concise Oxford Dictionary of Linguistics*. Reissued. Oxford: Oxford University Press, 2005. Print.
Mitterand, Henri. *Les mots français*. Paris : Presses universitaires de France, 1972. Print.
Le Micro Robert. Paris: Dictionnaires Robert, 1988. Print.
Napoli, Donna Jo. *Linguistics – An Introduction*. New York and Oxford: Oxford University Press, 1996. Print.
Pinker, Steven. *The Language Instinct – How Mind Creates Language*. London: Harper

Perennial, 1995. Print.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. *A Grammar of Contemporary English*. Sixth Impression. London: Longman, 1974. Print.
Schulte im Walde, Sabine, Anna Hätty, Stefan Bott & Nana Khvtisavrishvili. "G_hoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds." *Proceedings of the 10th International Conference on Language Resources and Evaluation LREC 2016*. Eds. Nicoletta Calzolari et al. Portorož: European Language Resources Association, 2016. 2285–2292. Print.
Smith, Neil & Deirdre Wilson. *Modern Linguistics: The Results of Chomsky's Revolution*. Hardmondsworth: Penguin, 1979. Print.
Škiljan, Dubravko. *Pogled u lingvistiku*. Zagreb: Školska knjiga, 1980.
The Penguin English Dictionary. Harmondsworth: Penguin Books Ltd., 1985. Print.
Trask, R. L. *The Penguin Dictionary of English Grammar*. London: Penguin Books, 2000. Print.
Vanderwende, Lucy. "SENS: The System for Evaluating Noun Sequences." *Natural Language Processing: The PLNLP Approach*. Eds. Karen Jensen, George E. Heidorn & Stephen D. Richardson. Boston: Kluwer Academic Publishers, 1993. 161-173. Print.
Vanderwende, Lucy. "Algorithm for automatic interpretation of noun sequences." *Proceedings of the Fifteenth International Conference on Computational Linguistics* (*COLING 1994*). Eds. Makoto Nagao & Yorick Wilks. Kyoto, 1994. 782-788. Print.
Vanderwende, Lucretia H. *The Analysis of Noun Sequences Using Semantic Information Extracted from On-Line Dictionaries*. *Technical Report MSR-TR-95-57*. Redmond, WA: Microsoft Research and Microsoft Corporation, 1995. Print.
Vanderwende, Lucy, Gary Kacmarcik, Hisami Suzuki & Arul Menezes. "MindNet: An automatically-created lexical resource." *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Eds. Donna Byron, Anand Venkataraman & Dell Zhang. Vancouver: The Association for Computational Linguistics, 2005. 8–19. Print.
Vujić, Jelena. „Morfološka analiza bipolarnih imeničkih složenica u kompjuterskom registru." *Lingvističke analize: Zbornik u čast 25 godina Instituta za strane jezike u Podgorici*. Ed. Slavica Perović & Vesna Bulatović. Podgorica: Institut za strane jezike, 2004. 155-156. Print.
Ziering, Patrick René. *Indirect Supervision for the Determination and Structural Analysis of Nominal Compounds*. Stuttgart: Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart, 2018. Print.

Miloš D. Đurić

SOME ASPECTS OF COMPOUNDS AND COMPOUND ANALYSIS SYSTEMS IN COMPUTATIONAL LINGUISTICS

Summary

English lexical items such as *electric circuit*, *ion path*, *interrupting capacity*, and French lexical items such as *circuit électrique*, *trajectoire ionique*, *intensité maximum de rupture* have been studied in the pertinent linguistic literature under the labels of *compounds*, *complex nominals*, *mots composés*, *complex noun phrases*, to name just a few. Diverse classifications and taxonomies have been proposed in the literature in order to identify semantic relations within the panoply of interpretations these terms are supposed to display. Nevertheless, these attempts resulted in proliferation of taxonomies that were aimed at capturing the syntax and semantics of compounds, which were at best arbitrary, which ultimately obfuscated rather than illuminated the phenomenon of compounds. With the advent of Levi's generative semantics framework for complex nominals and subsequent computational linguistics investigation these relations were described more precisely.

The first part of my paper is a descriptive exploration of the term 'compound'. The second part provides a concise overview of semantic classifications of compounds based on computational linguistic research data. The third part is a descriptive survey of the systems TANKA, HAIKU, SENS, SESEMI, MINDNET, SEMEVAL-2010 i G_hoSt-NN. The fourth part briefly illustrates the application of the sens system to the analysis of compounds in electrical engineering discourse. The fifth part provides certain relevant concluding remarks.

Key words: Compounds, Semantic Relations, Compound Analysis Systems, MINDNET, SENS, SESEMI, TANKA, HAIKU, SEMEVAL-2010, G_hoSt-NN, Computational Linguistics.