UNIVERZITET U BEOGRADU
FILOLOŠKI FAKULTET

Jelena B. Jaćimović

AUTOMATSKO PREPOZNAVANJE I NORMALIZACIJA VREMENSKIH IZRAZA U NESTRUKTURIRANIM NOVINSKIM I MEDICINSKIM TEKSTOVIMA NA SRPSKOM JEZIKU

doktorska disertacija

Beograd, 2016

UNIVERSITY OF BELGRADE
FACULTY OF PHILOLOGY

Jelena B. Jaćimović

AUTOMATIC RECOGNITION AND NORMALIZATION OF TEMPORAL EXPRESSIONS IN SERBIAN UNSTRUCTURED NEWSPAPER AND MEDICAL TEXTS

Doctoral Dissertation

Belgrade, 2016

Mentor:

Dr Cvetana Krstev, redovni profesor
Univerzitet u Beogradu, Filološki fakultet

Članovi komisije:

Dr Miloš Utvić, docent
Univerzitet u Beogradu, Filološki fakultet

Dr Rajna Dragićević, redovni profesor
Univerzitet u Beogradu, Filološki fakultet

Dr Đurica Grga, vanredni profesor
Univerzitet u Beogradu, Stomatološki fakultet

Dr Duško Vitas, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Mihajlu R. Iliću

Zahvalna sam profesorki Cvetani Krstev na tome što me je povela na pionirsko putovanje kroz prostore računarske obrade prirodnih jezika i na tom putu imala strpljenja, volje i snage da me podrži i motiviše kada posustanem. Svojim entuzijazmom inspirisala me je da istrajem u mom radu i hrabro krenem put novih ideja.

Automatsko prepoznavanje i normalizacija vremenskih izraza u nestrukturiranim novinskim i medicinskim tekstovima na srpskom jeziku

Rezime

Ljudi u svakodnevnom životu koriste vreme kao univerzalni referentni sistem, u okviru koga se događaji ili stanja nižu jedan za drugim, utvrđuje dužina njihovog trajanja i navodi kada se neki događaj desio. Značenje vremena i način na koji čovek poima vreme ogledaju se i u komunikaciji, pre svega, u jezičkim izrazima koji se učestalo koriste u svakodnevnom govoru. Vremenski izrazi, kao fraze prirodnog jezika koje na direktan način ukazuju na vreme, pružaju informaciju o tome kada se nešto dogodilo, koliko dugo je trajalo ili koliko često se dešava.

Uporedo s razvojem informatičkog društva, povećava se i količina slobodno dostupnih digitalnih informacija, što daje veće mogućnosti pronalaženja potrebnih informacija, ali i utiče na složenost ovog procesa, iziskujući korišćenje naprednih računarskih alata i moćnijih metoda automatske obrade tekstova prirodnih jezika. S obzirom na to da se značenje većine elektronskih informacija menja u zavisnosti od vremena iskazanog u njima, radi uspešnog razumevanja tekstova pisanih prirodnim jezikom, neophodno je korišćenje alata koji su sposobni da automatski označe i informacije koje referišu na vreme i omoguće uspostavljanje hronološkog sleda opisanih događaja. Stoga je potrebno razviti alate namenjene ekstrakciji vremenskih izraza, kod kojih su preciznost i odziv na visokom nivou i koji se brzo i jednostavno mogu prilagoditi novim zahtevima ili tekstovima drugog domena. Postojanje ovakvog sistema može u velikoj meri uticati na poboljšanje učinka primene mnogih drugih aplikacija iz oblasti jezičkih tehnologija (ekstrakcija informacija, pronalaženje informacija, odgovaranje na pitanja, rezimiranje teksta itd.), ali i doprineti očuvanju srpskog jezika u savremenom digitalnom okruženju.

Osnovni cilj ovog istraživanja jeste izgradnja alata koji vrši automatsko ekstrahovanje lingvističkih izraza vremenskog značenja u nestrukturiranim tekstovima standardnog srpskog jezika, sa postizanjem visokog nivoa odziva i preciznosti, a koji se na jednostavan način može prilagoditi novim i drugačijim zahtevima. Imajući to u vidu, drugi od postavljenih ciljeva odnosi se na procenu efikasnosti

primene kreiranog alata i u domenu medicinskih narativnih tekstova.

Elementi formalne strukture najčešćih oblika pojavljivanja vremenskih izraza standardnog srpskog jezika opisani su na osnovu analize primera identifikovanih u korpusu novinskih tekstova. Naprednim metodama ekstrakcije informacija definisana su pravila, organizovana u vidu složenih gramatika plitkog parsiranja, koje analizirajući lokalni kontekst potencijalnih vremenskih izraza vrše njihovu identifikaciju, određivanje opsega i tipa. Metodom konačnih stanja izvršena je i normalizacija vrednosti prepoznatih izraza, koje se predstavljaju u standardizovanom, strukturiranom obliku, nezavisnom od datog lingvističkog izraza i u skladu sa opšteprihvaćenom ISO-TimeML shemom.

Za evaluaciju uspešnosti sistema u automatskoj obradi vremenskih izraza srpskog jezika korišćena je kolekcija novinskih tekstova, koji nisu upotrebljavani u fazi obuke sistema. Procena uspešnosti sistema u prepoznavanju i normalizaciji vremenskih izraza novinskih tekstova srpskog jezika izvršena je na osnovu standardnih mera za procenu učinka sistema za pronalaženje i ekstrakciju informacija — preciznosti, odziva i *F* mere. U pogledu uspešnosti sistema u prepoznavanju vremenskih izraza, odnosno određivanju opsega i tipa, sistem je postigao veoma visoku preciznost (99%) u odnosu na odziv (80%), uz ukupnu F_1 meru od 88%. Postignuti rezultati u pogledu uspešnosti procesa normalizacije vremenskih izraza su izuzetno dobri (F_1 mera je 99,7%), imajući u vidu činjenicu da su vrednosti svih korektno prepoznatih vremenskih izraza ispravno normalizovane, dodeljivanjem svih odgovarajućih atributa neophodnih za njihovu interpretaciju.

Učinak kreiranog sistema u prepoznavanju vremenskih izraza medicinskih narativnih tekstova veoma je dobar, uz postignutu F_1 meru od 92%. Kada je reč o normalizaciji prepoznatih izraza, kao i u slučaju novinskih tekstova, identifikovane greške ukazuju na to da ovaj proces u potpunosti zavisi od uspešnosti prepoznavanja, odnosno ispravnog određivanja opsega i tipa vremenskih izraza, s obzirom na to da su vrednosti svih korektno identifikovanih i obeleženih izraza na ispravan način i normalizovane.

Osnovni rezultat ovog istraživanja je proizvod u vidu alata, koji bez prethodne pripreme i sa visokom preciznošću i odzivom, vrši automatsko prepoznavanje i normalizaciju vremenskih izraza u tekstovima srpskog jezika. Zahvaljujući postignutoj visokoj preciznosti, ovim alatom je omogućeno obeležavanje proizvoljne

količine tekstova, koji se mogu koristiti za obuku i testiranje mašinski zasnovanih metoda namenjenih automatskoj obradi vremenskih informacija, kao i za lingvističku analizu određenih vremenskih fenomena. Kreirani alat se može koristiti i jednostavno prilagoditi primeni u domenu medicine, za potrebe automatske obrade vremenskih izraza medicinskih narativnih tekstova.

Ključne reči: računarska lingvistika, obrada prirodnog jezika, imenovani entiteti, konačni transduktori, srpski jezik

Naučna oblast: Bibliotekarstvo i informatika

Uža naučna oblast: Pronalaženje informacija

UDK broj: 811.163.41'322.2:[070+61](043.3)

Automatic Recognition and Normalization of Temporal Expressions in Serbian Unstructured Newspaper and Medical Texts

Summary

People in everyday life use time as a universal reference system, within which, events or states are sequenced one after the other, it is established how long they lasted and it is stated when an event occurred. The meaning of time and the way humans perceive time is reflected in communication, most of all, in linguistic expressions frequently used in everyday speech. Temporal expressions, as natural language phrases which directly refer to time, provide information on when something happened, how long it lasted and how often it occurs.

Alongside with the information society development, the amount of freely available digital information has increased, which provides a greater possibility of finding the necessary information, but also affects the complexity of this process, by requiring the use of advanced computer tools and more powerful natural language text processing methods. Having in mind that the meaning of most electronic information can change depending on time expressed in them, it is essential to use tools which can both automatically mark the information related to time and enable the establishment of chronological order of described events. Therefore, it is necessary to develop tools for extraction of temporal expressions with high levels of precision and recall, which can be easily and quickly adapted to new demands and texts from different domains. The existence of such a system can, to a great extent, affect the effectiveness improvement in implementation of many other applications from the field of language technology (information extraction, information retrieval, question answering, text summarization, etc.), but also contribute to the preservation of the Serbian language in the contemporary digital environment.

The basic goal of this research is building a tool which performs automatic extraction of linguistic expressions with temporal meaning in unstructured texts of the standard Serbian language, which achieves a high level of recall and precision and is easily adaptable to new and different demands. Having that in mind,

the second of the set goals refers to the evaluation of efficiency of the created tool in the domain of medical narrative texts as well.

The elements of formal structure of the most common forms of the standard Serbian language temporal expressions are described on the basis of analysis of examples taken from newspaper texts. The rules are defined by using advanced information extraction methods and organized in the form of complex shallow parsing grammars, which, by analysing the local context of potential temporal expressions, perform their identification, range and type determination. By using the Finite-State methodology, normalization of recognized expressions' values was also performed and those values will be represented in a standardized, structured form, independent of the given linguistic expression and in accordance with the accepted ISO-TimeML scheme.

A collection of newspaper texts, which hadn't been used in the training phase of the system, was used for the evaluation of system's performance in automatic processing of the Serbian language temporal expressions. The evaluation of system's success in recognition and normalization of temporal expressions in newspaper texts of the Serbian language was performed based on the standard performance measures, used for evaluation of a system for finding and extracting information - precision, recall and *F* measure. Regarding the system's success in recognition of temporal expressions, that is, range and type determination, the system achieved very high precision (99%) regarding recall (80%), with the overall F_1 measure of 88%. The results achieved regarding the system's success in normalizing temporal expressions are exceptionally good with F_1 measure of 99,7%, having in mind the fact that the values of all the correctly recognized temporal expressions are accurately normalized, by assignation of all the corresponding attributes necessary for their their interpretation.

The effectiveness of the created system in recognition of temporal expressions of medical narrative texts is very good, with achieved F_1 measure of 92%. Speaking of normalization of recognized expressions, as well as in the case of newspaper texts, the identified errors point to the fact that this process is completely dependent on recognition success, that is, correct range and type determination of temporal expressions, having in mind that the values of all the correctly identified and marked expressions, are also accurately normalized.

The basic result of this research is a product in the form of a tool, which, without prior preparation and with high precision and recall, performs automatic recognition and normalization of temporal expression in texts of the Serbian language. Thanks to the high precision achieved, this tool enables annotation of an arbitrary quantity of texts, which can be used for training and testing of machine-learning methods designed for automatic processing of temporal information, as well as for linguistic analysis of certain temporal phenomena. The created tool can be used and easily adjusted to application in the field of medicine, for the purpose of automatic processing of temporal expressions of medical narrative texts.

Key words: computational linguistics, natural language processing, named entities, finite-state transducers, Serbian language

Research area: Library and Information Science

Research subarea: Information Retrieval

UDC number: 811.163.41'322.2:[070+61](043.3)

Sadržaj

Glava 1

Uvod

Počev od druge polovine prošlog veka, ostvarena dostignuća na području informacionih i komunikacionih tehnologija uslovila su pojavu digitalne revolucije i stvaranje informatičkog društva pred kojim su novi izazovi. Stalni rast količine pohranjenih informacija prevazilazi mogućnosti pojedinca da u obilju slobodno dostupnih digitalnih informacija pronađe one relevantne i u potpunosti iskoristi potencijal novih tehnologija. Sa porastom količine dostupnih podataka povećava se mogućnost, ali istovremeno i složenost procesa pronalaženja potrebnih informacija. Kako bi potencijal ovako bogatog izvora znanja bio iskorišćen u što većoj meri, potrebno je koristiti napredne računarske alate i moćnije metode automatske obrade dostupnih jezičkih izvora.

Računari trenutno uspešno obrađuju informacije date u strukturiranom ili polustrukturiranom obliku, poput baze podataka ili XML dokumenta, u okviru kojih je značenje relevantnih informacija kodirano, odnosno predstavljeno u standardnom nedvosmislenom obliku, jednostavnom za mašinsku obradu. Međutim, većina dostupnih informacija je danas ipak zabeležena u nestrukturiranom obliku prirodnim jezikom, odnosno jezikom koji ljudi koriste za svakodnevnu komunikaciju. Stoga je potreba za razvojem softverskih sistema koji su projektovani za rad sa prirodnim jezicima i koji su sposobni da transformišu informacije iskazane prirodnim jezikom u strukturiran oblik sve izraženija, o čemu svedoči obimna literatura iz oblasti jezičkih tehnologija.

Osnovni cilj razvoja sistema zasnovanih na metodama obrade prirodnih jezika (eng. *natural language processing*) i računarske lingvistike (engl. *computational linguistics*) jeste olakšana interakcija između ljudi i računara. Imajući u vidu višeznačnost koja karakteriše prirodne jezike, pravi je izazov formalizovati ih bez

gubitka informacija i omogućiti odvijanje nedvosmislene komunikacije. Jedan od preduslova uspešnog razumevanja tekstova pisanih prirodnim jezikom jeste i mogućnost utvrđivanja hronološkog sleda događaja opisanih nekim tekstom. Sistemi za automatsku obradu vremenskih informacija (eng. *temporal processing systems*) usmereni su na identifikaciju izraza koji ukazuju na vreme, događaje i vremenske odnose koji postoje među njima. Tokom poslednjih godina sve je izraženija potreba za alatima namenjenim ekstrakciji vremenskih informacija, kod kojih su preciznost i odziv na visokom nivou i koji se brzo i jednostavno mogu prilagoditi novim zahtevima ili tekstovima drugog domena.

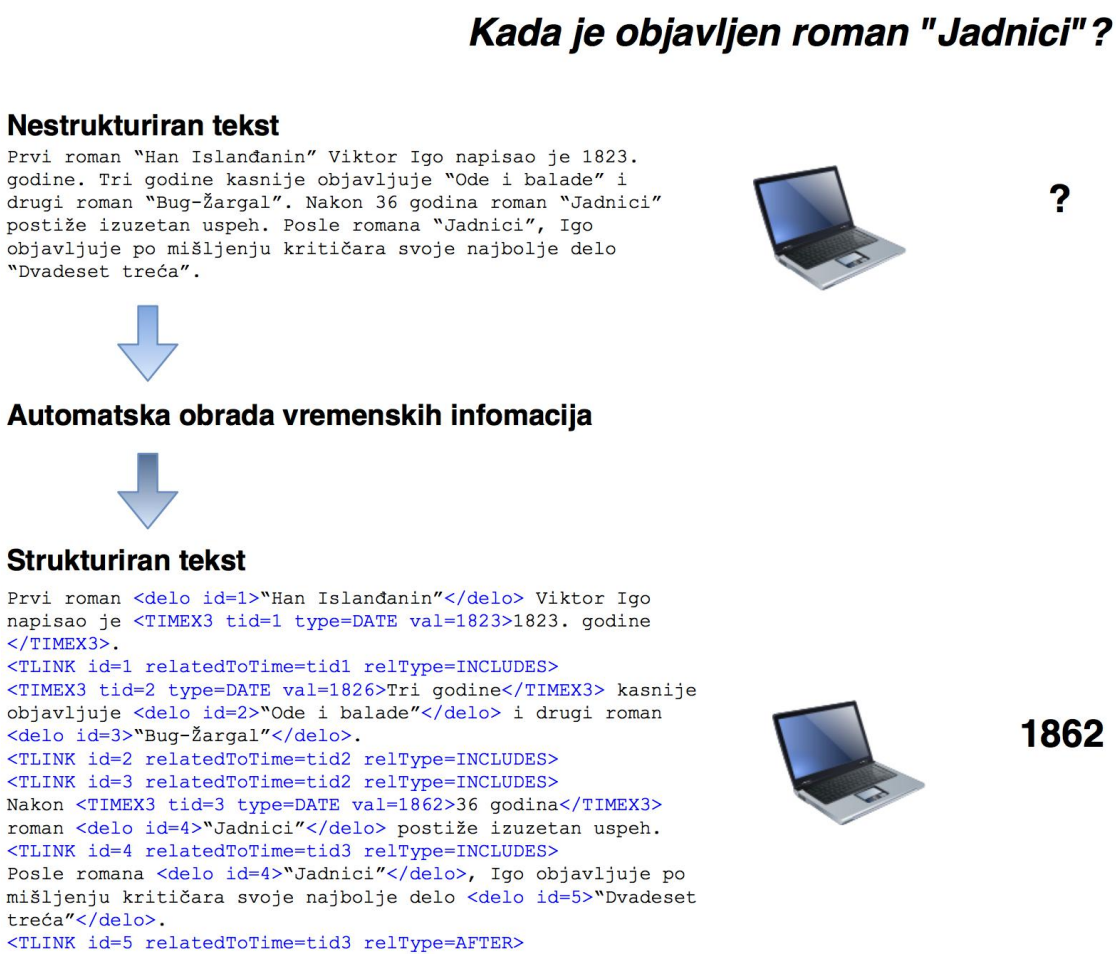
1.1 O automatskoj obradi vremenskih informacija

Vreme je oduvek bilo predmet rasprava u okviru religije, filozofije i raznih oblasti nauke (fizike, istorije, lingvistike, logike i psihologije). Na osnovna pitanja o prirodi pojma vremena, naučnici daju veoma različite odgovore, te je stoga veoma teško usvojiti jasnu definiciju vremena, koja bi bila primenljiva u okviru svih naučnih disciplina. Ipak, bez obzira na postojeće poteškoće oko definisanja ili tumačenja pojma vremena, ljudi sa lakoćom manipulišu vremenom u stvarnom svetu. Svet je dinamičan po svojoj prirodi, i upravo vreme nam omogućava da razumemo na koji se to način svet menja. Stvari koje se dešavaju i menjaju (*događaji*) ili situacije koje ostaju nepromenjene tokom određenog vremenskog perioda (*stanja*) međusobno su povezane upravo vremenom na koje ukazuju. Ljudi se koriste pojmom vremena kako bi nizali događaje ili stanja jedna za drugim, utvrdili dužinu trajanja nekog događaja ili stanja i naveli kada se neki događaj desio. Dakle, čini se da vreme igra ulogu univerzalnog referentnog sistema koji se koristi za povezivanje,ređanje, merenje i upoređivanje intervala događaja i stanja (Marşic 2011).

Oblast automatske obrade vremenskih informacija jeste tema koja privlači sve više pažnje tokom poslednjih godina. Osnovni cilj istraživanja u okviru ove oblasti jeste automatska identifikacija svih izraza koji ukazuju na vreme, kao i identifikacija i uspostavljanje hronologije događaja nekog teksta. Stoga je potrebno identifikovane vremenske izraze, događaje i vremenske relacije transformisati u informacije strukturiranog oblika, kako bi mogle kasnije da se koriste u okviru nekih složenijih aplikacija.

Osnovni motiv za sprovođenje automatske obrade vremenskih informacija ilustrovan je primerom na slici 1.1. U okviru datog teksta se govori o publikova-

nim romanima Viktora Igoa, koji predstavljaju niz događaja smeštenih u određeno vreme. Radi strukturiranog prikaza datih informacija, potrebno je uspostaviti veze koje postoje između pomenutih romana i datuma njihovog objavljivanja.



Slika 1.1: Primer primene automatske obrade vremenskih informacija

Složenost ovog zadatka se ogleda, pre svega, u višeznačnosti prirodnog jezika. Događaji (etiketa `<delo>`), izrazi koji ukazuju na vreme (etiketa `<TIMEX3>`), kao i odnosi koji postoje među njima (etiketa `<TLINK>`) često su jasno određeni, kao u rečenici *Prvi roman "Han Islanđanin" Viktor Igo napisao je 1823. godine*. Međutim, ove informacije mogu biti saopštene i posredno, poput *svoje najbolje delo* i *tri godine kasnije*. U pojedinim situacijama, vremenske informacije mogu da budu i neodređene, kao što je datum objavljivanja romana "Dvadeset treća" izostavljen iz teksta, te se zna samo da je objavljen posle romana "Jadnici". Dakle, zadatak automatske obrade vremenskih informacija može da podrazumeva jednostavno prepoznavanje određenih obrazaca (npr. *roman_objavljen_datuma*), ali i razumevanje značenja jezika koje ukazuje na to da je *najbolje delo* roman "Dvadeset treća", kao i da je roman "Ode i balade" objavljen *tri godine kasnije*, što ukazuje na godinu

datum publikovanja romana "Han Islanđanin" + tri godine = 1826.

Identifikovanje vremenskih informacija u tekstu nije tako jednostavno ni zbog činjenice da se informacije koje ukazuju na vreme mogu preneti, odnosno saopštiti putem više mehanizama, uključujući glagolska vremena, glagolski vid, kao i različite leksičke koncepte (Mani, Pustejovsky, and Gaizauskas 2005). Iako rečenice u primerima 1.1 i 1.2 imaju sličnu sintaksičku strukturu, hronologija opisanih događaja nije ista.

Primer 1.1.
Ana je pala. Petar ju je gurnuo.
Primer 1.2.
Ana je pala. Petar joj je pomogao da ustane.

U okviru datih izraza vremenske informacije su iskazane implicitno, dok opisani događaji nisu vezani ni za jednu preciznu tačku u vremenu, niti su hronološki poređani u odnosu na susedne događaje. Radi korektne interpretacije vremenskih informacija ovih izraza, potrebno je osloniti se na semantički sadržaj i postojeće uzročne veze. Bez obzira na sličnosti struktura rečenica, u prvom primeru događaj *padanje* vremenski je smešten u poziciju posle događaja *guranje*, dok se u drugom primeru događaj *padanje* desio pre događaja *pomaganje*.

Računarima je veoma teško da razumeju različite tipove semantičkih informacija, da ih postave na vremensku osu i utvrde relacije koje postoje među njima. Iz tog razloga je neophodno usmeriti pažnju na identifikovanje različitih mehanizama jezika, kojima se na eksplicitan ili implicitan način prenose vremenske informacije. Bilo koji sistem za automatsku obradu vremenskih informacija trebalo bi da je sposoban da u tekstu identifikuje ove mehanizme i da ih upotrebi za rešavanje sledećih zadataka: identifikaciju i normalizaciju vremenskih izraza, anotaciju događaja i identifikaciju vremenskih odnosa.

Razvijenost podrške jezičkih tehnologija se u velikoj meri razlikuje od jedne jezičke zajednice do druge. Veliki broj alata i resursa koji se koriste za automatsku obradu vremenskih informacija razvijen je za potrebe engleskog jezika, ali i nekih drugih, kao što su francuski, italijanski, španski, nemački, kineski itd. Kada je reč o srpskom jeziku, dosadašnji napori uloženi u razvoj jezičkih tehnologija (Krstev 2008; Krstev et al. 2011; Lecuit et al. 2009) pružaju neophodnu osnovu za

razvoj sistema za automatsku obradu vremenskih informacija. Kako se postavlja pitanje opstanka mnogih evropskih jezika u umreženom društvu i ističe opasnost od njihovog izumiranja na digitalnom tržištu (Vitas et al. 2012), potreba za razvijanjem ovakvih sistema koji će doprineti očuvanju srpskog jezika u digitalnom okruženju je sve veća. Osim toga, postojanje ovakvog sistema može u velikoj meri uticati na poboljšanje učinka primene mnogih drugih aplikacija iz oblasti jezičkih tehnologija (ekstrakcija informacija, pronalaženje informacija, odgovaranje na pitanja, rezimiranje teksta itd.), u opštem i mnogim drugim domenima. Na primer, i u oblasti medicine vreme se javlja kao jedan od suštinskih koncepata (Augusto 2005, Zhou and Hripcsak 2007, Reeves et al. 2013), jer je, između ostalog, tokom postavljanja dijagnoze veoma bitno znati vremenski redosled pojavljivanja određenih simptoma ili dužinu njihovog trajanja. Stoga bi bilo korisno razviti sistem koji se jednostavno može prilagoditi novim i drugačijim zahtevima i primeniti na tekstove iz nekog drugog vremenskog perioda ili domena.

1.2 Značaj obrade vremenskih informacija u oblasti obrade prirodnih jezika

Razvoj i evaluacija sistema namenjenih automatskoj obradi vremenskih informacija nije samo važna tema za istraživanje, već je i zadatak koji ima izuzetnu praktičnu primenu.

Sistemi namenjeni odgovaranju na pitanja (eng. *question answering*) obrađuju velike količine tekstova s ciljem pronalaženja kratkih izraza ili rečenica koje sadrže precizan odgovor na pitanje korisnika. Na osnovu rezultata primene sistema za obradu vremenskih informacija, sistemi namenjeni odgovaranju na pitanja mogli bi da daju odgovore na pitanja koja kao odgovor zahtevaju eksplicitnu vremensku informaciju (primer 1.3), ali i na pitanja čiji odgovor posredno zavisi od nekog vremena (primeri 1.4 i 1.5).

Primer 1.3.
Kada je počeo Prvi svetski rat?
Primer 1.4.
Da li je Vladimir Kostić trenutno predsednik SANU?
Primer 1.5.
Ko je bio dekan Filološkog fakulteta kada je oformljena Katedra za bibliotekarstvo i

informatiku?

Sistemi namenjeni automatskom rezimiranju teksta (eng. *automatic summarisation*) iz jednog ili više tekstova ekstrahuju najvažnije informacije. Automatski identifikovane vremenske informacije mogu poslužiti kao osnova za utvrđivanje hronološki povezane naracije o događajima koji se pominju u više različitih informacionih izvora. Proces automatskog rezimiranja može se praktično primeniti za generisanje biografija, ili kao pomoć prilikom sažimanja tekstova kliničkih beleški i sumiranja toka razvoja neke bolesti itd.

Pronalaženje informacija (eng. *information retrieval*) jeste oblast koja omogućava „pronalaženje materijala nestrukturirane prirode iz velikih kolekcija, koji zadovoljavaju informacione potrebe korisnika“ (Manning, Raghavan, Schütze, et al. 2008). Sa brzim porastom količine digitalnih informacija, koncept vremena kao dimenzije u okviru koje su informacije organizovane postaje veoma važan za ovu oblast. Pristup vremenskim informacijama koristan je prilikom rešavanja određenih zadataka pronalaženja informacija, kao što su grupisanje (eng. *clustering*) rezultatata pretraživanja prema različitim vremenskim atributima, pregledanje rezultata (eng. *browsing*) zasnovano na vremenu ili istraživanje rezultata pretraživanja korišćenjem vremenske linije.

S druge strane, sistemi za ekstrakciju informacija (eng. *information extraction*) jesu sistemi koji „selektivno strukturiraju i kombinuju podatke koji su nađeni, eksplicitno navedeni ili nagovešteni, u jednom ili više tekstova“ (Cowie and Wilks 2000). Osnovni cilj ove oblasti je prevashodno ekstrahovanje atributa entiteta (npr. profesija neke osobe) ili odnosa koji postoje između entiteta (npr. odnos „zaposlen_u“). U mnogim slučajevima su ekstrahovani atributi i među njima postojeći odnosi validni samo u okviru određenih vremenskih okvira, jer se entiteti i njihove osobine menjaju tokom vremena, te je stoga važno na pravilan način identifikovati i korektno interpretirati ove vremenske granice kako bi se unapredio proces ekstrakcije informacija.

Imajući u vidu pomenuti značaj procesa automatske obrade vremenskih informacija radi uspešnosti drugih složenijih aplikacija, kao i činjenicu da je, s obzirom na ogromne količine dostupnih podataka, nemoguće ručno dodavati vremenske oznake, potreba za postojanjem pouzdanih sistema za obradu vremena nikad nije bila veća. Na putu ka izgradnji kompletnog sistema namenjenog obradi vremen-

skih informacija, prvi od značajnih koraka jeste utvrđivanje i formalizacija vremenskih izraza, odnosno izgradnja sistema za automatsko prepoznavanje i normalizaciju vremenskih izraza.

1.3 Ciljevi i doprinos teze

U ovom radu izvršena je analiza elemenata formalne strukture različitih tipova vremenskih izraza u korpusima novinskih i medicinskih narativnih tekstova srpskog jezika. Vremenski izrazi, kao fraze prirodnog jezika koje se direktno odnose na vreme, pružaju informaciju o tome kada se nešto dogodilo, koliko dugo je trajalo ili koliko često se dešava. Na osnovu utvrđenih načina iskazivanja vremena, izvršena je izgradnja sistema za automatsko prepoznavanje i normalizaciju vremenskih izraza srpskog jezika.

Osnovni cilj sprovedenog istraživanja jeste automatsko obeležavanje vremenskih izraza u nestrukturiranim tekstovima srpskog jezika sa postizanjem visokog nivoa odziva i preciznosti, što pritom podrazumeva ostvarivanje sledećih zadataka:

1. kritički pregled postojećih pristupa za automatsku obradu vremenskih informacija, kao i standarda za njihovo obeležavanje;
2. prikupljanje i priprema korpusa narativnih (novinskih) tekstova;
3. identifikacija elemenata formalne strukture različitih tipova vremenskih izraza srpskog jezika;
4. određivanje, primena i vrednovanje metodologije za identifikovanje strukture i opsega vremenskih izraza;
5. određivanje, primena i vrednovanje metodologije za normalizaciju prepoznatih vremenskih izraza;
6. utvrđivanje ograničenja predložene metodologije i pružanje smernica za dalji rad.

Još jedan od postavljenih ciljeva jeste i procena efikasnosti metode za prepoznavanje vremenskih izraza u domenu medicinskih nestrukturiranih tekstova, kako bi se i u oblasti medicinskih narativnih tekstova omogućilo automatsko obeležavanje vremenskih izraza sa postizanjem visokog nivoa i odziva i preciznosti. U tom smislu potrebno je izvršiti:

1. pregled postojećih pristupa za automatsku obradu vremenskih informacija u medicinskom domenu;
2. prikupljanje i pripremu korpusa medicinskih narativnih tekstova;
3. primenu i vrednovanje definisane metodologije za identifikovanje i normalizaciju vremenskih izraza;
4. identifikovanje propusta radi njihovog otklanjanja i unapređivanja i prilagođavanja sistema medicinskom domenu.

Pod pretpostavkom da je metodom zasnovanom na pravilima, odnosno metodom konačnih automata moguće sa dovoljnom preciznošću i odzivom opisati strukture vremenskih izraza, bez sprovođenja sintaksne analize, očekivani su sledeći rezultati:

- kreiranje proizvoda koji će, bez dodatne pripreme, omogućiti automatsko obeležavanje vremenskih izraza srpskog jezika;
- izgradnja novih izvora za dalja istraživanja u oblasti obrade prirodnojezičkih tekstova korišćenjem statističkih metoda, odnosno korpusa (opšteg tipa i iz domena medicine) obeleženih dovoljnim brojem primera.

1.4 Kratak pregled teze

Ova disertacija je organizovana u osam poglavlja, u okviru kojih se problemu automatske obrade vremenskih izraza pristupa počevši od teorijskog aspekta vremena u jeziku (poglavlje 2), zatim pregleda primene postojećih metodologija za obradu vremenskih izraza (poglavlje 3), uz opis originalnog doprinosa u oblastima prepoznavanja (poglavlje 4) i normalizacije (poglavlje 5) vremenskih izraza, čija je uspešnost evaluirana u okviru poglavlja 6. Evaluacija i procena uspešnosti sistema u prepoznavanju i normalizaciji vremenskih izraza medicinskih narativnih tekstova izvršena je u okviru poglavlja 7, dok su opšti zaključci dati u poglavlju 8. Svako poglavlje usmereno je na ispunjavanje jednog ili više postavljenih zadataka.

U okviru **drugog poglavlja** obrađena su različita teorijska pitanja koja se tiču oblasti obrade vremenskih informacija. Predstavljeni su različiti mehanizmi jezika za izražavanje vremenskih informacija, uz poseban osvrt na opis struktura vremenskih izraza. Ovo poglavlje služi kao teorijska osnova ovog istraživanja, budući da pruža pregled načina prenošenja vremenskih informacija, što je neophodno radi

razvoja automatizovanog sistema čiji je cilj identifikacija različitih tipova vremenskih izraza.

Poglavlje 3 opisuje postojeće sheme za obeležavanje vremenskih izraza i do sad korišćene izvore i računarske pristupe za obavljanje različitih zadataka koji uključuju obradu vremenskih informacija. S ciljem izdvajanja najuspešnije metodologije u zadacima prepoznavanja i normalizacije vremenskih izraza, izvršen je kritički pregled i pažljivo su ispitane pozitivne i negativne strane korišćenih pristupa.

Poglavlja od 4 do 7 predstavljaju originalan doprinos ovog istraživanja.

Prva faza procesa anotacije vremenskih izraza, odnosno prepoznavanje vremenskih izraza opisano je u okviru **poglavlja 4**. Osim identifikacije vremenskih izraza, opisan je i proces otkrivanja opsega, odnosno niski teksta koje čine vremenski izraz prisutan u datom tekstu. Na samom početku poglavlja data je iscrpna klasifikacija najuobičajenijih tipova vremenskih izraza, koji se javljaju u tekstovima srpskog jezika. U daljem tekstu opisan je razvijeni alat koji sa visokom preciznošću identifikuje sve tipove vremenskih izraza, definisanih prethodno pomenutom klasifikacijom.

Proces normalizacije vrednosti prepoznatih vremenskih izraza, kao druga faza procesa anotiranja vremenskih izraza, dat je u **petom poglavlju**. U ovom koraku obrade vremenskih izraza identifikovane su vrednosti atributa koje je potrebno pripisati vremenskim izrazima. U nastavku ovog poglavlja detaljno je opisana metodologija, kao i način primene radi normalizacije vrednosti vremenskih izraza, identifikovanih u prethodnoj fazi prepoznavanja.

Poglavlje 6 je posvećeno evaluaciji kreiranog sistema, odnosno uspešnosti u sprovođenju zadataka prepoznavanja i normalizacije vremenskih izraza srpskog jezika, radi uočavanja propusta i ograničenja koja će biti uzeta u obzir prilikom kasnijeg unapređenja sistema.

Opis primene kreiranog sistema za automatsku obradu vremenskih izraza na nestrukturirane tekstove iz oblasti medicine dat je u **poglavlju 7**. Nakon izvršene pripreme korpusa medicinskih narativnih tekstova, primenjena su pravila definisana u prethodnim fazama istraživanja, a zatim je izvršena evaluacija uspešnosti

i identifikacija propusta radi njihovog otklanjanja i unapređivanja sistema, koji će se za potrebe automatske obrade vremenskih izraza koristiti u ovoj oblasti.

Na kraju, **poglavlje 8** rezimira ostvareni doprinos i postignute ciljeve ovog istraživanja, uz jasno definisane buduće pravce istraživanja.

Glava 2

Vreme u prirodnom jeziku

> *Dok me niko ne pita, ja znam; kad bi, pak, valjalo da to objasnim – ja ne znam.*
>
> — Sveti Avgustin Hiponski, *Ispovesti, knjiga 11*

Određenje pojma vremena jedan je od velikih problema i izvor brojnih nedoumica, koji je prisutan još od samih početaka naučne misli. Da li se o vremenu može razmišljati kao o toku, i ako je tako, da li ono teče iz budućnosti u prošlost, prolazeći mimo nas, ili plovi iz prošlosti u budućnost, noseći nas sa sobom? Može li ono proticati brže ili sporije? S druge strane, ako odbacimo metaforu vremenskog toka, ostaje pitanje kako shvatiti njegovo proticanje. Postoji li nešto što razlikuje sadašnjost od prošlosti i budućnosti, ili među njima nema objektivne razlike? Da li je vreme beskrajno deljivo do najmanjeg kvantuma ili komadića vremena? Mnoga od ovih pitanja po prvi put su postavljena u filozofiji antike, u Aristotelovoj *Fizici* (350. g. p.n.e.), u obliku paradoksa ili problema o samom postojanju vremena (Blekburn 2013). Aristotel vreme smatra merom promene, ističući da vreme samo po sebi nije promena, jer promena, za razliku od vremena, može da bude brža ili sporija.

Najdublji problem filozofije vremena odnosi se upravo na status vremenske promene (Oklander 1998). Da li događaji zaista prelaze iz budućnosti u sadašnjost i prošlost kao što su tvrdili A-teoretičari, ili je proticanje vremena mit i iluzija kao što su tvrdili B-teoretičari? Vremenska promena uključuje pojam proticanja, odnosno kretanja vremena ili događaja u vremenu s jednog položaja u A-nizu na drugi. Dakle, vremenska promena uključuje menjanje od budućnosti u sadašnjost i prošlost, što dovodi do pitanja kako je zapravo moguće da se vreme, koje je samo mera promene, menja. Razmatrajući pitanje da li naše subjektivno iskustvo vremenske promene (na primer, nešto će se desiti, dešava se ili se desilo)

i verovanje u nju (na primer, danas je utorak, sutra sreda, a juče je bio ponedeljak) uopšte odražavaju stvarnu prirodu vremena, argumente protiv A-teorije prvi put iznosi Mak-Tagart (1908), svrstavajući događaje kao međusobno ranije ili kasnije, bez odnošenja na njihovo mesto u prošlosti, sadašnjosti ili budućnosti.

Vreme posmatrano kao, pre svega, ontološka kategorija, ponajviše je obeležilo filozofska razmišljanja, u kojima ovo pitanje ima posebno značenje. Složenost odgovora na pitanje *Šta je vreme?* i njegova neizreciva priroda ogledaju se u mnoštvu različitih tumačenja i nemogućnosti usvajanja jasne definicije vremena, koja bi bila primenljiva u okviru svih naučnih disciplina. U drugoj polovini XVII veka, zahvaljujući radu Isaka Njutna, po prvi put je definicija vremena razmatrana, ne iz čisto filozofskog, već iz naučnog ugla. Njutn je verovao u apsolutno vreme, koje je nezavisno od prostora i svih procesa koji se odigravaju unutar njega. Radom Alberta Ajnštajna i teorijom relativnosti vreme i prostor su posmatrani zajedno, kao specifična struktura, tj. četvorodimenzionalna celina, u okviru koje su tačke definisane kao događaji. Moderno fizičko mišljenje podrazumeva vreme posmatrano kao suštinski uniformno s prostorom. Prema savremenim definicijama, vreme je shvaćeno kao apstrakcija tj. odraz promena fizičke stvarnosti u ljudskoj svesti i način na koji ljudski um poima i tumači događaje.

2.1 Vreme i jezik

Vreme kao jedna od osnovnih dimenzija čovekovog ispoljavanja poima se, u zapadnoj civilizaciji, kao „objektivni, jednodimenzionalni, homogeni entitet linearne prirode sa osobinom neograničenog, neprekidnog trajanja, ali i sa podatnošću segmentiranju na neograničeni broj odsečaka različite veličine, od onih koji impliciraju samo trenutak do onih koji impliciraju vremenski period neograničenog trajanja, i predstavom o stalnom kretanju i promeni“ (Antonić 2001). Vreme se shvata, i uobičajeno predstavlja, u vidu vremenske ose na kojoj se događaji raspoređuju u uređeni, sukcesivni niz.

Vreme se na različite načine izražava u jeziku, dok jezik odražava određene osobine vremena i načine njegove konceptualizacije, koje zavise od mišljenja i ukazuju na razliku između objektivnog i subjektivnog vremena (Antonić 2001). Kategorija vremena se, uz kategoriju prostora, smatra osnovnom semantičkom kategorijom u jeziku. Iz kategorije prostora se prvo uobličila upravo kategorija vremena, da bi te dve kategorije zatim postale osnova za razvoj ostalih složenijih

kategorija (Piper 1997). Povezanost vremena i prostora je u lingvistici od posebnog značaja, imajući u vidu savremeno shvatanje da vremenski odnosi u jeziku zapravo predstavljaju metaforizaciju prostornih odnosa (Antonić 2001). Odnos prema vremenu – uz onaj prema prostoru – bitno određuje čovekovo poimanje sveta, koje se ogleda u ljudskoj sposobnosti razlikovanja događaja koji se dešavaju u situaciji u kojoj se govori tj. sadašnjosti od onih koji su se događali pre tog trenutka tj. u prošlosti ili, pak, od onih koji će se tek dogoditi, odnosno za koje se očekuje da će se dogoditi u budućnosti (Badurina 2013).

Kada je reč o konceptualizaciji vremena u iskustvu govornih predstavnika većine jezika, kategorija vremena se u jeziku vezuje isključivo za predikaciju, budući da se samo predikacija može smestiti (locirati) u vreme, ili, pak, odmeravati (kvantifikovati) u vremenu. Predmeti, stvari, ličnosti se primarno smeštaju u prostor, što znači da ih je moguće određivati u pogledu vremena samo preko predikacije koja im se pripisuje. Antonić smatra da ova karakteristika verovatno proizilazi iz činjenice da se „jedino predikacija, po prirodi stvari, dovodi u direktnu vezu s pojmovima trajanja, proticanja, promene kao distinktivnim karakteristikama vremena, dok ostali pojmovi to bivaju samo preko predikacije“ (Antonić 2001).

Složenost pojma vremena se odražava i u raznolikim načinima izricanja vremenskih odnosa i u srpskom jeziku, jer oni mogu biti iskazani na morfološkom, sintaksičkom, leksičkom nivou, kao i na nivou teksta. Osnovni elementi teksta poseduju svoj unutarnji princip organizacije u vremenu, tj. afirmišu jedan primarni doživljaj vremena (Miljković 2013). Doživljaj vremena jeste, zapravo, princip odnosno način organizacije svih sadržaja koji stvara oblik i osobenu semantiku u strukturi teksta. Uspostavljanje poretka pojavljivanja u tekstu, u kome je dato vreme, karakteriše narativ kao „osnovni način na koji ljudska vrsta organizuje svoje shvatanje vremena“ (Abot 2009).

Kako bi se omogućilo automatsko „razumevanje“ temporalnosti nekog teksta, neophodno je ispitati na koje se sve načine jezik koristi za saopštavanje vremenskih informacija. Srpski jezik, kao i bilo koji drugi prirodni jezik, omogućava izražavanje vremenskih informacija putem više sredstava, grupisanih u tri kategorije, koje čine suštinu temporalne ontologije (Ašić 2005):

- događaji, koji su na osnovu Vendlerove taksonomije grupisani u okviru četiri tipa entiteta – stanja, aktivnosti, ostvarenja i dostignuća (Vendler 1957);
- vremenski izrazi, koji ukazuju na pojam vremenskog intervala i

- tri tipa vremenskih odnosa – sukcesivnost, inkluzija i preklapanje.

Upravo se zahvaljujući gramatičkoj, odnosno predikatskoj kategoriji vremena uspostavljaju vremenski odnosi u rečenici: odnosi između događaja označenih rečenicom i samog govornog čina (tzv. indikativno izražavanje vremena), ali i odnosi između vremena u kome se zbiva ono o čemu je reč u rečenici i vremena o kome se govori, i to nezavisno od vremena govora (tzv. konjuktivno ili vezano izražavanje vremena). Kategorija glagolskih vremena i funkcije njihovih oblika su najsloženija pitanja u oblasti sistema glagolskih oblika (Stevanović 1967). Osnovne funkcije glagolskih vremena jesu određivanje vremena: radnje ili stanja, odnosno osobine i zbivanja, odnosno njihovo vezivanje za jedan od tri vremenska perioda: za prošlost, sadašnjost i budućnost, van kojih ni po suštini ni gramatički nema nikakvog vremena. Problem vremenskih i drugih funkcija oblika glagolskih vremena dosta je složen, pa u slovenskim jezicima za iskazivanje glagolom označenih procesa i osobina svakom vremenskom periodu odgovara po jedan vremenski glagolski oblik: prošlosti – perfekat, sadašnjosti – prezent, a budućnosti – futur. Međutim, glagolska vremena mogu, osim funkcije vremenskog određivanja prema vremenu govorenja (indikativ), glagolom označeni proces da odrede i u odnosu na bilo koji drugi trenutak (relativ). Tako se, na primer, prezentom mogu označavati radnje ili osobine koje su vezane za prošlost, a i za budućnost.

S gramatičkom kategorijom vremena tesno je povezana semantička kategorija temporalnosti. Pod semantičkom kategorijom temporalnosti se može podrazumevati značenje veoma visokog stepena opštosti, koje može biti raznovrsno po obliku, a koje u izrazu nije ograničeno samo na jednu sintaksičku, morfološku, leksičko-gramatičku ili tvorbenu kategoriju (Piper et al. 2005). Temporalnost, odnosno temporalna determinacija podrazumeva određivanje rečenične predikacije u odnosu na vreme. Osnovni oblik ostvarivanja temporalnosti jeste *temporalna identifikacija*, odnosno izdvajanje odseka na vremenskoj osi (kao orijentira) i smeštanje rečenične predikacije tj. objekta lokalizacije unutar (interiorizacija) ili van njegovih okvira (eksteriorizacija). Budući da se vreme nužno ispoljava u trajanju, rečenična predikacija, osim lokalizacijom, može da bude određena i u pogledu kvantitativne dimenzije vremena. *Temporalna kvantifikacija* kao odmeravanje rečenične predikacije u vremenu ispoljava se kao odmeravanje dužine trajanja rečenične predikacije i kao odmeravanje učestalosti pojavljivanja rečenične predikacije. Odmeravanje dužine trajanja rečenične predikacije može se ostvariti kao longitudinalnost – u smislu *koliko dugo?*, kao ingresivnost – *koliko dugo?* uz identifikaciju početne tačke ili kao terminativnost – *koliko dugo?* uz identifikaciju krajnje

tačke. Odmeravanje učestalosti pojavljivanja rečenične predikacije predstavlja, zapravo, temporalnu frekvenciju (Antonić 2001).

Rečenična predikacija može biti temporalno determinisana s obzirom na momenat govora, kao apslolutna temporalna determinacija, i s obzirom na neku drugu referentnu tačku u vremenu, koja je različita od momenta govora tj. relativna temporalna determinacija. Oba ova tipa temporalne determinacije se realizuju različitim jezičkim sredstvima, i to, pre svega, vremenskim glagolskim oblikom u kojem stoji rečenična predikacija, kao i adverbijalnom, nominalnom ili sentencijalnom formom, koja u funkciji temporalnog determinatora, prati rečeničnu predikaciju.

Predmet ovoga rada jeste identifikacija i formalizacija leksičkih sredstava izražavanja temporalnih značenja. Iako je glagolima kao vrsti reči kategorija vremena jedna od najbitnijih odrednica, eksplicitno vremensko značenje se ređe izražava ovim oblicima, već se vremenska značenja prevashodno iskazuju imenicama, pridevima, predloško-padežnim i predloško-priloškim izrazima, kao i drugim brojnim izrazima kojima je u osnovi koncept vremena.

2.2 Izrazi vremenskog značenja

Vremenski izrazi su fraze prirodnog jezika koje direktno ukazuju na vreme, pružajući informaciju o tome kada se nešto dogodilo, koliko dugo je trajalo ili koliko često se dešava. Većina vremenskih izraza u srpskom jeziku ima sintaksičku ulogu priloških odredbi, koje izražavaju semantičku ulogu vremena. Najčešće označavaju kalendarske datume, vremena dana, trajanja i učestalosti ponavljanja vremena. Sintaksička osnova vremenskih izraza sadrži odgovarajući leksički okidač, odnosno reč ili numerički izraz sa značenjem jedinice ili koncepta vremena, i to:

- imenice: *vek*, *godina*, *mesec*, *dan*, *vikend*, *minut*, *sat*, *ponedeljak*, *mart*, *Božić*;
- pridevi: *prošli*, *budući*, *sadašnji*, *sledeći*, *mesečni*;
- prilozi: *sada*, *onda*, *nedeljno*, *juče*, *danas*, *noćas*;
- brojevi: *4*, *prvog*;
- posebni vremenski obrasci: *12:35*, *3.04.1999*.

Vremenski izrazi kojima se prenose vremenska značenja realizuju se u jeziku putem određenih gramatičkih oblika, koji, pre svega, podrazumevaju upotrebu imeničkih fraza koje se pojavljuju samostalno ili u vidu predloško-padežnih konstrukcija. Leksički eksponenti temporalnih padežnih konstrukcija su imenice sa primarno vremenskim značenjem (kada je u pitanju neposredno identifikovanje temporalnosti) ili imenice kojima je svojstveno obeležje „trajanja u vremenu“. Prema predlogu Milke Ivić (1955–1956) imenice sa primarno vremenskim značenjem mogu se podeliti na: (a) oznake jedinica vremenske mere (*dan*, *godina*); (b) oznake konstatno identifikovanih jedinica vremenske mere (*ponedeljak*, *januar*); (c) oznake pojmova-perioda (*noć*, *jesen*); kao i (d) oznake pojmova-praznika (*Đurđevdan*, *Božić*). Uz temporalne imenice postoje razne mogućnosti upotrebe atributskih reči, koje funkcionalno, semantički i sintaksički dopunjuju ili upotpunjuju dati iskaz (Stevanović 1979; Fekete 2009). Drugi čest način za određivanje vremenskih izraza predstavljen je određenim prilozima, pridevima ili njihovim odgovarajućim frazama (Stanojčić and Popović 2011).

Predloško-padežne konstrukcije kojima se izražavaju vremenska značenja organizovane su kao poseban sintaksičko-semantički podsistem, koji i u strukturi i u sredstvima kojima se ostvaruje pokazuje dosta sličnosti sa sistemom mesnih padeža (Ivić 1983). Većinom strukturu ovih modela čine predlog i određeni padežni oblik, dok se vremenska značenja znatno ređe mogu izraziti i bespredloškim padežnim oblikom, po pravilu sa obaveznim determinatorom (npr. *sreli su se* ***prvog*** *aprila*) (Piper 1997). Kao i u predloško-padežnim konstrukcijama sa prostornim značenjem, i u vremenskim konstrukcijama predlozi su ili jednočlani, bilo prosti, npr. *u sredu*, bilo izvedeni, npr. *usred leta*; ili dvočlani (tj. predloški izrazi, npr. *u toku nedelje*). Za iskazivanje vremenskih značenja koristi se više modela predloško-padežnih konstrukcija, kao i nekoliko slučajeva bespredloške upotrebe padežnih oblika, poput bespredloškog genitiva sa obaveznim determinatorom (npr. *drugog juna*) ili bespredloškog instrumentala (npr. *godinama*).

2.2.1 Izražavanje pozicije u vremenu

Vremenski izrazi koji impliciraju tačku u vremenu ukazuju na temporalnu lokaciju, odnosno odsek na vremenskoj osi koji je shvaćen kao tačka u vremenu (kalendarske pozicije), pa time preciziraju kada se nešto dogodilo. Vremenski lokalizator u tim situacijama je najčešće neka jedinica mere vremena (npr. *sekund*, *minut*, *godina*, *petak*) ili oznaka pojma-perioda (npr. *noć*, *jutro*, *zima*), uz koju je česta upotreba pridevskih zamenica (primer 2.1). Za izražavanje vremenske pozi-

cije najčešća je upotreba predloških konstrukcija, poput one u primeru 2.2, kao i korišćenje priloga za vreme (primer 2.3).

Primer 2.1.
*Ana ga je pozvala **tog jutra**.*

Primer 2.2.
*Posetili su nas **u sredu**.*

Primer 2.3.
***Sutra** putujemo u Italiju.*

Na poziciju u vremenu se može precizno ukazati upotrebom vremenskih tačaka (primer 2.4), ali i nedovoljno precizno izrazima koji označavaju vremenske periode ili intervale (primer 2.5). Vremenski izraz *pet sati popodne* ispred kog se nalazi predlog *u* u primeru 2.4 ukazuje na preciznu tačku u vremenu (kada su stigli), dok vremenski izraz *juče* u primeru 2.5 ukazuje na vremenski interval u kom se poziv desio.

Primer 2.4.
*Stigli su **u pet sati popodne**.*

Primer 2.5.
*Pozvali su nas **juče**.*

2.2.2 Izražavanje trajanja u vremenu

Vremenski izrazi koji se takođe ispoljavaju kao „čista“ semantička kategorija jesu izrazi kojima se ispoljava temporalna kvantifikacija, odnosno odmerava dužina trajanja u vremenu. Izrazi ovoga tipa ukazuju na odsek na vremenskoj osi koji je shvaćen kao kraći ili duži „prostor“ u vremenu i koji ukazuje na trajanje.

Značenje trajanja kao preciznog perioda vremena može biti implicirano i određenim tačkama u vremenu, odnosno značenjem ingresivnosti (identifikacija leve granične tačke) ili terminativnosti (identifikacija desne granične tačke). Vremenski izrazi koji ukazuju na trajanje obično su formirani pridruživanjem kvantifikatora (npr. *nekoliko*, *tri*, *mnogo*) jedinici vremena (npr. *godina*, *nedelja*, *mesec*, *sat*) (primeri 2.6 i 2.7).

Primer 2.6.
*Svirali smo **nekoliko sati**.*

Primer 2.7.
*Živela je u našoj kući **tri meseca**.*

2.2.3 Izražavanje učestalosti ponavljanja u vremenu

Vremenski izrazi ovoga tipa ukazuju na temporalnu frekvenciju (iteraciju), odnosno učestalost pojavljivanja u vremenu. Ovde se radi o posebnom vidu temporalnog kvantifikovanja koje se može ostvariti kao povremeno ponavljanje više puta u vremenu i regularno ponavljanje više puta u vremenu, i pri tom uvek u isto vreme.

Za izražavanje učestalosti često se koriste pridevi ili prilozi učestalosti (npr. pridev *mesečni* ili prilog *mesečno* kao u primeru 2.8), ili upotrebom pridevske zamenice *svaki* u vezi sa jedinicom vremena (npr. *svake godine*, *svakog utorka*, *svakog dana*) (primer 2.9), koja može biti i numerički kvantifikovana (primer 2.10).

Primer 2.8.
*Izveštaj nam dostavljaju **mesečno**.*

Primer 2.9.
*Slao nam je poziv **svake godine**.*

Primer 2.10.
*Sastanci se održavaju **svakog trećeg petka**.*

Osim činjenice da se vreme na različite načine izražava u jeziku, još jedan vid veze između vremena i jezika oslikava se i u tome da se jezik istovremeno ispoljava u vremenu, čime je podložan kretanju i promenama (Antonić 2001). Imajući u vidu ovu karakteristiku jezika, svaki pokušaj sveobuhvatne analize lingvističkih izraza vremenskog značenja u standardnom srpskom jeziku ne omogućava obuhvatanje svih postojećih načina za iskazivanje vremena. U ovom poglavlju opisana su samo neka od jezičkih sredstava kojima se prenose osnovna vremenska značenja, data u obliku vremenskih izraza, kao jednog od osnovnih entiteta temporalnosti. Detaljniji opis ovih formalnih jedinica, identifikovanih u narativnim

tekstovima srpskog jezika, kojima se prenose tri osnovna tipa vremenskog značenja, dat je u poglavlju 4. Nakon utvrđivanja i analize oblika najuobičajenijih jezičkih struktura ovog značenja, moguć je rad na razvoju sistema namenjenog automatskom prepoznavanju vremenskih izraza i normalizaciji njihovih vrednosti.

Glava 3

Pregled postojećih resursa i računarskih pristupa za obradu vremenskih izraza

U okviru ovog poglavlja biće dat pregled postojećih resursa i računarskih pristupa korišćenih za automatsku obradu vremenskih izraza u novinskim tekstovima. S ciljem da se pruži detaljniji opis pristupa i sistema koji su ostvarili značajan napredak u odnosu na prethodna dostignuća ili se izdvajaju svojom originalnošću, manje prostora biće posvećeno teorijskim predlozima ili metodama čija je namena ograničena isključivo na određene zadatke. Isto tako, sistemi koji se ne bave prepoznavanjem i klasifikacijom vremenskih izraza, već samo događaja ili vremenskih relacija, kao osnovnih entiteta temporalnosti, neće biti predstavljeni u ovom pregledu.

S obzirom na to da automatska obrada vremenskih izraza u tekstu počinje obeležavanjem ovih entiteta u skladu sa unapred definisanim pravilima, u okviru ovog poglavlja prvo će biti dat opis postojećih shema na osnovu kojih se vrši obeležavanje vremenskih izraza. Kako je definisanje ovih shema omogućilo obeležavanje različitih tekstualnih resursa oznakama vremenskih informacija, u sledećem delu će biti predstavljeni najznačajniji anotirani korpusi korišćeni za istraživanje u oblasti automatske obrade vremenskih informacija. Pregled najznačajnijih postojećih pristupa namenjenih prepoznavanju i normalizaciji vremenskih izraza biće dat u poslednjem delu ovog poglavlja.

3.1 Sheme za obeležavanje vremenskih izraza

Korišćenje neobeleženih tekstova omogućava sprovođenje površne lingvističke analize, dok je za potrebe izdvajanja određenog značenja tekstualnih podataka neophodno njihovo obeležavanje dodatnim lingvističkim informacijama (npr. sintaksičkim, semantičkim itd.). Dakle, osim potrebe za definisanjem semantičke prirode ciljane informacije koju je potrebno izdvojiti, u ovom slučaju vremenskih izraza, važno je prepoznati razliku među njima i precizirati i različite klase entiteta, kao i njihov obim. Ovakav postupak zahteva postojanje određene sheme za obeležavanje, na osnovu koje će biti definisane one instance određene klase nekog fenomena, izražene na jedan ili više načina, koje treba da budu obeležene u korpusima i na koji način. Shema takođe treba da sadrži i detaljno opisane smernice za njeno ispravno korišćenje.

Od trenutka kada je ekstrakcija vremenskih informacija prvi put uključena u kontekst Konferencija o razumevanju poruka (eng. *Message Understanding Conferences, MUC*) 1995. godine, više napora je uloženo u uobličavanje standarda za obeležavanje vremenskih informacija u tekstu. Postojeća naučna i stručna literatura koja se bavi pitanjima vremena i jezika poslužila je kao osnov za razvijanje shema za obeležavanje vremenskih informacija. Na samom početku, vremenske informacije obeležavane su SGML (eng. *Standard Generalized Markup Language*) (ISO 1986) etiketom TIMEX i bile određene samo jednim atributom (Grishman and Sundheim 1996). Od tada je razvijeno više shema za obeležavanje, ali su tri od njih u velikoj meri korišćene za razvoj resursa za obradu vremenskih informacija: TIDES (Ferro et al. 2000), STAG (Setzer and Gaizauskas 2000) i TimeML (Pustejovsky et al. 2003). U ovom delu su hronološkim redosledom predstavljene sheme za obeležavanje, najčešće korišćene za razvoj resursa za obradu vremenskih informacija, s ciljem da se naglase novine koje je svaka od njih donela u odnosu na prethodnu.

3.1.1 Prvi TIMEX (1995/97)

Među organizovanim projektima i konferencijama koje su bile posvećene imenovanim entitetima, poseban značaj, pre svega u pogledu definisanja standarda, pripada MUC konferencijama. Ove konferencije, koje je organizovala DARPA (eng. *Defence Advanced Research Projects Agency*), agencija za napredne istraživačke projekte odbrane američke vlade, imale su takmičarski karakter i služile su za razvijanje i ocenjivanje sistema za ekstrakciju informacija. Na samom početku, zadatak

takmičara bio je prepoznavanje određenih događaja u tekstu, koje su opisali stručnjaci američke vlade kao organizatora konferencija, i automatsko popunjavanje polja određenog obrasca (formulara), koja su označavala različite informacije o događaju (npr. tip događaja, odnosno radnje, vršilac radnje, vreme i mesto dešavanja, posledice itd.). Ovim automatskim popunjavanjem podrazumevala se identifikacija specifičnih odnosa koji postoje među vremenskim elementima, u ovom slučaju između vremena i događaja. Od sistema koji su učestvovali tražilo se da određenim tipovima događaja pripišu određena vremena dešavanja. Na primer, u slučaju događaja lansiranja rakete, obrazac ovog scenarija sadržao je polje LAUNCH_DATE koje je bilo vezano za odgovarajući vremenski entitet. Vremenski odnos između vremena i događaja nije dalje evaluiran, dok vremenske relacije koje postoje između događaja i drugih događaja nisu bile predmet zadatka. Tako je identifikacija vremenskih izraza isprva bila samo jedan od koraka neophodnih za popunjavanje polja koja se odnose na vreme dešavanja nekog događaja. Tek je u okviru poslednje dve MUC konferencije (MUC-6 i MUC-7) precizno definisan pojam imenovanih entiteta i postavljen zadatak njihovog automatskog prepoznavanja. Ovaj zadatak podrazumevao je identifikaciju i klasifikaciju različitih tipova imenovanih entiteta, kao što su: imena osoba, lokacija, organizacija (ENAMEX), datuma i vremena (TIMEX)[1] i brojčanih izraza (NUMEX). Vremenski izrazi su, dakle, prvi put uključeni kao klase imenovanih entiteta u okviru MUC-6 konferencije (Grishman and Sundheim 1996) i bili su prisutni do poslednje MUC-7 konferencije (Chinchor 1998).

Dakle, najraniji primer obeležavanja vremenskih informacija datira unazad 20 godina, kada je za potrebe održavanja MUC-6 konferencije razvijena prva shema za obeležavanje vremenskih informacija. Ova prva shema bila je veoma jednostavna i podrazumevala je prepoznavanje apsolutnih vremenskih izraza koji eksplicitno ukazuju na kalendarske datume (npr. *mart 2011. godine*) i vremena koja označavaju delove dana (npr. *20 časova*). Od učesnika u takmičenju traženo je da izgrade sisteme koji će u korpusu tekstova prepoznate vremenske izraze obeležiti SGML etiketom TIMEX koja ukazuje na njihov tip (DATE za datume ili TIME za izraze koji ukazuju na delove dana). Primer pronađenih entiteta obeleženih u tekstu odgovarajućim elementima sa odgovarajućim atributima dat je u primeru 3.1.

[1]Naziv etikete TIMEX izveden je iz engleskog izraza *time-expression*.

Primer 3.1.

```
Stigao je u Novi Sad <TIMEX type="DATE">25. marta 2005. godine
   </TIMEX>.
Sastanak počinje u <TIMEX type="TIME">13:30 h</TIMEX>.
```

Proizvedeni izlazi sistema takmičara bili su evaluirani u odnosu na ručno pripremljene očekivane rezultate, i to u smislu procene preciznosti i odziva sistema u određivanju opsega vremenskog izraza, dok se razrešavanje vrednosti prepoznatih vremenskih izraza nije zahtevalo od takmičara.

Dok je ovaj zadatak u okviru MUC-6 konferencije bio ograničen samo na identifikaciju apsolutnih vremenskih izraza, kasnije u okviru poslednje MUC-7 konferencije zadatak prepoznavanja TIMEX entiteta uključuje i relativne vremenske izraze (npr. *juče uveče*, *dve godine kasnije*).

Nakon 1997. godine, program Automatska ekstrakcija sadržaja (eng. *Automatic Content Extraction, ACE*) (ACE 1999) zamenio je MUC takmičenja, čime je uvećana složenost zadataka. U okviru programa ACE bilo je uključeno prepoznavanje više vremenskih izraza, a njihovo obeležavanje podrazumevalo je korišćenje daleko složenije TIMEX2 etikete, koja je detaljnije opisana u sledećem delu.

3.1.2 TIDES TIMEX2 (2000)

TIDES TIMEX2 shema (Ferro et al. 2000) razvijena je kao podrška istraživanjima sprovedenim u okviru ACE i još jednog od DARPA programa – Otkrivanje, ekstrakcija i sumarizacija višejezičkih informacija (eng. *Translingual Information Detection, Extraction and Summarization, TIDES*), da bi se obeležili vremenski izrazi u tekstovima na više jezika i predstavile njihove vrednosti u standardizovanom formatu, koji će govornicima engleskog jezika omogućiti da ih brzo pronađu i razumeju a da ne moraju da poznaju jezik kojim su zapisani. Kao veoma temeljno uputstvo za obeležavanje vremenskih izraza, TIDES TIMEX2 shema se nastavlja na MUC-7 TIMEX shemu, proširujući je u smislu povećanja broja različitih tipova vremenskih izraza koji treba da budu prepoznati, kao i uvođenja atributa koji će preciznije opisati njihovo značenje. Ovom novonastalom shemom TIMEX etiketa zamenjena je etiketom TIMEX2, u okviru koje je predstavljena njegova vrednost u formatu kompatibilnim sa ISO-8601[2] standardnim formatom (ISO 2004). Atribut

[2]ISO-8601 standardom se utvrđuje predstavljanje datuma po gregorijanskom kalendaru, predstavljanje sati i perioda vremena. Standard je namenjen za razmenu onformacija

koji je u okviru TIMEX etikete eksplicitno ukazivao na tip identifikovanog vremenskog izraza (`type`) je isključen, te je njegova vrednost uočljiva samo na osnovu formata kojim je izraz predstavljen.

Osnovne razlike, koje u isto vreme predstavljaju i prednosti TIMEX2 sheme u odnosu na prethodnu, su sledeće:

Osim kalendarskih datuma i delova dana kao izraza koji predstavljaju tačku na vremenskoj osi, razmatrani su i drugi tipovi vremenskih izraza koji ukazuju na trajanje (npr. *tokom 25 godina*) i učestalost ponavljanja nekog događaja (npr. *godišnje*);

Vrednosti prepoznatih apsolutnih (npr. *22. aprila 2007.*)[3] i relativnih (npr. *danas*, *juče*) vremenskih izraza, kao i perioda vremena (npr. *tri meseca*) i učestalosti (npr. *mesečno*) treba da budu normalizovane u okviru novog atributa VAL, koji je u skladu sa ISO-8601 standardom;

Osim atributa VAL, uključeni su i dodatni atributi koji preciznije opisuju značenje vremenskog izraza:

MOD – identifikuje reči koje na određeni način modifikuju značenje vremenskog izraza (npr. ***oko*** *tri sata*, ***početkom*** *2005. godine*);

ANCHOR_VAL – sadrži normalizovanu vrednost orijentira, odnosno datuma ili vremena za koji je neki vremenski izraz koji označava trajanje vezan (npr. izraz *trajaće dva meseca zaključno sa 15. novembrom 2011. g.* je vremenski izraz koji označava trajanje *dva meseca* i vezan je za datum *15. novembar 2011. g.*, što će biti vrednost ANCHOR_VAL atributa iskazana u formatu `2011-11-15`);

ANCHOR_DIR – koristi se radi pozicioniranja datuma ili vremena predstavljenog atributom ANCHOR_VAL u okviru perioda trajanja za koji je vezan (npr. u prethodnom primeru *trajaće dva meseca zaključno sa 15. novembrom 2011. g.* datum `2011-11-15` predstavlja kraj navedenog trajanja, te će vrednost atributa ANCHOR_DIR biti ENDING);

SET – koristi se za predstavljanje izraza koji ukazuju na vremena koja se ponavljaju, u pravilnim ili nepravilnim razmacima (npr. *svakog četvrtka*);

COMMENT – atribut koji anotatorima omogućava unošenje napomena, kao što je npr. razlog donošenja neke posebne odluke u slučaju višeznačnih izraza.

Primer teksta obeleženog u skladu sa TIDES TIMEX2 shemom dat je u primeru 3.2.

Primer 3.2.

```
Prvi ispit je položio <TIMEX2 VAL="1999-10">oktobra 1999. godine
   </TIMEX2>. Tokom <TIMEX2 VAL="P1Y" ANCHOR_VAL="2000">[4]sledeće
   godine</TIMEX2> dao je još dva ispita.
Ona ga posećuje <TIMEX2 SET="YES" VAL="XXXX-XX-XX">[5]svakog dana
   </TIMEX2>.
Ivana ga poznaje <TIMEX2 VAL="P1Y" MOD="LESS_THAN">manje od godinu
   dana</TIMEX2>.
```

Od kada je ova shema originalno razvijena tokom 2000. godine u okviru TIDES programa i prvi put dokumentovana u (Ferro et al. 2000), doživela je nekoliko revizija (Ferro et al. 2001, 2003), sa poslednjom verzijom koja je predstavljena 2005. godine (Ferro et al. 2005). Poslednje uputstvo za obeležavanje opisuje širok spektar vremenskih izraza koje je potrebno obeležiti, uključujući prevashodno tipove vremenskih izraza koji su predstavljeni u delu 2.

Ova shema za obeležavanje korišćena je za potrebe evaluacije sistema koji su učestvovali u ACE TERN (eng. *Temporal Expression Recognition and Normalisation*) (Ferro et al. 2004) i EVALITA-07[6] (eng. *Evaluation of NLP and Speech Tools for Italian*) programima. To je bila najvažnija shema (Negri and Marseglia 2005; Saquete, Muñoz, and Martínez-Barco 2006) sve dok TimeML shema nije usvojena kao standard. Međutim, kao jedan od osnovnih nedostataka ove sheme navodi se njeno ograničenje u smislu obeležavanja jedino vremenskih izraza, dok događaji i vremenski odnosi kao važni vremenski entiteti nisu obuhvaćeni.

[3]U primerima će biti ostavljana tačka ispred drugog znaka interpunkcije u onim situacijama kada predstavlja deo primera vremenskog izraza.

[4]S obzirom na to da se u prethodnoj rečenici ovog primera pominje 1999. godina, vrednost izraza *sledeće godine* biće 2000. godina.

[5]Dan, koji u navedenom primeru označava period u okviru koga se nešto dešava s određenom učestalošću, prikazan je kalendarskim oblikom. Kako tačna pozicija navedenog dana na vremenskoj osi nije poznata, za reprezentaciju njegove vrednosti koristi se karakter X.

[6]EVALITA predstavlja inicijativu za razvoj i evaluaciju alata namenjenih obradi tekstova i govora prirodnog italijanskog jezika. Dostupno na: http://www.evalita.it/2007.

3.1.3 STAG (2000/01)

Dalji radovi i bavljenje problemom obrade vremenskih infomacija teksta inicirali su nastanak naredne sheme. U okviru radionice vezane za obradu vremenskih i prostornih informacija (eng. *Workshop on Temporal and Spatial Information Processing*), održane 2001. godine u organizaciji Udruženja za računarsku lingvistiku (eng. *Association for Computational Linguistics*),[7] predstavljeni su i neki pristupi rešavanja ovog problema za engleski (Filatova and Hovy 2001; Katz and Arosio 2001) i nemački (Schilder and Habel 2001a) jezik, koji su izvršili značajan uticaj i ukazali na potrebu za proširenjem postojećih shema za obeležavanje vremenskih informacija. Baveći se izgradnjom sistema koji će omogućiti razumevanje vremenske dimenzije teksta, sugerisali su proširenje načina obeležavanja vremenskih informacija sa samo vremenskih izraza i na druge vremenske entitete, kao što su događaji i vremenski odnosi koji postoje među njima.

STAG (eng. *Sheffield Temporal Annotation Guidelines*) shema, koju su razvili (Setzer and Gaizauskas 2000; Setzer 2001), ponudila je rešenje ovog problema svojim kompletnijim pristupom obeležavanju vremenskih informacija. Njihov cilj je, osim pronalaženja vremenskih izraza, bio i identifikacija događaja, njihovo vezivanje za određeno vreme dešavanja radi njihovogređanja jednog u odnosu na drugi i uspostavljanja vremenskih odnosa koji postoje među njima. Iako je TIDES uputstvo za obeležavanje vremenskih izraza u određenoj meri prihvaćeno kao osnova, predložena STAG shema nije tako detaljna u ovom pogledu.

Na osnovu STAG sheme napravljena je razlika između jednostavnih (npr. *prošlog četvrtka*) i složenih (npr. *tri dana pre Božića*) vremenskih izraza. Pod pojmom jednostavnih vremenskih izraza podrazumevani su oni izrazi predstavljeni prilozima za vreme (npr. *juče*, *ujutru*) ili imeničkim jedinicama s predlogom koji im daje priloško vremensko značenje (npr. *pre Uskrsa*, *od 1. marta*). Događaji na koje ovi jednostavni vremenski izrazi referišu ne pripadaju ovim jedinicama. S druge strane, složeni vremenski izrazi se sastoje od jednog izraza koji označava vremensku distancu i drugog izraza koji identifikuje tačku vremena od koje se distanca meri (integrativno označavaju i distancu i događaj u odnosu na koji se distanca meri). Javljaju se u slučajevima kada je neki događaj sastavni deo vremenskog izraza, odnosno kada je vezan za leksičko jezgro jedinice koja predstavlja vremenski izraz (npr. *dva minuta nakon tvog dolaska*). Kako u navedenom primeru *dva minuta nakon tvog dolaska* iskazani vremenski interval (*dva minuta*) referiše (*nakon*)

[7] http://www.aclweb.org/website/

na događaj (*tvog dolaska*), i njegova vrednost bi trebalo da bude interpretirana u odnosu na vreme tog podređenog događaja.

I jednostavni i složeni vremenski izrazi trebalo bi da budu obeleženi pomoću SGML etikete TIMEX (nasleđeno iz MUC i TIDES shema), koja ima sledeće atribute:

- `tid`: jedinstveni identifikacioni broj vremenskog izraza u tekstu;
- `type`: tip vremenskog izraza (moguće vrednosti su DATE, TIME ili COMPLEX);
- `calDate`: izraz koji predstavlja kalendarski datum, u formatu koji je u skladu sa ISO-8601 standardom;
- `eid`: jedinstveni identifikacioni broj događaja za koji je vremenski izraz vezan;
- `signalID`: jedinstveni identifikacioni broj signala koji ukazuje na vremenski odnos između događaja i vremenskog izraza;
- `relType`: vremenski odnos koji postoji između vremenskog izraza i događaja (moguće vrednosti su BEFORE, AFTER, INCLUDES, IS_INCLUDED ili SIMULTANEOUS).

Atributi `eid`, `signalID` i `relType` se mogu primeniti samo na složene vremenske izraze, a njihova namena je da pruže informacije o tipu odnosa (`relType`) koji postoji između vremenskog izraza (`tid`) i događaja (`eid`), a koji je uspostavljen putem signala (`signalID`).

Osim TIMEX etikete, STAG shema koristi i etikete SIGNAL i EVENT. Primer teksta obeleženog u skladu sa STAG shemom dat je u primeru 3.3.

Primer 3.3.

```
Brzi voz <event eid="7" class="OCCURRENCE" tense="past"
   relatedToTime="4" timeRelType="included" signal="7"> je krenuo
   </event> iz Beograda za Suboticu <signal sid="7">u</signal>
   <timex tid="4" type="DATE" calDate="17022012">sredu</timex>.
```

U odnosu na prethodnu TIDES TIMEX2 shemu, STAG shema uvodi određene novine vezane za način obeležavanja vremenskih informacija teksta, i to:

Zahteva se i obeležavanje lingvističkog izraza događaja (tj. stvari koje se dešavaju) etiketom EVENT, u okviru koje su njihovi odnosi i druge osobine, kao što je npr. klasa događaja, precizirani kao atributi. Zbog svoje složenosti, događaji koji označavaju stanja bili su isključeni iz ove sheme.

Osim označavanja događaja, uvedeno je i obeležavanje vremenskih signala, kao elemenata koji ukazuju na postojanje odnosa između dva vremenska entiteta (vremenskog izraza i događaja, dva vremenska izraza ili dva događaja). Oni su označeni predlozima (npr. *u*, *posle*, *tokom*) i veznicima (npr. *dok*, *kada*). Recimo, u primeru 3.3 predlog *u* ukazuje na činjenicu da je događaj polaska voza povezan sa *sredom*.

Po prvi put su se zahtevale i identifikacija i predstavljanje postojećih vremenskih odnosa, ali ne putem zasebne etikete već u okviru određenih atributa EVENT etikete. Stoga, etiketa za događaj sadrži i atribut (`timeRelType`) koji ukazuje na vrstu odnosa nekog događaja sa drugim događajem ili vremenskim izrazom. Na primer, u primeru 3.3 događaj *je krenuo* uključen je, odnosno dešava se tokom vremenskog intervala predstavljenog *sredom*.

Iako je STAG shema prepoznata kao značajan napredak na polju obeležavanja vremenskih informacija uvođenjem događaja i vremenskih odnosa, ponuđena rešenja su proizvela određene probleme tokom njene kasnije primene na tekstove koji sadrže složene vremenske odnose. Stoga, činjenica da STAG shema isključuje identifikovanje događaja koji označavaju stanja, kao i nedovoljno detaljan opis vremenskih odnosa koji je uključen u etiketu EVENT, izazvali su veće interesovanje i podstakli širi krug istraživača da intenzivno nastave rad na rešavanju problema sveobuhvatnijeg obeležavanja vremenskih informacija.

3.1.4 TimeML (2002/03) i ISO-TimeML (2009)

Zainteresovanost naučne zajednice za izgradnju sistema koji će omogućiti automatsku vremensku analizu tekstova i rasuđivanje zasnovano na događajima manifestovana je brojnim važnim specijalizovanim radionicama i pratećim skupovima organizovanim u okviru konferencija, kao što su ACL (eng. *Association for Computational Linguistics*) konferencije, LREC (eng. *Language Resources and Evaluation Conference*) 2002, TERQAS (eng. *Time and Event Recognition for Question Answering Systems*) 2002, TANGO (eng. *TimeML Annotation Graphical Organizer*) radionica 2003, Dagstuhl seminar 2005, TIME 2006 - *International Symposium on Temporal Representation and Reasoning*, ARTE ACL COLING 2006 radionica. To-

kom ovih dešavanja postignut je značajan napredak koji je vodio ka osmišljavanju i pročišćavanju nove sheme poznate pod nazivom TimeML (eng. *Time Markup Language*). Iako integriše karakteristike i TIDES i STAG shema, TimeML (Pustejovsky et al. 2002, 2003) predstavlja formalni specifikacioni jezik za obeležavanje događaja, vremenskih izraza i njihovih odnosa u prirodnom jeziku, koji ima daleko opštiju svrhu u odnosu na prethodne sheme. Osim u obeležavanju vremenskih izraza i događaja, ova shema je bogatija i kompletnija posebno u pogledu prepoznavanja i označavanja vremenskih odnosa, vezivanja događaja za određeno vreme dešavanja, kao i u pogledu mogućnosti relativnog nizanja događaja jednog u odnosu na drugi.

U okviru TimeML sheme definisana je primena sledećih osnovnih anotacija, zapisanih u XML (skr. od *eXtensible Markup Language* - proširivi jezik za obeležavanje)[8] formatu: EVENT za obeležavanje događaja, TIMEX3 za obeležavanje vremenskih izraza, SIGNAL za označavanje elemenata teksta koji ukazuju na određenu vremensku relaciju, kao i TLINK, SLINK i ALINK koje identifikuju i označavaju različite tipove relacija. Primer 3.4 predstavlja rečenicu obeleženu u skladu sa TimeML shemom. Ova rečenica sadrži vremenski odnos (TLINK) između događaja (EVENT) *došao je* i vremenskog izraza (TIMEX3) *ponedeljak*, u kome učestvuje *u* kao signal (SIGNAL).

Primer 3.4.

```
Ivan <EVENT eid="e1" class="occurrence">je došao</EVENT>
   <SIGNAL sid="s1">u</SIGNAL> <TIMEX3 tid="t1" type="date"
   value="2011-08-17"> ponedeljak </TIMEX3><TLINK reltype="includes"
   relatedToEvent="e1" timeID="t1" signal="s1">
```

S obzirom na to da je predmet ovog rada obeležavanje jedino vremenskih izraza, u nastavku teksta detaljnije će biti opisana samo TIMEX3 etiketa, dok događaji i vremenski odnosi koji postoje među njima ostaju van opsega ovoga rada. Sâm naziv etikete ukazuje na razlike koje postoje između TIMEX3 i prethodnih etiketa: TIMEX etikete prisutne u MUC i STAG shemama, i TIMEX2 etikete definisane u okviru TIDES programa. Vremenski izrazi, kao lingvistički izraz informacija o tome kada se nešto dogodilo, koliko je trajalo ili koliko često se ponavlja, u okviru TimeML sheme su obeleženi etiketom TIMEX3, koja može da ima sledeće atribute:

- `tid`: jedinstveni identifikacioni broj vremenskog izraza u tekstu, čija je upo-

[8] http://www.w3.org/XML/

treba obavezna;

- `type`: takođe obavezan atribut kojim je preciziran tip prepoznatog vremenskog izraza (moguće vrednosti su DATE, TIME, DURATION i SET); neki primeri vremenskih izraza istaknuti su u primeru 3.5.

  Primer 3.5.
  (a) *Ivan je došao u* ***ponedeljak*** *i otići će* ***25. oktobra 2007. godine****.*
  (b) *Predstava počinje u* ***20 časova****.*
  (c) *Biće kod nas* ***tri nedelje****.*
  (d) *Ja odlazim u bioskop* ***mesečno****.*

  U primeru 3.5a *ponedeljak* predstavlja relativnu, a *25. oktobra 2007. godine* apsolutnu vremensku referencu na određeni kalendarski datum (DATE), dok izraz *20 časova* u primeru 3.5b izražava mnogo precizniju vremensku referencu sa granulacijom sata (TIME). Primer 3.5c uključuje imeničku frazu *dve nedelje* koja predstavlja period vremena (DURATION). Na kraju, primer 3.5d ilustruje vremenski izraz koji označava učestalost ponavljanja (SET).

- `value`: obavezan atribut u potpunosti preuzet iz TIDES TIMEX2 sheme; sadrži normalizovani oblik identifikovanog vremenskog izraza, koji je izveden iz ISO 8601 standardnog formata; neki primeri normalizovanih vrednosti vremenskih izraza dati su u primeru 3.6.

  Primer 3.6.
  (a) *25. oktobar 2008.* → `2008-10-25`
  (b) *tri nedelje* → `P3W`
  (c) *mesečno* → `XXXX-XX`
  (d) *sutra u pet popodne* → zavisi od npr. datuma kreiranja dokumenta; ako je datum kreiranja dokumenta *15. februar 2005. godine*, izraz će biti normalizovan u obliku `2005-02-16T17:00`

- `mod`: opcioni atribut, takođe u potpunosti izveden iz TIDES TIMEX2 sheme; identifikuje reči koje na određeni način modifikuju značenje vremenskog izraza (npr. ***oko*** *dva minuta*, ***krajem*** *2011. godine*);

- `temporalFunction`: binarni atribut koji ukazuje na potrebu za daljim razrešavanjem vrednosti identifikovanog vremenskog izraza; vrednost ovog atri-

buta biće pozitivna u onim slučajevima kada vremenski izraz ne sadrži sve informacije neophodne za određivanje apsolutne vrednosti (npr. *prošle nedelje*);

- `anchorTimeID`: opcioni atribut čija je vrednost uvek `tid`, odnosno jedinstveni identifikacioni broj vremenskog izraza za koji je identifikovani relativni izraz vezan i koji će poslužiti kao orijentir, odnosno referentno vreme u procesu računanja njegove normalizovane vrednosti;

- `valueFromFunction`: opcioni atribut relevantan za potrebe računanja normalizovane vrednosti;

- `functionInDocument`: atribut koji obezbeđuje vremensku referencu drugim vremenskim izrazima u tekstu; prevashodno se koristi za obeležavanje vremena nastanka teksta, vremena modifikovanja teksta, publikovanja itd.;

- `beginPoint` i `endPoint`: atributi koji se mogu koristiti kada je neko trajanje vezano za drugi vremenski izraz na osnovu koga se mogu precizirati početna i krajnja tačka identifikovanog intervala;

- `quant` i `freq`: atributi koji se koriste isključivo za opisivanje vremenskih izraza sa značenjem učestalosti; atribut `quant` je uglavnom doslovno preuzeta reč iz teksta koja precizira količinu (npr. za izraz *svake godine* atribut `quant` će imati vrednost "EVERY"); atribut `freq` sadrži ceo broj i jedinicu vremena koja predstavlja frekvenciju ponavljanja (npr. za izraz *dva puta svake nedelje* atribut `freq` će imati vrednost "2X", dok će atribut `quant` imati vrednost "EVERY").

Detaljniji opis metodologije obeležavanja vremenskih izraza u skladu sa TimeML shemom dat je u poglavlju 4, dok je metodologija normalizovanja njihovih vrednosti opširnije opisana u poglavlju 5.

Primena TimeML uputstva donela je sledeće novine u obeležavanje vremenskih informacija:

Iako je TimeML uputstvom specifikovano da TIMEX3 etiketa treba da bude primenljiva na većinu izraza koji se obeležavaju TIMEX2 etiketama, osnovna razlika među njima vezana je za ugnežđene izraze i određivanje opsega identifikovanih vremenskih izraza. Za razliku od TIDES TIMEX2 sheme, ugnežđeni vremenski izrazi više nisu dozvoljeni, već bi na osnovu TimeML

sheme trebalo da budu obeleženi kao dva vremenska izraza povezana signalom (primer 3.7). Osim toga, u slučaju složenijih vremenskih izraza koji pored iskazanog vremenskog intervala sadrže i jedinice sa značenjem događaja ili stanja, TimeML uputstvo definiše primenu TIMEX3 etikete bez njihovog uključivanja, dok bi one trebalo da budu obeležene etiketom EVENT (primer 3.8).

Primer 3.7.

```
<TIMEX3>tri nedelje</TIMEX3> <SIGNAL>posle</SIGNAL>
<TIMEX3>utorka</TIMEX3>
```

Primer 3.8.

```
<TIMEX3>četiri decenije</TIMEX3> <EVENT> iskustva </EVENT>
```

Za razliku od vremenskih izraza i događaja, koji su u tekstu predstavljeni nizom tokena, te mogu biti jasno označeni, vremenske relacije, predstavljajući asocijacije tj. veze između ovih entiteta, ne pružaju mogućnost jednostavnog i direktnog označavanja sekvenci tokena. TimeML shemom i uvođenjem TLINK, SLINK i ALINK etiketa omogućeno je detaljnije definisanje veza koje mogu da postoje između vremenskih entiteta.

U okviru tabele 3.1 sumirani su i upoređeni SGML/XML etikete i atributi, koji su propisani u prethodno opisanim shemama za obeležavanje vremenskih izraza.

Kombinovanjem i proširivanjem osobina prethodnih shema, TimeML uputstvo se izdvaja kao veoma moćan alat, koji je kao standard usvojila većina istraživača koji se bave automatskom obradom vremenskih informacija. Opšta prihvaćenost upotrebe ove sheme, kao i njen sveobuhvatan i u velikoj meri razrađen pristup sagledavanja vremenskih entiteta, doveli su do toga da TimeML bude predložen kao ISO standard (ISO 2007). Radi obezbeđivanja sistematičnog načina za ekstrakciju vremenskih informacija i olakšavanja njihove razmene, predložena TimeML shema je i usvojena kao međunarodni standard 2009. godine (ISO 2009). Iako je TimeML shema originalno razvijena za engleski jezik, više napora je uloženo u njeno prilagođavanje kako bi se mogla korektno primeniti i na druge jezike, kao što su francuski (Bittar 2009), italijanski (Caselli, Dell'Orletta, and Prodanof 2009), korejski (Im et al. 2009), rumunski (Forascu 2008), portugalski (Costa and Branco 2010), turski (Seker and Diri 2010), španski (Saurı, Saquete, and Pustejovsky 2009), kineski (Xue and Zhou 2010) itd.

Tabela 3.1: Poređenje shema za obeležavanje vremenskih izraza

Shema	**MUC-6&7 1995/97**	**TIDES 2000**	**STAG 2000/01**	**TimeML 2002/03**
Etiketa	**TIMEX**	**TIMEX2**	**TIMEX**	**TIMEX3**
Atributi	type		type	type
		val	calDate	value
		mod		mod
		anchor_dir		
			tid	tid
				temporalFunction
		anchor_val		anchorTimeID
				functionInDocument
				valueFromFunction
				beginPoint
				endPoint
				quant
				freq

U okviru ovog dela predstavljene su postojeće sheme za obeležavanje različitih aspekata vremenskih informacija u tekstovima prirodnog jezika. Činjenica da je TimeML shema usvojena kao ISO standard potvrđuje njen status najznačajnijeg jezika za obeležavanje veremenskih informacija. Međutim, u okviru sledećeg dela ćemo pokazati da je ipak potrebno uraditi još dosta toga kako bi se prevazišle sve greške i nesaglasnosti koje nastaju prilikom obeležavanja korpusa u skladu sa TimeML shemom. Radi što boljeg učinka bilo u ručnom ili automatskom obeležavanju, od izuzetne važnosti je da se shema doradi da bi se omogućila njena primena u brojnim aplikacijama koje zahtevaju pristup vremenskim informacijama sadržanim u tekstu.

Značajniji leksički resursi u vidu korpusa anotiranih u skladu sa opisanim shemama biće predstavljeni u sledećem delu.

3.2 Korpusi

Anotirani vremenski korpusi su posebna vrsta leksičkih resursa koji se najčešće primenjuju u rešavanju problema obrade vremenskih informacija. To su kolekcije

mašinski čitljivih nestrukturiranih tekstova koji su obeleženi oznakama odgovarajućeg vremenskog značenja. Zajedno sa postojećim shemama za obeležavanje vremenskih informacija, anotirani korpusi predstavljaju korisne izvore koji su dostupni široj zajednici, te se o njima može diskutovati, tj. mogu se analizirati i prerađivati u širim okvirima. Ručno obeležavanje teksta predstavlja vremenski zahtevan i monoton proces, ali može biti od koristi, pre svega, kao skup na osnovu koga će se omogućiti obučavanje i formiranje modela zasnovanih na mašinskom učenju, dok će, s druge strane, pružiti osnovu za evaluaciju performansi sistema koji sprovode automatsko obeležavanje vremenskih informacija. Anotirani tekstovi koji se koriste za evaluaciju su u literaturi poznati pod nazivom *zlatni standard* (eng. *gold standard*). Tako je prevashodno za svaku definisanu shemu za obeležavanje vremenskih informacija sastavljen po jedan korpus, s ciljem obezbeđivanja objektivne osnove za vrednovanje takmičarskih algoritama. Za evaluaciju učinka automatskog TIMEX2 obeležavanja u okviru ACE TERN takmičenja 2004. godine upotrebljen je tzv. TERN korpus, koji je sadržao tekstove na engleskom i kineskom jeziku. Iako se TERN korpus smatra do sada najpouzdanije obeleženim izvorom za obradu vremenskih informacija, njegova analiza vremenskih informacija ograničena je jedino na prepoznavanje i normalizaciju vremenskih izraza. TimeBank je ručno obeležen korpus koji, osim vremenskih izraza, sadrži i događaje i vremenske odnose, označene u skladu sa TimeML shemom.

Iako postoje i drugi manji resursi dostupni za proučavanje različitih vremenski osetljivih problema, oni zbog ograničenja prostora neće biti predstavljeni u ovom radu. U ovom delu biće opisani postojeći anotirani korpusi, koje su istraživači najčešće upotrebljavali za proučavanje različitih aspekata vremena.

3.2.1 TERN korpus anotiran u skladu sa TIDES TIMEX2 shemom

TERN korpus (Ferro et al. 2004) je ručno obeležen korpus korišćen za potrebe ACE TERN takmičenja 2004. godine (Ferro 2004), kako bi se omogućila evaluacija učinka sistema koji vrše automatsko TIMEX2 obeležavanje. U odnosu na klase vremenskih izraza definisane u okviru MUC konferencija, ACE TERN 2004 proširuje definiciju vremenskih izraza koje je potrebno identifikovati u tekstu, i uvodi primenu novih atributa koji opisuju njihove osobine. TERN korpus sadrži nestrukturirane podatke na engleskom i kineskom jeziku, obeležene u skladu sa TIDES TIMEX2 shemom za obeležavanje.

TERN podaci na engleskom jeziku, sastavljeni od vesti preuzetih iz informativnih programa, novinskih i servisnih izveštaja različitih izvora, sadržali su 767 dokumenata korišćenih za obuku i 192 dokumenta namenjena evaluaciji sistema. Za anotiranje korpusa upotrebljeni su alati *Alembic Workbench* (Day et al. 1997) i *Callisto* (Day et al. 2004), koji anotatorima omogućavaju kreiranje sopstvenih anotacionih shema. Ceo korpus su anotirala tri anotatora, nakon čega je izvršena procena slaganja među njima (eng. *inter-annotator agreement*). Usledio je proces diskusije i usaglašavanja sve dok nije postignuto značajno slaganje anotatora veće od 90% u pogledu prepoznavanja vremenskih izraza i određivanja njihove normalizovane vrednosti. Na osnovu postignutog slaganja anotatora, Lisa Fero, koautor TIMEX2 uputstva, proizvela je konačnu anotiranu verziju korpusa koja će kasnije biti korišćena kao zlatni standard za poređenje.

U odnosu na zlatni standard, pre svega uočeno je neslaganje anotatora usled čestog izostavljenog obeležavanja vremenskih izraza koji su predstavljeni prilozima i pridevima, dok je izuzetno slaganje postignuto u slučaju izraza predstavljenih imenicama i numeričkim obrascima. Kada je reč o opsegu identifikovanih vremenskih izraza, anotatori se u većini slučajeva nisu osvrtali na kontekst reči-okidača, te su često izostavljali one reči koje modifikuju značenje izraza (npr. ***skoro*** *dva meseca*). To je proizvelo i greške u obeležavanju MOD atributa, što se dešavalo i u situacijama kada anotatori nisu bili sigurni koju vrednost za ovaj atribut treba da odaberu (u slučaju navedenog primera *skoro dva meseca*, jedan anotator je za prilog *skoro* odabrao vrednost APPROX, dok je drugi smatrao je da je ispravno LESS_THAN). U slučaju neslaganja, odnosno grešaka nastalih u određivanju normalizovane vrednosti vremenskog izraza tj. vrednosti VAL atributa, anotatori su najčešće pravili greške prilikom kucanja ili računanja vrednosti i korišćenja kalendara. Još jedna vrsta grešaka se vezuje za obeležavanje atributa ANCHOR_DIR i ANCHOR_VAL, a nastala je jer su anotatori često zaboravljali da ih primene ili nisu obraćali dovoljno pažnje na sve informacije prisutne u tekstu, što sve govori u prilog automatskom obeležavanju vremenskih izraza. Važno je reći da su anotatori takođe imali problema sa razumevanjem uputstava za obeležavanje, kao i pamćenjem svih detalja specifikovanih uputstvima.

Ipak, TERN korpus se i danas smatra najpouzdanije obeleženim izvorom za obradu vremenskih informacija, sa značajnim slaganjem anotatora koje nije dostigao više ni jedan drugi izvor.

3.2.2 Korpusi anotirani u skladu sa TimeML shemom

TimeBank korpus

Tokom 2002. godine, u okviru TERQAS radionice, jedan od ciljeva koji je proizašao iz postavljenih zadataka bilo je definisanje opšteg meta-standarda za označavanje događaja, njihovih vremenskih referenci, kao i veza koje postoje među njima, kasnije poznat kao TimeML standard. Da bi se podržao predloženi koncept, za drugi cilj postavljena je izgradnja korpusa koji bi ručno označili anotatori vremenskim izrazima, događajima i vremenskim odnosima u skladu sa TimeML specifikacijama, s ciljem izrade zlatnog standarda u obeležavanju vremenskih informacija.

Razvoj TimeBank korpusa (Pustejovsky et al. 2006) započeo je odabirom 300 tekstova različitih medijskih informativnih izvora, uključujući tekstove DUC (eng. *Document Understanding Conference*) korpusa (sadrži biografije, opise jednog ili više događaja), tekstove ACE programa (prepise emitovanih vesti i servisnih informacija), kao i Propbank korpusa (novinski članci *Wall Street* časopisa). TimeBank podaci obeleženi su temeljnim i detaljnim oznakama događaja, vremenskih izraza, vremenskih signala i, što je najvažnije, oznakama veza koje ukazuju na postojeće vremenske relacije među ovim entitetima. Ovako efikasna reprezentacija vremena i vremenskih odnosa između događaja može da pruži pouzdanu osnovu za buduća empirijska istraživanja vezana za način iskazivanja i međusobnog povezivanja događaja u tekstu, istraživanje značenja gramatičkih kategorija glagolskog vida i glagolskih vremena, otkrivanje narativne strukture tekstova ili kao pomoć pri izgradnji sistema sposobnih za automatsko vremensko zaključivanje.

Veći deo TimeBank korpusa obeležilo je po pet anotatora koji su učestvovali u definisanju TimeML sheme. Nakon diskusije i usaglašavanja među anotatorima, preostalih 30% korpusa anotirali su studenti Univerziteta Brandeis, koji nisu imali nikakvo prethodno iskustvo u korišćenju TimeML sheme. Proces anotacije korpusa sproveden je u dve faze. U pripremnoj fazi, prvo je upotrebljen alat za automatsko prepoznavanje jednostavnih vremenskih izraza u skladu sa TimeML shemom, kako bi se smanjila količina ručnog rada i održala pažnja anotatora koju treba da usmere na prepoznavanje složenijih vremenskih izraza. Nakon toga, primenjena je modifikovana verzija *Alembic NLP* sistema (Day et al. 1997), s ciljem obeležavanja glagolskih fraza koje mogu ukazati na događaj. Tek nakon ovih procesa pripremne obrade, anotatori su uz pomoć *Alembic Workbench* alata za anotiranje označavali

postojeće događaje, vremenske izraze, signale i vremenske relacije parova entiteta koji su u međusobnom odnosu. Identifikovani entiteti su u tekstu bili obeleženi XML etiketama. Za razliku od TERN korpusa, nije preduzeta odgovarajuća kontrola kvaliteta anotacija, niti je uopšte pokušano postizanje slaganja anotatora od barem 90%.

Prva verzija ovog korpusa namenjena javnosti (TimeBank verzija 1.1) sastojala se od 186 anotiranih tekstova odabranih iz početne kolekcije od 300 tekstova. Druga verzija korpusa, TimeBank verzija 1.2, izdata je nakon revidiranja prethodne i sadržala je 183 dokumenta, odnosno nešto više od 61.000 reči, što je isuviše mala kolekcija podataka da bi mogla da bude od koristi za potrebe mašinskog učenja. Na osnovu 10 dokumenata ove poslednje verzije korpusa, koji su anotirala dva iskusna anotatora, izvršena je procena slaganja anotatora, merenog na osnovu postignute prosečne mere preciznosti i odziva. Zvanični podaci ukazuju na niže slaganje anotatora po pitanju određivanja klase događaja (78%) ili anotiranja vremenskih relacija (55%), što pokazuje da, usled izrazite neodređenosti u interpretaciji vremena i vremenskih odnosa koja postoji u prirodnom jeziku, ni ljudima vreme nije sasvim razumljiv fenomen i da anotiranje tekstova u skladu sa TimeML shemom i nije tako jednostavnan zadatak. O određenim postojećim problemima koji se tiču TimeML sheme i TimeBank korpusa govore sami autori koji u svojim radovima ukazuju na postojanje protivrečnosti u obeležavanju TimeBank korpusa (Boguraev and Ando 2005, 2006; Derczynski and Gaizauskas 2010a). S druge strane, uočeno je to da u okviru TimeBank korpusa vremenski izrazi koji ukazuju na vreme dana i učestalost nisu podjednako zastupljeni kao izrazi koji označavaju kalendarske datume i trajanja.

AQUAINT korpus

AQUAINT korpus (Graff 2002) je još jedan od korpusa anotiranih na osnovu TimeML sheme, odnosno istih uputstava korišćenih za anotiranje TimeBank 1.2 korpusa. Pripreman u organizaciji američkog Nacionalnog instituta za standarde i tehnologiju (NIST) kao zlatni standard programa AQUAINT (eng. *Advanced Question Answering for Intelligence*),[9] ovaj korpus dostigao je veličinu od oko 375 miliona reči i približno 3 GB podataka. Izabrani tekstovi potiču iz tri izvora: kineske novinske agencije *Xinhua*, kao i novinskih agencija *New York Times* i *Associated Press Worldstream*. Sam proces anotiranja sproveden je na sličan način kao i u slučaju TimeBank korpusa. Od 2012. godine dostupna je poslednja prečišćena i

[9]Dostupno na: http://www-nlpir.nist.gov/projects/aquaint/

unapređena verzija ovog korpusa (33.973 reči) i on je, kao i TimeBank korpus, korišćen kao zlatni standard u evaluaciji sistema učesnika TempEval-3 takmičenja.

TempEval korpusi

TempEval korpus (Verhagen et al. 2007), zasnovan na TimeBank 1.2 korpusu, kreiran je za potrebe TempEval takmičenja, koje je organizovano u okviru evaluacione kampanje SemEval-2007,[10] kada su se po prvi put zadaci obeležavanja vremenskih informacija pojavili kao deo SemEval izazova. S obzirom na to da je pomenuto TempEval-1[11] takmičenje bilo usmereno samo na identifikaciju vremenskih relacija koje postoje između događaja i vremenskih izraza i između glavnih događaja dve uzastopne rečenice, takmičarima je bio dostavljen korpus sa već unetim anotacijama za događaje i vremenske izraze (uključujući podatak o datumu kreiranja dokumenta), koje su doslovno preuzete iz korpusa TimeBank 1.2. Na osnovu iskustva u obeležavanju vremenskih odnosa u okviru TimeBank korpusa, takmičarima je postavljen zadatak automatskog određivanja pojednostavljenih TLINK etiketa TimeML sheme, što podrazumeva unapred definisane parove entiteta koji mogu biti u vremenskom odnosu. Organizatori TempEval-1 takmičenja su očekivali da će, zahvaljujući ovako pojednostavljenim zadacima, proces pripreme tekstova biti olakšan, te da će obezbediti postizanje određenog nivoa doslednosti u obeležavanju, i na taj način značajno uticati na postizanje veće saglasnosti među anotatorima. Zaista je anotatorima za anotiranje korpusa bilo potrebno mnogo manje vremena nego što je utrošeno za anotiranje TimeBank korpusa, ali je postignuto slaganje između anotatora ostalo daleko ispod očekivanih 90%. Po pitanju obeležavanja veza koje postoje između događaja i vremenskih izraza anotatori su postigli saglasnost od 72%, dok je postignuto slaganje anotatora u hronološkom nizanju događaja iznosilo 65%. Ovi pokazatelji su potvrdili činjenicu da je proces izbora odgovarajućih vremenskih relacija između dva entiteta još uvek izuzetno složen zadatak, što je uticalo na oblikovanje i razlaganje budućih zadataka, uz definisanje detaljnih uputstava za obeležavanje.

Sledeće takmičenje, održano u okviru SemEval-2010 kao TempEval-2 (Verhagen et al. 2010), omogućavalo je takmičarima da se, osim zadacima sa prethodnog takmičenja, pozabave i samo prepoznavanjem događaja ili prepoznavanjem i normalizacijom vremenskih izraza. Za razliku od TempEval-1 korpusa koji je sadržao

[10] Ovo je bila četvrta u nizu Senseval evaluacionih kampanja, čiji je cilj bio procena jačih i slabijih strana sistema koji se bave semantičkom analizom teksta.

[11] Takmičenje održano u okviru SemEval-2007 zapravo je poznato pod nazivom TempEval, ali ćemo koristiti naziv TempEval-1 kako bi se izbegle zabune.

samo tekstove na engleskom jeziku, korpus korišćen za TempEval-2 takmičenje sastojao se od tekstova na još pet jezika (kineskom, italijanskom, francuskom, korejskom i španskom jeziku). Na osnovu TimeML 1.2.1 uputstva za obeležavanje razvijene su detaljne smernice za obeležavanje vremenskih informacija i za preostalih pet jezika. Ipak, nisu svi korpusi sadržali podatke neophodne za rešavanje svih zadataka. Dok su se etikete za događaje i vremenske izraze nalazili u svim jezičkim varijantama korpusa, zadatak prepoznavanja svih unapred definisanih vremenskih relacija bio je omogućen samo učesnicima koji su razvijali sisteme za engleski, kineski i francuski jezik. Proces anotacije ovih korpusa sproveden je u dve faze. Po dva anotatora su obeležavala dokumente, nakon čega je usledila faza razrešavanja neslaganja među anotatorima, da bi treći anotator doneo konačnu odluku. Za većinu jezika anotatori su koristili *Brandeis Annotation Tool* (*BAT*) (Verhagen 2010) kao alat za obeležavanje korpusa. Engleska verzija TempEval-2 korpusa bila je u potpunosti zasnovana na TimeBank korpusu. Ipak, sve unete etikete koje označavaju događaje prošle su kroz ponovni proces provere kako bi se osigurala njihova usaglašenost sa poslednjom verzijom uputstava za obeležavanje. Osim toga, dodate su i oznake svih vremenskih relacija čije prepoznavanje je definisano zadacima TempEval-2 takmičenja.

S obzirom na to da obeležavanje vremenskih informacija u tekstu zahteva izuzetno mnogo ljudskog rada, za potrebe prethodnih TempEval takmičenja korišćeni su anotirani korpusi ograničenih veličina. Budući da je uočeno da određeni postojeći sistemi sprovode automatsko obeležavanje koje karakteriše pouzdanost koja je bliska saglasnosti među anotatorima, među organizatorima TempEval-3 (UzZaman et al. 2013) takmičenja rodila se ideja da proizvedu veći korpus koji će biti automatski obeležen i nad kojim će ljudi obaviti manje ispravke. Ovog puta u okviru SemEval-2013, a za potrebe TempEval-3 takmičenja, pripremljeni su anotirani korpusi engleskog i španskog jezika. Kada je reč o engleskom jeziku, kao zlatni standard upotrebljene su prečišćene i unapređene verzije TimeBank i AQUAINT korpusa od po 61.418 reči, odnosno 33.973 reči. Osim toga, sastavljene su još dve kolekcije tekstova, i to: jedna manja, ručno obeležena kolekcija potpuno novih tekstova (*platinasta*), i druga mašinski obeležena kolekcija, automatski sastavljena na osnovu rezultata više sistema koji sprovode automatsko obeležavanje vremenskih informacija (*srebrna*).

TempEval-3 *platinasti* korpus (6.375 reči) korišćen za evaluaciju su obeležili, a nakon toga i recenzirali organizatori takmičenja. Svaki dokument su nezavi-

sno anotirala dva anotatora, dok je treći bio zadužen za njihovo usaglašavanje i donošenje konačne odluke. Za proces anotiranja korišćeno je TimeML v1.2.1 uputstvo za obeležavanje (Sauri et al. 2006), dok su neki anotatori svoje odluke donosili i na osnovu sugestija datih u okviru TIPSem sistema (Llorens, Saquete, and Navarro-Colorado 2012), o kome će biti više reči u delu 3.3.6. Postignuta saglasnost među anotatorima, sračunata kao F1-mera,[12] po pitanju prepoznavanja vremenskih izraza i normalizacije njihovih vrednosti iznosi 87%, odnosno 88%.

TempEval-3 srebrni korpus, namenjen takođe za potrebe evaluacije, čine tekstovi ukupne veličine 666.309 reči. Ova kolekcija tekstova automatski je obeležena vremenskim informacijama korišćenjem sistema TIPSem, TIPSem-B (Llorens, Saquete, and Navarro 2010) i TRIOS (UzZaman and Allen 2010). Ovi sistemi zasnovani na mašinskom učenju ponovo su obučeni koristeći zlatni standard, odnosno prečišćene i unapređene verzije TimeBank i AQUAINT anotiranih korpusa. Ipak, prvi put upotrebljeni srebrni korpus nije se pokazao korisnim kada je reč o vremenskim izrazima i klasifikaciji vremenskih relacija, ali bi mogao biti od koristi u obučavanju sistema koji se bave ekstrakcijom događaja. Naime, neki autori su pokušali da utvrde u kojoj meri korišćenje srebrnog korpusa za obuku sistema može da utiče na poboljšanje performansi, pa su došli do zaključka da upotreba veće količine automatski obeleženih podataka ne postiže statistički značajnu razliku u rezultatu (Filannino, Brown, and Nenadic 2013; Kolomiyets and Moens 2013). S druge strane, evaluacija je pokazala da bi automatski obeležene veće kolekcije tekstova mogle biti od koristi u obučavanju sistema koji se bave ekstrakcijom događaja, jer je sistem obučen na srebrnom korpusu u zadatku prepoznavanja događaja postigao daleko bolji rezultat u odnosu na pstale verzije (Kolomiyets and Moens 2013).

3.3 Različiti pristupi projektovanju sistema za ekstrakciju vremenskih izraza

Na osnovu prethodnog prikaza specifičnosti obrade vremenskih informacija, odnosno opisa definicija postojećih shema za obeležavanje i njihove primene, može se uočiti da su načini prikazivanja vremenskih informacija postepeno bivali sve bogatiji, što je uslovilo i usložnjavanje procesa automatske analize. Za iz-

[12] Jedna od standardnih mera za procenu učinka sistema za pronalaženje i ekstrakciju informacija. Predstavlja harmonijsku sredinu između odziva i preciznosti, detaljno objašnjenim u delu 6.

gradnju sistema namenjenih automatskoj obradi vremenskih informacija, kao i za bilo koji drugi zadatak ekstrakcije informacija, najčešće su korišćena dva osnovna principa, i to: metode zasnovane na pravilima, odnosno lingvističkom znanju i statističke metode. U poslednje vreme sve češće se primenjuju hibridni pristupi koji kombinuju pomenute pristupe, kako bi se nedostaci jednog nadomestili prednostima drugog.

Metode zasnovane na pravilima (eng. *rule-based*) karakteriše usmerenost na specifičnosti jezika i primena lingvističkog znanja radi formiranja modela, odnosno skupova unapred definisanih pravila na osnovu kojih će se vršiti ekstrakcija informacija iz teksta. Razvoj sistema zasnovanih na ručno formiranim modelima zahteva više utrošenog vremena eksperata, koji će uz pomoć svojih znanja i veština, a nakon izvršene analize teksta, definisati pravila tj. formalne opise informacija koje je potrebno ekstrahovati iz teksta. U skladu sa analizama dobijenog izlaza, prvobitno napisana pravila moguće je modifikovati i popravljati više puta, sve dok se ne postigne što viši nivo i odziva i preciznosti.[13] Pravila za ekstrakciju, osim ručnom metodom, mogu biti formirana i metodom vođeni na podacima (eng. *data driven*), odnosno pomoću mašinskog učenja, koje iziskuje velike količine tekstova prethodno obeleženih informacijama koje je potrebno ekstrahovati. Nad ovako obeleženim tekstovima sistem pokreće algoritam za treniranje, koji uči i produkuje neku vrstu „pravila“ za ekstrakciju informacija na proizvoljnom tekstu. Sve propuste i greške korisnik može da uoči, ali ne i da ih koriguje, već sistem sam ponovnim procesom treniranja modifikuje svoja pravila s ciljem poboljšanja performansi.

S druge strane, statističke metode bitno se razlikuju od prethodnih u samom pristupu koji ne iziskuje znanje o jeziku. Rečenice su posmatrane kao niz tokena (pojedinačnih reči) ili većih segmenata (eng. *chunks*) kojima se pomoću modela (npr. posmatrajući učestalosti njihovog zajedničkog pojavljivanja u tekstu) dodeljuju kategorije, kao vrste entiteta koje je potrebno ekstrahovati. Nakon određivanja kategorija, svi tokeni ili segmenti iste kategorije proglašavaju se za entitete kojima ta kategorija odgovara i predstavljaju pomoću određenih unapred definisanih osobina.

U sledećem delu teksta će hronološkim redosledom biti predstavljeni najistak-

[13] Više reči o ovim standardnim merama, koje se koriste za procenu uspešnosti sistema za pronalaženje i ekstrakciju informacija, biće u šestom poglavlju.

nutiji sistemi koji se bave automatskim obeležavanjem vremenskih izraza u tekstovima prirodnog jezika.

3.3.1 Konferencije o razumevanju poruka (MUC konferencije)

Prvi sistemi koji su se bavili identifikacijom i klasifikacijom vremenskih izraza u tekstovima prirodnog jezika predstavljeni su u okviru poslednje dve MUC konferencije. Dok je zadatak automatskog prepoznavanja vremenskih izraza na MUC-6 konferenciji podrazumevao identifikaciju jedino apsolutnih vremenskih izraza, tokom poslednje u nizu MUC konferencija 1997. godine u ovaj zadatak uključeni su i relativni vremenski izrazi. Zbog ograničenog prostora, u daljem tekstu bićemo usmereni na prikaz najuspešnijih sistema MUC-7 konferencije, od kojih su se mnogi bavili problemom automatske ekstrakcije vremenskih izraza i u okviru prethodne MUC-6 konferencije.

Sistem koji je, u odnosu na ostale učesnike MUC-7 konferencije, postigao najbolji učinak u rešavanju zadatka ekstrakcije svih imenovanih entiteta razvijen je u okviru Grupe za jezičke tehnologije Univerziteta u Edinburgu (Mikheev, Grover, and Moens 1998). Nakon sprovedene tokenizacije teksta i označavanja vrstama reči (eng. *Part-of-Speech tagging, PoS*), LTG (eng. *Language Technology Group*) sistem za identifikaciju vremenskih izraza koristi, pre svega, pravila u obliku gramatika. S obzirom na to da je većina vremenskih izraza koji se javljaju u novinskim tekstovima predstavljena u izuzetno strukturiranom obliku, razumljiva je primena metode zasnovane na pravilima, koja na jednostavan način omogućava prepoznavanje i izdvajanje onih delova teksta koji mogu biti deo vremenskog izraza. Osim ovih pravila u vidu gramatika, za izdvajanje delova teksta koji mogu biti vremenski izraz upotrebljen je i popis određenih vremenskih entiteta (npr. imena dana, meseci itd.), pripremljen u vidu liste (eng. *gazetteer*). U poslednjoj fazi rada sistema, SGML transduktor vrši objedinjavanje prikupljenih informacija i njihovo obeležavanje TIMEX etiketama. U odnosu na ostale sisteme učesnike MUC-7 konferencije, LTG sistem postigao je najviše uspeha u prepoznavanju vremenskih izraza koji predstavljaju datume (F-mera 93,73%). Na osnovu mišljenja autora relativno nizak odziv (79%), postignut u slučaju vremenskih izraza koji označavaju delove dana, rezultat je nedovoljno precizno datih uputstava, koje su za učesnike pripremili organizatori konferencije.

Drugi sistem koji je takođe učestvovao u MUC-7 izazovu i postigao najbolji učinak u prepoznavanju vremenskih izraza koji označavaju delove dana (89,03%) bio

je FACILE (eng. *Fast and Accurate Categorization of Information by Language Engineering*) (Black, Rinaldi, and Mowatt 1998). Ovaj sistem zasnovan na pravilima, čiji formalizam u potpunosti podržava kontekstno osetljivo parcijalno parsiranje teksta, omogućava identifikovanje delova teksta koji se odnose, između ostalih entiteta, i na vremenske izraze četiri jezika, i to: engleskog, nemačkog, italijanskog i španskog jezika. U fazi predprocesiranja, odn. pripremnoj fazi, FACILE sprovodi tokenizaciju i morfološku analizu teksta, nakon čega sledi pronalaženje onih tokena koji su smešteni u bazu podataka i na taj način raspoređeni u određene semantičke klase. Koristeći obeležja izdvojena u okviru ovih faza, kao što su morfološke, sintaksičke i semantičke informacije, sistem kreira tzv. vektor svojstava (eng. *feature vector*). Atributi, koji su rezultat različitih nivoa analize, smešteni u vektor svojstava i na taj način povezani sa odgovarajućim tokenom, na jednostavan način se mogu biti upotrebiti kao sastavni delovi pravila za ekstrakciju imenovanih entiteta. Već u sledećoj razvojnoj fazi sistem FACILE (Ciravegna et al. 1999) za potrebe ekstrahovanja imenovanih entiteta koristi kaskade konačnih transduktora.

Sličan pristup primenjen je i na Univerzitetu u Šefildu za izgradnju sistema LaSIE-II (eng. *Large Scale Information Extraction*) (Humphreys et al. 1999), koji predstavlja noviju verziju LaSIE sistema kreiranog za potrebe učešća u MUC-6 izazovima (Gaizauskas et al. 1995). LaSIE-II sistem koristi različite tehnike, uključujući modele konačnih stanja za prepoznavanje leksičkih formi specifičnih za određene domene, kontekstno slobodne gramatike za parcijalno parsiranje, pojednostavljenu semantičku reprezentaciju svake rečenice teksta i formalnu reprezentaciju čitavog diskursa, koji služi za ekstrakciju informacija i razrešenje koreferenci. S te strane, LaSIE-II sistem se može posmatrati, ne samo kao odraz teorije sprovođenja ekstrakcije informacija, već kao pokušaj upotrebe različitih tehnika obrade prirodnog jezika i ispitivanja mogućnosti dobijenih njihovom interakcijom. Ovaj sistem integrisan je u GATE (eng. *General Architecture for Text Engineering*) platformu (Cunningham et al. 1997), koju čini niz modula namenjenih sprovođenju različitih procesa obrade teksta, obezbeđujući na taj način visoko modularan pristup obradi jezika. GATE upravlja informacijama o tekstu, koje su proizveli svi integrisani moduli, i obezbeđuje njihov grafički prikaz, kontrolišući tok pokretanja različitih kombinacija modula i primenu sistema za ekstrakciju informacija. Za potrebe prepoznavanja imenovanih entiteta, LaSIE-II primenjuje kaskadu gramatika, sastavljenih od približno 400 ručno pisanih pravila, koja koriste etikete sa oznakama vrste reči i semantičke informacije dobijene na osnovu postojeće liste imenovanih entiteta (eng. *gazetteer*). Kada je reč o identifikaciji i klasifikaciji vre-

menskih izraza, LaSIE-II postiže uspešnost u prepoznavanju izraza tipa DATE od 90,90%, dok za izraze koji označavaju delove dana F-mera iznosi 87,44%.

Najraniji pristupi, karakteristični za radove predstavljene u okviru MUC-6 konferencije, bili su zasnovani na ručno rađenim pravilima u vidu gramatika, što je omogućavalo podjednako uspešno pronalaženje kako datuma i vremena ustaljenog formata, tako i čitav asortiman izraza prirodnog jezika koji su definisani uputstvima. Kada je reč o metodama korišćenim u okviru MUC-7 konferencije, primenjivala su se oba rešenja, i ona zasnovana na transduktorima, kao što je opisano u (Mikheev, Grover, and Moens 1998; Krupka 1995), kao i druge tehnike poput skrivenih Markovljevih modela (eng. *Hidden Markov model, HMM*) u (Miller et al. 1998). Tokom ova poslednja dva MUC takmičenja, uspešnost najboljih sistema u prepoznavanju vremenskih izraza bila je velika (Krupka and Hausman 1998). Učesnici MUC-7 konferencije su ukazivali na posebnu težinu zadatka prepoznavanja vremenskih izraza zbog nedovoljno precizno datih uputstava za prepoznavanje relativnih izraza vezanih za datume i delove dana. Ipak je većina sistema koji su učestvovali u ovom izazovu, uključujući i prethodno pomenute, postigla visok nivo uspešnosti u opsegu od 87,82-93,73% za identifikaciju datuma i 86,36-89,03% za identifikaciju vremena kao delova dana (Chinchor 1998). S druge strane, (Mani and Wilson 2000) smatraju da je zadatak identifikacije vremenskih izraza bio olakšan činjenicom da je barem 30% datuma i vremena predviđenih za prepoznavanje imalo ustaljenu formu, jednostavnu za prepoznavanje korišćenjem malog broja pravila. U okviru MUC-7 konferencije evaluirana je postignuta preciznost sistema u obeležavanju vremenskih izraza, dok se razrešavanje njihovih vrednosti nije zahtevalo.

3.3.2 Mani i Vilson (2000)

Jedan od prvih sistema koji je izuzetno doprineo razvoju procesa normalizacije vremenskih izraza bio je TempEx (Mani and Wilson 2000), kasnije razrađen i upotrebljen za izgradnju GUTime sistema (Verhagen et al. 2005), o kome će biti više reči u delu 3.3.4. Za razliku od prethodnih radova predstavljenih u okviru MUC konferencija koji su se bavili isključivo prepoznavanjem vremenskih izraza, Mani i Vilson se usredsređuju na potrebu za razrešavanjem njihovih vrednosti. Pre svega, razmatrali su problem normalizacije, ne samo apsolutnih, već i onih vremenskih izraza koji su tokom MUC-7 konferencije posmatrani kao relativni vremenski izrazi i koji označavaju vremena zavisna od govornikovog vremena ili nekog vremena iskazanog drugim vremenskim izrazom (npr. *sada*, *danas*, *sutra*, *sledećeg utorka*, *pre*

dve nedelje, dva dana posle 15. juna itd.). U okviru tog rada predložena je i shema za obeležavanje vremenskih izraza, koja za reprezentaciju vremena u skladu sa ISO-8601 standardom predviđa upotrebu novog atributa VAL. Osim onih oblika za predstavljanje vremenskih izraza već definisanih ovim ISO standardom, autori predlažu i primenu liste tokena koji predstavljaju jedinice vremena često korišćene u vremenskim izrazima (npr. godišnje doba leto predstavljeno je tokenom SU od engleskog izraza *summer*). Svi predlozi vezani za upotrebu atributa VAL bili su prihvaćeni i kasnije uključeni u TIMEX2 shemu za obeležavanje vremenskih izraza. TempEx ubrzo počinje da koristi najnoviju verziju TIMEX2 standarda iz decembra 2001. godine, te tako postaje i prvi sistem koji je za obeležavanje vremenskih izraza koristio TIDES TIMEX2 shemu.

TempEx sistem je relativno jednostavan alat napisan u programskom jeziku Perl koji putem konačnih automata primenjuje nekoliko heurističkih pravila zasnovanih na morfološkim informacijama. Nakon sprovedene tokenizacije i označavanja teksta vrstama reči, ovaj sistem prvo koristi modul za identifikovanje vremenskih izraza, a zatim i modul koji razrešava vrednosti samostalnih, odnosno eksplicitnih vremenskih izraza. Po završetku ovih procesa, pokreće se modul za razrešavanje vrednosti onih vremenskih izraza, koje nisu eksplicitno sadržane u samom izrazu već zavise od datuma kreiranja dokumenta ili prethodno pomenutog vremenskog izraza. Navedeni moduli zasnovani su na ručno rađenim i naučenim pravilima. Učinak TempEx sistema evaluiran je u odnosu na ručno obeležen korpus koji se sastojao od 221 novinskog članka. Kada je reč o prepoznavanju vremenskih izraza, TempEx postiže F-meru od 96,2%. Većina grešaka nastalih u ovoj fazi prepoznavanja odnosi se na propuštene izraze, čiji oblici pojavljivanja nisu opisani pravilima (npr. godina napisana slovima). S druge strane, uspešnost TempEx sistema u određivanju vrednosti prepoznatih vremenskih izraza iznosila je 83,2%. Prilikom normalizacije ispravno prepoznatih vremenskih izraza najveći izvor grešaka bile su situacije kada su određenim izrazima koji su upotrebljeni u širem značenju dodeljene vrednosti užeg značenja (npr. prilog *danas* može imati uže značenje određenog dana, ali i šire referišući na današnje vreme, čije trajanje vremenski nije precizno određeno). Ovaj problem određivanja specifične nasuprot opštoj upotrebi pojedinih vremenskih izraza autori su pokušali da reše primenom metoda klasifikacije zasnovanih na mašinskom učenju. Za učenje pravila koja će poslužiti za razlikovanje opšte nasuprot specifičnoj upotrebi vremenskih izraza autori su upotrebili neke od najpopularnijih sistema za klasifikaciju, ali su samo najuspešnija pravila, naučena pomoću C4.5 (Quinlan 1993), inkorporirali u svoj

sistem.

Iako TempEx sistem predstavlja jedan od prvih pristupa koji otkriva nova značajna saznanja vezana za proces normalizacije vremenskih izraza, važno je pomenuti i njegova postojeća ograničenja. TempEx ne vrši automatsku identifikaciju određenih oblika apsolutnih i relativnih vremenskih izraza sa značenjem datuma ili vremena (npr. *krajem 1999. godine*, *juče ujutru*), kao ni onih vremenskih izraza koji ukazuju na trajanje, odnosno vremenski interval koji nije vezan za neko vreme u odnosu na koje se interval meri (npr. *dve godine*). Kada je reč o procesu normalizacije, razrešavane su samo vrednosti onih izraza koji referišu na datum ili vreme, dok su svi ostali tipovi vremenskih izraza bili isključeni (npr. *za nekoliko sati*). I pored ovih ograničenja, smatralo se da TempEx sistem može da pruži dobru osnovu za razvoj novih alata, te je tokom prve organizovane evaluacije sistema u automatskom prepoznavanju i normalizaciji vremenskih izraza svim učesnicima dat na raspolaganje kao eksterni izvor koji će poslužiti za otkrivanje određenih karakteristika teksta.

3.3.3 TERN zadatak

Zahvaljujući uspehu koji su postigle MUC konferencije, narednih godina pokrenut je veliki broj inicijativa (npr. *Text REtrieval Conference - TREC*,[14] *Cross-Language Evaluation Forum - CLEF*,[15] *Conference on Natural Language Learning - CoNLL*,[16] *Evaluation Exercises for the Semantic Analysis of Text - Senseval*)[17] za sprovođenje istraživanja i rad na izgradnji sistema koji će omogućiti automatsko razumevanje tekstualnih podataka. Već od 1999. godine, kada je održan prvi u nizu ACE programa,[18] uvedena su tri nova zadatka u smislu automatskog prepoznavanja entiteta, relacija i događaja nekog teksta. Evaluacija uspešnosti sistema u automatskom prepoznavanju i normalizaciji vremenskih izraza sprovedena je u okviru ACE programa 2004. godine (TERN 2004), dok je već naredne godine (ACE05) po prvi put ovaj izazov postavljen kao poseban zadatak - TERN zadatak. Osnovni cilj TERN zadatka bio je izgradnja sistema koji će vršiti automatsko prepoznavanje vremenskih izraza u tekstovima prirodnog jezika i normalizaciju njihovih vrednosti u skladu sa TIDES TIMEX2 shemom za obeležavanje. Kako bi se tokom evaluacije TERN 2004 omogućilo objektivno vrednovanje učinka takmičarskih algoritama,

[14] http://trec.nist.gov/
[15] http://www.clef-campaign.org/home.html
[16] http://ifarm.nl/signll/conll/
[17] http://www.senseval.org/
[18] http://www.itl.nist.gov/iad/mig/tests/ace/

kao zlatni standard upotrebljen je TERN korpus, opisan u delu 3.2.1. Pre svega, merena je uspešnost sistema u identifikaciji, odnosno otkrivanju barem delova vremenskih izraza koji su već bili obeleženi u TERN korpusu, te je izlazna etiketa sistema smatrana tačnom ako je imala makar minimalno preklapanje sa etiketom obeleženom u zlatnom standardu. Zatim, merena je i sposobnost sistema da za sve tačno prepoznate vremenske izraze ispravno odredi i njihov pun opseg, uz očekivano apsolutno poklapanje opsega vremenskog izraza sistema i zlatnog standarda. Na samom kraju, vrednovana je i uspešnost sistema u sprovođenju procesa normalizacije, te je očekivano da za sve ispravno prepoznate veremenske izraze budu određene i vrednosti atrbuta uključenih u TIMEX2 etiketu (VAL, MOD, SET, ANCHOR_VAL i ANCHOR_DIR), od kojih je svaki bio posebno evaluiran.

Pred učesnike takmičenja TERN 2004 postavljena su dva zadatka, uz mogućnost odabira problema kojim će se baviti. Prvi zadatak podrazumevao je samo prepoznavanje vremenskih izraza, te je uspešnost sistema koji su se bavili ovim problemom vrednovana u odnosu na njihovu sposobnost identifikovanja i određivanja opsega vremenskog izraza. U okviru drugog zadatka učesnicima je data mogućnost da se, osim prepoznavanjem vremenskih izraza, pozabave i problemom normalizacije njihovih vrednosti. Za razliku od prvih sistema koji su se bavili prepoznavanjem vremenskih izraza metodama zasnovanim na pravilima, svi učesnici ovog takmičenja, koji su rešavali samo problem prepoznavanja, gradili su sisteme bazirane na statističkim metodama.

Jedan od sistema učesnika, koji je koristeći statistički pristup obavljao automatsko otkrivanje i obeležavanje vremenskih izraza u tekstovima engleskog i kineskog jezika, razvijen je na Univerzitetu Kolorado pod nazivom ATEL (Hacioglu, Chen, and Douglas 2005). Proces ekstrakcije vremenskih izraza posmatran je kao problem nadgledanog etiketiranja, odnosno deljenja u skupove prema pripadnosti određenoj klasi. Nakon segmentacije teksta na rečenice pomoću programa MXTERMINATOR zasnovanog na modelu maksimalne entropije, ovaj sistem sprovodi tokenizaciju teksta u skladu sa neznatno modifikovanom verzijom *Penn Tree Bank* standarda,[19] na osnovu čega je svaka rečenica korišćenog TERN korpusa konvertovana u vertikalni prikaz na nivou tokena kojima su dodeljene određene etikete u zavisnosti od njihove pozicije u okviru vremenskog izraza. Dakle, postojeća XML reprezentacija vremenskih izraza TERN korpusa transformisana je u reprezentaciju vremenskih izraza korišćenjem zagrada (obeležavanje slično tzv. BIO

[19] http://www.cis.upenn.edu/~treebank/tokenization.html

formatu),[20] koja ukazuje na to da li je token na početku („(*“), kraju („*)“), unutar („(*)“) vremenskog izraza, ili mu pak ne pripada („()“). Nakon izvršene odgovarajuće reprezentacije podataka, definisano je više specifičnih osobina koje mogu biti pridružene svakom tokenu i koje su grupisane u četiri opšte klase: leksičke (npr. sam token, oznaka vrste reči, njegova frekventnost u rečniku), sintaksičke (npr. osnovni segmenti fraza), semantičke (npr. osnovno značenje reči, odnosi zavisnosti između tokena i osnovnog značenja reči) i eksterne osobine. Pod pojmom eksternih osobina podrazumevale su se osobine određene TempEx sistemom, programom za etiketiranje vremenskih izraza zasnovanog na pravilima koji je organizator takmičenja TERN 2004 distribuirao učesnicima. Koristeći na ovaj način predstavljene podatke i njima pridružene osobine, autori obučavaju klasifikatore zasnovane na vektorima podrške tj. podržavajućim vektorima (eng. *Support Vector Machines, SVM*) radi identifikacije i klasifikacije ugnježđenih[21] vremenskih izraza. ATEL je evaluacijom na test podacima TERN korpusa pokazao uspešnost u otkrivanju vremenskih izraza za engleski i kineski jezik od 93,5%, odnosno 90,5%, dok su rezultati za određivanje opsega iznosili 87,8% za engleski i 78,6% za kineski jezik.

Dok je u okviru TERN 2004 takmičenja za rešavanje zadatka prepoznavanja vremenskih izraza prevashodno bio usvojen statistički pristup (Hacioglu, Chen, and Douglas 2005; Carpenter 2004; Ittycheriah et al. 2003), kombinovani zadatak koji je podrazumevao i prepoznavanje i normalizaciju bio je rešavan korišćenjem pristupa zasnovanih na znanju. Najdetaljnije opisan sistem u objavljenoj literaturi jeste sistem Chronos (Negri and Marseglia 2005), razvijen u italijanskom Institutu za naučna i tehnološka istraživanja Zavoda za kulturu Trentino (ital. *Istituto Trentino di Cultura, Instituto per la Ricerca Scientifica e Tecnologica, ITC-irst*). Dizajniran kao višejezički sistem, Chronos sprovodi automatsko prepoznavanje i normalizaciju vremenskih izraza engleskog i italijanskog jezika metodom zasnovanom na pravilima. U osnovi strukture ovog sistema nalaze se dve komponente: prva je namenjena prepoznavanju, a druga normalizaciji. Nakon tokenizacije teksta, etiketiranja vrstama reči i prepoznavanja različitih oblika složenih reči na osnovu

[20] U oblasti računarske lingvistike BIO format je jedan od uobičajenih formata za obeležavanje tokena teksta i njihovu klasifikaciju. Prefiks B (Beginning) ispred postavljene etikete ukazuje na to da označeni token predstavlja početak, dok prefiks I (Inside) označava token koji je deo segmenta koji je potrebno prepoznati. O (Outside) etiketa služi za obeležavanje tokena koji ne pripadaju ni jednom segmentu koji je potrebno prepoznati.

[21] Ugnježđeni vremenski izrazi se sastoje od elemenata različitih granularnosti. Na primer, u izrazu *leta prošle godine* element *leta* bi trebalo da bude obeležen etiketom, koja će biti uključena u etiketu kojom je označen ceo izraz.

automatski ekstrahovane liste iz WordNet-a, približno 1.000 ručno rađenih pravila upotrebljeno je za otkrivanje svih mogućih vremenskih izraza prisutnih u tekstu, određivanje njihovog opsega, kao i za prikupljanje dodatnih informacija iz konteksta, koje će poslužiti u kasnijoj fazi normalizacije. Osim ovih osnovnih pravila, na sledećem nivou upotrebljena su i pravila za razrešavanje dvosmislenosti nastalih usled mogućeg višestrukog označavanja. U ovoj fazi procesa, pored atributa definisanih TIMEX2 shemom, uvedena su u upotrebu još četiri dodatna atributa, čije će vrednosti sistem iskoristiti u sledećoj fazi normalizacije za računanje konačne normalizovane vrednosti vremenskog izraza. Dakle, u okviru prve faze su na osnovu lingvističke analize teksta sprovedeni identifikacija i određivanje opsega otkrivenih vremenskih izraza, što za rezultat daje neku vrstu posredne anotacije koja sadrži sve informacije neophodne za sledeći proces normalizacije. Druga komponenta, zadužena za interpretaciju vremenskih izraza, upotrebiće isti skup pravila korišćen za identifikaciju vremenskih izraza i ovu posrednu anotaciju kako bi se TIMEX2 atributima svakog otkrivenog vremenskog izraza pripisale tačne vrednosti. Ovaj proces izvodi se u tri koraka, i to: određivanje orijentira (eng. *anchor*) neophodnog za razrešavanje vrednosti relativnih vremenskih izraza, zatim normalizacija datuma tj. unošenje vrednosti atributa VAL, i na kraju uklanjanje privremenih atributa koji su bili upotrebljeni za računanje normalizovanih vrednosti. Na taj način je, po završetku ovih procesa, stvarana konačna verzija teksta koji je obeležen u skladu sa TIMEX2 shemom. Ovakvim pristupom Chronos je nadmašio sve sisteme učesnike takmičenja TERN 2004 u pogledu određivanja onih atributa koji oslikavaju uspešnost procesa normalizacije (VAL: 87,2%, MOD: 77%, ANCHOR_DIR: 76% i ANCHOR_VAL: 72%).

Ipak, kada je reč o otkrivanju i određivanju opsega vremenskih izraza najbolji učinak postigao je Aerotext (Cassel et al. 2006), sistem razvijen pod okriljem kompanije *Lockheed Martin*, koja je poznata po proizvodnji u oblastima aerosvemirske industrije, odbrane, sigurnosti i naprednih tehnologija. Ovaj sistem usmeren na ekstrakciju informacija iz tekstova prirodnog jezika, prepoznaje i vremenske izraze i normalizuje njihove vrednosti pomoću ručno pisanih pravila. U slučaju relativnih vremenskih izraza, normalizovane vrednosti su prvo prikazane kao intervali koji postoje između vremena navedenog u izrazu i orijentira (u ovom slučaju, vremena kreiranja dokumenta). Pomoću drugog skupa pravila ovi intervali su prevedeni u normalizovane oblike, propisane zahtevima takmičenja TERN 2004.

Još jedan sistem koji je u okviru evaluacije TERN 2004 metodom zasnovanom

na pravilima pristupio rešavanju problema prepoznavanja i normalizacije vremenskih izraza jeste TimexTag sistem (Ahn, Fissaha Adafre, and De Rijke 2005). Skup ručno rađenih pravila, organizovanih kao konačni automat, upotrebljen je za obeležavanje vrsta reči i pronalaženje onih segmenata teksta u čijoj osnovi se nalaze leksički okidači (eng. *trigger words*), odnosno reči vremenskog značenja. Za potrebe normalizacije su većim delom iskorišćena pravila za prepoznavanje vremenskih izraza, dopunjena obrascima koji će omogućiti ekstrahovanje onih elemenata izraza koji su neophodni za računanje normalizovanih vrednosti, kao i funkcijama koje izvršavaju računanje u odnosu na vreme kreiranja dokumenta. Kada je reč o uspešnosti TimexTag sistema, zvanični rezultati TERN evaluacije beleže postignutu *F* meru od 69,6% za zadatak identifikacije vremenskih izraza, 58,4% za određivanje opsega i 69,9% za uspešnost u određivanju normalizovane vrednosti, odnosno vrednosti VAL atributa. TimexTag sistem je bio jedan od učesnika i kasnije organizovanog takmičenja ACE TERN 2007, kada su autori, ne bi li uticali na poboljšanje performansi sistema, za potrebe prepoznavanja i normalizacije vremenskih izraza ovoga puta primenili metode zasnovane na podacima, o čemu će biti više reči u sledećem delu teksta.

3.3.4 Pre TempEval takmičenja

Osim takmičenja ACE TERN druge veće evaluacije u međunarodnim okvirima, koje bi omogućile direktno poređenje performansi sistema namenjenih automatskom prepoznavanju i normalizaciji vremenskih izraza, nisu bile organizovane, već su autori izrađivali sopstvene evaluacione okvire za merenje postignute uspešnosti. Određena istraživanja vezana za probleme obrade vremenskih informacija bila su predstavljena u okviru raznih konferencija i radionica (COLING 2002, ACL 2001, LREC 2002, TERQAS 2002, TANGO 2003, Dagstuhl 2005) sve do pojave kasnije organizovanih TempEval izazova (Verhagen et al. 2007, 2010; UzZaman et al. 2013). Neki od najznačajnijih sistema ovog perioda biće predstavljeni u ovom delu.

GUTime sistem (Verhagen et al. 2005), razvijen na Univerzitetu Džordžtaun, rezultat je rada uloženog u dalji razvoj ranije opisanog TempEx sistema (Mani and Wilson 2000) i njegovo prilagođavanje zahtevima za prepoznavanje elemenata definisanih TimeML shemom. Primenjen je kao jedan od ukupno pet modula sistema TARSQI (eng. *Temporal Awareness and Reasoning Systems for Question Interpretation*), koji predstavlja kompletan TimeML sistem za obeležavanje vremenskih izraza, događaja i odnosa koji postoje među njima u novinskim tekstovima. Autori

su smatrali da se automatsko obeležavanje ovih entiteta najuspešnije može sprovesti kaskadnom primenom modula, koji tokom obrade dokumentu sukcesivno dodaju sve više i više TimeML anotacija. U tom smislu, GUTime sistem je korišćen kao modul za otkrivanje vremenskih izraza i umetanje TIMEX3 etiketa. U odnosu na TempEx sistem, pokrivenost različitih oblika vremenskih izraza mnogo je veća. Kada je reč o normalizaciji prepoznatih vremenskih izraza, ovaj sistem koristi pristup kojim se prvo identifikuju funkcije koje je potrebno sprovesti (npr. u slučaju relativnog izraza *juče* funkcija bi bila „minus jedan dan“), a zatim se u sledećoj fazi na osnovu ove funkcije računaju prave vrednosti izraza. EVITA je modul zadužen za prepoznavanje događaja, dok Slinket automatski identifikuje odnose koji postoje između događaja i koji su u TimeML shemi predstavljeni SLINK etiketom. Vremenski odnosi koji postoje između vremenskih izraza i događaja obeležavaju se korišćenjem GUTenLINK modula, koji koristi sintaksičke i leksičke informacije. Na samom kraju, SputLink je komponenta koja zaokružuje čitav proces automatskog anotiranja pronalaženjem novih potencijalnih odnosa povezujući poznate relacije koje postoje između vremenskih izraza i događaja. Ovaj sistem prevashodno zasnovan na pravilima, jedino za prepoznavanje događaja koristi statističke tehnike. Smatra se da je osnovni nedostatak TARSQI sistema sama struktura koja na više nivoa koristi omotače (eng. *wrapper*), te je analiza pravila otežana, kao i njihovo jednostavno dodavanje i menjanje.

Osim GUTime sistema, još jedan sistem koji koristi TimeBank korpus za implementaciju modula za prepoznavanje vremenskih izraza opisan je u (Boguraev and Ando 2005). Primenjujući hibridni pristup, autori u jednom delu koriste kaskadu konačnih gramatika za prepoznavanje i normalizaciju vremenskih izraza, dok u drugom putem tehnika mašinskog učenja, na osnovu velikih količina neobeleženih podataka, sprovode identifikaciju i klasifikaciju događaja i odnosa koji postoje među vremenskim entitetima. Dakle, kada je reč o vremenskim izrazima, metodom zasnovanom na pravilima prepoznaju se vremenski entiteti (npr. *unit*, *point*, *period*, *relation* itd.), koji dalje mogu biti razloženi na manje jedinice (npr. dan, mesec, godina za *unit* ili interval, trajanje za *period*). Na osnovu dalje analize izvršeno je izdvajanje više svojstava ovih entiteta, odnosno atributa, kao što su granularnost,[22] kardinalnost,[23] pravac referentnog vremena[24] itd., koji su neophodni

[22]Granularnost se odnosi na nivo detaljnosti vremenske jedinice (npr. izraz *pet dana* je granularnosti dan).

[23]Kardinalnost predstavlja meru broja vremenskih entiteta (npr. izraz *pet dana* ima kardinalnost 5).

[24]Pravac referentnog vremena ukazuje na operaciju koja će biti izvršena (npr. oduzimanje od ili sabiranje sa vrednošću referentnog vremena).

za proces normalizacije i izvođenje vrednosti vremenskog izraza. Iako je u okviru evaluacije TERN 2004 utvrđeno da se proces prepoznavanja vremenskih izraza gotovo podjednako uspešno može sprovoditi i metodama zasnovanim na pravilima i statističkim metodama, neophodno je napraviti razliku između procesa prepoznavanja opsega vremenskog izraza i procesa predstavljanja njegovih atributa. Ukoliko je prepoznavanje posmatrano samo kao proces određivanja opsega, onda se može reći da ne postoji ništa što bi dalo prednost metodama zasnovanim na pravilima u odnosu na statističke metode. Međutim, kako su ovi atributi, jednostavni za otkrivanje putem pravila u vidu gramatika, nevidljivi za metode mašinskog učenja (npr. korpus TimeBank ne sadrži tako detaljne oznake), prepoznavanje sprovedeno metodama zasnovanim na pravilima omogućiće otkrivanje nekih dodatnih podataka koji su neophodni za uspešnu interpretaciju vrednosti vremenskog izraza u okviru normalizacije. S druge strane, kaskada konačnih gramatika omogućava i parcijalnu sintaksnu analizu, tretirajući vremenske izraze kao lingvističke jedinice, za razliku od mnogih drugih dotašanjih sistema, koji ih posmatraju jednostavno kao leksičke jedinice tj. reči koje ukazuju na vreme. Autori ističu da ovakav pristup u velikoj meri može olakšati zadatak identifikovanja događaja, jer je na osnovu sintaksičkih osobina predloško-padežne konstrukcije *tokom zimovanja na Kopaoniku* moguće zaključiti da imenski argument *zimovanje* vremenskog predloga *tokom* najverovatnije imenuje događaj. Ovu vrstu informacije koju je jednostavno uhvatiti putem parsiranja, odnosno sintaksne analize koja se ne ograničava samo na TimeML komponente, autori će upotrebiti za izvođenje svojstava potrebnih klasifikatorima čiji je zadatak pronalaženje događaja i relacija. Dakle, vremenski izrazi obeleženi su korišćenjem konačnih gramatika, kojima su na posredan način označeni i signali, događaji i relacije. Na kraju, njihovu konačnu identifikaciju sprovode klasifikatori zasnovani na prinicipu minimizacije empirijskog rizika (eng. *empirical risk minimization*) putem analize TimeBank korpusa i drugih velikih neobeleženih korpusa.

Na osnovu analize publikovanih radova jasno je da su se i tokom narednog perioda očuvala dva glavna metodološka pravca u rešavanju problema prepoznavanja i normalizacije vremenskih izraza. Neki autori (Mazur and Dale 2007) usvajaju pristup zasnovan na pravilima i, u pokušaju da daju sopstveni doprinos na polju predstavljanja vremenskih izraza, na osnovu prethodnog rada uloženog u razvoj GATE sistema (Cunningham et al. 1997) grade sopstveni sistem poznat pod nazivom DANTE (eng. *Detection and Normalization of Temporal Expressions*). Ovaj sistem, namenjen obeležavanju vremenskih izraza engleskog jezika vrši njihovo

prepoznavanje i normalizaciju u skladu sa TIMEX2 standardom. Nakon tokenizacije teksta, segmentacije na rečenice i označavanja vrstama reči, pokreće se deo za prepoznavanje imenovanih entiteta, zatim deo za prepoznavanje vremenskih izraza i, na kraju, deo koji sprovodi interpretaciju vremenskih izraza. Autori koriste termin interpretacija jer smatraju da je to ono što se zapravo dešava tokom procesa normalizacije – interpretacija vremenskih izraza u kontekstu ostalih delova dokumenta. Rezultat faze prepoznavanja dat je u vidu posrednog formata, koji nosi lokalnu semantiku identifikovanih vremenskih izraza, odnosno interpretaciju onih informacija sadržanih u samom izrazu. Ovaj deo sistema zadužen za prepoznavanje vremenskih izraza implementiran je pomoću gramatike JAPE, čija se pravila pokreću nad dokumentom u pet faza. Svaka naredna faza koristi anotacije uvedene tokom prethodne obrade teksta (npr. tokenizacija, etiketiranje vrstama reči) i prethodnim fazama gramatike JAPE. JAPE pravila su tradicionalna pravila „obrazac-akcija“, što znači da leva strana pravila sadrži obrazac koji je potrebno pronaći, a desna akciju koju je potrebno sprovesti nakon njegovog pronalaženja. Osim pravila, sistem u fazi prepoznavanja upotrebljava i liste, koje sadrže popis segmenata najčešće korišćenih u izrazima za datume i vremena, brojeve napisane slovima, zatim nazive dana, meseci i vremenskih zona. Ovaj modul razvijen je na osnovu TIMEX2 smernica i u njima sadržanih primera vremenskih izraza. Korpus upotrebljen tokom evaluacije ACE05 autori su iskoristili za identifikaciju nekih složenijih primera vremenskih izraza, težih za automatsko prepoznavanje. Rešavanje ovog problema predstavljalo je drugi stadijum razvoja DANTE sistema. Drugi deo sistema, koji vrši normalizaciju, prolazi kroz dokument rečenicu po rečenicu. Za sve one izraze kojima je potreban kontekst za određivanje vrednosti, lokalna semantika zabeležena posrednim formatom transformisana je u globalnu semantičku reprezentaciju, odnosno značenje koje ima na nivou dokumenta. Ovaj modul je u potpunosti implementiran u programskom jeziku Java i sadrži funkcije za računanje datuma i vremena. U zavisnosti od interpretacije vremenskog izraza (npr. potpuno ili nedovoljno precizna tačka na vremenskoj osi, relativni izraz, trajanje, učestalost itd.), obavljane su različite operacije, i to: izjednačavanje sa nekim referentnim vremenom i dodavanje ili oduzimanje određenog broja jedinica od referentnog datuma. Isto tako je u zavisnosti od tipa prepoznatog vremenskog izraza donošena i odluka o tome koji će TIMEX2 atributi (osim atributa VAL) biti generisani. Zahvaljujući modularnoj strukturi i delu za normalizaciju koji ne zavisi od jezika, autori smatraju da je prilagođavanje sistema za primenu na druge jezike olakšano. DANTE sistem učestvovao je u evaluaciji ACE07 i na osnovu zvaničnih rezultata postigao je ukupnu *F* meru od 57,2%. Kada je reč o

prepoznavanju vremenskih izraza, postignuta uspešnost sistema je 56,1%. Bez obzira na postojeće sličnosti DANTE i Chronos sistema, koje se ogledaju u upotrebi pravila u vidu gramatika namenjenih identifikaciji vremenskih izraza, učinak DANTE sistema u određivanju opsega vremenskih izraza je niži, uz postignutu *F* meru od 75,9%. Iako se DANTE sistem pokazao kao veoma uspešan u vršenju normalizacije (*F* mera iznosila čak 98,9%), opšta uspešnost sistema bila je na nižem nivou u odnosu na ostala tri učesnika ACE07 izazova (Aerotext, IBM, TimexTag).

Za razliku od DANTE sistema, čiji autori problem obrade vremenskih izraza rešavaju metodom zasnovanom na pravilima, drugi autori kao što je Ahn (Ahn, Adafre, and Rijke 2005; Ahn, Rantwijk, and Rijke 2007), nastavljaju sa unapređivanjem svog ranijeg pristupa (Ahn, Fissaha Adafre, and De Rijke 2005), ali ovog puta koristeći alternativne tehnike mašinskog učenja. Kako su rezultati evaluacije TERN 2004 pokazali da se za zadatak prepoznavanja vremenskih izraza gotovo podjednako uspešno mogu primeniti obe metode, autori su rešili da ispitaju u kojoj se meri mogu poboljšati performanse oba procesa ako se u fazi prepoznavanja upotrebe neke od metoda zasnovanih na podacima (Ahn, Adafre, and Rijke 2005). Upotrebili su savremeniju metodu baziranu na uslovnim slučajnim poljima (eng. *Conditional Random Fields, CRF*) (Lafferty, McCallum, and Pereira 2001) za ekstrakciju vremenskih izraza nekog teksta. Nad TERN korpusom obeleženim vremenskim izrazima putem XML etiketa, sistem je obučen za kreiranje pravila ili obrazaca koji će poslužiti za prepoznavanje novih instanci. Zadatak prepoznavanja fraza sa značenjem vremenskih izraza sveden je na obeležavanje reči jednom od etiketa B, I ili O (XML format za obeležavanje transformisan je u BIO format). Nakon upoređivanja postignutih rezultata sistema zasnovanih na različitim principima, autori zaključuju da se za prepoznavanje primenom metoda mašinskog učenja može postići i premašiti uspešnost sistema zasnovanih na pravilima (bez obzira na činjenicu da primena metoda mašinskog učenja utiče na povećanje odziva i smanjenje preciznosti), ali da se glavni problem preciznog određivanja opsega identifikovanog vremenskog izraza i dalje uspešnije rešava metodama zasnovanim na pravilima. Kada je reč o procesu normalizacije, autori su došli do jednostavnog i očekivanog rezultata, koji potvrđuje pretpostavku da se uspešnijim prepoznavanjem neposredno utiče i na poboljšanje uspešnosti procesa normalizacije, bez obzira na primenjenu metodu tokom faze prepoznavanja. Autori ističu da se proces normalizacije, nesumnjivo, može uspešno rešiti jedino primenom metoda zasnovanih na pravilima, ali da i u ovoj fazi postoje određeni problemi, čijem rešavanju u velikoj meri mogu doprineti metode zasnovane na podacima. Gotovo

polovina grešaka nastalih u fazi normalizacije odnosi se na pogrešno određivanje pravca (prošlost, sadašnjost ili budućnost) u računanju vrednosti nedovoljno preciznih ili anaforičkih izraza (npr. *utorak*), što autori vide kao klasifikacioni problem koji se na najjednostavniji način može rešiti primenom tehnika mašinskog učenja. Tokom kasnije organizovanog takmičenja ACE07, (Ahn, Rantwijk, and Rijke 2007) prepoznavanje razmatraju kao binarni problem klasifikacije fraza (ne pojedinačnih reči), odnosno sintaksičkih konstituenata koji jesu ili nisu vremenski izraz. Nakon identifikovanja vremenskih izraza, vrši se njihova dalja klasifikacija na različite semantičke tipove (npr. učestalost, trajanje, tačka u vremenu ili višeznačni primeri) putem SVM klasifikatora. Sledeći zadatak normalizacije vremenskih izraza razložen je u više faza: pronalaženje vremenskog orijentira za razrešavanje vrednosti relativnih vremenskih izraza, praćenje referentnog vremena za razrešavanje anaforičkih vremenskih izraza, određivanje magnitude i pravca relativnih vremenskih izraza (opet pomoću SVM, koristeći isti skup osobina, ali uzimanjem u obzir i kategorije vremena okolnih glagola kao osobinu), rešavanje problema uže nasuprot opštoj upotrebi nekih izraza, kao i krajnje računanje tj. kombinovanje rezultata prethodnih koraka kako bi se dobila konačna vrednost. Problem praćenja referentnog vremena razrešen je korišćenjem dva modela: jedan koji koristi vreme kreiranja dokumenta kao referencu za sve neodređene vremenske izraze, i drugi koji koristi najbliži vremenski izraz kao referentno vreme za anaforičke vremenske izraze. Opet je fokus samo na uspešnosti u određivanju vrednosti atributa VAL, sa rezultatima koji su obećavajući (najbolji model postiže 0,77% *F* meru za VAL atribut).

Osim ove dve metode, po prvi put se u radu (Poveda, Surdeanu, and Turmo 2009) primenjuje samorazvijajuća metoda (eng. *bootstraping*) kojom se iz velikih neobeleženih korpusa ekstrahuju vremenski izrazi. Reč je o slabo nadgledanom sistemu koji koristi mali inicijalni skup obrazaca ekstrakcije i neobeleženi korpus za obuku. Na osnovu ovog malog broja ručno obeleženih primera, ponavljanjem se vrši ekstrakcija obrazaca i njihovo svrstavanje, a zatim i ponovna primena na korpus radi ekstrahovanja novih primera koji će biti dodati inicijalnom skupu pravila. Dakle, sistem od malog početnog skupa obrazaca ekstrakcije razvija kolekciju pravila i na taj način sam gradi sistem. Dalja interakcija korisnika i sistema ne postoji, tj. korisnik nema mogućnost intervenisanja ili podešavanja pravila ekstrakcije. Iako nov i interesantan, ovaj pristup postigao je rezultate koji su daleko lošiji u odnosu na druge metode (u najboljem scenariju sistem postiže *F* meru od 60,59%).

U pogledu daljeg unapređenja istraživanja na polju obrade vremenskih informacija, od velikog značaja biće serija zadataka TempEval, organizovanih u okviru takmičenja SemEval nakon 2007. godine.

3.3.5 TempEval-1

U okviru izazova SemEval-2007 po prvi put je, kao nov zadatak, uključen zadatak iz oblasti obrade vremenskih informacija, poznat pod nazivom zadatak TempEval (Verhagen et al. 2007). Iako je osnovni cilj istraživanja na ovom polju automatska identifikacija vremenskih izraza, događaja i vremenskih odnosa koji postoje u tekstu, kao što je specifikovano TimeML uputstvom za obeležavanje (Sauri et al. 2006), organizatori su smatrali da bi prvi evaluacioni izazov trebalo da bude usmeren samo na automatsku kategorizaciju vremenskih odnosa (događaj – vremenski izraz, događaj – datum kreiranja dokumenta i relacije među glavnim događajima uzastopnih rečenica). Činilo se da bi, i iz perspektive organizatora koji treba da pripreme dokumenta za razvoj i testiranje sistema, kao i iz perspektive učesnika, rešavanje svih TimeML zadataka bilo isuviše zahtevno. S obzirom na to da su u to vreme bili dostupni samo TimeML podaci na engleskom jeziku, samo sistemi razvijeni za engleski jezik mogli su da učestvuju.

Tokom ovog takmičenja korišćen je korpus TimeBank. U prethodnom periodu su korpusi TimeML i TimeBank već korišćeni kao osnova u rešavanju zadataka automatskog obeležavanja vremenskih izraza, događaja i relacija u brojnim istraživačkim projektima (Mani et al. 2006; Boguraev and Ando 2006). Za potrebe takmičenja TempEval-1, iz korpusa TimeBank v1.2 doslovno su preuzete TIMEX3 etikete kao pomoć prilikom rešavanja zadatka pronalaženja vremenskih relacija.

3.3.6 TempEval-2

Drugi u nizu izazova TempEval (Verhagen et al. 2010), zasnovan na prethodnom takmičenju TempEval-1, pred učesnike postavlja šest zadataka. Osim zadataka definisanih prethodnim izazovom, koji se odnose na određivanje relacija, ovoga puta se od učesnika očekuje i automatsko obeležavanje vremenskih izraza i događaja. Kada je reč o vremenskim izrazima, zadatak je podrazumevao određivanje opsega vremenskih izraza postavljanjem TIMEX3 etiketa u skladu sa TimeML shemom i vrednosti atributa TYPE i VAL. Učesnici su imali mogućnost da odluče hoće li se baviti svim postavljenim zadacima ili će se fokusirati na samo neke od

njih. Za potrebe ovog takmičenja učesnicima su, osim na engleskom, bili obezbeđeni podaci na još pet jezika: španskom, italijanskom, francuskom, korejskom i kineskom jeziku. Ipak, važno je naglasiti da nisu svi korpusi sadržali podatke neophodne za rešavanje svih zadataka takmičenja (npr. korpus na korejskom jeziku omogućavao je učestvovanje samo u zadacima obeležavanja vremenskih izraza i događaja). U okviru TempEval-2 takmičenja učestvovalo je osam timova, sa ukupno 18 različitih verzija sistema, prevashodno namenjenih engleskom jeziku. Svi timovi su rešavali problem prepoznavanja i normalizacije vremenskih izraza engleskog jezika, dok su za španski učestvovala samo dva tima sa ukupno tri prijavljena sistema. Uspešnost prijavljenih sistema u određivanju opsega vremenskih izraza engleskog jezika, odnosno postignuta *F* mera kretala se od 26% do 86%. Ispostavilo se da je zadatak određivanje tipa vremenskih izraza bio lakši od normalizacije, jer je sistem TERSEO postigao uspešnost tj. *F* meru u određivanju tipa od 98%, dok je HeidelTime bio najuspešniji u normalizaciji sa *F* merom od 85%.

Sistem koji je postigao najbolju uspešnost u određivanju opsega i normalizovane vrednosti vremenskih izraza bio je HeidelTime (Strötgen and Gertz 2010). Ovaj sistem baziran na znanju predstavlja deo sistema UIMA (eng. *Unstructured Information Management Architecture*) namenjenog automatskoj obradi audiovizuelnih i tekstualnih dokumenata. Nakon segmentacije teksta na rečenice, tokenizacije i označavanja vrstama reči HeidelTime sistem vrši određivanje opsega vremenskih izraza, njihovog tipa i normalizovane vrednosti. Skup ručno pisanih pravila, grupisanih na osnovu četiri tipa vremenskih izraza koji su definisani TimeML uputstvom za obeležavanje, upotrebljen je za identifikaciju, određivanje opsega i tipa vremenskih izraza, kao i za definisanje funkcija na osnovu kojih će se izvršiti normalizacija. S obzirom na to da su eksplicitni vremenski izrazi u potpunosti precizirani, njihova vrednost se direktno prenosi korišćenjem odgovarajuće funkcije pravila za normalizaciju. Kada je reč o implicitnim vremenskim izrazima, poput naziva prazika (npr. *Božić*), neophodno je poznavanje njihovog značenja u kontekstu radi ispravnog određivanja normalizovane vrednosti. Prilikom rešavanja jednog od najizazovnijih zadataka, odnosno normalizacije relativnih vremenskih izraza, kao referentno vreme HeidelTime sistem koristi datum kreiranja dokumenta, što ima smisla kad se radi o novinskim tekstovima, ili prethodno pomenut datum. Za razrešavanje dvosmislenih situacija upotrebljavano je i lingvističko znanje, odnosno podaci o vrstama reči i upotrebljenom glagolskom vremenu u rečenici. HeidelTime, kao najuspešniji sistem izazova TempEval-2, postigao je najbolje rezultate u određivanju opsega vremenskih izraza (86%), kao

i u zadatku normalizovanja njihovih vrednosti (85% normalizovanih vrednosti je određeno ispravno).

Najbolji rezultat u određivanju tipa vremenskih izraza postigao je još jedan sistem zasnovan na znanju - sistem TERSEO+T2T3 transduktor (Saquete 2010). Ovde se zapravo radi o kombinaciji dva nezavisna sistema, i to: TERSEO sistema (Saquete, Muñoz, and Martínez-Barco 2006), namenjenog obeležavanju vremenskih izraza u skladu sa TIDES TIMEX2 shemom, i T2T3 transduktor koji vrši konverziju TIMEX2 u TIMEX3 etikete definisane TimeML shemom. TERSEO sistem je razvijen na Univerzitetu Alikante kao sistem zasnovan na znanju, namenjen prepoznavanju događaja i njihovog redosleda dešavanja u tekstovima španskog jezika, ali je automatski proširen na engleski, italijanski i katalonski jezik. Radi automatske obrade događaja, TERSEO vrši i identifikaciju i normalizaciju vremenskih izraza obeležavajući ih TIMEX2 etiketama. Autori su sa ovim sistemom učestvovali na TERN 2004 takmičenju, postigavši uspešnost u određivanju opsega i normalizovane vrednosti vremenskih izraza od 86,2%, odnosno 69,8%. Za potrebe TempEval-2 izazova, na dobijeni TIMEX2 izlaz primenjen je T2T3 transduktor, razvijen zahvaljujući saradnji univerziteta Alikante i Brandajs. Ovaj transduktor primenjuje skup pravila s ciljem usaglašavanja shema za obeležavanje TIDES i TimeML, ali samo za engleski jezik. Budući da u pogledu određivanja opsega vremenskih izraza između shema TIDES TIMEX2 i TimeML postoje određene razlike, prvo su primenjena pravila za adaptaciju već određenog TIMEX2 opsega. Kako se dve prethodno pomenute sheme međusobno razlikuju i u pogledu korišćenih atributa, bilo je neophodno primeniti i pravila koja na osnovu formata vremenskog izraza određuju vrednost atributa TYPE (npr. ako vrednost vremenskog izraza počinje slovom P, onda se sa sigurnošću može zaključiti da se radi o vremenskom izrazu koji ukazuje na trajanje). Iako najbolji u određivanju tipa vremenskog izraza (98%), sistem TERSEO+T2T3 transduktor nije postigao značajnije rezultate u pogledu procesa normalizacije (65%).

TRIPS i TRIOS (UzZaman and Allen 2010) je sistem koji je učestvovao u svim zadacima TempEval-2 takmičenja, a po uspešnosti u zadacima određivanja tipa i normalizovane vrednosti vremenskih izraza osvojio je drugo mesto. Autori su za rešavanje svih zadataka koristili hibridni pristup, kombinujući rešenja zasnovana na znanju o jeziku i klasifikatore mašinskog učenja. Na samom početku, upotrebljen je TRIPS parser, koji iz neobeleženog teksta ekstrahuje logičke forme kao potencijalne vremenske izraze. Na dobijeni rezultat se, zatim, primenjuje oko

100 ručno pisanih pravila za ekstrakciju. Kako učinak TRIPS parsera koji sprovodi identifikaciju vremenskih izraza nije postigao zadovoljavajuće rezultate, autori primenjuju i tradicionalni CRF klasifikator, koji vremenske izraze predstavlja na nivou tokena u BIO formatu. Vremenski izrazi koje je predložio sistem zasnovan na CRF klasifikatoru prolaze kroz fazu filtriranja (deo sistema poznat pod nazivom TRIPS i TRIOS), tako da se svaki vremenski izraz za koji je u ovoj fazi moguće ekstrahovati normalizovanu vrednost i tip prihvata kao validan. Za sve ostale identifikovane vremenske izraze se u sledećem koraku tehnikom zasnovanom na pravilima određuju tip i normalizovana vrednost na osnovu datuma kreiranja dokumenta. Ovaj sistem postigao je *F* meru, odnosno uspešnost u određivanju opsega, tipa i normalizovane vrednosti vremenskih izraza od 85%, 94% i 76%.

TIPSem (eng. *Temporal Information Processing based on Semantic information*) je hibridni sistem za ekstrakciju vremenskih informacija iz tekstova engleskog i španskog jezika, koji je takođe učestvovao i u svim ostalim zadacima TempEval-2 izazova (Llorens, Saquete, and Navarro 2010). Zadatak ekstrakcije vremenskih izraza sproveden je u tri odvojene faze: prepoznavanje, klasifikacija i normalizacija vrednosti. Prepoznavanje se odnosi na određivanje opsega vremenskih izraza, a klasifikacija podrazumeva određivanje atributa TYPE, odnosno tipa vremenskog izraza. Na kraju, normalizacija predstavlja određivanje vrednosti vremenskog izraza. Tokom svake od ovih faza koriste se određeni skupovi svojstava neophodnih za učenje modela obeležavanja, grupisanih u dva podskupa. S jedne strane, koristili su opšta svojstva, odnosno morfološke i sintaksičke karakteristike u velikoj meri upotrebljavane u različitim oblastima obrade prirodnih jezika. S druge strane, kao novina, koriste se semantička svojstva. Kako bi izmerili značaj primene semantičkih svojstava, autori su eksperimentisali i upoređivali performanse TIPSem sistema zasnovanog na semantičkim svojstvima sa osnovnom TIPSem-B verzijom ovog sistema, implementiranog bez primene semantičkih svojstava. U fazi prepoznavanja su svojstva, neophodna za obučavanje modela klasifikacije zasnovanog na CRF, prikupljana na nivou tokena. Osim opštih svojstava tokena, korišćene su i njihove semantičke uloge (npr. uloga koju token ima u odnosu na glagol od koga zavisi) i leksička značenja na ontološkom nivou (WordNet (Fellbaum 1998)). Na ovaj način, uz dobijenu listu tokena i njihovih svojstava, obučeni model za prepoznavanje će svaki token obeležiti jednom od etiketa BIO formata za obeležavanje (npr. B-TIMEX3, I-TIMEX3 ili O). Svojstva korišćena za klasifikaciju, odnosno određivanje tipa vremenskog izraza su zapravo ista ona koja su korišćena za prepoznavanje. Ipak, osnovna razlika je u tome što ona ovog puta

nisu prikupljena na nivou tokena već na TIMEX3 nivou, odnosno posmatrajući kontekst čitavog opsega prepoznatog vremenskog izraza. Prateći ovaj opis, modeli za klasifikaciju će pripisati svakom elementu jednu od validnih klasa (DATE, TIME, DURATION, SET). Kao i u fazi određivanja tipa vremenskog izraza, svojstva potrebna za normalizaciju su dobijena na nivou TIMEX3 izraza. Svaki broj TIMEX3 izraza zamenjen je izrazom NUM, a svaka jedinica vremena izrazom TUNIT (npr. u izrazu *3 dana* broj *3* će biti zamenjen izrazom NUM, a *dan* izrazom TUNIT). U prvom koraku, CRF model obučen na osnovu ovih svojstava, određuje tip normalizacije koji je potrebno sprovesti, i zatim pomoću pravila proizvodi normalizovanu vrednost (npr. period *3 dana* biće konvertovan u P_NUM_TUNIT i normalizovan kao P3D).

Ovaj sistem je postigao najveću preciznost u određivanju opsega vremenskih izraza (92%), dok je po uspešnosti u određivanju opsega vremenskih izraza odmah iza HeidelTime sistema sa *F* merom od 85%. Vezano za uspešnost u određivanju tipa vremenskih izraza, TIPSem je postigao *F* meru čija je vrednost malo niža u odnosu na najbolje rezultate TempEval-2 izazova (92%). Kada je reč o normalizaciji, jedinom delu koji nije čisto statistički, postignut je lošiji rezultat od samo 65%. Korišćenje semantičkih informacija uticalo je na povećanje sposobnosti modela za učenje, kao i na uočljivo povećanje odziva (za 33%).

Sistem koji je u okviru izazova TempEval-2 učestvovao u zadacima prepoznavanja vremenskih izraza i događaja bio je Edinburgh-LTG sistem (Grover et al. 2010). Ovaj sistem, upotrebljen za ekstrakciju vremenskih izraza metodom zasnovanom na pravilima, zapravo je komponenta sistema Geoparser, koji vrši georeferenciranje dokumenata i koji je evaluiran na SpatialML korpusu (Tobin et al. 2010). Iako ovaj sistem za rešavanje problema prepoznavanja imenovanih entiteta kombinuje metode zasnovane na pravilima sa metodama mašinskog učenja, kada je reč o vremenskim izrazima primenjuje se isključivo metod zasnovan na pravilima. Kao osnovnu motivaciju za donošenje ove odluke autori navode neophodnost izvođenja procesa normalizacije, odnosno određivanja vrednosti vremenskih entiteta, čija pravila najprirodnije i na najjednostavniji način mogu biti implementirana kao elaboracija pravila za prepoznavanje. S obzirom na to da je ovaj sistem učestvovao u MUC-7 izazovu, u okviru koga je i postigao najbolje rezultate, prepoznati vremenski izrazi su više odgovarali MUC-7 TIMEX nego TIMEX3 entitetima. Stoga je za potrebe TempEval-2 takmičenja bilo neophodno dodavanje određenih pravila za prepoznavanje primera koji nisu bili obuhvaćeni prethodnom verzijom sistema

(npr. *prošlih godina*, *10 minuta* itd.), kao i pravila koja će omogućiti prikupljanje informacija neophodnih za računanje normalizovanih vrednosti predstavljenih u određenom formatu (npr. P1W). Osim toga, ovaj sistem je pomoću više atributa prikazivao vrednost nekog vremenskog izraza, što je za potrebe TimeML obeležavanja bilo potrebno konvertovati u jedan atribut. Kada je reč o određivanju opsega vremenskih zraza, Edinburgh-LTG sistem je ostvario veoma izbalansiran odnos postignutih preciznosti i odziva, i na trećem je mestu u rešavanju ovog zadatka. Za određivanje tipa vremenskih izraza postigao je uspešnost tj. *F* meru od 84%, a u određivanju normalizovane vrednosti 63%.

Sistem koji je, kao i Edinburgh-LTG, bio podjednako uspešan u rešavanju zadatka određivanja opsega vremenskih izraza bio jeste KUL sistem (Kolomiyets and Moens 2010). Za obradu vremenskih izraza KUL koristi hibridni pristup, tretirajući zadatak prepoznavanja izraza kao klasifikacioni problem fraza tj. delova rečenica, dok je normalizacija sprovedena pomoću metode zasnovane na pravilima. U pripremnoj fazi se pomoću OpenNLP paketa vrši segmentacija teksta na rečenice, tokenizacija, obeležavanje vrstama reči i parsiranje. Sistem je implementiran u Java okruženju i sastoji se iz dva procesa. Prvi deo koji pronalazi vremenske izraze u tekstu uzima rečenicu kao ulaz i na osnovu sintaksičke analize u okviru nje traži potencijalne vremenske izraze, uzimajući u obzir samo one segmente koji sadrže sledeće leksičke kategorije: imenice, imena dana i meseci, imeničke fraze, prideve, priloge, priloške fraze i brojeve. Identifikovane fraze biće obeležene u stablu za parsiranje koje pripada jednom od tipova vremenskih izraza. Izdvojeni segmenti rečenica su zatim upoređivani sa postojećim anotacijama TimeBank korpusa, da bi oni kojima se opseg poklapa bili uzimani kao pozitivni primeri vremenskih izraza. U sledećoj fazi su kreirani vektori svojstava, koje će klasifikatori koristiti za obuku radi prepoznavanja vremenskih izraza. Nakon više eksperimenata sa različitim tehnikama mašinskog učenja izabran je klasifikator maksimalne entropije.

Kako oskudnost obeleženih korpusa, u smislu nedovoljnog broja raznovrsnih primera, predstavlja najveći problem za bilo koju tehniku nadgledanog mašinskog učenja, autori su, ne bi li rešili ovaj problem, pokušali da uvećaju raznovrsnost primera pridruživanjem reči sličnog značenja, odnosno sinonima poznatih reči, kojih u korpusu za treniranje nema. Cilj im je bio da izbegnu naivnu selekciju sinonimnih reči iz npr. popisa temporalnih okidača koji ukazuju na vreme i da automatski pronađu i nauče reči, koje će biti upotrebljene u skladu sa gramatičkim pravilima i leksičkim kontekstom vremenskog izraza u tekstu. U tu svrhu upotrebljen

je jezički model za potencijalne reči (eng. *latent word language model, LWLM*) (Deschacht and Moens 2009), koji koristi SVM za procenu njihovih svojstava, kao i WordNet (Miller 1995) kao izvor koji može da pruži najkompletniji skup novih reči sličnog značenja. Budući da korišćenje WordNet-a nije tako jednostavno zbog polisemije, KUL sistem na osnovu novih reči dobijenih putem LWLM modela bira sinset koji ima najveće preklapanje. U drugom delu KUL sistema, kao deo procesa normalizacije, vrše se određivanje tipa i standardizovane vrednosti vremenskog izraza. Kako su se sistemi zasnovani na pravilima pokazali kao daleko uspešniji u ovom zadatku, autori odlučuju da u ovoj fazi primene i ručno pisana pravila. Za određivanje tipa vremenskog izraza upotrebljen je klasifikator, koji koristi isti skup svojstava kao i prethodni, i koji vremenske izraze razvrstava na DATE, TIME, DURATION i SET. Na osnovu značenja vremenski relevantnih informacija i jednostavne sintakse (redni brojevi, osnovni brojevi, imena meseca i dana u nedelji, nazivi godišnjih doba, delovi dana, modifikatori itd.) određene su kategorije reči koje mogu da čine vremenski izraz, pa je za svaku od ovih kategorija ručno napravljen i rečnik, u okviru koga su za svaku odrednicu precizirani parametri u vidu njene vrednosti ili metoda primene navedenog parametra. Dakle, u ovoj fazi se tokeni koji čine vremenski izraz obeležavaju oznakama kategorija na osnovu rečnika, dok se za tokene koji nisu pronađeni u rečniku kao oznaka koristi etiketa vrste reči (PoS etiketa). U sledećem koraku se na osnovu svih ovih prikupljenih informacija vrši procena vrednosti vremenskog izraza. U slučaju onih izraza za koje ne može direktno da se sračuna vrednost, sistem se oslanja na informacije pronađene u prethodnoj fazi prepoznavanja, što uključuje semantički tip izraza, tip diskursa ili vremenske informacije iz konteksta (npr. datum kreiranja dokumenta ili prethodno pomenut vremenski izraz). S obzirom na to da KUL sistem vrednosti relativnih izraza u većini slučajeva računa u odnosu na datum kreiranja dokumenta, postignut je niži nivo uspešnosti u procesu normalizacije (55%), te autori posebno ukazuju na značaj pronalaženja odgovarajućeg vremenskog izraza koji će poslužiti kao orijentir.

USFD2 (Derczynski and Gaizauskas 2010b) je sistem razvijen na Univerzitetu u Šefildu, koji je u TempEval-2 izazovu učestvovao u prepoznavanju i normalizaciji vremenskih izraza i relacija koje postoje između događaja i vremena u istoj rečenici i između dva događaja (TLINK). Ovaj sistem je naslednik USFD sistema (Hepple, Setzer, and Gaizauskas 2007). Problem automatske obrade vremenskih izraza su rešavali pomoću skupa pravila. Deo koji prepoznaje vremenske izraze prvo gradi skup *n-grama* iz podataka koji treba da budu obeleženi ($1 \leq n \leq 5$).

Svaki *n-gram* upoređen je sa ručno pisanim skupom regularnih izraza, jer se ovaj pristup pokazao kao uspešan, uz postizanje visoke preciznosti i odziva koji je u skladu sa veličinom skupa pravila (Han, Chodorow, and Leacock 2006; Mani and Wilson 2000; Ahn, Adafre, and Rijke 2005). Pronalaženjem najdužih mogućih sekvenci reči koje bi mogle da budu jedan vremenski izraz se završava proces identifikacije i određivanja opsega, koji je postigao *F* meru od 82%. U sledećoj fazi određivanja tipa vremenskog izraza ponovo su upotrebljena pravila (npr. ako se do tri reči ispred identifikovanog izraza nalaze reči *za* ili *tokom*, a sam izraz je u množini, onda se može reći da je u pitanju vremenski izraz tipa DURATION). Međutim, ovaj sistem razlikuje samo vremenske izraze tipa DATE i DURATION, i za uspešnost u ovom delu je ostvario *F* meru od 90%. U poslednjoj fazi se normalizovana vrednost računa samo za dva prethodno pomenuta tipa vremenskih izraza. Izrazi *danas* i *sada* normalizovani su izrazom PRESENT_REF, što je na primeru novinskih tekstova tačno u 90% slučajeva (Ahn, Adafre, and Rijke 2005). Ipak je određivanje precizne tačke na vremenskoj osi za relativne izraze koji ukazuju na datume i trajanja bilo najteži zadatak, pa je sistem postigao uspešnost od samo 17% za normalizaciju.

JU CSE TEMP (Kolya, Ekbal, and Bandyopadhyay 2010) je hibridni sistem koji se bavio svim zadacima izazova TempEval-2. Za obradu vremenskih izraza i događaja korišćena su ručno rađena pravila zasnovana na morfološkim i sintaksičkim informacijama, dok je za identifikaciju vremenskih odnosa korišćen dobro poznati statistički algoritam CRF, uz skup svojstava zasnovanih na atributima vremenskih izraza i događaja. Regularni izrazi za identifikaciju vremenskih izraza zasnovani su na entitetima koji ukazuju na imena meseca, godine, dane u nedelji i različite numeričke izraze, kao i na listi ključnih reči koje ukazuju na određene vremenske izraze. Vrednosti različitih atributa računate su putem jednostavnog algoritma. Koristili su samo ona pravila koja sa velikom sigurnošću pronalaze validne vremenske izraze, i u sledećoj fazi razvoja sistema autori planiraju rad na otkrivanju robustnijih pravila. Ovaj sistem postigao je najlošiju *F* meru od 26% za zadatak određivanja opsega vremenskih izraza, dok je uspešnost u određivanju tipa vremenskog izraza i normalizaciji bila ravna nuli.

Osim korpusa na engleskom jeziku, TempEval-2 izazov uključivao je korpuse i na pet drugih jezika. Ipak, bilo je zainteresovanih učesnika samo za engleski i španski jezik, dok za preostala četiri jezika (možda reći koja) ni jedan sistem nije bio prijavljen. Osim sistema TIPSem, samo još jedan sistem je učestvovao u

zadatku obrade vremenskih izraza španskog jezika: UC3M sistem (Vicente-Díez, Schneider, and Martínez 2010), zasnovan je na skupu morfoloških i sintaksičkih pravila dobijenih proučavanjem najfrekventnijih vremenskih izraza španskog jezika.

3.3.7 TempEval-3

U okviru evaluacije SemEval-2013 organizovano je po treći put TempEval-3 takmičenje (UzZaman et al. 2013), koje ovoga puta obuhvata samo engleski i španski jezik. U odnosu na prethodna dva takmičenja, TempEval-3 se razlikuje od njih po:

- veličini korpusa, koji je 12 puta veći nego korpus korišćen za prethodna dva takmičenja;
- zadatak klasifikacije vremenskih relacija sprovodi se na neobeleženim tekstovima i od učesnika se očekuje da prvo ekstrahuju sopstvene događaje i vremenske izraze, a onda da odluče koje od njih će povezati i kojim tipom relacije;
- ovoga puta se od učesnika zahteva pronalaženje svih relacija opisanih u TimeML uputstvu, a ne samo nekih, kao što je to bio slučaj u ranijim TempEval takmičenjima;
- za ovo takmičenje razvijen je i novi skup tekstova, koji su ručno obeležili eksperti i koji je poznat pod nazivom platinasti korpus za testiranje;
- za evaluaciju učinka u određivanju vremenskih relacija ovoga puta se koristi drugačiji sistem bodovanja, kako bi se sistemi rangirali na osnovu jednog rezultata, odn. bodovnog sistema.

TempEval-3 usmeren je na rešavanje tri osnovna zadatka, i to: ekstrakcija i normalizacija vremenskih izraza, ekstrakcija i klasifikacija događaja i obeležavanje vremenskih relacija. Prvi zadatak podrazumevao je određivanje tipa i opsega vremenskog izraza, kao i njegove normalizovane vrednosti. Sve identifikovane vremenske izraze trebalo je obeležiti TimeML TIMEX3 etiketom. Za rešavanje ovog zadatka prijavilo se 9 sistema, sa ukupno 21 verzijom. Za zadatak ekstrakcije vremenskih izraza, odnosno određivanje opsega i tipa, učesnici su koristili i metode zasnovane na pravilima i na mašinskom učenju, kao i hibridne pristupe. Ipak, za normalizaciju vremenskih izraza, odnosno identifikovanje atributa `value`,

svi učesnici su koristili metode zasnovane na pravilima.

Kada je reč o identifikaciji vremenskih izraza, bez neophodnosti striktnog poklapanja opsega, svi sistemi učesnici, zasnovani na različitim strategijama, postigli su bliske rezultate u opsegu od 1%. Najvišu *F* meru od 90,32% postigao je NavyTime sistem (Chambers 2013), koji sprovodi obeležavanje događaja i vremenskih izraza, kao i identifikaciju relacija koje postoje među njima. Što se tiče zadatka prepoznavanja vremenskih izraza, autori su upotrebili sistem koji se pokazao kao najuspešniji tokom LREC 2012. godine i to je SUTime sistem (Chang and Manning 2012), zasnovan na pravilima koji ekstrahuje fraze i normalizuje ih kao TimeML vreme. Sistem SUTime su unapredili nekim svojim specifičnim pravilima do kojih su došli proučavanjem TimeBank korpusa. Autori sve zasluge daju SUTime sistemu, ali smatraju da su njihova dodata pravila unapredila NavyTime sistem, koji je zahvaljujući tome, postigao bolje rezulate (i u određivanju opsega, tipa, kao i u normalizaciji; u ovom poslednjem zadatku su bili drugoplasirani, odmah posle HeidelTime sistema).

Sistem koji je postigao najbolji uspeh u preciznom određivanju opsega i tipa vremenskih izraza bio je ClearTK sistem (Bethard 2013), koji je takođe učestvovao u svim zadacima TempEval-3 takmičenja, od kojih je svaki rešavan kao klasifikacioni problem mašinskog učenja. Identifikacija opsega vremenskog izraza je bila oblikovana kao BIO zadatak razdvajanja na tokene, gde je svaki token u tekstu klasifikovan kao onaj koji je na početku, unutar ili van vremenskog izraza. Za karakterizaciju tokena su korišćena svojstva poput okolnog teksta, osnova tokena, POS itd. Identifikacija tipa vremenskih izraza modelovana je kao višeklasifikacioni zadatak, gde je svaki vremenski izraz klasifikovan kao DATE, TIME, DURATION ili SET. Za to su korišćenea sledeća svojstva: tekst svih tokena vremenskog izraza, tekst poslednjeg tokena i dr. Kada je reč o normalizaciji vrednosti vremenskih izraza, korišćen je sistem za normalizaciju TimeN (Llorens et al. 2012).

U određivanju najvažnijeg atributa, odnosno normalizovane vrednosti vremenskog izraza najuspešniji je bio HeidelTime sistem (Strötgen, Zell, and Gertz 2013). Ovaj sistem učestvovao je samo u rešavanju prvog zadataka vezanog za vremenske izraze engleskog i španskog jezika. Sistem je javno dostupan i koristio se za mnoge jezike. Da bi omogućili rad na višejezičkom pronalaženju informacija, ovaj sistem su usmerili na podršku i integraciju drugih jezika (Strötgen and Gertz 2012). Za ovaj poslednji izazov TempEval doradili su izvore za engleski jezik i razvili nove za

španski jezik. I ovoga puta HeidelTime je u vrhu u identifikaciji vremenskih izraza, kao i u normalizaciji. Ovaj sistem inače može da radi i sa tekstovima nemačkog i holandskog jezika. HeidelTime je sistem zasnovan na pravilima u okviru koga su striktno razdvojeni izvorni kod i izvori koji zavise od jezika i koji sadrže dokumente sa obrascima, normalizacijom i pravilima. Obrasci sadrže reči i fraze, koje se obično koriste za izražavanje vremenskih izraza (npr. imena meseci). Dokumenti za normalizaciju sadrže informacije o obrascima, npr. vrednost imena određenog meseca. Na kraju, dokumenti za pravila sadrže pravila za datum, vreme, trajanje i izraze koji iskazuju učestalost.

Sva pravila se sastoje iz dela koji se odnosi na ekstrakciju i dela koji je vezan za normalizaciju. Delom za ekstrakciju definisani su izrazi koji treba da budu pronađeni u dokumentu, dok deo za normalizaciju pomoću izvora za normalizaciju normalizuje sadržaj izraza koji je nezavistan od konteksta. Normalizacija nedovoljno preciznih (npr. *novembar*) i relativnih (npr. *danas*) izraza se sprovodi posle faza ekstrakcije i normalizacije eksplicitnih vremenskih izraza. Kada je reč o novinskim tekstovima, HeidelTime koristi datum kreiranja dokumenta kao referentno vreme.

Sistem koristi TreeTagger (Schmid 1994) za potrebe pripreme, odnosno za segmentaciju teksta na rečenice, tokenizaciju i obeležavanje vrste reči. Njihov cilj je prevashodno bio da razviju sistem koji će postići rezultate normalizacije visokog kvaliteta, što znači da su hteli da ekstrahuju vremenske izraze koji mogu da budu normalizovani tačno sa velikom verovatnoćom (zato im je odziv niži). Ovaj sistem postiže najbolju F meru za normalizaciju 77,61%, dok je drugoplasirani sistem NavyTime daleko iza.

SUTime sistem (Chang and Manning 2013) je dostupan kao deo Stanford CoreNLP koji se koristi za obeležavanje dokumenata. Za TempEval-3 koristili su svoj standardni skup pravila SUTime sistema, koji inače prepoznaje i neke vremenske izraze koji nisu specifikovani TIMEX3 uputstvom. Sistem je zasnovan na pravilima koja su izgrađena kao obrasci regularnih izraza nad tokenima. Još sistem FASTUS (Hobbs et al. 1997) je pokazao da kaskada konačnih automata može uspešno da se primeni na ekstrahovanje informacija iz teksta, pa i SUTime primenjuje istu ovu etapnu strategiju, gde prvo gradi obrasce reči za pronalaženje numeričkih izraza, zatim obrasce reči i numeričkih izraza za pronalaženje jednostavnih vremenskih izraza, i na kraju formira složene obrazace nad otkrivenim vremenskim izrazima. Osim TIMEX3, SUTime prepoznaje i ugnježđene vremenske izraze i opsege tra-

janja. Za prepoznavanje vremenskih izraza koriste se tri tipa pravila: 1. pravila u formi regularnih izraza (regex) koja na osnovu jednostavnih regularnih izraza nad karakterima ili tokenima pronalaze reprezentacije vremena, 2. pravila za sastavljanje, koja koriste regularne izraze nad segmentima (tokenima i vremenskim objektima), i 3. pravila za filtriranje, koja uklanjaju dvosmislene izraze koji ne treba da budu obeleženi (npr. *fall*). Pravila za sastavljanje se primenjuju u više navrata sve dok se ne stabilizuje konačna lista vremenskih izraza. Nakon što su prepoznati, svi vremenski izrazi su udruženi sa vremenskim objektom. Svaki vremenski objekat je razrešen u odnosu na referentni datum korišćenjem pravila heuristike. U ovom koraku, relativna vremena su konvertovana u apsolutna vremena, a složeni vremenski objekti su što je više moguće pojednostavljeni. Mogućnosti razrešavanja vrednosti relativnih vremenskih izraza su ograničene zbog upotrebe jednostavnih pravila. Sistem trenutno radi samo za engleski jezik, uz mogućnost uključivanja pravila i za druge jezike. Autori smatraju da je NavyTime sistem, iako koristi SUTime, postigao bolje rezultate jer je bio podešen prema TimeBank anotacijama, za razliku od SUTime sistema. Ovaj sistem, kao i NavyTime, ima najviši odziv u otkrivanju vremenskih izraza. SUTime je, kao i HeidelTime, zasnovan na pravilima, što ukazuje na efikasnost primene pravila u ovom domenu.

Još jedan od sistema vrhunskih performansi jeste ManTime sistem (Filannino, Brown, and Nenadic 2013) koji koristi hibridni pristup kombinujući CRF, pristup mašinskog učenja za identifikovanje vremenskih izraza sa pravilima za normalizaciju njihovih vrednosti. Pravila koja se koriste u fazi normalizacije deo su sistema NorMA (Filannino 2012), koji je slobodno dostupan. Istražujući u kojoj meri različita svojstva sistema utiču na njegove performanse u identifikovanju vremenskih izraza, autori su pokazali da korišćenje svojstava uzetih iz WordNet-a negativno utiče na sveukupnu uspešnost sistema. Takođe su pokazali da nema statistički značajne razlike u korišćenju različitih morfoloških svojstava (listi, plitkog parsiranja i etiketa iskaznih imenskih fraza). Na iznenađenje autora, pokazalo se i da korišćenje srebrnog korpusa (samog ili zajedno sa zlatnim korpusom) nije unapredilo performanse. Sistem je obučen na ručno obeleženim (TimeBank i AQUAINT korpusi) i srebrnim podacima, koje su obezbedili organizatori izazova TempEval-3. Autori smatraju da uspeh primenjene metode mašinskog učenja uglavnom zavisi od korišćene sheme za označavanje (BI, BIO, BIOE ili BIOEU), kao i kvaliteta korišćenih svojstava. Koristili su BIO format za sve eksperimente tokom ovog istraživanja. Sistem sprovodi tokenizaciju svakog dokumenta u korpusu i ekstrahuje 94 svojstva, grupisana u okviru sledećih kategorija:

1. morfološka svojstva, koja podrazumevaju niz svojstava karakterističnih za zadatke prepoznavanja imenovanih entiteta, kao što su sama reč, lema, koren reči (stem), prva tri karaktera, poslednja tri karaktera, prvo veliko slovo, glagolska vremena itd. Za lematizaciju i etiketiranje vrstama reči koristili su TreeTagger, dok su za vremenske izraze korišćena svojstva u obliku regularnih izraza.

2. sintaksička svojstva, kao što su segmenti (eng. *chunks*) i iskazne imenske fraze, ekstrahovane pomoću programa za plitko parsiranje MBSP.

3. popisi tj. imenici (eng. *gazetteers*), koji se sastoje od popisa muških i ženskih imena, američkih gradova, nacionalnosti, imena svetskih festivala i standardizovanih skraćenica naziva zemalja. Zahvaljujući mogućnosti obuhvatanja i izraza dužih od jedne reči, ovi popisi su kreirani korišćenjem BIO formata.

4. WordNet: svakoj reči je pridružen određeni broj značenja (leme, antonimi, hiperonimi, hiponimi itd.), definisan kao odvojena karakteristika.

Korišćenje sematičke mreže WordNet je uticalo negativno na sprovođenje klasifikacije, po mišljenju autora stoga što mnogi tokeni nisu imali pridružen smisao iz WordNet-a. Iako je već CRF omogućio postizanje dosta dobrih performansi, koristili su i dodatne module u kasnijoj obradi (npr. modul za korekciju zasnovan na verovatnoći (probabilistički), BIO fixer i modul koji menja granične vrednosti etiketa). Najbolje rezultate su postigli treniranjem sistema na ručno obeleženim podacima. Srebrni podaci nisu poboljšali performanse, ali autori smatraju da je korpus za testiranje bio premali da bi ovi rezultati mogli da se uopšte.

ATT1 sistem (Jung and Stent 2013) postigao je najveću preciznost u identifikaciji vremenskih izraza. Osim vremenskim izrazima, sistem se bavi i prepoznavanjem i klasifikacijom događaja. Na prethodnom takmičenju autori su pokazali da bogate sintaksičke i semantičke karakteristike mogu da dovedu do dobrih performansi u prepoznavanju događaja i vremenskih izraza (Llorens, Saquete, and Navarro 2010; UzZaman and Allen 2010), pa su ovoga puta hteli da pokažu da korišćenje samo leksičkih karakteristika može da radi podjednako dobro. Koristili su Stanford CoreNLP alate za tokenizaciju, lematizaciju i PoS tagiranje. Etiketiranje semantičkim ulogama sprovodili su pomoću SENNA alata (Collobert and Weston 2011) otvorenog pristupa. Za ekstrakciju vremenskih izraza obučili su BIO klasifikatore, dok su za računanje standardne vrednosti vremenskih izraza korišćeni sistemi TIMEN (Llorens et al. 2012) i TRIOS (UzZaman and Allen 2010).

Na osnovu postignutih rezultata, autori su zaključili da uspešnost sistema opada bez korišćenja semantičkih i sintaksičkih karakteristika. Suprotno očekivanjima autora, verzija sistema koja za ekstrakciju vremenskih izraza koristi isključivo leksički kontekst postigla je najbolju preciznosti.

JU_CSE sistem (Kolya et al. 2013) je i ovog puta učestvovao u rešavanju sva tri zadatka. U okviru ovog takmičenja su za identifikaciju vremenskih izraza koristili CRF tehniku mašinskog učenja, odnosno preciznije klasifikator CRF++0.57 otvorenog koda. U prvom koraku su sve rečenice promenjene u vertikalni token po token nivo sekvencijalne strukture za reprezentaciju pomoću BIO kodiranja, korišćenjem prevashodno leksičkih svojstava, poput informacija o vrsti reči ili informacija koje ukazuju na imena meseci, godine ili različite numeričke izraze, razlikujući tokene koji ukazuju na temporalnu lokaciju od onih koji ukazuju na temporalnu kvantifikaciju. Za određivanje tipa vremenskog izraza i njegove normalizovane vrednosti primenili su kombinovanu tehniku korišćenjem sopstvenih ručno rađenih pravila i alata za anotaciju Stanford CoreNLP.

Još jedan sistem koji je učestvovao u svim zadacima jeste sistem KUL (Kolomiyets and Moens 2013), koji za prepoznavanje koristi nadgledanu tehniku mašinskog učenja kojom se obrađuje svaki segment, odn. fraza izvedena iz stabla za parsiranje. Vremenski izrazi su otkrivani pomoću modela kao segmenti fraza sa odgovarajućim opsegom u stablu. Vremenskim izrazima koji su predstavljeni jednim tokenom su dodeljivane posebne etikete. Autori koriste klasifikator logičke regresije, čiji se izlaz uvećava malim brojem ručno rađenih pravila s ciljem povećanja odziva. Normalizacija, kao i u prethodnoj verziji ovog sistema, kreirana metodom zasnovanom na pravilima (Kolomiyets and Moens 2010). Evaluacija postignutih rezultata je pokazala da korišćenje veće količine obeleženih podataka nije uticalo na poboljšanje uspešnosti u prepoznavanju vremenskiih izraza kao što je očekivano. S druge strane, korišćenje veće količine obeleženih podataka je uticalo na bolji rezultat u prepoznavanju događaja.

FSS-TimEx (Zavarella and Tanev 2013) je sistem koji je učestvovao u zadacima prepoznavanja i normalizacije vremenskih izraza, kao i prepoznavanja događaja. Razvijen je kao deo Nexus višejezičkog sistema (Tanev, Piskorski, and Atkinson 2008), namenjenog ekstrakciji događaja. Sastoji se od pravila u vidu kaskade konačnih automata. Ovaj sistem, ranije upotrebljen u medicinskom domenu, za potrebe ovog takmičenja prilagođen TimeML standardu. Trenutno se koristi za

francuski, engleski i italijanski jezik. Pošto je trebalo da ekstrahuju visoko strikturirane informacije iz otkrivenih vremenskih izraza, koje će biti korišćene u sledećem procesu normalizacije, primenili su metod zasnovan na pravilima koristeći kaskadu konačnih automata pre nego tehnike mašinskog učenja. Postojeći modul za prepoznavanje vremenskih izraza engleskog jezika prilagođen je prepoznavanju izraza španskog jezika. Nakon tokenizacije i segmentacije teksta na rečenice, upotrebljeni su rečnici određenih domena i sprovedena je morfološka analiza. Prva faza obrade vremenskih izraza, odnosno prepoznavanje sprovedeno je kaskadom ručno rađenih pravila, čiji je rezultat reprezentacija strukturiranih karakteristika, koje će biti upotrebljene u fazi normalizacije kada se računaju precizne vrednosti vremenskih izraza, u skladu sa TimeML standardom. Za izraze engleskog jezika su koristili kaskadu od oko 90 pravila koja određuju opseg, tip vremenskog izraza, kao i atribute značajne za kasniji proces normalizacije. Proces normalizacije je potpuno nezavisan od jezika i sprovodi se na osnovu karakteristika koje su rezultat faze prepoznavanja. Prilikom razrešavanja vrednosti relativnih vremenskih izraza, kao orijentir su korišćeni datum kreiranja dokumenta ili najbliži prethodno pomenut vremenski izraz. Postignuti rezultati u pogledu uspešnosti u identifikaciji vremenskih izraza nisu obećavajući, posebno kada je reč o određivanju preciznog opsega izraza (često su signali bili uključeni u izraz, što nije dozvoljeno TimeML shemom). Postignuti odziv je takođe dosta niži u odnosu na najbolji sistem ovog takmičenja.

Većina pomenutih istraživanja usmerena je na automatsku obradu vremenskih izraza engleskog jezika, dok postoje i sistemi, od kojih ćemo pomenuti samo neke, namenjeni prepoznavanju i normalizaciji vremenskih izraza francuskog jezika (Bittar 2009; Vazov 2001), španskog (Black, Rinaldi, and Mowatt 1998; Strötgen and Gertz 2010; Llorens, Saquete, and Navarro 2010; Strötgen, Zell, and Gertz 2013), nemačkog (Schilder and Habel 2001b; Strötgen and Gertz 2011), italijanskog (Negri and Marseglia 2005; Caselli, Dell'Orletta, and Prodanof 2009; Robaldo et al. 2011), kineskog (Hacioglu, Chen, and Douglas 2005; Li, Wong, and Yuan 2001; Li et al. 2004; Zhang et al. 2008) itd.

3.3.8 Zaključak

U pogledu automatske obrade vremenskih informacija i osnovnih entiteta temporalnosti tj. vremenskih izraza, događaja i vremenskih odnosa, postoje sistemi koji su usmereni na obradu jednog, više ili svih pomenutih entiteta. Većina ovih sistema se oslanja na morfološke i sintaksičke karakteristike jezika, analizirajući

lokalni kontekst pojavljivanja vremenskih entiteta. U poslednje vreme sve je veći broj sistema za ekstrakciju informacija višeg semantičkog nivoa, koji pokušavaju da analiziraju veće delove teksta (npr. cele rečenice ili, pak, nekoliko rečenica koje pripadaju istom diskursu).
Najrazvijeniji postojeći sistemi za prepoznavanje i normalizaciju vremenskih izraza u tekstovima prirodnih jezika koristili su nekoliko računarskih pristupa za rešavanje ovih zadataka. Kao i za bilo koji drugi zadatak ekstrakcije informacija, i prilikom identifikacije vremenskih izraza upotrebljene su metode zasnovane na pravilima (odnosno znanju) ili metode zasnovane na podacima (statističke metode). Pokazalo se da obe tehnike mogu biti uspešno primenjene, ali samo pod uslovom da postoje dovoljno velike kolekcije konkretnih primera neophodnih za obuku sistema zasnovanih na podacima. Sistemi zasnovani na ručno izrađenim pravilima postižu veoma visoku preciznost, dok se za postizanje dobrog odziva zahteva ulaganje značajnih napora lingvista uključenih u razvoj pravila. Međutim, kada je reč o procesu normalizacije prepoznatih vremenskih izraza, statističke metode ne mogu rešiti ovaj problem ni približno uspešno kao metode zasnovane na pravilima (Kolomiyets and Moens 2010; Negri and Marseglia 2005; Strötgen and Gertz 2010; Jijkoun et al. 2008). Sistemi zasnovani na pravilima su popularni među onima koji se bave samo obradom vremenskih izraza (npr. HeidelTime i TERSEO+T2T3). To je verovatno zato što su pravila mnogo prikladnija za zadatke gde je potrebna i normalizacija. Sistemi zasnovani na podacima i hibridni sistemi su mnogo popularniji među sistemima koji se bave automatskom obradom događaja, u čemu su postigli bolje rezultate nego sistemi zasnovani na pravilima (Kolomiyets and Moens 2013; Jung and Stent 2013).

U okviru tabele 3.2 dat je sumiran prikaz podataka o najrazvijenijim sistemima koji se bave automatskom obradom vremenskih izraza.

Tabela 3.2: Računarski pristupi za obradu vremenskih izraza

Metod	Model	Sistem	Shema	Prepoznavanje	Normalizacija	Evaluacija	Jezik
pravila	ručno	LTG (Mikheev, 1998)	TIMEX	✓	✗	MUC-7	eng
pravila	ručno	FACILE (Black, 1998)	TIMEX	✓	✗	MUC-7	eng, nem, ital, špan
pravila	ručno	LaSIE-II (Humphreyes et al., 1999)	TIMEX	✓	✗	MUC-7	eng
pravila	hibridni	TempEx (Mani and Wilson, 2000)	TIMEX(2)	✓	✓	samostalno	eng
statistički	BIO, SVM	ATEL (Hacioglu, 2005)	TIMEX2	✓	✗	ACE04	eng, kines
pravila	ručno	Chronos (Negri, 2005)	TIMEX2	✓	✓	ACE04	eng, ital
pravila	ručno	Aerotext (Cassel, 2006)	TIMEX2	✓	✓	ACE04	eng
pravila	ručno	TimexTag (Ahn, 2005c)	TIMEX2	✓	✓	ACE04	eng
pravila	ručno	TTK ili TARSQUI (Verhagen et al., 2005)	TIMEX3	✓	✓	samostalno (TERN)	eng
pravila	hibridni	(Boguraev & Ando, 2005)	TIMEX3	✓	✓	samostalno	eng
pravila	ručno	DANTE(Mazur & Dale, 2007)	TIMEX3	✓	✓	ACE07	eng
statistički	CRF, BIO, SVM	TimexTag (Ahn, 2007)	TIMEX2	✓	✓	samostalno (TERN)	eng
pravila	ručno	HeidelTime (Strötgen & Gertz, 2010)	TIMEX3	✓	✓	TempEval-2,3	eng, špan
pravila	ručno	TERSEO+T2T3 (Saquete, 2010)	TIMEX3	✓	✓	ACE04, TempEval-2	eng
hibridni		TRIOS (UzZaman & Allen, 2010)	TIMEX3	✓	✓	TempEval-2	eng
hibridni		TipSem (Llorens, 2010)	TIMEX3	✓	✓	TempEval-2	eng, špan

Metod	Model	Sistem	Shema	Prepoznavanje	Normalizacija	Evaluacija	Jezik
pravila	ručno	Edinburgh-LTG (Grover et al., 2010)	TIMEX3	✓	✓	TempEval-2	eng
hibridni		KUL (Kolomiyets & Moens, 2010, 2013)	TIMEX3	✓	✓	TempEval-2,3	eng
pravila	ručno	USFD2 (Derczynski & Gaizauskas, 2010)	TIMEX3	✓	✓	TempEval-2	eng
pravila	ručno	JU_CSE TEMP (Kolya et al., 2010, 2013)	TIMEX3	✓	✓	TempEval-2, 3	eng
statistički		ClearTK (Bethard, 2013)	TIMEX3	✓	✗(TimeN)	TempEval-3	eng
pravila	ručno	HeidelTime (Strötgen, 2013)	TIMEX3	✓	✓	TempEval-3	eng, špan
pravila	ručno	SUTime (Chang, 2013)	TIMEX3	✓	✓	TempEval-3	eng
hibridni		ManTime (Filannino, 2013)	TIMEX3	✓	✓	TempEval-3	eng
hibridni		ATT1 (Jung & Stent, 2013)	TIMEX3	✓	✗(TimeN)	TempEval-3	eng
pravila	ručno	FSS-TimEx (Zavarella & Tanev, 2013)	TIMEX3	✓	✓	TempEval-3	fran, eng, ital

Glava 4

Prepoznavanje vremenskih izraza

Osnovna tri tipa vremenskih entiteta, koja bi trebalo da budu uključena u bilo koji pokušaj prepoznavanja temporalne dimenzije tekstova prirodnog jezika, predstavljena su sa teorijskog aspekta u drugom poglavlju. U trećem poglavlju dat je pregled postojećih izvora i računarskih pristupa koji se koriste za identifikaciju vremenskih izraza u novinskim člancima. Naredna dva poglavlja (četvrto i peto poglavlje) posvećena su rešavanju problema obeležavanja vremenskih izraza srpskog jezika, odnosno detaljnom opisu metodologije usvojene u okviru ovog istraživanja radi formiranja modela kojim se vrši njihova automatska obrada.

Proces automatskog obeležavanja vremenskih izraza uključuje dve faze obrade. Prva faza tiče se identifikacije onih fragmenata teksta koji nose vremensko značenje i koji predstavljaju pun opseg vremenskih izraza prisutnih u tekstovima, kao i određivanja tipa, što upućuje na proces prepoznavanja vremenskih izraza. Druga faza obeležavanja vremenskih izraza jeste normalizacija njihovih vrednosti, iskazanih eksplicitno ili implicitno u tekstu. U okviru ovog poglavlja opisan je pristup upotrebljen u ovom radu za identifikaciju vremenskih izraza, dok je u poglavlju 5 fokus na procesu normalizacije.

Na početku ovog poglavlja biće reči o klasifikaciji vremenskih izraza kojima se ovo istraživanje bavi. U delu 4.2 detaljno su opisani najuobičajeniji tipovi vremenskih izraza koji se mogu sresti u tekstovima prirodnog jezika. Nakon upoznavanja sa ciljanim entitetima, ostatak ovog poglavlja i poglavlje 5 daju dalji opis automatskog procesa obeležavanja. Metodologija usvojena za proces prepoznavanja vremenskih izraza detaljno je predstavljena u delu 4.5.

4.1 Određivanje vrsta izraza koje je potrebno obeležiti

Predmet ovoga rada jeste automatsko prepoznavanje lingvističkih izraza vremenskog značenja u standardnom srpskom jeziku. Izrazi takvoga tipa predstavljaju jezičku strukturu reprezentovanu različitim formalnim jedinicama, kojima se prenose tri osnovna tipa vremenskog značenja: KAD (pravo vreme), KOLIKO DUGO (kvantitativnost u vremenskom smislu) i KOLIKO ČESTO (iterativnost). Iako temporalnost, kao semantička kategorija kojom se vremenski lokalizuje iskazana situacija, poseduje izrazito široke mogućnosti gramatikalizacije (Pavlović, 2006), u ovome radu prati se realizacija ovog značenja u sistemu padežnih konstrukcija i priloga (bez analize njihovih komponenata kao gramatičkih fenomena), dok realizacija na predikatskom planu tj. putem vremenskih glagolskih oblika ostaje van opsega ovog istraživanja.

Vremenski izrazi, koji bi trebalo da budu prepoznati i obeleženi, mogu da ukažu na kalendarske datume, vremena dana, trajanja i učestalosti ponavljanja vremena. Njihova sintaksička osnova mora da sadrži odgovarajući leksički okidač, odnosno reč ili numerički izraz sa značenjem jedinice ili koncepta vremena. Neophodno je, dakle, da postoji mogućnost da se označeni izraz lokalizuje na vremenskoj liniji ili da barem može da se orijentiše u odnosu na neko vreme (prošlo, buduće, sadašnje). Radi ilustracije, u okviru tabele 4.1 dati su primeri različitih termina koji se u ovom radu posmatraju kao okidači.

Tabela 4.1: Leksički okidači

Vrsta reči	**Leksički okidač**
Imenice	sekund, minut, sat, dan, vikend, nedelja, mesec, godina, decenija, vek, podne, noć, ponedeljak, mart, proleće
Vlastite imenice	Božić, Nova godina, Uskrs
Posebni vremenski obrasci	12:35, 3.04.1999., 11/30/2005, 1998, 1970-tih
Pridevi	prošli, tekući, sledeći, mesečni
Prilozi	mesečno, dnevno, nedeljno, večeras, danas, juče, noćas, jesenas, zimi, devedesetih
Broj	4, dva, prvog, 5.

Osnovu vremenskih izraza koji će biti uzeti u razmatranje prevashodno čine imenice vremenskog značenja: jedinice vremenske mere (npr. *dan* (24 sata), *minut*, *godina*, imena dana i meseci); pojmovi-periodi, odnosno reči koje označavaju određeno vremensko trajanje, pa se mogu upotrebiti i kao reči koje izriču jedinice vremenske mere (npr. *jutro*, *dan* (svetli deo perioda od 24 sata u trajanju od približno 12 sati), *veče*, *leto*); i pojmovi-praznici (npr. *Božić*, *Nova godina*) (Ivic). Ove jedinice vremenske mere u svom osnovnom značenju „sadrže potpunu određenost koja proizilazi iz njihove pripadnosti sistemu (za svaku se jedinicu zna koliko traje, npr. sat traje 60 minuta, godina 12 meseci) i koja je objektivna, univerzalna, konvencijom utvrđena i upotrebom potvrđivana“ (Znika 1979). Iako navedene forme predstavljaju konvencionalno utvrđene tačke u vremenu, u situacijama kada se upotrebljavaju izvan sistema kome pripadaju, sintaksički je obavezna odredba pomoću atributa. Uz jedinice mere može da stoji broj (npr. *2010. godine*, *tri sata*) ili im može biti pridružena neka druga reč kojom se sužava značenje jedinice i vrši subjektivno preciziranje, odnosno određenje količine vremena (npr. *prošle godine*, *nekoliko dana*). Važno je napomenuti da se neki pridevi smatraju leksičkim okidačima samo u situacijama kada su upotrebljeni za vremensko određenje; npr. pridev *sledeći* nije okidač kada se, umesto u temporalnom smislu (npr. *sledećeg dana*), koristi u smislu prostornog određenja (npr. *sledeća osoba u redu*).

4.1.1 Izrazi koji se ne obeležavaju

Osim prethodno opisanih izraza koje je potrebno obeležiti, postoje i izrazi koji, iako nose određeno vremensko značenje i govore kada se nešto dogodilo, koliko dugo će trajati, ili sa kojom učestalošću će se ponavljati, pripadaju klasi onih koji nedovoljno precizno ukazuju vreme i kojima nije moguće smeštanje na vremenskoj osi, te ih nije potrebno obeležavati (npr. *u međuvremenu*, *jednom*, *nekad*, *odmah*, *smesta*, *stalno*, *često*, *uvek*, *ponekad*, *retko*, *nikad* i sl.) (primer 4.1).

Primer 4.1.
*Ta razlika je **u međuvremenu** znatno smanjena, pa je putarina za strana vozila u proseku veća za oko 30 odsto.*
*U Jatu kažu da je nakon pada aviona Turkiš erlajnsa njihova letelica **odmah** preusmerena ka Briselu, gde je sletela bez ikakvih problema.*
*Porodica i prijatelji su **stalno** uz njega.*

Naporedni veznici (sastavni: *i, pa*, *te*, *ni*, *niti*; rastavni: *ili*; suprotni: *a*, *ali*, *no*, *nego*, *već*) i predlozi spadaju u grupu vrsta reči koje nikada neće biti posmatrane

kao leksički okidači.

4.2 Semantičke klase vremenskih informacija

Utvrđivanje tipologije semantičkih obeležja i semantičkih tipova vremenskih izraza koji se pojavljuju u tekstovima prirodnog jezika neophodno je za razvoj automatskog sistema, koji će, u određenoj meri poput ljudi, biti sposoban da tumači izraze koji ukazuju na vreme. S tim ciljem su definisana razna uputstva za obeležavanje i normalizovanje vrednosti vremenskih izraza, o kojima je bilo više reči u poglavlju 3.1. U tom smislu, polazna tačka ovog istraživanja bile su specifikacije date u izuzetno korišćenoj TimeML shemi, kao i osnovna tipologija vremenskog značenja data u radu (Ivić 1955–1956). Naime, autorka tada po prvi put ističe da u okviru vremenskog značenja treba praviti razliku između tri osnovna značenjska tipa: prvo, značenje tačke u vremenu koje odgovara na pitanje *kada?* se nešto dogodilo (**pozicija u vremenu**); drugo, značenje vremenske mere koje odgovara na pitanje *koliko dugo?* je nešto trajalo (**trajanje**); i treće, značenje ponavljanja koje odgovara na pitanje *koliko često?* se nešto dešava (**učestalost**). Druga važna distinkcija napravljena je između izraza čije vrednosti mogu biti normalizovane na osnovu njih samih (ovi izrazi poznati su kao **potpuno precizna**, **kontekstno nezavisna** ili **apsolutna vremena**) i izraza koji zahtevaju vrednost drugog vremenskog izraza kao orijentira u procesu normalizacije (ovi izrazi poznati su kao **nedovoljno precizni**, **kontekstno zavisni** ili **relativni vremenski izrazi**). Jedan primer potpuno preciznog vremenskog izraza je izraz *trinaest časova 3. januara 2007. godine* (primer 4.2) koji sadrži unutar sebe sve informacije neophodne za temporalnu identifikaciju, odnosno smeštanje na vremenskoj osi i normalizaciju njegove vrednosti. Nasuprot tome, postoje i nedovoljno precizni izrazi, kao što je npr. *sledećeg dana* (primer 4.3), kojima je potreban drugi potpuno precizan vremenski izraz koji će poslužiti kao orijentir u odnosu na koji će se identifikovati vremenska tačka. U slučaju da se ta dva primera jave kao dva uzastopna izraza u tekstu, potpuno precizni vremenski izraz *trinaest časova 3. januara 2007. godine* će biti orijentir za izraz *sledećeg dana* i omogućiće određivanje kalendarske pozicije nedovoljno preciznog vremenskog izraza tj. *4. januara 2007. godine*.

Primer 4.2.
Petar je stigao u grad u ***trinaest časova 3. januara 2007. godine****.*

Tabela 4.2: Slovni karakteri koji predstavljaju granularnost vremenskog izraza

Granularnosti	**Kod granularnosti**
milenijum	ML (*millennium*)
vek	CE (*century*)
decenija	DE (*decade*)
godina	Y (*year*)
mesec	M (*month*)
nedelja/sedmica	W (*week*)
dan	D (*day*)
sat/čas	H (*hour*)
minut	MIN (*minute*)
sekunda	S (*second*)

Primer 4.3.
*Marija je otputovala **sledećeg dana**.*

Dakle, na osnovu prethodno iznetih zapažanja, u okviru ovog istraživanja napravljena je razlika između sledećih grupa vremenskih izraza:

1. vremenski izrazi koji ukazuju na poziciju u vremenu (kalendarski datumi i vremena kao delovi dana),
2. vremenski izrazi koji ukazuju na kvantitativnost (trajanja) i
3. vremenski izrazi koji ukazuju na iterativnost.

4.2.1 Vremenski izrazi koji impliciraju tačku u vremenu

Izrazi ovoga tipa ukazuju na temporalnu lokaciju, odnosno odsek na vremenskoj osi koji je shvaćen kao tačka u vremenu (kalendarske pozicije). Oni mogu biti specifikovani sa određenom granularnošću dimenzije vreme, odnosno do određenog nivoa detaljnosti (npr. vremenski izraz *petak* iskazan je na nivou dana, dok su izrazi *2005. godine* ili *tri godine kasnije* na novou godine). Radi prikazivanja granularnosti vremenskih izraza u procesu normalizacije koriste se određeni slovni kodovi, prikazani u tabeli 4.2, o čemu će biti više reči u sledećem poglavlju.

Vremenski izrazi koji impliciraju tačku u vremenu mogu dodatno biti klasifikovani na potpuno precizne, odnosno apsolutne vremenske izraze i relativne vre-

menske izraze.

Potpuno precizni vremenski izrazi dalje mogu biti razvrstani na osnovu njihove granularnosti:

- Nivo milenijuma (primer 4.4)

Potpuno precizan vremenski izraz na nivou milenijuma može biti izražen pomoću rednih brojeva (napisanih arapskim ili rimskim ciframa ili slovima) i leksičkog okidača *milenijum* ili njegovih sinonima poput reči *tisućleće* (npr. *2. milenijum*, *prvo tisućleće*, *III milenijum*).

Primer 4.4.
*Za **2. milenijum** su karakteristične demografske, ekonomske, političke i kulturne promene koje su po svojoj brzini i dalekosežnosti nadmašile promene u ranijim periodima pisane istorije, ...*

- Nivo veka (primer 4.5)

Potpuno precizan vremenski izraz na nivou veka može biti izražen pomoću rednih brojeva (napisanih arapskim ili rimskim ciframa ili slovima) i leksičkih okidača *vek*, *stoleće* (npr. *12. vek*, *dvadeseto stoleće*, *XVII vek*).

Primer 4.5.
*Od **XIII veka** postepeno se menjao život u gradovima, pa su i oni poprimali drugačiji izgled.*

- Nivo decenije (primer 4.6)

Vremenski izrazi na nivou decenije mogu biti izraženi na više načina, i to: pomoću brojeva (napisanih arapskim ciframa ili slovima) i leksičkih okidača *decenija*, *godina* (npr. *prvoj deceniji*, *1970-tih godina*), kao i izrazima iz kojih je jedinica mere vremena izostavljena. Značenje vremenskog izraza *1960-tih* takođe je iskazano i izrazom koji čini prilog *šezdesetih*.

Primer 4.6.
*Neki predmeti umetničkog odeljenja Gradskog muzeja u Subotici, **1970-tih** godina su obnavljani u Matici srpskoj u Novom Sadu.*

- Nivo godine (primer 4.7)

Potpuno precizni vremenski izrazi na nivou godine mogu biti izraženi bilo korišćenjem samo godine (npr. *1999, '99*), bilo da im prethodi ili sledi reč *godina* (npr. *1998. godine, godine 1998.*). Kontekst može da odredi kojim bi tipom vrednosti godine vremenski izraz trebalo da bude obeležen. Osim jednostavnog tipa godine koji je najčešće u upotrebi (npr. *2002*), može se naići i na školske i finansijske godine (npr. *školske 2005/06*, *finansijske godine 1997-1998.*), ili godine pre početka ove ere (npr. *600. godine pre nove ere*, *20. godine pre rođenja Hrista*).

Primer 4.7.
Napadač Reala Raul najverovatnije će se povući ***2011. godine****, ali je priznao da bi razmotrio da još godinu dana posle toga igra u Sjedinjenim Američkim Državama.*

- Nivo meseca (primer 4.8)

Sa granularnošću meseca, potpuno precizni vremenski izrazi redovno su izraženi korišćenjem bilo numeričkih obrazaca (npr. *11.2005.*) ili punih ili skraćenih naziva meseci (npr. *januara 2009. godine, veljače*[1] *2005, Dec. 2008.*). Osim toga, uobičajeno i navođenje rednog broja meseca ili prideva *poslednji* uz leksički okidač *mesec* (npr. *prvog meseca 2008*, *peti mesec 2007. godine*, *poslednji mesec 1998*). Isto tako ovom tipu vremenskih izraza pripadaju i oni izrazi u kojima se pominje deo godine, i to korišćenjem forme godišnjeg doba (npr. *leta 2009. godine*), kvartala (npr. *prvi kvartal 2002.*), trećine (npr. *druga trećina 1996. godine*) ili polovine godine (npr. *druga polovina 2003. godine*).

Primer 4.8.
U ***februaru 2009.*** *dostigle su rekord od neverovatnih 57 odsto godišnje.*

- Nivo nedelje (primer 4.9)

Vremenski izrazi iskazani na nivou nedelje prilično se retko javljaju u tekstovima u svom potpuno preciznom obliku, ali kad god se pojavljuju prevashodno uključuju redni broj nedelje iza koga slede leksički okidači *nedelja* ili *sedmica*, a zatim mesec i godina (npr. *druga nedelja avgusta 2014. godine*) ili samo godina (primer 4.9) kojoj navedena sedmica pripada.

Primer 4.9.
Ovaj Uskrs pada u ***17. nedelju 2009.***

[1] Kako bi sistem bio što potpuniji, osim naziva meseci srpskog jezika, u prepoznavanje su uključeni i nazivi meseci koji pripadaju hrvatskom jeziku.

- Nivo dana (primer 4.10)

Kada je reč o potpuno preciznim vremenskim izrazima na nivou dana, uočljiv je veoma širok spektar različitih oblika kojima se oni ostvaruju. Često se nailazi na različite numeričke obrasce (npr. *17.02.2010.*, *17.2.2010.*, *2/17/2010*, *17-02-2010*), koji omogućavaju jednostavno pronalaženje tačnog datuma, osim u slučajevima kada je upotrebljen format različit od onog koji je karakterističan za evropsko područje i koji može izvati problem pravilne interpretacije (npr. izraz *02/03/2007* može da ukaže na 3. februar 2007. ili 2. mart 2007. godine). Potpuno precizni izrazi sa granularnošću dana takođe mogu biti izraženi korišćenjem rednih brojeva za oznaku dana, potpunog ili skraćenog imena meseca i numeričke vrednosti za oznaku godine (npr. *3. aprila 1999. godine*, *2001. godine prvog oktobra*). Ispred izraza ovoga tipa može se naći i potpuno ili skraćeno ime odgovarajućeg dana u nedelji (npr. *utorak*, *6. maj 2008. godine*).

Primer 4.10.
*Rođendan je proslavila **18. januara 2014. godine**.*

- Nivo sata (primer 4.11)

Za dobijanje potpuno preciznih vremenskih izraza na nivou sata mogu se koristiti različiti obrasci u kombinaciji sa ili bez leksičkih okidača *sat*, *čas*, *h* (npr. *13 časova*, *pet sati popodne*, *15 h*, *Dolazim u **pet*** itd.).

Primer 4.11.
*Časovi počinju u **15 časova 21. januara 2003. godine**.*

- Nivo minuta (primer 4.12)

Potpuno precizni izrazi na nivou minuta mogu se javiti kao posebni numerički obrasci, ili u kombinaciji brojeva i leksičkih okidača (npr. *17:25*, *7 sati i 20 minuta*, *petnaest do tri popodne*, *pola osam ujutru*, *8.30 uveče* itd.).

Primer 4.12.
*Iz Beograda smo krenuli u **7:30 1. jula 2004. godine**.*

- Nivo sekunde (primer 4.13)

Nešto je ređe pojavljivanje vremenskih izraza koji su specifikovani do nivoa sekundi. Granularnost vremena ovoga nivoa najčešće se nalazi u vremenskim oznakama formiranim korišćenjem čisto numeričkih obrazaca (kao u primeru 4.13),

što nije obavezno za sve situacije pojavljivanja (npr. *01:43:35 09. februara 2010. godine*).

Primer 4.13.
Vest je objavljena tačno u ***05:37:39, 19.01.2010.***

Vremenski izrazi koji ukazuju na delove dana (tj. sa granularnošću sata, minuta ili sekunde) mogu da sadrže i referencu na časovnu/vremensku zonu (npr. *12:00 UTC*, *8 časova po srednje-evropskom vremenu* itd.).

Osim navedenih nivoa granularnosti, postoje i uži nivoi granularnosti (npr. stotinka, mikrosekunda i sl.) i širi (npr. era, doba), ali njima se ovde nećemo baviti jer pripadaju specifičnim domenima, koji nisu predmet ovog rada.

Klasa relativnih vremenskih izraza uključuje izraze čija vrednost nije eksplicitno iskazana, te im je neophodan potpuno precizan vremenski izraz koji će poslužiti kao orijentir u odnosu na koji će se računati njihova krajnja vrednost i pozicija na vremenskoj osi. Temporalna determinacija ovih izraza može se vršiti na dva različita načina, i to (1) s obzirom na momenat govora, i (2) s obzirom na neku drugu referentnu tačku u vremenu, različitu od momenta govora. Kada je reč o novinskim tekstovima, momentom govora najčešće se smatra vreme kreiranja dokumenta (eng. *Document Creation Time – DCT*). Više reči o načinu računanja normalizovanih vrednosti relativnih vremenskih izraza biće u poglavlju 5. Svi izrazi kojima nedostaje barem vrednost za godinu, mogu se smatrati za relativne vremenske izraze (npr. *ponedeljak*, *19. april*, *marta ove godine*, *popodne 13. decembra*, *petak kasno uveče*).

Među relativnim vremenskim izrazima karakteristična je grupa izraza koji su predstavljeni deiktičkim prilozima *danas*, *juče*, *sutra* i *sutradan* i koji radi identifikacije eksplicitne vrednosti zahtevaju kao referentnu tačku neki potpuno precizan izraz na nivou dana. Na primer, vremenska pozicija koju ispunjava izraz *juče* može se dobiti oduzimanjem jednog dana od vrednosti orijentira, što je predmet sledećeg poglavlja.

Deiktički izrazi formirani korišćenjem pridevske zamenice *ovaj* i jedinice vremenske mere (tj. *milenijum*, *vek*, *decenija*, *godina*, *mesec*, *nedelja*, *dan*, *čas*, *minut*, ime dana) pripadaju grupi relativnih vremenskih izraza čija se vrednost računa u

odnosu na momenat govora (npr. *ove godine*, *ovog petka*). Izrazi zavisni od konteksta diskursa, koji su vezani za najbliže referentno vreme, a ne momenat govora, jesu oni izrazi formirani korišćenjem pridevskih zamenica *taj* ili *onaj* koje su praćene jedinicom vremena (npr. *godina*, *mesec*, *ponedeljak*, *mart*, *leto*, *jutro* itd.).

Uz ranije pomenute jedinice vremenske mere, temporalne vlastite imenice (npr. *Nova godina*) i temporalne zajedničke imenice (npr. *jutro*, *veče*), kao leksičke okidače, uobičajena je i upotreba prideva *poslednji*, *prošli*, *prethodni/predstojeći*, *naredni*, *sledeći*, čime se formiraju izrazi koji pripadaju posebnoj grupi relativnih izraza.

U grupi deiktičkih izraza koji bi trebalo da se referišu u odnosu na momenat govora nalaze se izrazi formirani kombinovanjem kvantifikovanih ili nekvantifikovanih jedinica vremena sa prilozima *ranije* ili *kasnije* (npr. *dva meseca ranije*, *godinama ranije*, *sat kasnije*). Kombinovanjem numeričkih kvantifikatora i jedinica vremena determinisana je dužina trajanja u vremenu tj. značenje kvantifikacije, ali je u okviru jezičke forme sa prilozima *ranije* ili *kasnije* zadržano i značenje identifikacije, odnosno smeštanje u vreme. Na osnovu primera jedinica vremena u množini (npr. *godinama ranije*) i neparametričkih kvantifikatora (npr. *nekoliko godina ranije*, *par dana kasnije*) jasno je da kvantifikacija ne mora nužno da sadrži ideju broja.

Određena pojavljivanja zamenničkih priloga *onda* i *tada* takođe bi trebalo da budu obeležena kao relativni vremenski izrazi zavisni od najbliže referentne tačke u vremenu. Takav slučaj deiktičkog adverba *onda* prisutan je u primeru 4.14, i njegov orijentir dat je izrazom *februara meseca 2010*.

Primer 4.14.
Pisao mi je ***februara meseca 2010.*** *i od onda mi se više nije javljao.*

Praznici

Izrazi ovoga tipa ukazuju na imena praznika i drugih prilika čije je ime poznato u određenim zajednicama (npr. *Nova godina*, *Božić*, *Uskrs*, *Dan zahvalnosti*). Na osnovu uputstava TimeML sheme takvi izrazi bi trebalo da budu obeleženi u tekstu kao vremenski izrazi, ali bez pripisivanja normalizovane vrednosti, osim ako je ona eksplicitno data u kontekstu. Ovakva praksa je preporučena, pre svega, zbog postojanja tzv. pokretnih praznika, čiji se kalendarski datum proslavljanja iz

godine u godinu menja, kao što je to u slučaju Uskrsa. S druge strane, iako se određeni praznici proslavljaju precizno određenog dana u godini, njihova pridružena vrednost u vidu kalendarskog datuma može da zavisi od zajednice u čijem kontekstu se pominju (npr. pravoslavni hrišćani Božić slave 7. januara, dok se kod katoličkih hrišćana ovaj isti praznik proslavlja 25. decembra).

4.2.2 Vremenski izrazi koji impliciraju trajanje

Vremenski izrazi koji se takođe ispoljavaju kao „čista“ semantička kategorija jesu izrazi kojima se ispoljava temporalna kvantifikacija, odnosno odmerava dužina trajanja u vremenu. Izrazi ovoga tipa ukazuju na odsek na vremenskoj osi koji je shvaćen kao kraći ili duži „prostor“ u vremenu i koji ukazuje na trajanje. Značenje trajanja kao preciznog perioda vremena može biti implicirano i određenim tačkama u vremenu, odnosno značenjem ingresivnosti (identifikacija leve granične tačke) ili terminativnosti (identifikacija desne granične tačke), kao što je u primeru od *5-10. aprila* implicirano trajanje od 6 dana. U ovom radu će ovakvi primeri biti posmatrani i obeleženi kao tačke u vremenu, što je u skladu sa smernicama TimeML uputstva za obeležavanje vremenskih izraza. Isto tako značenje trajanja može biti implicirano i kalendarskim datumima, ako se ispred datuma nalazi neki od oblika imenice *tok* (npr. tokom *leta 2007. godine*, u toku *2005*, tokom *aprila 2011. g.*).

Najčešći primeri izraza kojima se iskazuje značenje apsolutnog trajanja formirani su pridruživanjem numeričkih kvantifikatora (npr. *pet*, *desetak*, *pola*) i jedinica vremenske mere (npr. *godina*, *mesec*, *dan*, *sat*) (primer 4.15).

Primer 4.15.
*Episkop niški Irinej Gavrilović biće uveden u tron srpskih patrijaraha posle **pola veka** monaškog života.*

Nivo granularnosti je isti kao i u slučaju vremenskih izraza koji ukazuju na tačku u vremenu. Neki nivoi detaljnosti ilustrovani su primerom 4.16.

Primer 4.16.
*. . . a godinu kasnije i za episkopa niške eparhije, gde je ostao pune **tri decenije**.*
*. . . država će u izgradnju nove topionice, za šta je potrebno **tri godine**, uložiti oko 130 miliona dolara.*
*Istražni sudija Višeg suda u Beogradu odredio je pritvor do **mesec dana** četvoročlanoj*

kriminalnoj grupi osumnjičenoj . . .
*Komisija je dužna da u roku od **dve nedelje** odabere najpovoljnijeg ponuđača.*

Značenje apsolutnog trajanja može biti iskazano i jedinicom vremenske mere u akuzativu jednine iza koje sledi predlog ili imenica *vreme* (npr. *sat nakon*, *sat vremena*). Određeni izrazi mogu se javljati kao rekurzija dva ili više tipičnih izraza koji ukazuju na trajanje (npr. *dve nedelje i tri dana*).

S druge strane, relativne vremenske izraze čine nenumerički kvantifikatori (npr. *nekoliko*, *najviše*, *par*) i jedinice vremenske mere u genitivu množine (npr. *godina*, *mesec, dan*, *sat*), što je ilustrovano primerom 4.17.

Primer 4.17.
*Akcija će trajati **nekoliko meseci**.*

Uz pomenute vremenske izraze sa značenjem kvantifikacije vremena uobičajena je i upotreba prideva *poslednji*, *prošli*, *prethodni/predstojeći*, *naredni*, *sledeći* (npr. *poslednjih pet sati*, *idućih nekoliko dana*).

Izrazi koji označavaju starost obično su formirani kombinovanjem trajanja sa rečima *star* i *starost* (npr. *starosti 35 godina*), ili prisvojnih zamenica sa skraćenim oblicima izražavanja decenija (npr. *njene 70-te, njihove šezdesete*). Isto tako se mogu sresti slučajevi upotrebe konstrukcije *starosti od* (npr. *starosti od 34 godine*).

U okviru ovog istraživanja u grupu vremenskih izraza koji ukazuju na trajanje svrstani su i izrazi koji ukazuju na obeležavanje godišnjica (npr. *25-godišnjica njihovog venčanja*, *dvadesetpetogodišnjica*, *dvadeset peta godišnjica*, *25. godišnjica*).

4.2.3 Vremenski izrazi koji impliciraju učestalost ponavljanja vremena

Vremenski izrazi ovoga tipa ukazuju na temporalnu frekvenciju (iteraciju), odnosno učestalost pojavljivanja u vremenu. Ovde se radi o posebnom vidu temporalnog kvantifikovanja koje se može ostvariti kao povremeno (više puta u vremenu) i regularno (više puta u vremenu, i pri tom uvek u isto vreme) ponavljanje.

Povremeno pojavljivanje u vremenu (više puta u vremenu) se, pre svega, može

izraziti kombinacijom numeričkih kvantifikatora ili priloga *jednom*, *dvaput*, *triput* i priloga učestalosti (npr. *dva puta mesečno*, *triput nedeljno*) (primer 4.18).

Primer 4.18.
Letovi će se obavljati ***šest puta nedeljno****, svakog dana osim subotom.*

Izrazi koji ukazuju na učestalost ponavljanja u vremenu takođe mogu biti iskazani na svim nivoima granularnosti koji se javljaju i u slučaju prethodno opisana dva tipa vremenskih izraza.

Regularna ponavljanja u vremenu (više puta u vremenu, uvek u isto vreme) mogu biti izražena bilo korišćenjem prideva ili priloga učestalosti (npr. pridev *mesečni* ili prilog *mesečno* kao u primeru 4.19), ili upotrebom pridevske zamenice *svaki* u vezi sa jedinicom vremena (npr. *svake godine*, *svakog utorka*, *svakog dana*) (primer 4.20).

Primer 4.19.
Grupa od 450 radnika koji obavljaju takve poslove ***mesečno*** *će se rotirati sa drugom grupom od 450 radnika ...*

Primer 4.20.
Protesti se održavaju ***svakog dana*** *u 18 sati.*

Pridevska zamenica *svaki* može se javiti i u kombinaciji sa numerički kvantifikovanim jedinicama mere vremena (npr. *svakih jedanaest godina, svakih 13 dana*, *svakih 15 godina*, *svakog drugog dana*, *svakog treceg petka*, *svakog petog oktobra*) (primer 4.21).

Primer 4.21.
Uredbom o cenama naftnih derivata precizirano je da se presek cena radi ***svakih 15 dana****, pri čemu se prate cene nafte ...*
Srpska reprezentacija, ***svakog drugog dana*** *ima identičan raspored.*

Uz pridevsku zamenicu *svaki* i jedinice mere vremena mogu biti upotrebljeni i pridevi (npr. *svakog sledećeg sata*, *svakog narednog petka*, *svakog idućeg januara*) (primer 4.22).

Tabela 4.3: Opseg vremenskih izraza izvučen iz konteksta

Kontekst	**Opseg vremenskog izraza**
pre petka	petka
u 8 časova	8 časova
ili četvrtak	četvrtak
tokom poslednje dve godine	poslednje dve godine

Primer 4.22.
Sada haški operativci pokušavaju da otkriju ko je, ***svake sledeće godine****, odobravao produženje članstva.*

Množine vremenskih imenica koje označavaju dane u nedelji i mesece (npr. *avgustima*, *utorcima*) mogu takođe da izraze učestalost vremena.

4.3 Određivanje opsega vremenskih izraza

Leksički kontekst koji okružuje otkrivene reči-okidače pruža uvid u informacije relevantne za ispravno određivanje punog opsega vremenskih izraza, te kasniji proces normalizacije njihovih vrednosti. U nastavku teksta biće više reči o pravilima koja određuju početak i kraj svakog vremenskog izraza tj. njegov opseg.

Prilikom automatskog obeležavanja vremenskih izraza pun opseg izraza obuhvaćenog etiketom obeležavanja mora da bude jedna od gramatičkih kategorija:

- imenica (npr. *danas*, *petak*)
- imenička sintagma (npr. *sredu uveče*, *prošle godine*)
- pridev (npr. *današnji*)
- prilog (npr. *letos*, *mesečno*)
- pridevska/priloška sintagma (npr. *tih devedesetih*, *rano jutros*).

Na osnovu smernica datih TimeML uputstvom, predlozi i veznici koji se nalaze ispred vremenskih izraza ne treba da uđu u pun opseg koji je potrebno obeležiti, te se za fraze navedene u tabeli 4.3 dobijaju sledeći obeleženi izrazi:

Za razliku od predloga i veznika, reči ili grupe reči koje na određeni način modifikuju ili kvantifikuju vremenski izraz biće uključene u pun opseg koji je potrebno obeležiti (npr. ***početkom*** *godine*, ***manje od*** *dva sata*). Primeri navedeni u primeru 4.23 predstavljaju pun opseg vremenskih izraza koji bi trebalo da se obeleži:

Primer 4.23.
tog *dana*
sledećeg *dana*
kasno *sinoć*
oko *15 h*
više od *mesec dana*
ne manje od *60 dana*
samo *godinu dana*

Izraz se posmatra kao nedeljiva sintaksička jedinica u sledećim situacijama:

- Dva izraza iskazuju vrednosti koje se odnose na jednu te istu jedinicu vremena (npr. *12 sati noću*, gde i *12 sati* i *noću* izražavaju vrednosti za jedinicu dela dana);
- Dva izraza iskazuju vrednosti za jedinice koje su hijerarhijski povezane (npr. *petak uveče*, gde je jedinica dana veća od jedinice dela dana; ili *januara 2005. godine*, gde je jedinica meseca u hijerarhiji neposredno ispod jedinice godine).

Primeri navedeni u primeru 4.24 predstavljaju pun opseg vremenskih izraza koji bi trebalo da se obeleži.

Primer 4.24.
dvanaest sati noću
subota veče
15 h petak
subotu 16-og
maja 2009.
jesen 2008.

Bez obzira na to da li je upotrebljena i zapeta, izrazi ovoga tipa biće posmatrani kao jedan izraz, te obeleženi jednom etiketom (primer 4.25).

Primer 4.25.
petak, 27. marta 2005. godine
danas, 18. marta

U slučajevima kada su izrazi iste ili bliske granularnosti poređani jedan u odnosu na drugi, uz upotrebu predloga ili veznika (kao što su *od*, *pre*, *posle*, *nakon*, *i*, *ili* i sl.), biće obeleženi posebnim etiketama, kao dva zasebna izraza (primer 4.26).

Primer 4.26.
Koncert je zakazan za [subotu] u [osam uveče].

4.4 Format za obeležavanje vremenskih izraza

Proces klasifikovanja vremenskih informacija u semantičke klase jeste zapravo proces njihovog strukturiranja, koji će u ovom radu biti izvršen obeležavanjem teksta, odnosno umetanjem XML oznaka[2] koje opisuju značenje delova teksta koji je prepoznat kao vremenska informacija. U skladu sa TimeML uputstvom, svaki prepoznati vremenski izraz biće obeležen umetanjem `<TIMEX3>` etikete u okviru koje će biti definisan atribut `type`, kojim se specifikuje semantička klasa prepoznatog izraza. U zavisnosti od tipa vremenskog izraza, ovom atributu se može pripisati jedna od sledećih vrednosti: DATE, TIME, DURATION ili SET.

Izrazi koji ukazuju na temporalnu lokaciju u vidu kalendarskog vremena (primer 4.27) biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `type` biti DATE.

Primer 4.27.

Gospodin Petrović je otišao	*u* **petak 14. aprila 2007. godine**
	drugog decembra
	juče
	oktobra 2002.
	leta 2005.
	u **utorak 18-tog**
	leta ove godine
	prošle nedelje

[2] http://www.w3.org/

```
Petar je otišao <TIMEX3 type="DATE">juče</TIMEX3>.
```

S druge strane, izrazi koji ukazuju na temporalnu lokaciju u vidu vremena kao dela dana (primer 4.28) biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `type` biti TIME.

Primer 4.28.

Gospođa Ivanović je stigla	*u* **deset do tri popodne**
	u **pet minuta do osam**
	u **12 i dvadeset**
	pola sata posle ponoći
	u **jedanaest ujutru**
	u **9 h 14. oktobra 2009. godine**
	uveče 11. januara
	kasno sinoć
	prošle noći

```
Petar je otišao u <TIMEX3 type="TIME">18 časova</TIMEX3>.
```

Vremenski izrazi koji ukazuju na trajanje (primer 4.29) biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `type` biti DURATION.

Primer 4.29.

Gospodin Petrović je boravio na planini	***dva meseca***
	48 sati
	tri nedelje
	ceo mesec
	nekoliko dana

```
Koncert je trajao <TIMEX3 type="DURATION">dva sata</TIMEX3>.
```

Vremenski izrazi koji ukazuju na temporalnu frekvenciju, odnosno učestalost pojavljivanja u vremenu (primer 4.30) biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `type` biti SET.

Primer 4.30.

Gospođa Ivanović vežba	***dva puta nedeljno***
	svaka 2 dana
	svakog petka
	mesečno

```
Petar pliva <TIMEX3 type="SET">svakog petka</TIMEX3>.
```

U fazi prepoznavanja se, osim atributa `type`, svakom izrazu dodeljuje i atribut `temporalFunction`. Ovi je binarni atribut koji ukazuje na to da vrednost vremenskog izraza treba da bude određena putem određene računarske operacije. Vrednost ovog atributa će biti pozitivna u onim slučajevima koji ne sadrže sve informacije potrebne za popunjavanje pozicija višeg reda (leva strana) vrednosti datuma. Primer 4.31 ilustruje neke slučajeve:

Primer 4.31.

```
<TIMEX3 type="DATE" temporalFunction="true">31.  januara</TIMEX3>
<TIMEX3 type="TIME" temporalFunction="true">kasno sinoć</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true">nekoliko dana</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true">prošlih meseci</TIMEX3>
```

S druge strane, ako su pozicije višeg reda u datumu popunjene, onda atributu `temporalFunction` treba da se pripiše negativna vrednost. Takvi su primeri u 4.32:

Primer 4.32.

```
<TIMEX3 type="TIME" temporalFunction="false">13 časova 15.  januara
   2003.  godine</TIMEX3>
<TIMEX3 type="DATE" temporalFunction="false">leta 2007.</TIMEX3>
<TIMEX3 type="DATE" temporalFunction="false">petak 19.  oktobra
   1999.</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false">50 godina</TIMEX3>
```

Atribut `tid` je obavezan atribut, koji bi trebalo da bude dodeljen svakom vremenskom izrazu kao njegov jedinstveni identifikator. Međutim, u ovoj fazi rada se neće vršiti automatsko dodeljivanje identifikatora. O ostalim atributima, odnosno elementima etikete `<TIMEX3>` koji su relevantni za proces normalizacije prepoznatih vremenskih izraza, biće reči u petom poglavlju.

4.5 Metod za prepoznavanje vremenskih izraza zasnovan na konačnim transduktorima

Automatsko prepoznavanje vremenskih izraza je zadatak ekstrakcije informacija, odnosno preciznije jedan od zadataka prepoznavanja imenovanih entiteta, čiji je cilj automatsko ekstrahovanje onih segmenata teksta koji na direktan ili indirektan način prenose neku od vremenskih informacija. Kao i za bilo koji drugi zadatak ekstrakcije informacija, najčešće su korišćena dva osnovna principa, i to: metode zasnovane na pravilima, odnosno znanju i statističke metode (detaljna diskusija o ovome data je u delu 3.3). Pravila za ekstrakciju, osim ručnom metodom, mogu biti formirana i metodom vođeni na podacima (eng. *data driven*), odnosno pomoću mašinskog učenja, koje iziskuje velike količine tekstova prethodno obeleženih informacijama koje je potrebno ekstrahovati. Samo za zadatak prepoznavanja vremenskih izraza, obe tehnike mogu biti uspešno upotrebljene sve dok je dovoljna količina podataka potrebnih za obuku dostupna, i svaka tehnika ima svoje prednosti i nedostatke. Ipak, problem preciznog određivanja opsega identifikovanog vremenskog izraza i dalje se uspešnije rešava metodama zasnovanim na pravilima (Ahn, Fissaha Adafre, and De Rijke 2005). Isto tako, kada je reč o procesu normalizacije prepoznatih vremenskih izraza, statističke metode i metode zasnovane na mašinskom učenju ne mogu rešiti ovaj problem ni približno uspešno kao metode zasnovane na ručno pisanim pravilima (Kolomiyets and Moens 2010; Negri and Marseglia 2005; Jijkoun et al. 2008; Strötgen and Gertz 2010). S obzirom na to da je cilj ovoga rada razvijanje sistema koji vrši i identifikaciju i normalizaciju vremenskih izraza, usvojen je pristup zasnovan na ručno pisanim pravilima. Osim toga, budući da ne postoji prethodno obeleženi korpus, koji bi poslužio za obuku sistema u prepoznavanju vremenskih izraza, nije bilo moguće primeniti neku od metoda zasnovanih na mašinskom učenju. Kako kombinovanje i istovremeno rešavanje problema prepoznavanja i normalizacije u okviru jednog istog procesa utiče na složenost i efikasnost sistema, u okviru ovoga rada usvojen je metod koji u dve vremenski potpuno odvojene faze sprovodi prepoznavanje, a potom i normalizaciju vremenskih izraza. Iako odvojene, ove faze nisu nezavisne jedna u odnosu na drugu, jer su informacije prikupljene u prvoj fazi neophodne za sprovođenje druge faze. U fazi identifikacije se, osim opsega i tipa vremenskih izraza, identifikuju i one eksplicitno iskazane informacije relevantne za sledeći proces normalizacije vrednosti vremenskih izraza, poput semantičkih oznaka koje ukazuju na relativne izraze i potrebu za računanjem njihove vrednosti ili modifikatora koji direktno utiču na konačnu interpretaciju vrednosti prepoznatih vre-

menskih izraza. Osim toga, pravila za normalizaciju se na najjednostavniji način mogu implementirati kao elaboracija pravila korišćenih za prepoznavanje vremenskih izraza.

U daljem tekstu ovog poglavlja biće predstavljeni procesi uključeni u fazu prepoznavanja vremenskih izraza.

4.5.1 Opis korpusa

Kao tekstualni resurs korišćen je deo Korpusa savremenog srpkog jezika (Utvić 2014),[3] odnosno kolekcija novinskih tekstova prikupljenih tokom 2005-2012. godine iz više različitih izvora na srpskom jeziku (Glas javnosti, Blic, Večernje novosti, Srpski nacional, Politika, Danas, B92, Beta, Tanjug, FoNet). Više podataka o veličini korišćenog korpusa nalazi se u tabeli 4.4. Korpus novinskih tekstova sastojao se od ukupno 49.146 rečenica. Podela teksta na rečenice izvršena je automatski, u fazi prethodne obrade teksta pomoću transduktora konačnih stanja, organizovanih u vidu posebne gramatike, čija pravila opisuju situacije u kojima znakovi interpunkcije treba da budu shvaćeni kao separatori rečenica.

Tabela 4.4: Kvantitativni podaci o elektronskom korpusu

Period	Broj reči	Broj rečenica
2005-12.	544.642	49.146

S obzirom na to da posebno formiranog korpusa na srpskom jeziku koji bi bio namenjen istraživanju vremenskih informacija nema, za potrebe ovog rada upotrebljeni su novinski tj. informativni tekstovi koji predstavljaju bogat izvor primera vremenskih izraza. Ova vrsta tekstova, kao danas dominantan medij realizacije standardnog jezika (Antonić 2001), kako na planu jezičkog sadržaja, tako i na planu jezičkog izraza, je dovoljna za iscrpno istraživanje vremenskih izraza koji nose bitne informacije o sadržajima tekstova.

4.5.2 Tehnike, alati i resursi korišćeni za otkrivanje i obeležavanje vremenskih informacija u tekstu

Nakon jasnog i preciznog utvrđivanja vrste entiteta koji su nosioci vremenskih infromacija, kao i definisanja strukture izlaznih podataka, potrebno je generisati

[3] http://korpus.matf.bg.ac.rs/

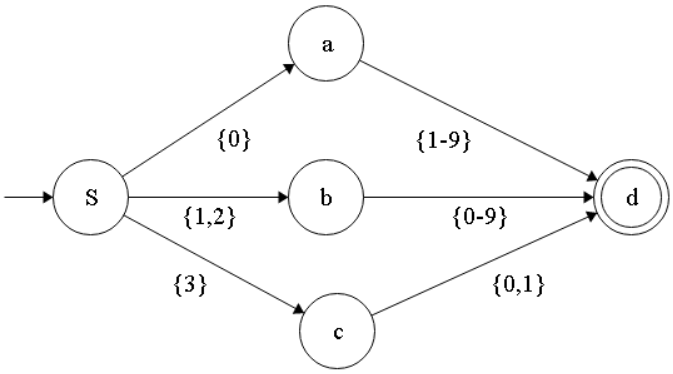
pravila na osnovu kojih će se vršiti ekstrakcija vremenskih izraza. Izvori za automatsku obradu srpskog jezika razvijeni su putem metode konačnih stanja, kao što su uveli Moris Gros i tim LADL laboratorije (Gross 1993), te je i za zadatak prepoznavanja vremenskih izraza u ovom radu usvojen isti metod (Friburger and Maurel 2004).

Konačni automati, a posebno transduktori, su u velikoj meri korišćeni u automatskoj obradi jezika, a izuzetno uspešno za rešavanje zadataka koji uključuju plitko parsiranje (eng. *shallow parsing*) ili segmentiranje (eng. *chunking*) (Abney 1996). Kako se zadatak prepoznavanja vremenskih izraza može posmatrati kao zadatak plitkog parsiranja, moguće je postići dobre rezultate korišćenjem pravila u obliku konačnih automata (Negri and Marseglia 2005; Ahn, Adafre, and Rijke 2005). Roche i Schabes (1997) čak smatraju da je modelovanje pomoću konačnih transduktora neophodno, jer, za razliku od pravila u vidu regularnih izraza, transduktori omogućavaju bavljenje mnogo složenijim formalizmima, poput kontekstno slobodnih gramatika.

Pojam lokalne gramatike, kao načina za opisivanje sintaksičke strukture grupe pojedinačnih elemenata koji su međusobno povezani, a čije sličnosti se ne mogu lako izraziti, prvi je upotrebio Gros (1993) kako bi opisao izraze za datume i vreme. Cilj takve gramatike je, na šta ukazuje i sam naziv lokalna, opis određenih lokalnih uslova i ograničenja u odnosu na susedne niske iskaza, a ne opis sintaksičke strukture cele rečenice. Osnovna ideja primene lokalnih gramatika jeste definisanje najčešćih konstrukcija teksta, tako da se ponovo mogu iskoristiti u opisu onih složenijih. Primeri lokalnih gramatika koje opisuju izraze za prepoznavanje imenovanih entiteta u srpskom mogu se naći u (Gucul-Milojević, Radulović, and Krstev 2008; Krstev, Vitas, and Gucul 2005; Krstev et al. 2014).

Svaki transduktor je zapravo lokalna gramatika. Za razliku od konačnih automata, koji definišu formalni jezik na osnovu koga se utvrđuje da li određena niska pripada jeziku opisanom tim automatom ili ne, konačni transduktori definišu relacije između dva skupa niski karaktera, odnosno transformišu zadatu nisku u drugu nisku nad istom ili nekom drugom azbukom, pa se iz tog razloga često nazivaju i konačnim automatima prevodiocima (Vitas 2006). Na primer, automat *A* prikazan na slici 4.1 definiše jezik *L={01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13 ... 31}*, odnosno jezik svih reči[4] sastavljenih od arapskih cifara (0-9) dužine *n=2*, a

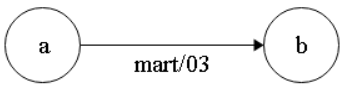
[4]Reč u smislu formalnih jezika – niske nad alfabetom.



Slika 4.1: Konačni automat koji prepoznaje dane u mesecu napisane ciframa

koje mogu da predstavljaju dan u mesecu. U automatu sa slike *S* je početno stanje koje odgovara praznoj reči, a *d* je završno stanje koje odgovara rečima iz skupa *L*. Prelazak iz početnog stanja *S* po luku 0 u stanje *a* moguće je kada automat pročita ulazni simbol 0. Da bi stigao u završno stanje automat zatim treba da pročita neki od simbola predstavljenih lukom iz čvora *a* u čvor *d* (1-9).

S druge strane, konačni transduktori su mašine koje prepoznaju jednu nisku karaktera, a generišu neku drugu. Moguće ih je predstaviti na sličan način kao i konačne automate, pa bi tako deo grafa za slučaj jednog meseca koji prepoznaje naziv meseca napisan slovima i proizvodi njegovu odovarajuću numeričku oznaku[5] izgledao kao na slici 4.2:



Slika 4.2: Prelazak iz stanja *a* u stanje *b* nakon što je pročitana, odnosno prepoznata reč *mart*, pri čemu se generiše reč *03*

Konačni automati i transduktori se mogu primeniti na više načina. Jedan primer je kaskadna primena transduktora umesto kontekstno slobodne gramatike (Abney 1996), što podrazumeva uzastopnu primenu serije transduktora na tekst preciznim redosledom kako bi se transformisao tekst ili ekstrahovali obrasci iz teksta. Primera radi, transduktor T_i analizira tekst L_{i-1} i proizvodi tekst L_i. Zatim transduktor T_{i+1} analizira tekst L_i i proizvodi tekst L_{i+1}, i tako dalje redom. Transduktor ne omogućava rešavanje svih lingvističkih fenomena, ali svaki transduktor u nizu modelira deo problema koji se rešava na lokalnom nivou. Stoga su sistemi zasnovani na kaskadama transduktora robusni i precizni. U (Friburger and Ma-

[5]Npr. mart mesec je treći mesec u godini, pa bi trebalo da bude predstavljen vrednošću 03.

urel 2004) prikazan je način kombinovanja konačnih transduktora u serije, radi ekstrakcije imenovanih entiteta iz novinskih članaka za francuski jezik.

Automati su čitljivi i jednostavni za korišćenje i lingvistima. Grafički prikaz automata omogućava njihovu vizuelizaciju i jednostavno modifikovanje, što olakšava korigovanje i kompletiranje gramatika. Za razvoj i primenu automata i transduktora u ovom radu korišćen je softverski alat Unitex (Paumier 2016), koji je namenjen lingvističkoj obradi teksta na prirodnim jezicima. Ovaj sistem otvorenog koda (engl. *open source*), projektovan je tako da može da podržava različite prirodne jezike. Kolekcija programa kojima se vrši analiza teksta pisani su na programskim jezicima C/C++, dok je grafički korisnički interfejs pisan u programskom jeziku Java. Analiza teksta u okviru Unitex-a omogućena je zahvaljujući postojećim lingvističkim resursima, koji se sastoje od elektronskih rečnika i gramatika, prvobitno razvijenih za francuski jezik, a potom i za engleski, grčki, portugalski, ruski, tajlandski, korejski, španski, italijanski, nemački, norveški, arapski i druge, uključujući i srpski jezik (Krstev 2008).

Obrada teksta pre identifikacije relevantnih entiteta je od izuzetnog značaja, pa se u Unitex-u izvode određene operacije, čije je sprovođenje takođe omogućeno konačnim mašinama. Rad u Unitex-u i počinje odabirom jezika, od čega će zavisiti i odabir alfabeta reči, kao i elektronskih rečnika i gramatika koje će se koristiti u fazi prethodne obrade. Nakon što je tekst koji će biti analiziran snimljen u odgovarajućem formatu (Unicode Little Endian 16bit format - UTF16LE),[6] pristupa se njegovoj pripremi prvo u vidu normalizacije teksta (normalizacija separatora teksta, uklanjanje znakova interpunkcije i slično), procesa koji, u zavisnosti od važnosti za kasniji proces ekstrakcije informacija i od samog teksta koji se obrađuje, može biti i izostavljen. Posle ove faze sledi faza tokenizacije teksta, u okviru koje se vrši razdvajanje teksta na tokene, odnosno niske nad alfabetom jezika. Najčešće se pod tokenima smatraju reči nekog jezika, dok se kod niski nealfabetskih karaktera svaki karakter smatra za token. Lista dobijenih tokena se koristi kasnije u procesu leksičke i morfološke analize. Osim ova dva procesa, vrši se i podela teksta na rečenice pomoću transduktora koji dodaje simbol {S} kao graničnik između rečenica i opisuje u kom kontekstu znakovi interpunkcije treba da budu shvaćeni kao separatori rečenica.

Nakon ovih procesa, na ovako pripremljen tekst primenjuju se elektronski reč-

[6]The Unicode Standard. Dostupno na: http://www.unicode.org

nici, koji omogućavaju sprovođenje leksičke i morfološke analize. U okviru leksičke analize, tokeni dobijeni u prethodnoj fazi pronalaze se u rečnicima, na osnovu čega se formira rečnik teksta, pa se svakoj reči iz teksta pridružuju moguće interpretacije: vrsta reči, gramatičke kategorije i sl. Obeležavanje teksta sa morfološkog i sintaksičkog gledišta putem rečnika omogućava pridruživanje odgovarajućih informacija rečima u tekstu, kao što su leme, gramatičke kategorije (imenica, glagol, pridev itd.) i semantičke odlike (zajedničke imenice, imena mesta, lična imena, skraćenice itd.). Zahvaljujući pridruženim lemama izbegava se opisivanje svih flektivnih oblika reči u transduktorima koji ih otkrivaju, dok zahvaljujući pridruženim sintaksičkim (npr. ADV) i semantičkim informacijama (Temp) može da se olakša npr. otkrivanje nekih vremenskih izraza koji se sastoje od različitih oblika vremenskih priloga. Postojeće elektronske morfološke rečnike za srpski jezik kreirali su u DELA (*Dictionnaires Electroniques du LADL - LADL electronic dictionaries*) formatu (Silberztein 1993) za korišćenje u Unitex-u C. Krstev i D. Vitas (Krstev 2008; Vitas et al. 2003; Krstev, Vitas, and Savary 2006). U okviru sistema Unitex omogućena je i primena više rečnika na tekst, uz jasno definisan redosled njihove primene.

Za razliku od drugih sistema zasnovanih na pravilima, čiji proces prepoznavanja vremenskih izraza počinje fazom prethodne obrade teksta (Negri and Marseglia 2005; Ahn, Adafre, and Rijke 2005), sistem razvijen u okviru ovog istraživanja proces započinje obradom teksta, koji nije podvrgnut analizi prethodne obrade, i primenom transduktora, koji vrši prepoznavanje i obeležavanje najčešćih numeričkih formi datuma. Ova odluka doneta je s ciljem da bi se izbegli problemi koji mogu nastati, pre svega, u procesu segmentacije teksta na rečenice (Ahn, Adafre, and Rijke 2005). Na primer, pogrešno umetanje separatora rečenice {S} često može dovesti do nepronalaženja ili pogrešnog prepoznavanja određenih obrazaca vremenskih izraza (kao u *05.{S}07.2010. godine*).

4.5.3 Kaskada transduktora za prepoznavanje vremenskih izraza

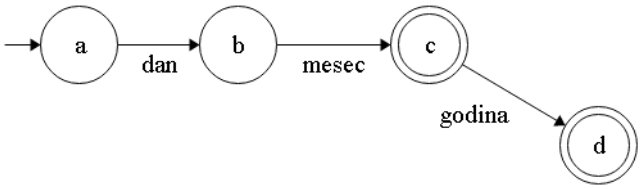
Sistem za prepoznavanje vremenskih izraza srpskog jezika zasniva se na kaskadi transduktora – CasSys (Friburger 2002; Friburger and Maurel 2004), koja je integrisana u Unitex sistem. Kaskada za prepoznavanje se sastoji od 16 transduktora, čija je uloga identifikacija izraza, kao i određivanje opsega i tipa svakog otkrivenog izraza (detaljno opisanih u delu 4.2), a u skladu sa TimeML shemom

(DATE, TIME, DURATION i SET). Kreirani grafovi ove opširne gramatike u vidu serije transduktora dizajnirani su s ciljem prepoznavanja izraza koji ukazuju na kalendarske datume, vremena dana, trajanja i učestalosti vremena koja se ponavljaju. Pomoću grafičkog korisničkog interfejsa Unitex-a i radnog okruženja koje ovaj sistem obezbeđuje izvodi se ceo proces, od prethode obrade teksta, preko kreiranja pravila prepoznavanja, do samog obeležavanja, odnosno izdvajanja vremenskih informacija.

Na osnovu analize primera vremenskih izraza i njihovih konteksta pojavljivanja u okviru korišćenog korpusa, definisana su prvo visoko pouzdana pravila ekstrakcije koja obuhvataju većinu slučajeva, odnosno najčešće oblike pojavljivanja vremenskih informacija. Daljim razvojem kreirana su nova pravila i modifikovana već postojeća, kako bi se obuhvatili i ređi slučajevi pojavljivanja vremenskih izraza. Na primer, algoritam koji koristi transduktor prikazan na slici 4.3 izdvaja neke moguće oblike kalendarskih datuma, koji se sastoje od oznake za dan (napisane ciframa ili slovima) iza koje sledi oznaka za mesec (napisana slovima, ciframa ili rimskim brojevima), iza kojih može, a ne mora da sledi oznaka za godinu (napisana arapskim ciframa). Samo ona sekvenca koja odgovara putanji definisanoj u transduktoru biće prepoznata kao vremenski izraz, na osnovu koga će se generisati izlaz u vidu prepoznatog izraza obeleženog leksičkim etiketama, koje će se koristiti tokom primene sledećeg transduktora u nizu. Semantički markeri koji se pridružuju prepoznatim vremenskim izrazima pružaju korisne informacije o tipu prepoznatog imenovanog entiteta (+*time* – vremenski izraz), kao i tipu vremenskog izraza (+*date*, +*hour*,[7] +*duration*, +*set*), kao što je dato u primeru 4.33. Dodatne informacije koje se odnose na tip vremenskog izraza date su putem semantičkih markera +*abs* i +*rel*, koji ukazuju na apsolutne i relativne vremenske izraze.

Primer 4.33.
{*13. juna 2008. godine,.NE+time+date+abs*}
donji put u transduktoru sa slike 4.3
{*petog maja,.NE+time+date+rel*}
gornji put u transduktoru sa slike 4.3

[7]Ovaj semantički marker označava vremenski izraz koji se odnosi na vreme kao deo dana i u okviru TimeML sheme je definisan kao TIME. U ovom trenutku obrade vremenskih informacija koristi se shema koja odgovara shemi korišćenoj u okviru rada na razvoju višejezične ontologije vlastitih imena (Maurel, 2007).



Slika 4.3: Jedna putanja transduktora koji prepoznaje određene potpuno precizne i neprecizne kalendarske datume

Pravila u vidu transduktora se primenjuju kaskadno, odnosno u nekoliko uzastopnih faza, pri čemu izlaz jedne faze predstavlja ulaz za sledeću. Prioritet je dat transduktorima koji pronalaze najduže obrasce kako bi se izbegli slučajevi pogrešnog obeležavanja. Na primer, izraz *15. mart 2014. godine* može biti prepoznat putem više transduktora, i to:

- transduktor *Godina* će prepoznati i obeležiti kao apsolutni vremenski izraz sekvencu *2014. godine*;
- transduktor *Datum relativni* će prepoznati i obeležiti kao relativni vremenski izraz sekvencu *15. mart*;
- transduktor *Datum prošireni* će prepoznati i obeležiti kao apsolutni vremenski izraz sekvencu *15. mart 2014. godine*.

Kako bi se omogućilo ispravno prepoznavanje i obeležavanje vremenskih izraza ove transduktore je neophodno primeniti određenim redosledom: prvo *Datum prošireni*, zatim *Godina* i *Datum relativni*.

Već je rečeno da će se dobijene leksičke etikete koristiti u narednim transduktorima, omogućavajući otkrivanje mnogo složenijih izraza, poput onih koji označavaju vremenski period (primer 4.34) ili predstavljaju kombinaciju kalendarskih datuma i vremena kao delova dana (primer 4.35).

Primer 4.34.
(a) *od {{8. marta,.NE+time+date+rel} do {7. aprila,.NE+time+date+rel},. NE+time+date+period}*
(b) *između {{18 i 30,.NE+time+hour+abs} i {19 h,.NE+time+hour+abs},. NE+time+hour+period}*

Primer 4.35.
{{*15. marta,.NE+time+date+rel*} *oko* {*2 sata,.NE+time+hour+abs*},.
NE+time+hour}

Kaskadnom primenom transduktora se u velikoj meri može uticati na poboljšanje efikasnosti procesa i povećanje preciznosti, ali i na smanjenje dvosmislenosti. Na primer, sposobnost sistema da odredi prioritete primenjivanja pravila omogućava rešavanje dvosmislenih situacija u slučaju izraza koji mogu da ukažu na tačku u vremenu, kao i na trajanje. Recimo, izraz *od 12 časova* može da implicira precizno vreme dana ili trajanje. Za oba primera je definisano više pravila, koja treba da se primene određenim redosledom kako bi se obezbedila što veća preciznost sistema. Prvo se primenjuje transduktor koji koristi levi kontekst vremenskog izraza (npr. ***u trajanju*** *od 12 časova*) i koji će 1*2 časova* obeležiti kao izraz koji ukazuje na trajanje (+*time*+*duration*). Nakon toga, sledeći transduktor u kaskadi sa sigurnošću može ostala pojavljivanja iste sekvence obeležiti kao vreme dana (+*time*+*hour*). Osim toga, veliki je broj pojavljivanja brojeva u tekstu koji ne moraju nužno da ukažu na npr. vreme dana, pa nam kaskadna primena transduktora pomaže u razrešavanju dvosmislenosti. Na primer, brojevi koji se pojavljuju u kontekstu nekih već prepoznatih vremenskih izraza koji ukazuju na vreme dana mogu pouzdano da ukažu na vreme dana, iza kojeg se ne nalazi oznaka jedinice mere vremena (npr. *sat*, *čas*, *h*), kao što je dato u primeru 4.34b (početak vremenskog perioda).

Leksičke etikete koje proizvode transduktori predstavljaju izlaz koji je prikladan za korišćenje u kaskadnoj primeni transduktora, ali nije koristan za druge aplikacije, te će na kraju procesa prepoznavanja vremenskih izraza sve leksičke etikete (primer 4.36a) biti konvertovane u dve vrste XML etiketa: jedne su u skladu sa primenjenom shemom (primer 4.36b), a druge su u skladu sa TimeML shemom (primer 4.36c).

Primer 4.36.
(a) {*hiljadu i 200 godina,.NE+time+duration+abs*}
(b) `<time.duration.abs>hiljadu i 200 godina</time.duration.abs>`
(c) `<TIMEX3 type="DURATION">hiljadu i 200 godina</TIMEX3>`

Kaskada se sastoji od sledećih 16 transduktora, od kojih se prvi koristi pre leksičke analize:

1. *Kalendarski datum* (prepoznaje najčešće oblike potpunih kalendarskih datuma napisanih ciframa);

2. *Datum prošireni* (prepoznaje i obeležava apsolutne i relativne datume, kao i trajanja u obliku kalendarskih datuma);

3. *Datum skraćeni* (prepoznaje i obeležava apsolutne i relativne datume kraće forme od prethodno prepoznatih; na kraju se prepoznaju oni vremenski izrazi iskazani samo danom ili nekom jedinicom vremenske mere);

4. *Period datuma* (prepoznaje i obeležava periode datuma, odnosno složene izraze sastavljene od kalendarskih datuma; u okviru ovog transduktora se prepoznaju i neki datumi koji nisu mogli sa preciznošću da budu otkriveni u prethodnim koracima, već samo uz pomoć prethodno prepoznatih datuma);

5. *Učestalost* (prepoznaje vremenske izraze koji označavaju učestalost);

6. *Trajanje modifikovano* (prepoznaje izraze koji ukazuju na trajanja koja su modifikovana na određeni način);

7. *Trajanje u kontekstu* (prepoznaje apsolutna trajanja svih nivoa granularnosti pomoću precizno definisanog levog ili desnog konteksta);

8. *Vreme dana* (u okviru ovog transduktora se prepoznaju izrazi granularnosti manje od dana, odnosno vremenski izrazi koji označavaju delove dana; u ovom koraku se prepoznaje i trajanje vremena iskazanog na nivou manjem od dana);

9. *Vreme dana u kontekstu* (transduktor prepoznaje vreme dana pomoću precizno definisanog levog ili desnog konteksta);

10. *Vreme dana skraćeno* (transduktor koristi prethodno otkrivene izraze radi pronalaženja novih izraza iskazanih na nivou manjem od dana);

11. *Period vremena dana* (prepoznaje periode vremena iskazanih jedinicama granularnosti manjim od dana; u ovom koraku koriste se i prethodno prepoznati izrazi ove granularnosti);

12. *Trajanje osnovno* (prepoznaje kraće oblike vremenskih izraza sa značenjem apsolutnih trajanja);

13. *Trajanje relativno* (prepoznaje najkraće oblike vremenskih izraza sa značenjem relativnih trajanja);

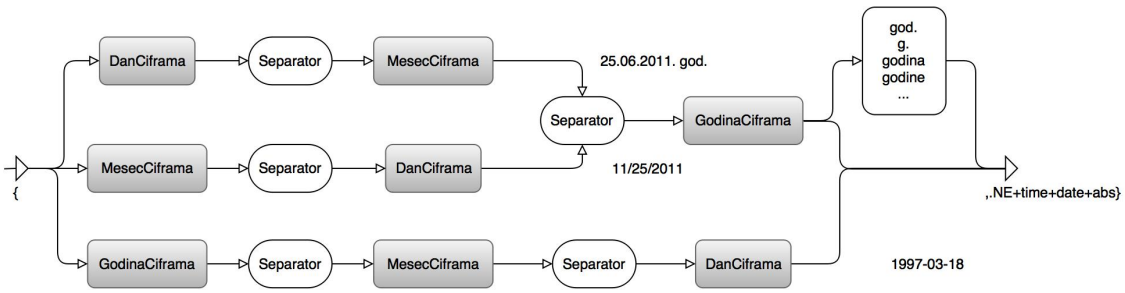
14. *Datum relativni* (prepoznaje kraće oblike vremenskih izraza sa značenjem relativnih datuma);

15. *Period trajanja* (prepoznaje periode trajanja, kao i najkraće oblike vremenskih izraza sa značenjem relativnih datuma);

16. *Vreme* (poslednji transduktor koji koristi sve ranije prepoznate izraze za otkrivanje složenih oblika koji se sastoje iz svih nivoa granularnosti).

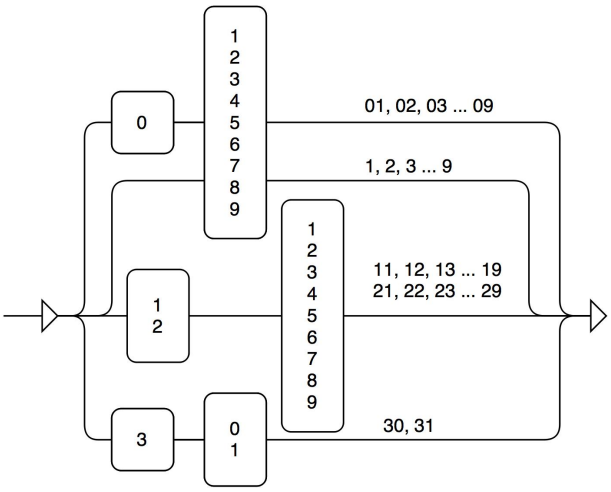
1. Kalendarski datum

Neposredno pred procese tokenizacije i segmentacije teksta na rečenice, kao i sprovođenje leksičke i morfološke analize, potrebno je primeniti transduktor *Kalendarski datum* koji prepoznaje najčešće oblike kalendarskih datuma napisanih ciframa. Ukoliko bi ovaj transduktor bio primenjen na tekst nad kojim je izvršena segmentacija teksta na rečenice, broj prepoznatih vremenskih izraza ovoga tipa bi bio daleko manji u odnosu na rezultate koje bismo dobili primenom transduktora na čist tekst (115 u odnosu na 5.397 kalendarskih datuma ispravno prepoznatih u ovoj fazi obrade u korišćenom korpusu zbog prethodno objašnjenih razloga). Kalendarski datumi koji predstavljaju potpuno precizne vremenske izraze na nivou dana najčešće su iskazani numeričkim obrascima za koje je kreirano posebno pravilo ekstrakcije. Na slici 4.4 prikazan je transduktor *Kalendarski datum* koji vrši identifikaciju i obeležavanje ovih vremenskih izraza.



Slika 4.4: Transduktor *Kalendarski datum.grf* za prepoznavanje i obeležavanje kalendarskih datuma iskazanih numeričkim obrascima

Prikazani transduktor poziva podgrafove *DanCiframa*, *MesecCiframa* i *GodinaCiframa*, koji su na slici označeni sivom bojom. Podgraf *DanCiframa* prepoznaje dane u mesecu predstavljene arapskim ciframa. Budući da se mesec sastoji od najviše 31 dana, omogućeno je prepoznavanje brojeva od 1 do 31. Kada je reč o



Slika 4.5: Podgraf *DanCiframa.grf* koji prepoznaje dane u mesecu iskazane arapskim brojevima

predstavljanju prvih devet dana, radi postizanja što većeg odziva u prepoznavanju kalendarskih datuma, omogućeno je prepoznavanje i onih primera kada su oni iskazani dvema ciframa (npr. *01*, *02*, *03* itd.), što nije u skladu sa pravopisinim pravilima srpskog jezika, ali je u čestoj upotrebi. Ovaj podgraf prikazan je na slici 4.5.

S obzirom na to da se godina sastoji od 12 meseci, na isti način je organizovan i funkcioniše i podgraf *MesecCiframa*, pomoću koga se identifikuju brojevi vrednosti od 1 do 12. Kako se u informativnim tekstovima koji su korišćeni za razvoj ovog sistema ne pojavljuju u ovoj formi kalendarski datumi s početka nove ere, u ovoj fazi obrade podgraf *GodinaCiframa* prepoznaje samo one godine koje pripadaju drugom i trećem milenijumu. Prepoznavanje svih godina stare i nove ere biće omogućeno u narednim fazama, i to uključivanjem modula tj. podgrafa *GodinaBC* koji u određenom kontekstu sa visokom preciznošću može prepoznati i ovakve oblike (npr. *25.10.354. godine pre n.e.*, *12. aprila 1876. g. p.n.e.*). Upravo ova mogućnost jednostavnog korišćenja određenih pravila ekstrakcije u različitim kontekstima oslikava modularnost čitavog sistema. Tako će i podgrafovi *DanCiframa*, *MesecCiframa* i *GodinaCiframa* biti ponovo upotrebljavani tokom dalje obrade u okviru sledećih transduktora.

Transduktor *Kalendarski datum*, dakle, prepoznaje kalendarske datume sa različitim pozicijama dana, meseci i godina, između kojih se mogu naći znaci interpunkcije, kao što su tačka, crtica ili kosa crta. Ako se iza godine nađu tačka ili oznaka za godinu (imenica *godina* u bilo kom padežu jednine[8] ili njen skraćeni

[8]S obzirom na to da u ovom trenutku obrade elektronski rečnici još uvek nisu primenjeni,

oblik), biće identifikovane kao sastavni deo vremenskog izraza. Tako ovaj transduktor prepoznaje, između ostalih, sledeće izraze: *12.07.1967. godine*, *11/29/2009*, *1995-09-25* itd. Prepoznate sekvence biće obeležene leksičkim etiketama, u okviru kojih se nalaze podaci o tome da je reč o imenovanom enitetu – *NE* (eng. *named entity*), koji je u grupi apsolutnih vremenskih izraza koji ukazuju na temporalnu lokaciju u vidu kalendarskog datuma *+time+date+abs* (npr. {*13.06.2008. godine,.NE+time+date+abs*}). Nad ovako obeleženim tekstovima se sa mnogo većom preciznošću može izvršiti umetanje separatora rečenica i, nakon toga, primena elektronskih rečnika. Prepoznati i obeleženi izrazi će u kasnijoj fazi obrade poslužiti za otkrivanje složenijih vremenskih izraza, o čemu će biti više reči u daljem tekstu.

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Kalendarski datum* nalaze se u primeru A.1, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.1.

2. Datum prošireni

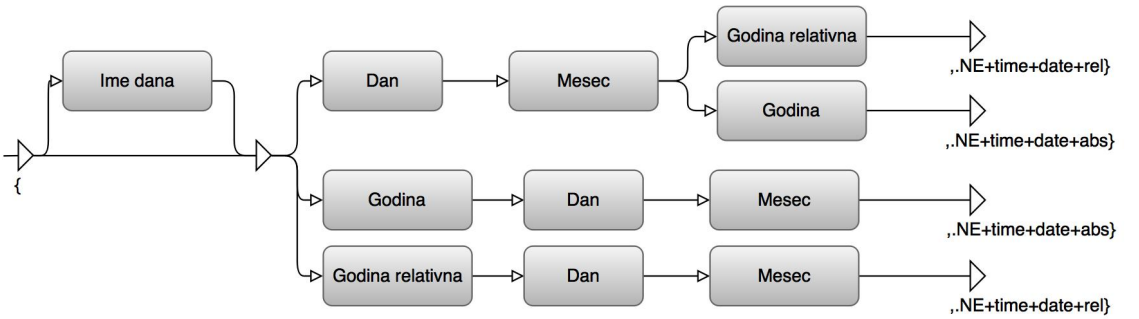
Prethodno korišćeni podgrafovi - *DanCiframa*, *MesecCiframa*, *GodinaCiframa*

Novi podgrafovi – *DanSlovima*, *MesecRimskiSlovima*, *Godina relativna*, *Jedinice mere vremena*, podgrafovi za prepoznavanje kraćih oblika datuma (mesec-godina, dan-mesec, samo mesec, samo dan), *ImeDana*, *ADJ_Date*

Sledeći transduktor u nizu prepoznaje najduže forme kalendarskih datuma, koji mogu biti u obliku dan-mesec-godina ili godina-dan-mesec (slika 4.6). Prepoznaje i apsolutne i relativne datume granularnosti na nivou godine. U ovom koraku se, pre svega, koriste prethodno opisani podgrafovi *DanCiframa*, *MesecCiframa* i *GodinaCiframa*, kao i novi koji prepoznaju mesece napisane slovima ili rimskim brojevima (*MesecRimskiSlovima*) i dane napisane slovima (*DanSlovima*), koji služe za prepoznavanje apsolutnih vremenskih izraza u obliku kalendarskih datuma. Ako se u kontekstu nalaze ime dana u nedelji (podgraf *ImeDana*), imenica *dan* ili prilozi (*danas*, *juče*, *sutra*, *sutradan*) oni će ući u opseg izraza. U svim slučejevima imenica godina može da bude navedena, ali i izostavljena. Dakle, ovaj transduktor prepoznaje datume i bez navedenog leksičkog okidača.

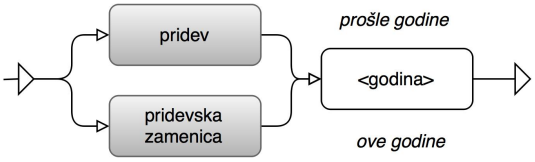
Za prepoznavanje najdužih formi relativnih izraza u obliku kalendarskih datuma, osim podgrafova za dane i mesece koji su upotrebljavani, umesto podgrafa

nije moguće upotrebiti leksičku masku <godina>, koja bi omogućila prepoznavanje svih oblika imenice *godina*.



Slika 4.6: Neke putanje transduktora *Datum prošireni*

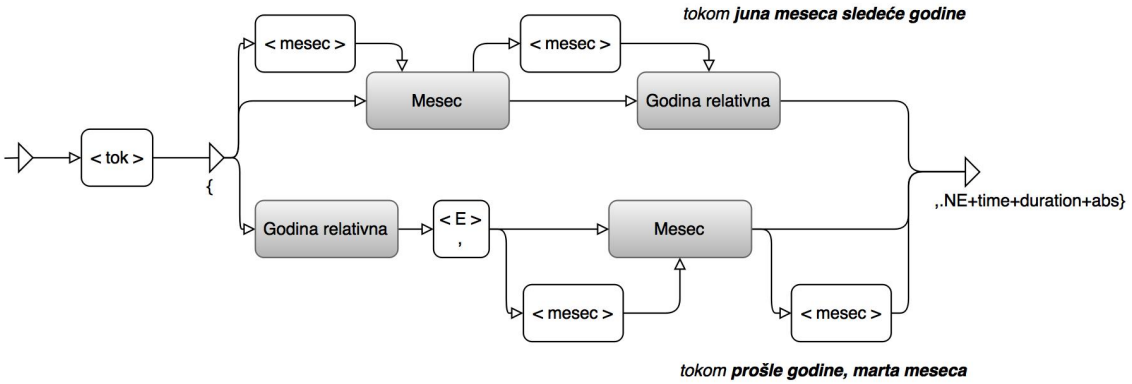
Godina koristi se podgraf *Godina relativna*, koji prepoznaje imeničke sintagme čije je leksičko jezgro imenica *godina*, a mogući zavisni konstituenti pridev (*prošli*, *sledeći*) ili pridevska zamenica (*ovaj*, *taj*) (slika 4.7).



Slika 4.7: Podgraf *Godina relativna.grf*

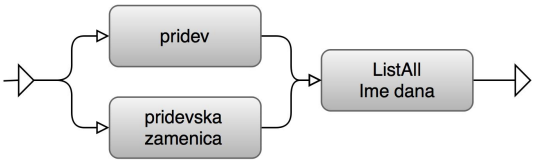
U okviru ovog transduktora koriste se i grafovi za prepoznavanje kraćih oblika kalendarskih datuma, ali ovoga puta u kontekstu u kome oni nose značenje vremenskog trajanja, kao što je u primerima *tokom* ***aprila meseca 2005. godine*** ili *tokom* ***sledeće godine***. Različiti predlozi formirani od imenice *tok* ispred apsolutnog ili relativnog vremenskog izraza koji implicira tačku u vremenu omogućavaju nam otkrivanje apsolutnih vremenskih izraza sa značenjem trajanja, u slučaju navedenih primera od po mesec i godinu dana. Vremenski izrazi koji se prepoznaju u ovoj fazi u opisanom kontekstu spadaju u grupu apsolutnih (npr. *marta 2009. g.*, *leta 1995*, *2011*) i relativnih vremenskih izraza (npr. *juna prošle godine*, *sledeće godine maja meseca*, *jeseni ove godine*, *petka 12. februara*, *petog marta*, *avgusta meseca*, *10-og*, *ove godine*, *sedmice*), detaljno opisanih u delu 4.2.1. Za prepoznavanje svakog od navedenih formata kalendarskih datuma koriste se već postojeći grafovi namenjeni identifikaciji različitih oblika kojima mogu biti predstavljeni dani, meseci i godine. Na slici 4.8 ilustrovana je jedna od putanja grafa *MesecGodina* koji u određenom kontekstu (ako se ispred nalazi neki od oblika imenice tok) prepoznaje relativni vremenski izraz u formi kalendarskog datuma i obeležava ga kao apsolutni vremenski izraz koji ukazuje na trajanje.

Među najkraćim oblicima kalendarskih datuma nalaze se oni oblici opisani pravilima u grafu *ADJ_Date* (slika 4.9) i koji se sastoje od imeničke sintagme sa pride-



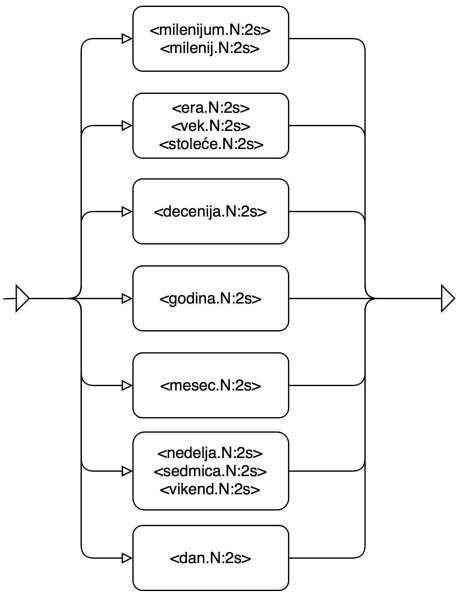
Slika 4.8: Jedna od putanja grafa *MesecGodina.grf*

vom ili pridevskom zamenicom i jedinice vremenske mere (*tokom **prošle godine**, u toku **tog utorka***).



Slika 4.9: Graf *ADJ_Date.grf*

Jedinice vremenske mere od najvišeg nivoa do nivoa granularnosti dana (one mogu označavati kalendarske datume) prepoznaju se grafom *ListAll*, koji je prikazan na slici 4.10.



Slika 4.10: Graf *ListAll.grf*

Na kraju ostaju relativni datumi koji se sastoje samo od imenice koja označava

jedinicu vremenske mere. Pošto se radi o kalendarskom datumu, najuža granularnost je na nivou dana. Ovde se koristi podgraf *ListAll*, ali modifikovan u tom smislu da obuhvata samo jedinice vremenske mere u genitivu jednine (npr. *godine*, *nedelje*, *veka*), ali ispred kojih mora da se nađe modifikator. To znači da se u ovom trenutku prepoznaju modifikovani izrazi poput *krajem godine*, *prvoj polovini godine*, *sredinom meseca*, *krajem decenije*.

U ovoj fazi se prepoznaju i izrazi koji ukazuju na trajanje, a dati su u obliku godine napisane ciframa iza koje nije navedena imenica *godina* (npr. *2009*). Za ove potrebe definisan je širi levi kontekst koji sadrži glagol u pasivnoj konstrukciji (npr. *rođen je*) iza koga sledi *tokom*, dok je u okviru desnog konteksta isključena mogućnost pojavljivanja nekog od oblika imenice *godina*.

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Datum prošireni* nalaze se u primeru A.2, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.2.

3. Datum skraćeni

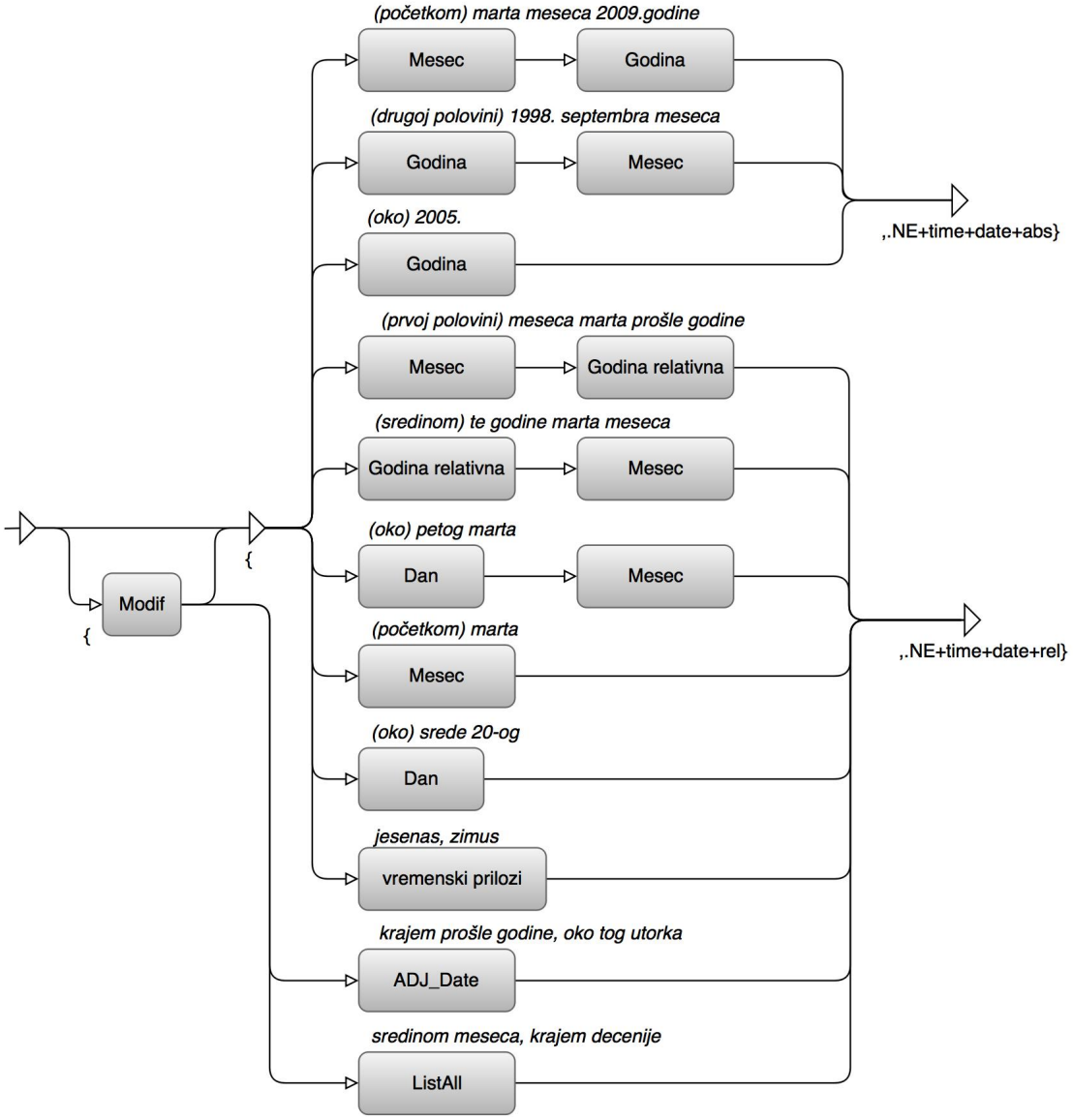
Prethodno korišćeni podgrafovi – *Dan*, *Mesec*, *Godina*, *DanSlovima*, *MesecRimskiSlovima*, *Godina relativna*, podgrafovi za prepoznavanje kraćih oblika datuma (mesec-godina, dan-mesec, samo mesec, samo dan), *ImeDana*, *ADJ_Date*

Prethodno korišćeni podgrafovi u modifikovanom obliku – *ListAll*

Novi podgrafovi – *Modif*

Ovaj transduktor poziva iste podgrafove koji su korišćeni i za prepoznavanje apsolutnih i relativnih datuma sa značenjem trajanja, ali ovoga puta u kontekstu u kome oni imaju značenje tačke u vremenu (slika 4.11).

U ovoj fazi se prepoznaju najduži oblici vremenskih izraza koji ukazuju na kalendarski datum, u smislu uključivanja i modifikatora koji može biti ispred vremenskog izraza. Prvo se prepoznaju apsolutni vremenski izrazi koji se sastoje od podataka o mesecu i godini, a ako se ispred nalazi modifikator i on se uključuje u opseg (npr. *početkom marta meseca 2009. godine*, *meseca aprila 1998.*). U slučaju da izraz nije modifikovan, transduktor obeležava prepoznati vremenski izraz. Zatim, sledi prepoznavanje relativnih vremenskih izraza, kao i modifikatora koji se mogu naći ispred njih (npr. *početkom meseca marta ove godine*, *krajem marta meseca te godine*, *krajem novembra prošle godine*, *sredinom te godine meseca marta*,

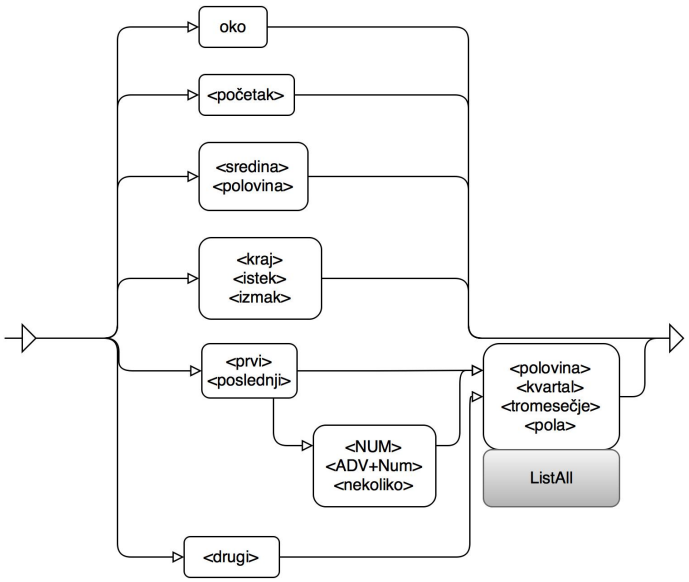


Slika 4.11: Uprošćeni prikaz transduktora *Datum skraćeni*

drugoj polovini aprila ove godine, *januara sledeće godine*).

Modif podgraf, prikazan na slici 4.12, obuhvata sve identifikovane modifikatore. Za njihovo prepoznavanje se koriste leksičke maske (npr. za pridev u jednini koji stoji uz imenicu koja ima ulogu modifikatora vremenskog izraza biće upotrebljena maska `<drugi.A:s>` za izraz ***drugoj polovini*** *meseca marta prošle godine*).

U okviru ovog transduktora se prepoznaju i apsolutni i relativni vremenski izrazi koji ukazuju na kalendarski datum na nivou godine, iskazan samo podatkom o godini. U slučaju da je prepoznati izraz modifikovan, modifikator će biti uključen u opseg izraza (npr. *krajem 2005.*, *2005. godine*, *krajem prošle godine*, *sledeće*



Slika 4.12: *Modif.grf* podgraf

godine). Situacije kada je izraz iskazan podatkom o godini iza koje se ne nalazi neki od oblika imenice godina, neophodno je koristiti informacije date levim ili desnim kontekstom (npr. **rođen je** *2005*, **(***2005.***)** i sl.).

Zatim sledi prepoznavanje relativnih vremenskih izraza u obliku kalendarskog datuma, koji se sastoje samo od podataka o danu i mesecu (npr. *5. marta*, *petog marta*, *petka, 12. februara*, *sredu 11. februara*), samo mesecu (npr. *marta meseca*, *meseca jula*, *početkom marta*) ili samo danu (npr. *10-og*, *petak, prvog*, *juče*). U ovom koraku se podgrafom *ADJ_Date* prepoznaju i modifikovane imeničke sintagme, poput *krajem prošle godine*, *oko tog utorka*, kao i oni vremenski izrazi sa značenjem tačke u vremenu, koji mogu biti iskazani vremenskim prilozima (npr. *jesenas*, *zimus*, *proletos*, *letos*).

Osim navedenih oblika, ovim transduktorom se vrši prepoznavanje i relativnih vremenskih izraza koji ukazuju na kalendarski datum, a iskazani su nekim oblikom imenice koja označava jedinicu vremenske mere. Pošto se radi o kalendarskom datumu, najuža granularnost iskazivanja vremena je na nivou dana. Za ove potrebe koristi se podgraf *ListAll*, modifikovan u tom smislu da obuhvata jedinice vremenske mere u genitivu jednine (npr. *godine*, *nedelje*, *veka*), u čijem se levom kontekstu obavezno nalazi modifikator (npr. *krajem godine*, *prvoj polovini godine*, *sredinom meseca*, *krajem decenije*).

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Datum*

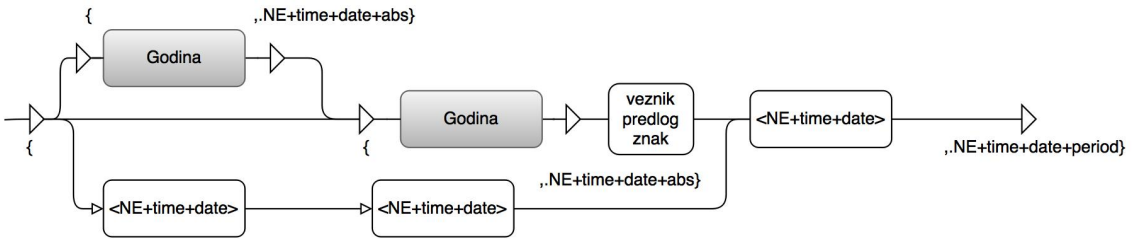
skraćeni nalaze se u primeru A.3, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.3.

4. Period datuma

Prethodno korišćeni podgrafovi - *Dan*, *Mesec*, *Godina*, *MesecSlovima*, *ImeDana*

Novi podgrafovi – osim već korišćenih podgrafova, u okviru ovog transduktora se još koriste leksičke maske kojima su obeleženi već prepoznati vremenski izrazi

Pomoću transduktora *Period datuma* prepoznaju se složeni vremenski izrazi sastavljeni od kalendarskih datuma, koji označavaju period nekog vremena ili predstavljaju nizanje više vremenskih izraza koji su u određenom odnosu (slika 4.13). Ovde je veoma bitna upotreba leksičkih maski, kojima su obeleženi vremenski izrazi identifikovani u prethodnim koracima.



Slika 4.13: Neke od putanja transduktora *Period datuma*

Za prepoznavanje ove vrste izraza koristi se leksička maska, kojom se identifikuju izrazi obeleženi kao vremenski izraz sa značenjem tačke u vremenu, a koji je dat u obliku kalendarskog datuma (npr. `<NE+time+date>`). Osim za identifikovanje perioda iskazanih dvema kalendarskim tačkama, leksičke maske se koriste i za otkrivanje složenih izraza koji se sastoje od više datuma koji su u nekom međusobnom odnosu (npr. *14. decembra 2007. i 16. januara 2008.*; *14. decembra 2007. ili 16. januara 2008.*; *sa 5. marta na 10. mart iste godine*; *5. aprila, 7. juna i 10. avgusta*).

Dalje, na osnovu već prepoznatih vremenskih izraza, moguće je i otkrivanje novih izraza, čiji kontekst pojavljivanja nije dovoljan za njihovu preciznu identifikaciju. Neki od primera, označeni podebljanim tekstom, dati su u 4.37.

Primer 4.37.
1989*, *1999 *i 2000. godine*
7*, *8. *i 9. aprila prošle godine*

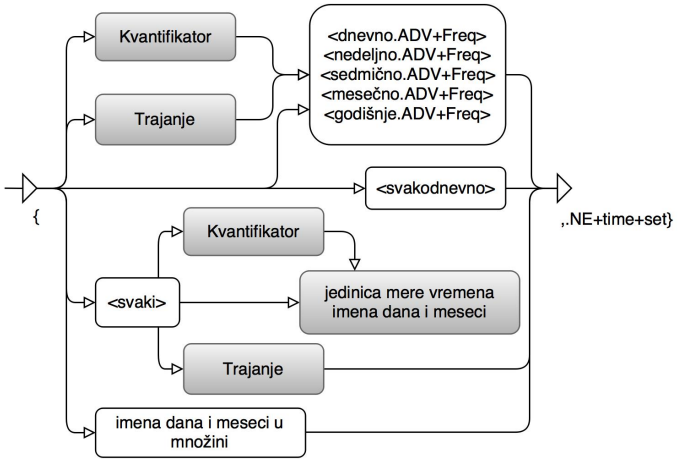
***2007**/2008. godine*
od ***11**. do 22. marta*
***11**. ili 12. februara 2005.*
između *petka **7.** i subote 8. novembra 2011.*

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Period datuma* nalaze se u primeru A.4, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.4.

5. Učestalost

Prethodno korišćeni podgrafovi - *MesecSlovima*, *ImeDana*, *ListAll*
Novi podgrafovi – podgraf *Trajanje*

Peti u nizu transduktora prepoznaje vremenske izraze koji ukazuju na učestalost pojavljivanja u vremenu (slika 4.14).



Slika 4.14: Neke od putanja podgrafova *Povremeno.grf* i *Regularno.grf* koje poziva transduktor *Učestalost*

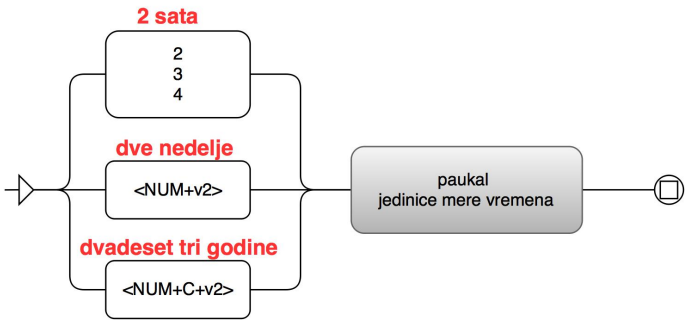
Za identifikaciju izraza koji ukazuju na povremena ponavljanja koristi se podgraf *Povremeno* koji izdvaja i obeležava izraze koji se sastoje od numeričkih kvantifikatora (za to se koristi leksička maska *<NUM>* koja prepoznaje brojeve zapisane na bilo koji način) ili konstrukcije od priloga *jednom*, *dvaput*, *triput* i reči *puta*, sa prilozima učestalosti (npr. *dnevno*, *nedeljno*, *sedmično*) (primer 4.38).

Primer 4.38.
dva puta mesečno
triput nedeljno

deset puta godišnje
3 puta dnevno

Osim toga, uz priloge učestalosti se po prvi put u okviru kaskade koristi i podgraf koji će se koristiti i kasnije za otkrivanje izraza koji ukazuju na trajanje - podgraf *Trajanje* (slika 4.15). Podgraf *Trajanje* prepoznaje imeničke sintagme koje se sastoje od broja (napisanog ciframa ili slovima) (podgraf *Kvantifikator*) i jedinice vremenske mere, ovoga puta svih postojećih granularnosti (dakle, ne samo do nivoa dana, kao što je to bilo u slučaju kalendarskih datuma, već i one uže granularnosti). Delovi izraza sa značenjem učestalosti, koji su prepoznati podgrafom Trajanje, u primeru 4.39 su obeleženi podebljanim tekstom.

Primer 4.39.
11 sati *dnevno*
dva dana *nedeljno*
pet noći *mesečno*
3 meseca *godišnje*



Slika 4.15: Neke od putanja podgrafa *Trajanje*, koji je zadužen za prepoznavanje kvantifikovanih jedinica mere vremena

Osim navedenih izraza učestalosti, u okviru podgrafa *Povremeno* moguće je prepoznavanje i onih oblika izraza, koji se sastoje samo od priloga učestalosti, a označavaju regularna ponavljanja (npr. *dnevno*, *nedeljno*, *sedmično*, *mesečno*, *godišnje*).

Za prepoznavanje vremenskih izraza koji ukazuju na regularna ponavljanja koristi se podgraf *Regularno*. Ovi izrazi se sastoje od priloga *svakodnevno* ili pridevske zamenice *svaki* i jedinice vremena. S obzirom na to da jedinica vremena može da bude kvantifikovana, za prepoznavanje i ovih oblika se upotrebljava podgraf *Trajanje* (npr. *svakih jedanaest godina*, *svakih 13 dana*, *svake pete godine*). Ovde su opet

svi nivoi granularnosti iskazivanja vremena uključeni u prepoznavanje. Budući da vremenski izrazi ovoga tipa mogu da sadrže i imena dana i meseci, za njihovo prepoznavanje se koriste podgrafovi upotrebljavani u prethodnim transduktorima (npr. *svakog trećeg utorka*, *svakog drugog prosinca*).

Jedinice mere, kojima se iskazuje period u okviru koga se vrši regularno ponavljanje, nisu uvek kvantifikovane, pa tako iza pridevske zamenice *svaki* može da se nađe samo imenička sintagma, koju čine imenica kao jedinica vremenske mere i pridev, poput *prošli*, *sledeći*, ili sama imenica. Ilustracija nekih primera ovoga tipa data je u 4.40.

Primer 4.40.
svakog sledećeg četvrtka
svaki naredni mesec
svake godine
svakog petka

U okviru ovog podgrafa vrši se i identifikacija i označavanje i onih izraza koji su predstavljeni samo oblikom množine imenica koje označavaju imena dana i meseci (npr. *utorcima*, *aprilima*).

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Učestalost* nalaze se u primeru A.5, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.5.

6. Trajanje modifikovano

Prethodno korišćeni podgrafovi - *Trajanje*, *ListAll*, *Kvantifikator*
Novi podgrafovi –

Ovim transduktorom se započinje prepoznavanje apsolutnih vremenskih izraza koji ukazuju na trajanje, a nisu iskazani u formi kalendarskih datuma (ovi oblici su već prepoznati u okviru transduktora *Datum prošireni*). Pre svega, prepoznaju se oni oblici izraza koji su na neki način modifikovani i u čijem se levom kontekstu nalaze, na primer, prilozi za količinu u obliku pozitiva, ali i komparativa i superlativa *skoro*, *gotovo*, *više*, *manje*, *preko*, *najviše* itd. Za rešavanje ovog zadatka koristi se ponovo podgraf *Trajanje*, koji omogućava prepoznavanje modifikovanih izraza u obliku jedinice vremenske mere svih granularnosti, koja može biti i kvan-

tifikovana. Nekoliko navedenih primera u 4.41 ilustruje oblike koji se prepoznaju transduktorom *Trajanje modifikovano*.

Primer 4.41.
najviše 11 minuta
skoro 2 sata
najmanje 45 godina
gotovo pet i po sati
duže od dva veka
više od godinu i po dana
najviše sat vremena

Transduktor *Trajanje modifikovano* prepoznaje i vremenske izraze sa značenjem trajanja, a koji se sastoje od jedinice vremenske mere, koja može biti kvantifikovana (podgraf *Trajanje*) i prideva ili pridevskih zamenica (podgraf *ADJ_Date*). Ilustracija ovih izraza data je u primeru 4.42.

Primer 4.42.
prošlih 11 minuta
sledeća 2 meseca
ovih 15 nedelja
jučerašnjih pet i po sati
tih desetak noći
sledećih godinu i po dana
zadnjih sat i po vremena

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Trajanje modifikovano* nalaze se u primeru A.6, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.6.

7. Trajanje u kontekstu

Prethodno korišćeni podgrafovi - *Trajanje*, *ListAll*, *Kvantifikator*
Novi podgrafovi –

Transduktor *Trajanje u kontekstu* namenjen je, kao i prethodni transduktor u nizu, prepoznavanju vremenskih izraza sa značenjem trajanja. S obzirom na to da je u prethodnom transduktoru posmatran levi kontekst u okviru koga su određeni

prilozi, koji označavaju približnu količinu jedinice mere vremena ili pridevi i pridevske zamenice, koji ulaze u opseg vremenskog izraza, ovim transduktorom se prepoznaju kraći oblici izraza sa značenjem trajanja na osnovu levog i desnog konteksta, koji sadrži različite vrste reči i izraza, koje neće ući u opseg prepoznatog izraza. Za prepoznavanje ovih izraza korišćen je podgraf *Trajanje*, kao i podgraf *ListAll*. Neki primeri levog i desnog konteksta, na osnovu kojih su izdvojeni vremenski izrazi sa značenjem trajanja, označeni su podebljanim tekstom u primeru 4.43.

Primer 4.43.
***u roku od** 11 minuta*
***trajaće** 2 sata*
***tokom** 45 godina*
*osam godina **zatvora***
*25 godina **staža***
*89 minuta **igre***

Kako bi se izbegla pogrešna prepoznavanja izraza koji ukazuju na trajanje (na primer, izraz *oko 12 sati* može da predstavlja i izraz koji ukazuje na vreme dana, kao i na trajanje), ostali mogući oblici biće identifikovani u kasnijim koracima, nakon prepoznavanja vremenskih izraza koji ukazuju na vreme dana.

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Trajanje u kontekstu* nalaze se u primeru A.7, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.7.

8. Vreme dana

Prethodno korišćeni podgrafovi - *Kvantifikator*
Novi podgrafovi – *Minut*, *Sat*, *advHour*, *Zone*

Transduktorom *Vreme dana* započinje identifikacija vremenskih izraza koji ukazuju na vreme dana.

Jedan od podgrafova koje poziva transduktor *Vreme dana* jeste podgraf *Sat_num*, koji prepoznaje uobičajene forme za iskazivanje vremena dana, date u numeričkom obliku. Vremenski izraz može da sadrži i leksički okidač, poput *sat*, *minut*, *h*, *čas*. Podgraf *Sat* prepoznaje sate, dok podgraf *Minut* služi za prepoznavanje

minuta i sekundi. U opseg vremenskih izraza ovog tipa mogu da uđu i priloške sintagme, poput *kasno sinoć* ili *rano jutros*. S obzirom na to da navedeni prilozi pripadaju izrazima iste granularnosti kao i vreme dana iskazano numeričkim oblikom, biće obeleženi kao jedan vremenski izraz. Za prepoznavanje pomenutih priloških, kao i imeničkih sintagmi koje označavaju vreme dana, zadužen je podgraf *advHour* (npr. noć, jutros, kasno prošle noći). Ako se u vremenskom izrazu pominju vremenske zone, biće prepoznate podgrafom *Zone*. Primeri izraza koji se prepoznaju ovim podgrafom dati su u primeru 4.44.

Primer 4.44.
09:06:38
10:41:06 h
oko 2 sata i 50 minuta po lokalnom vremenu
oko 15 časova
kasno sinoć oko 23 h
oko 6 sati ranom zorom

Vremenski izrazi koji ukazuju na vreme dana, a iskazani su alfabetskim karakterima, prepoznaju se podgrafom *Sat_alfa*. U prvom koraku prepoznavanja ovih izraza obavezan je levi kontekst, koji može da sadrži neki od predloga ili znakova interpunkcije (npr. *od*, *u*, *do*, *za*, *pre*, *oko* i sl.). Primeri izraza koji se prepoznaju ovim podgrafom ilustrovani su primerom 4.45.

Primer 4.45.
dva sata prošle noći
osam i po časova uveče
dva sata i 35 minuta
tri sata i 15 minuta popodne
osam sati uveče
tri sata iza ponoći

Podgraf *PodnePonoć* prepoznaje vremenske izraze koji označavaju delove dana i koji se sastoje od imenica *podne* i *ponoć*, ispred kojih se mogu naći predlozi, poput *u podne*, *od ponoći*, *pre ponoći*, *oko podneva* (primer 4.46). I u okviru ovog grafa podgraf *Zone* beleži i vremensku zonu ako je pomenuta.

Primer 4.46.
podne
minut do ponoći
pet minuta do podneva
25 minuta posle ponoći
pola sata iza ponoći

Osim vremena dana kao tačke u vremenu, transduktorom *Vreme dana* se prepoznaju i ovi oblici koji nose značenje trajanja. Kao i u slučaju kalendarskih datuma sa značenjem trajanja, koristi se odgovarajući levi kontekst. Na primer, različiti predlozi formirani od imenice *tok* ispred vremenskog izraza koji ukazuje na vreme dana omogućavaju nam otkrivanje vremenskih izraza sa značenjem trajanja (npr. tokom *prošle noći*, tokom *ovog popodneva*, u toku *večeri*).

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Vreme dana* nalaze se u primeru A.8, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.8.

9. Vreme dana u kontekstu

Prethodno korišćeni podgrafovi - *Minut*, *Sat*, *advHour*, *Zone*
Novi podgrafovi – *Sat_bezH*

U okviru ovog transduktora prepoznaju se apsolutna vremena dana, ali uvek sa nekim obaveznim levim kontekstom. Prepoznaju se izrazi koji se sastoje od podataka o satu i minutu, ispred kojih se mogu naći neki od predloga (npr. *u*, *od*, *do*, *za*, *pre*, *posle*). U opseg ovih izraza mogu da uđu i imeničke i priloške sintagme, kao i izrazi koji ukazuju na vremensku zonu.

U ostalim situacijama se prepoznaju i izrazi koji sadrže podatak o satu, bilo da je iskazan alfabetskim ili numeričkim karakterima, iza koga se ne nalazi leksički okidač *sat*, *čas* ili *h*. S tim ciljem neophodno je korišćenje i levog i desnog konteksta vremenskog izraza. Primeri izraza koji se prepoznaju ovim podgrafom dati su u 4.47.

Primer 4.47.
oko 15.35
12:45

11 i 15
osam sinoć
7:30 jutros

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Vreme dana u kontekstu* nalaze se u primeru A.9, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.9.

10. Vreme dana skraćeno

Prethodno korišćeni podgrafovi - *Sat*, *advHour*, *Zone*, upotreba već prepoznatih vremenskih izraza, odnosno leksičkih maski za identifikaciju već prepoznatih izraza

Novi podgrafovi –

U okviru ovog podgrafa se pomoću prethodno postavljenih etiketa otkrivaju novi izrazi koji su iskazani na nivou iste granulacije (npr. *kasno sinoć u 22 h*). Osim toga, već prepoznati vremenski izrazi se koriste i za otkrivanje onih koje nije bilo moguće pouzdano otkriti u prethodnim koracima (npr. *u **pet** umesto u osam časova*).

Transduktor *Vreme dana skraćeno* se istovremeno koristi i za otkrivanje relativnih izraza, kao što su oni koji se sastoje od jedinica mere vremena granulacije uže od dana ili priloga i priloških sintagmi (npr. *kasno sinoć*, *rano prošlog jutra*, *noći*, *jutros*, *večeras*).

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Vreme dana skraćeno* nalaze se u primeru A.10, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.10.

11. Period vremena dana

Prethodno korišćeni podgrafovi - *Sat*, *advHour*, upotreba već prepoznatih vremenskih izraza, odnosno leksičkih maski za identifikaciju već prepoznatih izraza

Novi podgrafovi –

U okviru transduktora *Period vremena dana* koriste se prethodno obeleženi izrazi za otkrivanje onih koji nisu mogli da budu pouzdano prepoznati u okviru

ranijih faza, kao i neki složeniji izrazi koji označavaju period vremena, odnosno sastoje se od dva ili više izraza, koji su u nekom međusobnom odnosu. Primeri izraza koji se prepoznaju ovim podgrafom su u 4.48 obeleženi podebljanim tekstom.

Primer 4.48.
***8** i 9 časova jutros*
***sedam** do 18 časova*
***sedam** do podneva*
***10** do 18 sati*
8 časova jutros do *17 časova popodne*

Još neki primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Period vremena dana* nalaze se u primeru A.11, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.11.

12. Trajanje osnovno

Prethodno korišćeni podgrafovi - *Trajanje*, *ListAll*
Novi podgrafovi –

Transduktor *Trajanje osnovno* se koristi za prepoznavanje kraćih oblika apsolutnih trajanja. Duži oblici ovih izraza (npr. modifikovani) su već prepoznati u prethodnim transduktorima kaskade, te je omogućeno pouzdanije prepoznavanje i kraćih oblika ovih izraza. Za identifikaciju ovih izraza koriste se isti podgrafovi koji su korišćeni u okviru transduktora *Trajanje modifikovano* i *Trajanje u kontekstu*, samo uz drugačije definisan levi kontekst (npr. *15 dana*, *dve i po godine*). Na samom kraju se prepoznaju apsolutni izrazi koji ukazuju na trajanje i koji su iskazani samo kvantifikovanom jedinicom mere vremena (primer 4.49).

Primer 4.49.
tri i po godine
2 sata i 53 minuta
nedelju i po dana
godinu dan
11 meseci
pedesetak minuta
pola godine
godinu kasnije

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Trajanje osnovno* nalaze se u primeru A.12, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.12.

13. Trajanje relativno

Prethodno korišćeni podgrafovi - *ListAll*, *advHour*, *ADJ_Date*, upotreba već prepoznatih vremenskih izraza, odnosno leksičkih maski za identifikaciju već prepoznatih izraza

Novi podgrafovi –

Transduktor *Trajanje relativno* je zadužen za prepoznavanje relativnih vremenskih izraza koji ukazuju na trajanje, a koji se sastoje od množine jedinica mere vremena kvantifikovane neparametričkim kvantifikatorima (npr. *nekoliko*, *par*). Rezultat primene ovog transduktora su izrazi poput *proteklih nekoliko nedelja*, *sledećih par dana*, *više nedelja*. U grupu izraza koji su identifikovani ovim transduktorom takođe spadaju izrazi koje čini samo neki od oblika množine jedinice mere vremena, kao što je to slučaj u izrazima *već* ***nedeljama***, *nakon godina* i sl.

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Trajanje relativno* nalaze se u primeru A.13, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.13.

14. Datum relativni

Prethodno korišćeni podgrafovi - *ListAll*, *ImeDana*, *ADJ_Date*

Novi podgrafovi –

Transduktor *Datum relativni* prepoznaje kraće oblike relativnih datuma, iskazanih nekim oblikom jedinice mere vremena u jednini ili vremenskim prilogom (npr. *prošle godine*, *prošli petak*, *sledeći mesec*, *danas*, *juče*, *ove zime*, *petak*, *prepodne*).

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Datum relativni* nalaze se u primeru A.14, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.14.

15. Period trajanja

Prethodno korišćeni podgrafovi - *ListAll*, *Kvantifikator*, upotreba već prepoznatih vremenskih izraza, odnosno leksičkih maski za identifikaciju već prepoznatih izraza

Novi podgrafovi –

Poput transduktora *Period datuma* i *Period vremena dana*, transduktor *Period trajanja* je, pre svega, zadužen za otkrivanje izraza koji ukazuju na trajanje, a nalaze se u određenom odnosu. S tim ciljem se koriste već prepoznati i u prethodnim koracima obeleženi izrazi, koji takođe omogućavaju i pouzdano prepoznavanje izraza sa značenjem trajanja, koji nisu mogli da budu prepoznati u prethodnim koracima s visokom preciznošću. U okviru primera izrazi identifikovani u ovom koraku su obeleženi podebljanim tekstom.

Primer 4.50.
*dan-**dva***
*godinu-**dve***
***dva**-tri sata*
8 do *10 godina*
tri dana i *tri noći*

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Period trajanja* nalaze se u primeru A.15, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.15.

16. Vreme

Prethodno korišćeni podgrafovi - upotreba već prepoznatih vremenskih izraza, odnosno leksičkih maski za identifikaciju već prepoznatih izraz

Novi podgrafovi –

U poslednjem koraku, transduktorom *Vreme* vrši se prepoznavanje složenih vremenskih izraza, koji se sastoje od kalendarskog datuma i vremena dana, i koji bi na osnovu korišćenog TimeML uputstva trebalo da budu obeleženi kao jedan vremenski izraz sa značenjem vremena dana. Za identifikaciju ove vrste izraza koriste se već prepoznati i obeleženi vremenski izrazi (primer 4.51).

Primer 4.51.
1. marta 2005. oko 9.45 h
16 časova, 2. marta
14:35 h, od utorka do petka
17. jula, između 14 i 15 sati
1. do 5. septembra između 14-16 časova

Primeri izraza obeleženih leksičkim etiketama pomoću transduktora *Vreme* nalaze se u primeru A.16, dok su isti ovi izrazi obeleženi XML etiketama dati u primeru B.16.

Glava 5

Normalizacija vremenskih izraza

Proces anotiranja, odnosno obeležavanja vremenskih izraza uključuje dve faze, i to: fazu identifikacije vremenskih izraza u tekstu i fazu normalizacije njihovih vrednosti. U okviru četvrtog poglavlja detaljno je opisana prva faza čiji je rezultat niz identifikovanih vremenskih izraza, predstavljenih precizno utvrđenom strukturom koja nosi i informacije o značenju sadržaja prepoznatog vremenskog izraza. Informacije ekstrahovane u fazi identifikacije vremenskih izraza upotrebljavaju se u fazi normalizacije radi pronalaženja vrednosti koja je označena određenim izrazom (eksplicitno ili implicitno) i koja može da zavisi bilo od unutrašnjeg semantičkog značenja samog izraza ili, pak, konteksta u kome se nalazi.

Ovo poglavlje usmereno je na detaljan opis procesa normalizacije, kao druge faze automatskog obeležavanja vremenskih izraza. Normalizacija, odnosno temporalna rezolucija u smislu određivanja odseka na vremenskoj osi kome izraz pripada, predstavlja proces koji se sprovodi kako bi se otkrile konačne vrednosti atributa pripisanih nekom vremenskom izrazu. Ovi atributi zavise, pre svega, od sheme za obeležavanje vremenskih izraza koja je upotrebljena. Tokom procesa normalizacije vrednosti atributa mogu biti ekstrahovane iz samog izraza ili sračunate pomoću vrednosti atributa drugih vremenskih izraza koji služe kao orijentiri.

Rezultat normalizacije vremenskih izraza vidljiv je kroz različite upotrebljene atribute koji opisuju dati vremenski izraz u skladu sa izabranom shemom za obeležavanje vremenskih izraza. Za potrebe ovog istraživanja usvojena je TIMEX3 shema za obeležavanje vremenskih izraza, koja je deo TimeML standarda.

5.1 Proces normalizacije vremenskih izraza

Automatsko obeležavanje vremenskih izraza jeste proces koji uključuje prepoznavanje ovih izraza, a potom i interpretaciju njihovih vrednosti, datih u nekom standardizovanom obliku, poput npr. oblika kalendarskih datuma definisanih međunarodnim standardom ISO 8601. Normalizacija, kao zadatak interpretacije vremenskih izraza, usmerena je na utvrđivanje apsolutne vrednosti vremenskog izraza, bez obzira na jezički oblik kojim je iskazan. Tako će dva različita vremenska izraza data u primeru 5.1 biti normalizovana i predstavljena istom vrednošću, budući da se radi o semantički ekvivalentnim izrazima.

Primer 5.1.
(a) *Rođen je **3. aprila 1999. godine.*** → 1999-04-03
(b) *Rođen je **3. IV 1999.*** → 1999-04-03

Složenost ovog zadatka ogleda se u raznovrsnosti jezičkih mogućnosti za izražavanje vremena, ali i kroz činjenicu da postoje i oni vremenski izrazi čija precizna interpretacija zavisi prevashodno od odlika datih u kontekstu ili samom jeziku. Za interpretaciju apsolutne vrednosti izraza datog u primeru 5.2a potrebno je utvrditi vreme kreiranja dokumenta, kao i upotrebljeno glagolsko vreme predikacije rečenice kojoj izraz pripada. U drugom slučaju (primer 5.2b) za razrešenje normalizovane vrednosti nekog vremenskog izraza može poslužiti neki prethodno pomenuti apsolutni vremenski izraz (ako postoji).

Primer 5.2.
(a) *utorak*
(b) *dva dana kasnije*

Razumevanje vremenskih izraza veoma je značajno za procese poput precizne analize diskursa, ali i drugih programa iz oblasti automatske obrade jezika, kao što su odgovaranje na pitanja (Saquete et al. 2009), sumarizacija teksta (Daniel, Radev, and Allison 2003) ili pronalaženje informacija (Alonso, Gertz, and Baeza-Yates 2007). Postizanje ovog cilja dugogodišnji je problem kojim su se bavili razni istraživači za više prirodnih jezika (Mani and Wilson 2000; Verhagen et al. 2010). Dok zadatak prepoznavanja vremenskih izraza veoma uspešno može biti rešen tehnikama zasnovanim na mašinskom učenju (Llorens, Saquete, and Navarro-Colorado 2013), bilo koji pristup normalizaciji vrednosti vremenskih izraza zahteva primenu ručno pisanih pravila (Llorens et al. 2012). Dosadašnji pristupi za normalizaciju

vremenskih izraza koristili su pravila kojima je preciznost sistema u razrešavanju apsolutnih vrednosti dostizala 60-90%. Prvi značajan sistem koji je postavio temelj procesa normalizacije jeste TempEx sistem (Mani and Wilson 2000), o kojem je bilo više reči u trećem poglavlju, a koji je kasnije razrađen i uključen u razvoj GU-Time sistema (Verhagen et al. 2005). U okviru ovog ranog pristupa definisana je i prva shema za obeležavanje vremenskih izraza, iz koje su bili isključeni izrazi koji ukazuju na trajanje, kao i relativni vremenski izrazi. Sistem je vršio normalizaciju i eksplicitnih i implicitnih apsolutnih vremenskih izraza, s tim da se razrešavanje vrednosti implicitnih vremenskih izraza obavljalo pomoću vremena kreiranja dokumenta ili prethodno pomenutog vremenskog izraza. Tokom evaluacije TERN 2004. godine, po uspešnosti u obavljanju zadatka normalizacije vremenskih izraza izdvojio se sitem Chronos (Negri and Marseglia 2005). Autori su za obeležavanje vremenskih izraza i normalizaciju njihovih vrednosti usvojili TIMEX2 shemu, razlikujući apsolutne od relativnih vremenskih izraza. TERSEO sistem (Saquete, Muñoz, and Martínez-Barco 2006) je još jedan od sistema koji za obeležavanje vremenskih izraza koristi TIMEX2 shemu i u okviru kog je razvijena metodologija za automatsko proširenje pravila normalizacije koja se mogu primeniti na više jezika. Još neki značajniji sistemi iz ovog perioda, poput DANTE (Mazur and Dale 2007) i TimexTag (Ahn, Fissaha Adafre, and De Rijke 2005) sistema, detaljnije su predstavljeni u trećem poglavlju. TimeML shema za obeležavanje vremenskih izraza postaje standard u vreme organizovanja TempEval-2 izazova, kada su najuspešniji sistemi u sprovođenju zadatka normalizacije bili zasnovani takođe na pravilima.

5.2 Format za obeležavanje atributa značajnih za proces normalizacije

Normalizacija vremenskih izraza je proces koji se vrši radi identifikovanja vrednosti atributa koji se dodaju svakom prepoznatom vremenskom izrazu i koji će omogućiti utvrđivanje apsolutne vrednosti vremenskog izraza. U skladu sa TimeML uputstvom, u prethodnoj fazi prepoznavanja svaki identifikovan vremenski izraz srpskog jezika obeležen je umetanjem `<TIMEX3>` etikete, u okviru koje su navedeni atributi `type` i `temporalFunction`, s ciljem definisanja semantičke klase kojoj izraz pripada, kao i informacije o tome da li je njegova vrednost eksplicitno iskazana samim izrazom ili ju je potrebno naknadno izračunati (primer 5.3).

Primer 5.3.
(a) `<TIMEX3 type="DATE" temporalFunction="false">12.09.2014.</TIMEX3>`
(b) `<TIMEX3 type="DATE" temporalFunction="true">juče</TIMEX3>`

U okviru faze normalizacije, jedan ili više atributa definisanih TimeML uputstvom, koji se direktno odnose na konačnu interpretaciju vrednosti identifikovanih vremenskih izraza, biće uključeni u `<TIMEX3>` etiketu. Svi mogući atributi određeni TimeML shemom, uključujući i dva opisana u četvrtom poglavlju (`type` i `temporalFunction`), dati su u tabeli 5.1, kao i njihove funkcije i primeri upotrebe.

Tabela 5.1: TIMEX3 atributi

Atribut	**Funkcija**	**Primer**
`tid`	Jedinstveni identifikacioni broj pridružen svakom vremenskom izrazu	`tid="t0"`
`type`	Tip vremenskog izraza: DATE, TIME, DURATION, SET	`type="DATE"`
`value`	Normalizovani oblik vrednosti vremenskog izraza	`value="P3Y"`
`mod`	Reči koje kvantifikuju ili modifikuju značenje vremenskog izraza	`mod="APPROX"`
`temporalFunction`	Binarni atribut koji ukazuje na to da li je potrebno na neki način izračunati apsolutnu vrednost vremenskog izraza	`temporalFunction="true"`
`anchorTimeID`	Identifikacioni broj vremenskog izraza koji će služiti kao orijentir u računanju vrednosti drugog vremenskog izraza	`anchorTimeID="t1"`

Atribut	Funkcija	Primer
`valueFromFunction`	Opcioni atribut relevantan za potrebe izračunavanja normalizovane vrednosti	`valueFromFunction="-1Y"`
`functionInDocument`	Funkcija vremenskog izraza u dokumentu (npr. vreme kreiranja dokumenta)	`functionInDocument="NONE"`
`beginPoint`, `endPoint`	Identifikacioni broj vremenskog izraza koji predstavlja početnu ili završnu tačku nekog izraza koji ukazuje na trajanje koje nije eksplicitno dato u tekstu	`beginPoint="t1"` `endPoint="t2"`
`quant`	Reč kojom je vremenski izraz koji ukazuje na učestalost kvantifikovan	`quant="EVERY"`
`freq`	Regularna učestalost ponavljanja u nekom periodu vremena	`freq="2D"`

Svaki od atributa prikazanih u okviru tabele 5.1 može sadržati samo one vrednosti koje su precizno određene TimeML shemom. U daljem tekstu će detaljno biti opisani svi oni atributi koji se u okviru ovog istraživanja tokom faze normalizacije dodaju prepoznatim vremenskim izrazima i koriste za interpretaciju njihovih vrednosti.

5.2.1 Atribut `value`

Atribut `value` je obavezni atribut, kojim se reprezentuju konačne vrednosti vremenskih izraza u standardizovanom obliku. TimeML shema se u velikoj meri oslanja na međunarodni standard za beleženje kalendarskih datuma, vremena dana i perioda vremena ISO 8601. Za reprezentaciju vremenskih izraza iskazanih na nivou različitih granularnosti (o čemu je bilo više reči u prethodnom poglavlju) standard ISO 8601 predviđa upotrebu određenih formi (npr. kalendarski datum

granularnosti meseca *aprila 2007.* biće predstavljen oblikom *2007-04*) i kodova (npr. za jedinicu mere *sedmica* predviđena je upotreba koda W, što potiče od engleske imenice *week*). S obzirom na to da postoje i primeri jedinica mere vremena čija reprezentacija nije standardizovana pomoću pomenutog standarda (npr. milenijum), za potrebe ovog rada prihvaćeni su oni oblici i kodovi predloženi TimeML shemom, što će biti objašnjeno u daljem tekstu.

Oblik normalizovane vrednosti vremenskog izraza kojim će biti iskazan atribut `value` zavisi od semantičke klase kojoj prepoznati vremenski izraz pripada. Iz tog razloga neophodno je napraviti razliku u interpretaciji vremenskih izraza koji impliciraju tačku u vremenu, izraza koji ukazuju na trajanje tj. interval vremena, kao i izraza kojima je iskazana učestalost ponavljanja u intervalima.

Oblik atributa `value` u reprezentaciji vremenskih izraza koji ukazuju na tačku u vremenu

U slučaju vremenskih izraza koji impliciraju tačku u vremenu, kao što je već opisano u prethodnom poglavlju, razlikujemo kalendarske datume i doba dana.

Kalendarski datumi, kao apsolutni vremenski izrazi, mogu biti iskazani kompletnom formom, kao što je oblik YYYY-MM-DD, gde je Y oznaka za godinu iskazanu četvorocifrenim brojem, M oznaka za mesec iskazan dvocifrenim brojem meseca u godini i D oznaka za dan u mesecu, takođe kao dvocifreni broj. U funkciji separatora, standard ISO 8601 predviđa upotrebu crtice. Dakle, datum *25. decembar 2011. godine* biće predstavljen u obliku datom u primeru 5.4.

Primer 5.4.
25. decembar 2011. godine → YYYY-MM-DD → `value="2011-12-25"`

Nekompletni kalendarski datumi, koji takođe spadaju u grupu apsolutnih vremenskih izraza koji impliciraju tačku u vremenu, i kod kojih je navedena samo godina ili godina i mesec, biće predstavljeni u vidu forme smanjene preciznosti, kao što je ilustrovano primerom 5.5.

Primer 5.5.

2011. godine	YYYY	`value="2011"`
decembra 2011.	YYYY-MM	`value="2011-12"`

Kalendarski datumi takođe mogu biti iskazani i na nivou šire granularnosti od dana, čija forma nije prilagođena standardu ISO 8601. Stoga, TimeML shema predviđa sledeći način reprezentovanja ovih izraza. S obzirom na to da je među-

narodnim standardom godina u osnovi predstavljena ustaljenim oblikom koji se sastoji od četiri komponente (YYYY), izrazi na nivou granularnosti decenija, vekova ili milenijuma biće predstavljeni skraćivanjem forme za godinu. Tako se prva pozicija smatra za podatak o milenijumu, druga za podatak o veku, dok se treća odnosi na decenije. Četvrta komponenta je rezervisana za reprezentaciju godine, kao što je već opisano. Dakle, vrednost atributa value u slučaju vremenskih izraza šire granularnosti može da se sastoji od samo jednog (primer 5.6a), dva (primer 5.6b) ili tri broja (primer 5.6c).

Primer 5.6.

(a)	*3. milenijum*	Y	`value="3"`
(b)	*XIX vek*	YY	`value="19"`
(c)	*osamdesete*	YYY	`value="198"`

Vremenski izrazi koji ukazuju na neko godišnje doba spadaju u grupu izraza koji mogu imati različita značenja za različite ljude. Na primer, zima za nekoga može da predstavlja hladni deo godine, dok bi isto tako u nekim situacijama mogla da bude interpretirana i kao period godine koji počinje zimskom kratkodnevnicom, a završava se prolećnom ravnodnevnicom. Iz tog razloga se u okviru atributa value za standardizovanu reprezentaciju ovih izraza, umesto numeričkih vrednosti (primer 5.7), koriste posebni alfabetski kodovi predstavljeni tabelom 5.2.

Primer 5.7.

proleća 2009. → YYYY-SP → `value="2009-SP"`

Tabela 5.2: Kodovi za reprezentaciju godišnjih doba

Godišnje doba	**Kod za reprezentaciju**
jesen	FA (eng. *fall*)
zima	WI (eng. *winter*)
proleće	SP (eng. *spring*)
leto	SU (eng. *summer*)

Iako se ređe javljaju u svakodnevnom govoru, i izrazi koji upućuju na određene nedelje u godini ili vikende u nedelji takođe su zastupljeni u okviru međunarodnog standarda, koji omogućava reprezentacije ilustrovane primerom 5.8. Da bi se u potpunosti predstavilo značenje koje stoji iza izraza vikend, potrebno je prvo utvrditi kojoj nedelji u godini pripada taj vikend. Zatim je potrebno na poziciju

dana u ISO formatu datuma postaviti token WE.

Primer 5.8.

(a)	*druge nedelje 2005.*	YYYY-Wn	`value="2005-W2"`
(b)	*vikend 28. nedelje 1999. godine*	YYYY-Wn-WE	`value="1999-W28-WE"`

U prethodno opisanim slučajevima reprezentacija umanjene preciznosti vremenskih izraza šire granularnosti (npr. milenijum, vek, decenija) predviđeno je izostavljanje vrednosti, odnosno podataka koji se nalaze sa desne strane. Isto tako, standardom je omogućeno i skraćivanje vrednosti kalendarskih datuma s leva (eng. truncation). Kako bi u tom slučaju reprezentacije predstavljenih vrednosti bile nedvosmislene, upotrebljava se karakter X, za svaku nepopunjenu poziciju kompletnog kalendarskog datuma s leve strane. U slučaju relativnih vremenskih izraza koji impliciraju tačku u vremenu, a odnose se na kalendarski datum iz koga je izostavljen podatak o godini (primeri 5.9a i 5.9b), ili godini i mesecu (primer 5.9c), biće upotrebljen karakter X za reprezentovanje nepoznatih vrednosti. Tako će izrazi ovoga tipa biti normalizovani kao što je dato u primeru 5.9.

Primer 5.9.

(a)	*22. novembra*	YYYY-MM-DD	`value="XXXX-11-22"`
(b)	*juna meseca*	YYYY-MM-DD	`value="XXXX-06"`
(c)	*petog u mesecu*	YYYY-MM-DD	`value="XXXX-XX-05"`

S obzirom na to da izostavljena vrednost predstavljena karakterom X ukazuje na nepoznatu ili nedovoljno preciznu vrednost, isti princip biće upotrebljen i za reprezentaciju relativnih vremenskih izraza čije se apsolutne vrednosti mogu računati na osnovu drugog vremenskog izraza kao orijentira. Različiti primeri ovog tipa izraza i oblika njihove normalizovane interpretacije ilustrovani su primerom 5.10.

Primer 5.10.

sledeći milenijum	YYYY-MM-DD	`value="X"`
prošlom veku	YYYY-MM-DD	`value="XX"`
ove decenije	YYYY-MM-DD	`value="XXX"`
sledeće godine	YYYY-MM-DD	`value="XXXX"`
22. novembra prošle godine	YYYY-MM-DD	`value="XXXX-11-22"`
novembra sledeće godine	YYYY-MM-DD	`value="XXXX-11"`
ovog vikenda	YYYY-MM-DD	`value="XXXX-WXX-WE"`
petak	YYYY-MM-DD	`value="XXXX-WXX-05"`*
juče	YYYY-MM-DD	`value="XXXX-XX-XX"`

* Za reprezentaciju dana u nedelji koriste se redni brojevi tih dana u nedelji.

Izrazi koji ukazuju na temporalnu lokaciju u vidu kalendarskog vremena biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa value biti predstavljena kao što je dato u primeru 5.11.

Primer 5.11.

Gospodin Petrović je otišao	*u* ***petak 14. aprila 2007. godine***	`2007-04-14`
	drugog decembra	`XXXX-12-02`
	juče	`XXXX-XX-XX`
	oktobra 2002.	`2002-10`
	leta 2005.	`2005-SU`
	u ***utorak 18-tog***	`XXXX-XX-18`*
	leta ove godine	`XXXX-SU`
	prošle nedelje	`XXXX-WXX`

* S obzirom na to da je izrazom dat redni broj dana u mesecu, ime dana (npr. *utorak*) se ne upotrebljava za normalizovanu reprezentaciju.

```
Petar je otišao <TIMEX3 type="DATE" temporalFunction="true"
   value="XXXX-XX-XX">juče</TIMEX3>.
```

U klasi vremenskih izraza koji ukazuju na tačku u vremenu nalaze se i izrazi koji ukazuju na **vreme dana**. Oblik njihove normalizovane vrednosti, koja treba da bude sadržaj atributa `value`, takođe je predviđen međunarodnim standardom ISO 8601. Ispred vrednosti vremenskog izraza ovog tipa unosi se karakter T (skr. od eng. ***time***), dok se u funkciji separatora radi odvajanja podataka o satu, minutu i sekundi, koristi dvotačka. Tako će kompletna reprezentacija apsolutnog izraza koji ukazuje na vreme dana biti u obliku hh:mm:ss (primer 5.12).

Primer 5.12.
14 časova, pet minuta i 35 sekundi → Thh:mm:ss → `value="T14:05:35"`

Kao i u slučaju kalendarskih datuma, kada je reč o nekompletnim izrazima omogućena je reprezentacija umanjene preciznosti izraza (primer 5.13).

Primer 5.13.

osam i trideset uveče	Thh:mm	`value="T20:30"`
ponoć	Thh	`value=" T24"`

S obzirom na to da relativni izrazi koji ukazuju na vreme dana, poput *jutro*, *popodne*, *noć* i sl. predstavljaju periode dana koji takođe mogu biti predmet individualne interpretacije, za reprezentaciju njihovih vrednosti TimeML shema predviđa upotrebu određenih tokena, odnosno alfabetskih karaktera na poziciji sata u okviru standardizovane vrednosti izraza. Kodovi koji se koriste za reprezentaciju ovih izraza dati su u tabeli 5.3, dok je njihova primena ilustrovana primerom 5.14.

Tabela 5.3: Kodovi za reprezentaciju relativnih izraza koji ukazuju na vreme dana

Vreme dana	**Kod za reprezentaciju**
jutro, jutros, zora	MO (eng. *morning*)
dan	MI (eng. *mid-day*)
popodne	AF (eng. *afternoon*)
uveče, večeras	EV (eng. *evening*)
noć, noćas	NI (eng. *night*)
dan (svetli deo dana)	DT (eng. *daytime*)

Primer 5.14.

jutros	Thh	`value="TMO"`
noćas	Thh	`value="TNI"`

Upotreba ovih tokena je ograničena samo na one izraze u okviru kojih nije navedeno precizno vreme dana. Na primer, u izrazu *u **jedanaest ujutru***, vrednost atributa `value` će jednostavno biti T11, i *ujutru* neće biti anotirano odvojeno jer se radi o izrazu iste granularnosti kao i dati apsolutni izraz (primer 5.15).

Primer 5.15.
jedanaest časova uveče → Thh → `value="T23"`

Kompletna reprezentacija vremenskog izraza na nivou granularnosti sata, minuta ili sekunde može sadržati i podatak o kalendarskom datumu i tada oblik normalizovane vrednosti treba da bude reprezentovan oblikom poput onog datog u primeru 5.16.

Primer 5.16.
15. časova 30. decembra 2014. godine YYYY-MM-DDThh
`value="2014-12-30T15"`

Izrazi koji ukazuju na temporalnu lokaciju u vidu vremena kao dela dana biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa value biti predstavljena kao što je dato u primeru 5.17.

Primer 5.17.

Gospođa Ivanović je stigla	*u* **deset do tri popodne**	`T14:50`
	u **pet minuta do osam**	`T07:55`
	u **12 i dvadeset**	`T12:20`
	pola sata posle ponoći	`T00:30`
	u **jedanaest ujutru**	`T11`
	u **9 h 14. oktobra 2009. godine**	`2009-10-14T09`
	uveče 11. januara	`XXXX-01-11TEV`
	kasno sinoć	`TEV`
	prošle noći	`TNI`

```
Petar je otišao u <TIMEX3 type="TIME" value="T18">18 časova</TIMEX3>.
```

Oblik atributa `value` u reprezentaciji vremenskih izraza koji ukazuju na trajanje

Kada je reč o vremenskim izrazima koji impliciraju trajanje, odnosno izrazima koji opisuju interval ili period vremena, ukazujući eksplicitno na to koliko dugo je nešto trajalo (npr. ***tri sata*** *dugo predavanje*), njihova vrednost treba da bude izražena atributom value u obliku koji je naveden u standardu ISO 8601: PnYnMnDnnHnMINnS ili PnW.

Oznaka P (skr. od eng. *period*) koristi se u obeležavanju izraza koji ukazuju na trajanje, odnosno neki period vremena. Za označavanje jedinica mere vremena koriste se slovni kodovi ilustrovani tabelom 5.4, kao i kodovi ranije opisani u ovom poglavlju, dok n označava broj jedinica vremenske mere.

Neki od primera normalizovanih vrednosti apsolutnih vremenskih izraza koji ukazuju na trajanje dati su u primeru 5.18.

Tabela 5.4: Kodovi za reprezentaciju jedinica mere vremena

Jedinica mere vremena	**Kod za reprezentaciju**
era	BC*
milenijum	L**
vek, stoleće	C**
decenija	E**
godina	Y (eng. *year*)
mesec	M (eng. *month*)
nedelja, sedmica	W (eng. *week*)
vikend	WE (eng. *weekend*)
dan	D (eng. *day*)
sat	H (eng. *hour*)
minut	MIN
sekund	S (eng. *second*)

* kodovi predloženi u okviru ovog rada

** kodovi predloženi TimeML uputstvom

Primer 5.18.

20 minuta	PnMIN	`value="P20MIN"`
pola sata	PnMIN	`value="P30MIN"`
9 meseci	PnM	`value="P9M"`
celu deceniju	PnE	`value="P1E"`
dva milenijuma	PnL	`value="P2L"`
dve noći	PnNI	`value="P2NI"`

Osim opisanih apsolutnih izraza koji ukazuju na trajanje i koji treba da budu reprezentovani odgovarajućim oblikom sa oznakom P ispred normalizovane vrednosti, postoje i izrazi koji takođe ukazuju na period vremena koji treba da bude reprezentovan oblikom kalendarskih datuma, što je jasno definisano TimeML uputstvom. S obzirom na to da često kontekst, a ne sâm vremenski izraz, implicira značenje izraza, neophodno je obratiti posebnu pažnju i primere poput izraza *tokom* ***marta meseca 2005.*** obeležiti u skladu sa TimeML uputstvom, koje nalaže da se u situacijama kada su moguće reprezentacije bilo tačka u vremenu (u slučaju navedenog primera atribut `value` bi mogao da ima vrednost `2005-03`) ili trajanje (vrednost atributa `value` bi mogla da bude iskazana i kao `P31D`), upotrebi uvek prvo onu koja implicira tačku u vremenu. Izraz ovoga tipa biće normalizovan u

okviru atributa `value` na način ilustrovan primerom 5.19.

Primer 5.19.

```
Petar je otišao tokom <TIMEX3 type="DURATION" value="2005-03"> marta
   meseca 2005</TIMEX3>.
```

Kad je reč o relativnim vremenskim izrazima koji ukazuju na neodređeno trajanje, upotrebljava se karakter X za označavanje nepoznate kvantifikacije jedinice vremenske mere (primer 5.20).

Primer 5.20.

poslednjih nekoliko godina	PnY	`value="PXY"`
sledećih par dana	PnD	`value="PXD"`
mesecima	PnM	`value="PXM"`
nakon godina	PnY	`value="PXY"`

Vremenski izrazi koji ukazuju na trajanje biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa value biti predstavljena kao što je dato u primeru 5.21.

Primer 5.21.

Gospodin Petrović je boravio na planini	***dva meseca***	`P2M`
	48 sati	`P48H`
	tri nedelje	`P3W`
	ceo mesec	`P1M`
	nekoliko dana	`PXD`

```
Koncert je trajao <TIMEX3 type="DURATION" value="P2H">dva sata</TIMEX3>.
```

Oblik atributa `value` u reprezentaciji vremenskih izraza koji ukazuju na učestalost

Da bi vremenski izrazi koji ukazuju na učestalost ponavljanja bili u potpunosti anotirani, osim atributa `value`, neophodna je i upotreba atributa `quant` ili `freq`, o kojima će biti više reči kasnije u tekstu. Sledećim primerom (5.22) ilustrovan je samo početni deo anotacije ovih izraza, koji se odnosi na vrednost atributa `value`.

Primer 5.22.

(a) *dva puta nedeljno*	PnW	`value="P1W"`
(b) *svaka dva dana*	PnD	`value="P2D"`
(c) *svakog oktobra*	YYYY-MM	`value="XXXX-10"`

U primerima 5.22a i 5.22b vrednost atributa `value` ima oblik vrednosti koja ukazuje na period jer se zapravo i radi o učestalosti ponavljanja u okviru nekog perioda (npr. *nedeljno* ili *dva dana*), koje može biti povremeno (npr. ***dva puta*** *nedeljno*) ili regularno (npr. ***svaka*** *dva dana*). Poslednji primer 5.22c koristi reprezentaciju kalendarskog datuma u okviru atributa `value`, ne bi li se sačuvala kalendarska informacija data sâmim izrazom. U slučaju vremenskih izraza koji ukazuju na učestalost ponavljanja, atribut `value` će uvek imati oblik vrednosti koji se koristi za izraze sa značenjem trajanja, osim ako precizana kalendarska informacija (npr. *oktobar* ili *utorak*) nije data izrazom.

Vremenski izrazi koji ukazuju na temporalnu frekvenciju, odnosno učestalost pojavljivanja u vremenu biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa value biti predstavljena kao što je dato u primeru 5.23.

Primer 5.23.

Gospođa Ivanović vežba	***dva puta nedeljno***	`P1W`
	svaka 2 dana	`P2D`
	svakog petka	`XXXX-WXX-5`
	mesečno	`P1M`

```
Petar pliva <TIMEX3 type="SET" value="XXXX-WXX-5">svakog petka</TIMEX3>.
```

5.2.2 Atribut `mod`

Atribut `mod` je neobavezni atribut, koji se može dodavati samo vremenskim izrazima koji ukazuju na tačke u vremenu ili trajanja. Zadatak ovog atributa jeste prikazivanje značenja vremenskih izraza koji su na neki način kvantifikovani (npr. ***približno*** *2 sata*, ***ne više od*** *10 minuta*) ili modifikovani (npr. ***početkom*** *2005. godine*, ***krajem*** *aprila meseca*). Potrebno je obratiti posebnu pažnju na činjenicu da predlozi koji se nalaze ispred vremenskog izraza ne utiču na njegovo značenje (izraz *pre petka* se ne smatra modifikovanim izrazom), te ne treba da uđu u opseg vremenskog izraza, što je detaljno objašnjeno u prethodnom poglavlju. Tabelom 5.5 su ilustrovane moguće vrednosti atributa `mod`, u zavisnosti od izraza kome će biti dodeljene.

Modifikovani vremenski izrazi koji ukazuju na tačku u vremenu ili trajanje biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `mod` biti predstavljena kao što je dato u primeru 5.24.

Tabela 5.5: Kodovi za reprezentaciju atributa `mod`

Tip vremenskog izraza	Token	Primeri izraza
Trajanje	LESS_THAN	*manje od* dva sata, *skoro* 5 minuta
	MORE_THAN	*više od* 2 sata
Tačka u vremenu i trajanje	START	*početkom* 2005, *prva polovina* godine
	MID	*sredinom* meseca
	END	*krajem* aprila 2006.
	APPROX	*oko* dva sata, *oko* 5. maja

Primer 5.24.

Gospođa Ivanović odlazi	***početkom 2017.***	`value="2017" mod="START"`
	sredinom marta	`value="XXXX-03" mod="MID"`
	krajem zime 2018.	`value="2018-WI" mod="END"`
	oko 28-og	`value="XXXX-XX-28" mod="APPROX"`

```
Petar spava <TIMEX3 type="DURATION" value="P2H" mod="MORE_THAN">više od
dva sata</TIMEX3>.
```

5.2.3 Atribut `valueFromFunction`

Atribut `valueFromFunction` značajan je za kasniji proces računanja apsolutne vrednosti izraza pomoću drugih izraza koji će poslužiti kao orijentir. Vrednost ovog atributa sastoji se od oznake za računsku operaciju koju je potrebno primeniti (plus, minus ili znak jednakosti), kao i za količinu ($n \geq 0$) jedinica vremenske mere koja treba da bude dodata ili oduzeta od vrednosti vremenskog izraza koji je uzet za orijentir. Na primer, apsolutna vrednost izraza datog u primeru 5.25 biće razrešena u odnosu na broj (u ovom slučaju 1), magnitudu, odnosno veličinu koja je izražena jedinicom vremenske mere (godina) i pravac na vremenskoj osi (*prošle* ukazuje na računsku operaciju oduzimanja). Tako će konačna vrednost ovog izraza biti sračunata oduzimanjem jedne godine od vrednosti apsolutnog vremenskog izraza koji služi kao orijentir.

Primer 5.25.

25. marta prošle godine → YYYY-MM-DD →
```
value="XXXX-03-25"
valueFromFunction="-1Y"
```

Relativni vremenski izrazi koji ukazuju na tačku u vremenu ili trajanje biće obeleženi `<TIMEX3>` etiketom, čija će vrednost atributa `valueFromFunction` biti predstavljena kao što je dato u primeru 5.26.

Primer 5.26.

Gospođa Ivanović	*o*dlazi **sledeće godine**	`value="XXXX"`
		`valueFromFunction="+1Y"`
	*j*e otišla **prošlog meseca**	`value="XXXX-XX"`
		`valueFromFunction="-1M"`
	*d*olazi **ove nedelje**	`value="XXXX-WXX"`
		`valueFromFunction="=1W"`
	*s*tiže **idućeg dana**	`value="XXXX-XX-XX"`
		`valueFromFunction="+1D"`

```
Petar je otišao <TIMEX3 type="DATE" temporalFunction="true"
value="XXXX-XX-XX" valueFromFunction="-1D">juče</TIMEX3>.
```

5.2.4 Atributi `quant` i `freq`

Atributi `quant` i `freq` koriste se radi upotpunjavanja obeležavanja vremenskih izraza koji ukazuju na učestalost ponavljanja u vremenu. Vrednost atributa `value` ovog tipa izraza već je opisana u delu 5.2.1. Atribut `quant` se prevashodno koristi za reprezentaciju vremenskih izraza koji ukazuju na regularnu učestalost ponavljanja perioda, a njegova vrednost je iskazana engleskom reči *every* kojim se kvantifikuje vrednost (primer 5.27).

Primer 5.27.

```
<TIMEX3 type="SET" value="P2D" quant="EVERY">svaka 2 dana</TIMEX3>
<TIMEX3 type="SET" value="XXXX-10" quant="EVERY">svakog
   oktobra</TIMEX3>
```

Vrednost atributa `freq` je koliko puta se nešto ponavlja (zadato celim brojem) i u kom vremenskom intervalu (zadato kodom nivoa granularnosti) (primer 5.28).

Primer 5.28.

```
<TIMEX3 type="SET" value="P1W" freq="2X">dva puta nedeljno</TIMEX3>
```

Vremenski izrazi koji ukazuju na učestalost ponavljanja biće obeleženi `<TIMEX3>` etiketom, čije će vrednosti atributa `quant` i `freq` biti predstavljene kao što je dato u primeru 5.29.

Primer 5.29.

Gospodin Petrović pliva	***svakih pet dana***	`value="P5D" quant="EVERY"`
	svakog jula	`value="XXXX-07" quant="EVERY"`
	6 puta mesečno	`value="P1M" freq="6X"`

```
Petar kliza <TIMEX3 type="SET" value="P1W" quant="EVERY" freq="3D">
   3 dana svake nedelje</TIMEX3>
```

Atributi čija je upotreba takođe predviđena TimeML shemom, a koji u ovoj fazi istraživanja neće biti uključeni u obeležavanje vremenskih izraza, su `anchorTimeID`, `functionInDocument`, `beginPoint` i `endPoint`. S obzirom na to da su vrednosti ovih atributa predstavljene u obliku jedinstvenog identifikacionog broja, koji bi trebalo da bude dodeljen svakom prepoznatom vremenskom izrazu (na primer, `anchorTimeID`), ili pak direktno zavise od njega (npr. `functionInDocument`, `beginPoint` i `endPoint`), neophodno je kreirati i primeniti spoljnji program zadužen za automatsku dodelu rednih brojeva, koji će služiti kao jedinstveni identifikator svakog prepoznatog vremenskog izraza.

Opcioni atribut `anchorTimeID` koristi se u opisu relativnih vremenskih izraza, radi upućivanja na vremenski izraz koji će poslužiti kao orijentir prilikom računanja normalizovane vrednosti relativnog vremenskog izraza. Stoga će njegova vrednost uvek biti neki identifikacioni broj, odnosno identifikacioni broj onog vremenskog izraza koji predstavlja orijentirnu tačku.

Još jedan atribut koji ukazuje na referentno vreme, neophodno za računanje vrednosti relativnih vremenskih izraza, jeste atribut `functionInDocument`. Njegova vrednost može označavati npr. vreme kreiranja, izmene ili publikovanja dokumenta.

Upotreba opcionih atributa `beginPoint` i `endPoint` moguća je takođe tek nakon uvođenja identifikacionih brojeva svih vremenskih izraza. Ovi atributi se koriste kako bi se potpomoglo obeležavanje vremenskih izraza koji ukazuju na trajanje u vremenu, čije leva i desna granična tačka odseka na vremenskoj osi mogu biti predstavljene posebnim etiketama, kao što je ilustrovano primerom 5.30.

Primer 5.30.

```
Petar je počeo da predaje <TIMEX3 tid="t1" type="DURATION"
   temporalFunction="false" value="P1W" beginPoint="t2" endPoint="t3">
   nedelju dana</TIMEX3> posle <TIMEX3 tid="t2" type="DATE"
   temporalFunction="true" value="XXXX-09-15">15.  novembra</TIMEX3>
   <TIMEX3 tid="t3" type="DATE" value="XXXX-09-22"
```

```
temporalFunction="true" anchorTimeID="t1"/>.
```

5.3 Lokalna gramatika za normalizaciju vrednosti vremenskih izraza

Sistem zadužen za normalizaciju prepoznatih vremenskih izraza razvijen je takođe putem metode konačnih stanja. Kao i u procesu prepoznavanja vremenskih izraza, za kreiranje kolekcije transduktora kojima se opisuju pravila normalizacije i njihovu primenu korićen je softverski alat UNITEX. Za razliku od procesa prepoznavanja vremenskih izraza, u okviru koga se primena pravila zadatih konačnim transduktorima vrši kaskadno određenim redosledom, pri čemu jedan transduktor koristi rezultate prethodno primenjenih, u fazi normalizacije vremenskih izraza primenjuju se gotovo ista pravila, ali ovoga puta u vidu lokalne gramatike konačnih transduktora, čiji je cilj interpretacija vrednosti prepoznatih izraza. Budući da oblik interpretacije vrednosti vremenskog izraza direktno zavisi od tipa vremenskog izraza, lokalna gramatika namenjena normalizaciji sastoji se od četiri glavna transduktora, od kojih svaki odgovara postojećim semantičkim klasama vremenskih izraza, detaljno opisanim unutar poglavlja 4.2. Stoga gramatiku čine sledeći transduktori:

1. *Datum* (transduktor koji je zadužen za normalizaciju vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju tačku u vremenu i koji su dati u obliku kalendarskog datuma);

2. *Vreme dana* (transduktor koji je zadužen za normalizaciju vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju tačku u vremenu i koji su dati u obliku vremena dana);

3. *Trajanje* (transduktor koji je zadužen za normalizaciju vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju trajanje);

4. *Učestalost* (transduktor koji je zadužen za normalizaciju vrednosti prepoznatih vremenskih izraza koji impliciraju regularnu ili povremenu učestalost ponavljanja).

5.3.1 Transduktori za normalizaciju vrednosti vremenskih izraza koji impliciraju tačku u vremenu

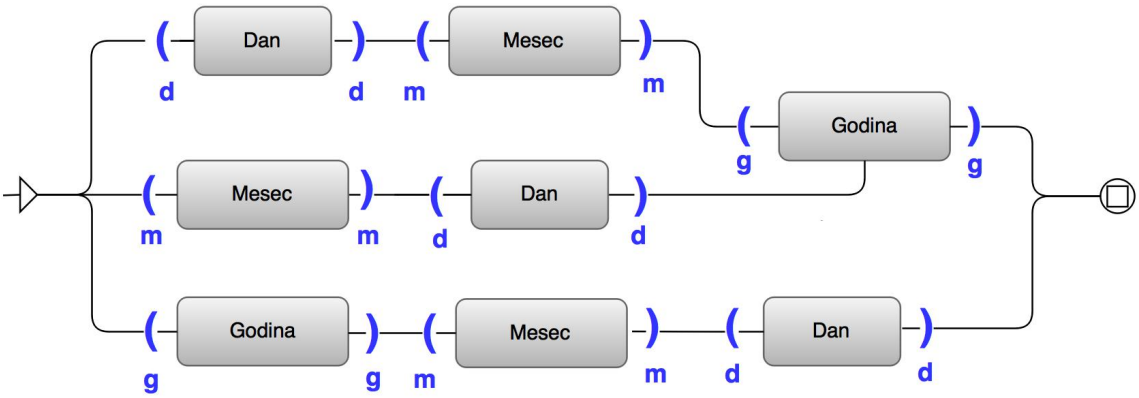
Vremenski izrazi ovoga tipa ukazuju na temporalnu lokaciju, odnosno odsek na vremenskoj osi koji je shvaćen kao tačka u vremenu i koji može biti iskazan određenim nivoom detaljnosti. Budući da se oblik normalizovane vrednosti vremenskih izraza datih u vidu kalendarskih datuma razlikuje od oblika predviđenog za interpretaciju izraza koji ukazuju na vreme dana (detaljno opisano u 5.2.1), kreirana su dva glavna grafa *Datum* i *Vreme dana*.

Transduktori za normalizaciju vrednosti vremenskih izraza datih u vidu kalendarskih datuma

Graf *Datum* poziva dva podgrafa, nazvana *Datum apsolutni* i *Datum relativni*, što oslikava razliku u procesima koje je potrebno sprovesti radi interpretacije vrednosti apsolutnih i relativnih izraza ovoga tipa.

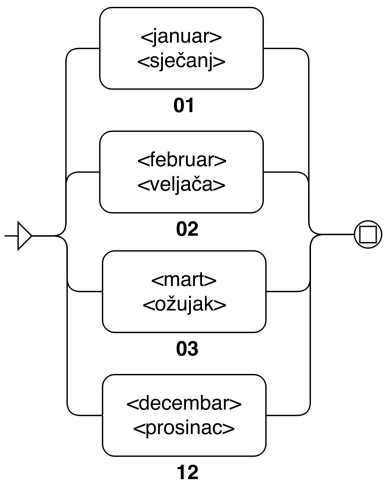
U fazi normalizacije vrednosti apsolutnih vremenskih izraza datih u vidu kalendarskih datuma koriste se ista pravila koja su upotrebljena radi njihove identifikacije u prethodnoj fazi. S obzirom na to da se u ovom koraku koriste već postavljene etikete kojima je precizno definisan tip prepoznatog vremenskog izraza, pravila koja opisuju oblike pojavljivanja preciznih kalendarskih datuma nije neophodno primenjivati određenim redosledom. Tako transduktor *Datum apsolutni* poziva sve one transduktore namenjene prepoznavanju ovog tipa kalendarskih datuma, bez obzira na nivo detaljnosti kojim su iskazani (npr. *II milenijum*, *sedamdesetih*, *2011*, *aprila 1999.* ili *25. mart 2007. godine*). Kako se proces normalizacije odnosi na interpretaciju vrednosti prepoznatih vremenskih izraza, u ovoj fazi je neophodno definisati vrednosti koje predstavljaju osnovne elemente datuma (npr. dan, mesec, godina) i koje će se naći u normalizovanom obliku, kao elementi atributa `value`. Delovi teksta koji odgovaraju osnovnim elementima datuma smeštaju se u izlazne promenljive, u transduktoru definisane pomoću plavih zagrada. Na slici 5.1 ilustrovan je primer definisanja osnovnih elemenata kompletnog kalendarskog datuma kao promenljivih koje će biti upotrebljene u izlazu za interpretaciju normalizovane vrednosti. Vrednosti prepoznate kao element godine smeštaju se u izlaznu promenljivu *g*, dok se vrednosti koje se odnose na mesec i dan smeštaju u izlazne promenljive *m* i *d*.

Korišćenje izlaznih promenljivih omogućava upotrebu izlaza koji proizvodi gramatika. Na primer, u slučaju meseca iskazanog alfabetskim karakterima (npr. *februar* ili *veljača*) vrednost odgovarajuće izlazne promenljive *m*, koja će biti upo-



Slika 5.1: Pojednostavljeni primer transduktora koji prepoznaje kompletne kalendarske datume i izdvaja osnovne elemente koji će biti upotrebljeni za normalizovanu vrednost

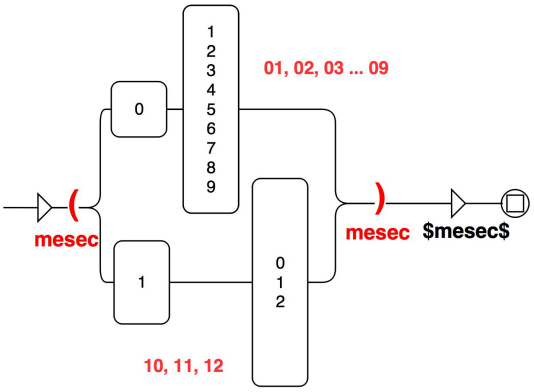
trebljena za interpretaciju normalizovanog oblika, neće biti niska koju je sravnila gramatika (u ovom slučaju npr. neki od oblika imena meseca *februar*), već će njena vrednost biti ono što je definisano izlazom koji proizvodi gramatika (npr. 02) (slika 5.2). Na isti način se i mesecima iskazanim rimskim brojevima u izlazu transduktora pridružuju ove vrednosti koje će se upotrebiti za standardizovanu interpretaciju elementa meseca u okviru atributa `value` (izrazu *I* se pridružuje izlazna vrednost 01, izrazu *II* vrednost 02 itd.).



Slika 5.2: Primer definisanja nekih izlaza koje proizvodi gramatika za normalizaciju vrednosti meseca iskazanog alfabetskim karakterima

S druge strane, u određenim situacijama je potrebno koristiti u izlazu upravo one niske koje je sravnila gramatika. Kada je reč npr. o mesecima ili danima iskazanim dvema arapskim ciframa, što u potpunosti odgovara obliku koji se upotrebljava kao normalizovana interpretacija, kao vrednost izlazne promenljive biće upotrebljena upravo ona niska koja predstavlja deo ulaznog teksta. Na primer, deo

prepoznate sekvencije zapamćen kao ulazna promenljiva *mesec* (u transduktoru definisana pomoću crvenih zagrada) (slika 5.3) biće upotrebljen u nepromenjenom obliku kao vrednost elementa mesec atributa `value`.

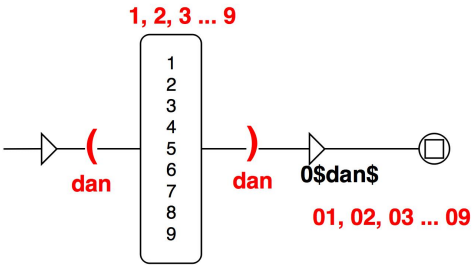


Slika 5.3: Primer definisanja izlaza koji pamti deo prepoznate sekvencije

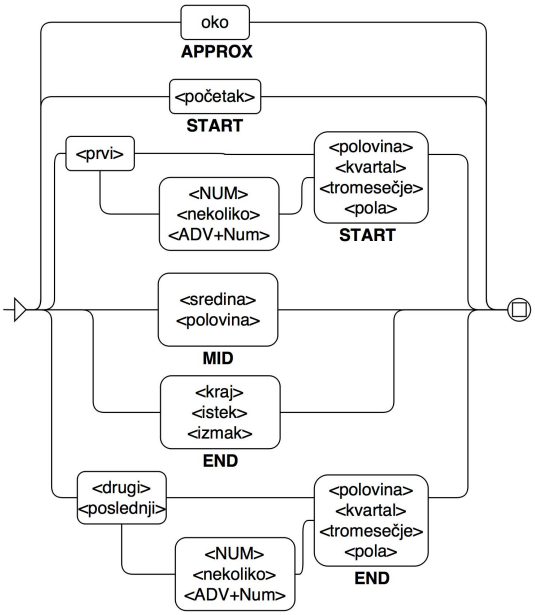
Međutim, kada je reč o korišćenju izlaznih promenljivih koje proizvodi gramatika, u određenim situacijama je neophodno umetanje određenih vrednosti onim niskama koje je gramatika sravnila. U slučaju prvih devet dana u mesecu ili prvih devet meseci u godini, koji mogu biti iskazani jednom cifrom (npr. *1*, *2*, *3* itd.), potrebno je proizvesti izlaz koji omogućava interpretaciju ovih elemenata dvema ciframa, što je u skladu sa standardom ISO 8601. Slika 5.4 ilustruje primer transduktora koji prepoznaje dane u mesecu iskazane jednocifrenim arapskim brojem i proizvodi izlaz u kome je ovaj element datuma interpretiran dvocifrenom vrednošću. Prepoznate sekvencije se smeštaju u ulaznu promenljivu *dan*, kojoj će u izlazu biti dodata 0 (npr. za prepoznatu nisku 1 proizvodi se izlaz u obliku 01 itd.).

Prethodno opisane ulazne i izlazne promenljive se mogu koristiti na globalnom nivou, što znači da se promenljiva definisana u okviru jednog grafa može upotrebiti u izlazu bilo kog drugog grafa koji upućuje na nju. Tako će u izlazu transduktora *Datum apsolutni* (slika 5.1), koji vrši normalizaciju vrednosti apsolutnih kalendarskih datuma, biti upotrebljene upravo one promenljive definisane u okviru podgrafova zaduženih za prepoznavanje osnovnih elemenata datuma (*Dan*, *Mesec*, *Godina*).

U slučaju modifikovanih apsolutnih vremenskih izraza u obliku kalendarskih datuma (npr. *oko 5. aprila 2015*, *početkom 2009.* i sl.) se takođe korišćenjem izlaznih promenljivih definišu i vrednosti koje će biti upotrebljene kao vrednosti atributa `mod`. Graf *Modif* za određene predloge, imenice i imeničke sintagme, upotrebljene u funkciji modifikatora vremenskog izraza, kreira izlaz koji odgovara vrednosti ovog atributa, koja je propisana TimeML uputstvom za obeležavanje vremenskih izraza (slika 5.5).



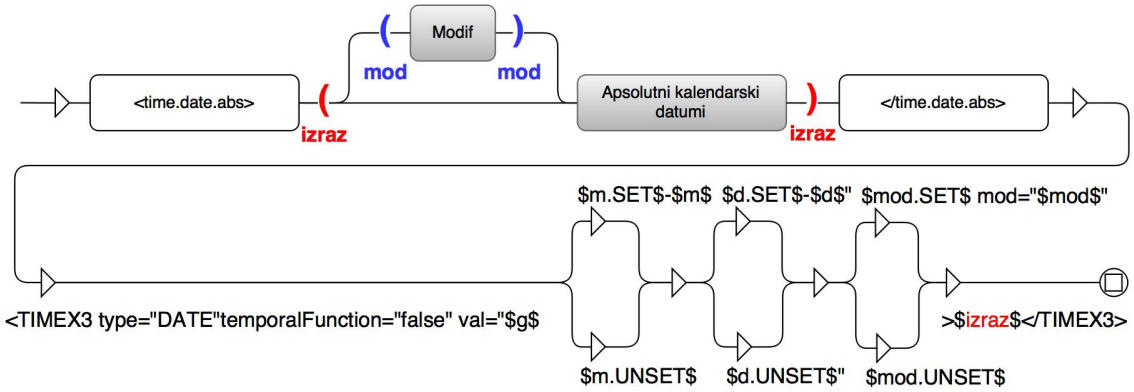
Slika 5.4: Primer definisanja izlaza koji dodaje određene vrednosti na prepoznatu sekvenciju



Slika 5.5: Neke od putanja grafa *Modif* zaduženog za kreiranje izlaza kao atributa `mod`

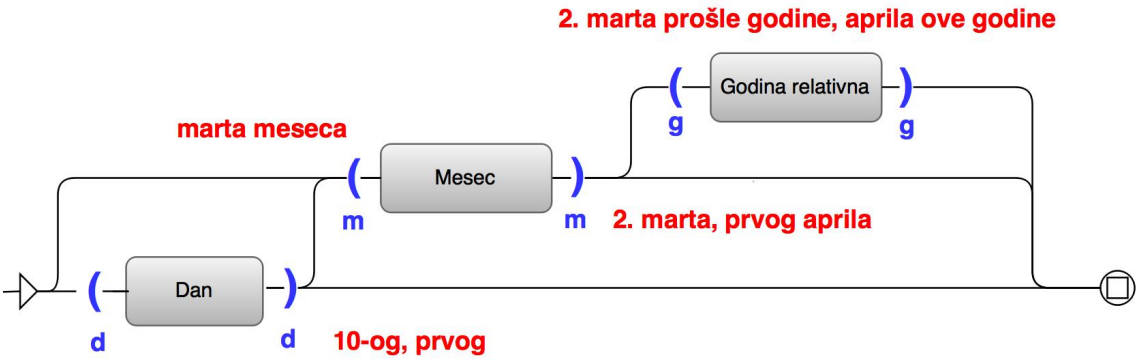
Nakon definisanja osnovnih elemenata i njihovih vrednosti koje će biti upotrebljene za normalizovanu interpretaciju izraza i popunjavanje atributa `value` i `mod`, transduktor *Datum apsolutni* proizvodi izlaz, koji predstavlja konačnu reprezentaciju vremenskog izraza obeleženog u skladu sa TimeML shemom (slika 5.6).

Podgraf *Apsolutni kalendarski datum* poziva sve grafove namenjene prepoznavanju apsolutnih vremenskih izraza u obliku kalendarskih datuma, u okviru kojih su osnovni elementi datuma, neophodni za popunjavanje atributa `value`, definisani u okviru izlaznih promenljivih *g*, *m* i *d* (neke od putanja jednog podgrafa date su na slici 5.1). Vrednosti modifikatora, definisane u okviru podgrafa *Modif*, smeštaju se u izlaznu promenljivu `mod`, kako bi bile upotrebljene u izlazu kao vrednost atributa `mod`. Prepoznata niska, koju čini u prethodnoj fazi obeležen vremenski izraz (npr. `<time.date.abs>25. mart 2007.</time.date.abs>`), biće u potpunosti zamenjena niskom definisanom u izlazu transduktora *Datum apso-*



Slika 5.6: Uopšten prikaz transduktora *Datum apsolutni*

lutni, a koja predstavlja prepoznat vremenski izlaz obeležen u skladu sa TimeML uputstvom. Početna i završna etiketa, dodeljene u fazi prepoznavanja, biće zamenjene `<TIMEX3>` etiketama, koje sadrže podatak o semantičkoj klasi izraza iskazanog atributima `type` i `temporalFunction`. Sledeći atribut, koji se u okviru početne `<TIMEX3>` etikete nalazi odmah iza atributa `temporalFunction`, jeste atribut `value`. Budući da se u ovom slučaju radi o apsolutnim kalendarskim datumima, na prvoj poziciji ovog atributa će se obavezno naći podatak o godini, definisan u izlazu kao izlazna promenljiva `$g$`. S obzirom na to da apsolutni kalendarski datum, osim podatka o godini, može a ne mora nužno da sadrži i elemente meseca i dana, potrebno je primeniti operaciju testiranja promenljivih, i omogućiti sravnjivanje ulaznog teksta i u slučajevima da vrednost promenljivih `$m$` i `$d$` nije definisana, odnosno nepostojanja podataka o mesecu i danu. Ako u okviru vremenskog izraza, osim godine, postoji podatak o mesecu, odnosno ako je njegova vrednost definisana promenljivom `$m$` (što je u grafu sa slike 5.6 predstavljeno sekvencijom `$m.SET$`), u okviru atributa `value` biće omogućen unos vrednosti promenljive `$m$` u sledećem obliku `val="$g$-$m$"` (npr. normalizovana vrednost izraza *marta 2009.* biće predstavljena u obliku `val="2009-03"`). S druge strane, u slučaju vremenskog izraza koji sadrži samo podatak o godini (npr. *2009*), sekvencijom `$m.UNSET$` omogućeno je popunjavanje atributa `value` samo onom vrednošću koja se odnosi na element godine i dalje sravnjivanje ulaznog teksta. U sledećem koraku se na isti način testira postojanje postavljene vrednosti promenljive `$d$`, koja predstavlja poslednji mogući element atributa `value`, datog u obliku `val=$g$-$m$-$d$`. Testiranjem promenljive `$mod$` je u izlazu ovog transduktora omogućeno postavljanje atributa `mod` i unošenje vrednosti koja modifikuje prepoznat vremenski izraz. Rezultat primene transduktora *Datum apsolutni*, koji je zadužen za normalizaciju apsolutnih vremenskih izraza u obliku kalendarskih datuma, ilustrovan je u okviru priloga C primerom C.1.



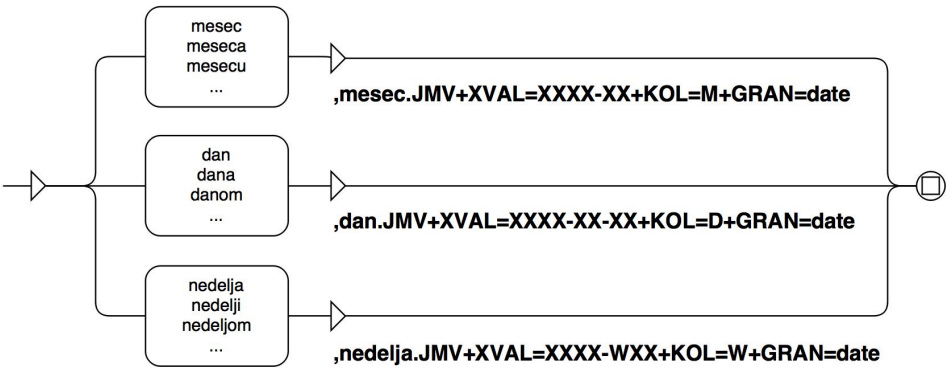
Slika 5.7: Neke od putanja transduktora *Relativni kalendarski datum*

Transduktor *Datum relativni* poziva sve one transduktore namenjene prepoznavanju vremenskih izraza koji spadaju u grupu relativnih vremenskih izraza datih u vidu kalendarskih datuma (npr. *13. juna prošle godine*, *sledeće godine maja meseca*, *jeseni ove godine*, *prošle godine*, *petka 12. februara*, *petog marta*, *avgusta meseca*, *10-og*, *juče*), detaljno opisanih u delu 4.2.1. Kao i u procesu normalizacije apsolutnih kalendarskih datuma, i u slučaju relativnih izraza neophodno je definisati vrednosti koje predstavljaju osnovne elemente datuma i koje će se naći u normalizovanom obliku, kao elementi atributa `value`. S obzirom na to da ovu klasu izraza, čija vrednost nije eksplicitno iskazana u tekstu, karakteriše odsustvo padatka o godini, godini i mesecu, ili, pak, svih elemenata kalendarskog datuma, samo oni osnovni elementi datuma koji su identifikovani u tekstu biće smešteni u izlazne promenljive `$g$`, `$m$` i `$d$`, u transduktoru označene pomoću plavih zagrada. Na slici 5.7 predstavljene su neke od putanja transduktora *Relativni kalendarski datum*, koji u okviru izraza izdvaja identifikovane osnovne elemente datuma smeštajući ih u odgovarajuće izlazne promenljive. Prikazanim putanjama omogućeno je, pre svega, prepoznavanje relativnih datuma koji sadrže samo element dana (npr. *10-og*), smešten u izlaznu promenljivu `$d$`, a zatim i datuma koji sadrže samo element meseca (npr. *aprila*), definisan izlaznom promenljivom `$m$`. Osim toga, definisanim putanjama se prepoznaju i relativni datumi iz kojih je izostavljen podatak o godini (npr. *2. marta*) ili je nedovoljno precizno dat u vidu imeničke sintagme čije je leksičko jezgro imenica *godina* (npr. *prošle godine*, *ove godine*). U tom slučaju se postojeći podatak o godini smešta u izlaznu promenljivu `$g$`.

Vrednosti izlaznih promenljivih, koje se odnose na eksplicitno date elemente datuma (kao što su u ilustrovanom primeru podaci o danu i mesecu), definišu se na isti način kao i u slučaju prethodno opisanih apsolutnih vremenskih izraza. Međutim, proces normalizacije izostavljenih ili nedovoljno precizno datih podataka podrazumeva drugačiji pristup i upotrebu karaktera X u reprezentaciji njihovih

vrednosti. U slučaju izostavljenog podatka o godini, na poziciji godine u okviru atributa `value` će se naći vrednost XXXX (npr. vrednost izraza *2. marta* biće normalizovana oblikom `value="XXXX-03-02"`). Kada je podatak o godini dat u vidu nepreciznog izraza (npr. *2. marta prošle godine*), osim vrednosti XXXX koja će biti upotrebljena na poziciji godine u okviru atributa `value`, potrebno je odrediti i vrednosti koje će biti upotrebljene u okviru atributa `valueFromFunction` (opisan u delu 5.2.3). Ovim atributom godina, kao jedinica vremenske mere koja treba da bude dodata ili oduzeta od vremenskog izraza koji će poslužiti kao orijentir u procesu računanja konačne vrednosti, treba da bude predstavljena karakterom Y koji ukazuje na period trajanja ove jedinice. Kako bi za potrebe normalizacije bilo omogućeno istovremeno predstavljanje jedinica vremenske mere različitim oblicima (npr. oblik kalendarskog datuma, oblik trajanja), kreiran je rečnički graf *JMV*. Ovaj graf je zapravo transduktor koji se primenjuje radi pronalaženja jedinica mere vremena u tekstu (npr. minut, dan, mesec, godina, decenija, vek itd.), kojima će u izlazu biti dodeljeni određeni semantički kodovi, korisni u procesu normalizacije. Neke od putanja transduktora *JMV* date su na slici 5.8. Na primer, transduktorom *JMV* se prepoznaju različiti oblici imenice *mesec*, koja predstavlja jednu od jedinica mere vremena, i u izlazu joj se dodeljuju sledeći kodovi:

1. JMV (oznaka za jedinicu mere vremena),
2. XVAL=XXXX-XX (oznaka za standardizovani kalendarski oblik prikazivanja meseca, čija vrednost nije poznata),
3. KOL=M (oznaka za standardizovani oblik koji ukazuje na period trajanja – mesec),
4. GRAN=date (oznaka za nivo granularnosti, čija vrednost ukazuje na to da mesec pripada izrazima sa značenjem kalendarskog datuma).



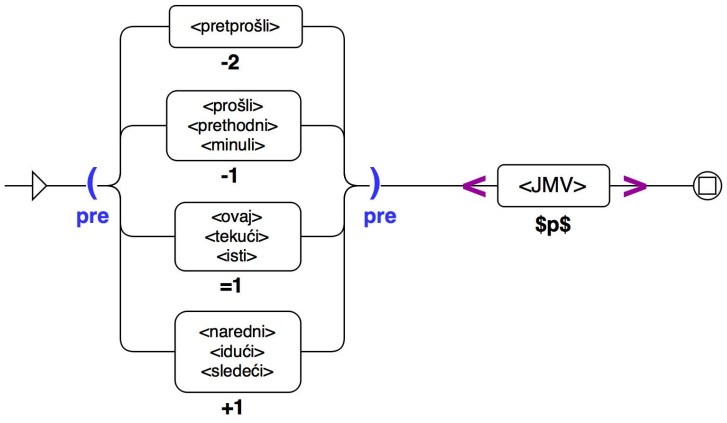
Slika 5.8: Neke od putanja rečničkog grafa *JMV*

Sekvencije proizvedene ovim transduktorom predstavljaju validne DELAF odrednice srpskog elektronskog rečnika (primer 5.31).

Primer 5.31.

```
meseca,mesec.JMV+XVAL=XXXX-XX+KOL=M+GRAN=date
```

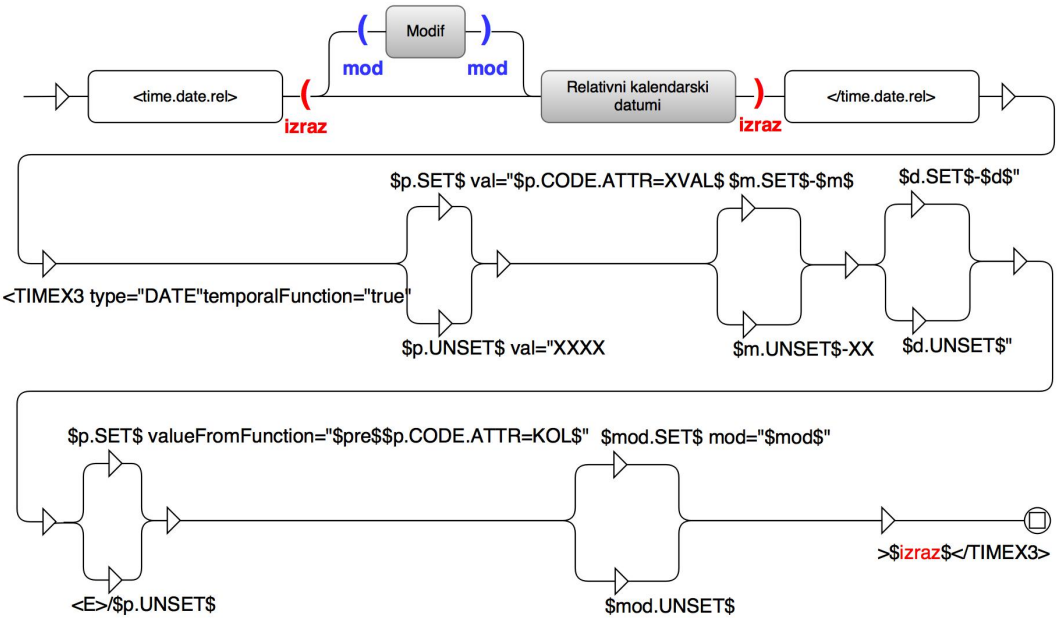
Nakon kompilacije ovog transduktora u odgovarajući oblik (*JMV.fst2*), moguće je uključiti ga u listu rečnika koji se primenjuju nad tekstom. S obzirom na to da jedinice mere vremena spadaju u grupu imenica koje se prepoznaju regularnim rečnicima, neophodno je ovaj rečnički graf primeniti sa višim prioritetom u odnosu na regularne rečnike, kako bi jedinice mere vremena u rečniku teksta bile obeležene informacijama neophodnim za proces normalizacije. Postojanje ovakvog tipa informacija u rečnicima omogućava da se u okviru transduktora, unutar morfološkog režima, koriste leksičke maske (slika 5.9), koje prepoznaju bilo koji oblik neke od jedinica mere vremena i uvode promenljivu (npr. `$p$`) koja će biti upotrebljena u okviru nekog drugog transduktora za referisanje na određene kodove prepoznatog oblika.



Slika 5.9: Neke od putanja grafa *Relativni kalendarski datum*

U okviru transduktora *Datum relativni* podgraf *Relativni vremenski izraz* poziva sve grafove namenjene prepoznavanju relativnih vremenskih izraza u obliku kalendarskih datuma, u okviru kojih su svi postojeći osnovni elementi datuma, neophodni za popunjavanje atributa `value`, definisani u okviru izlaznih promenljivih `g`, `m` i `d` (neke od putanja jednog podgrafa date su na slici 5.7). Kao i u slučaju apsolutnih kalendarskih datuma, vrednosti modifikatora koje mogu da se nađu u okviru relativnog datuma smeštaju se u izlaznu promenljivu `mod`. U izlazu transduktora *Datum relativni* prepoznata niska se menja izrazom, koji predstavlja konačnu reprezentaciju relativnog vremenskog izraza obeleženog u skladu sa Ti-

meML shemom. Na slici 5.10 prikazana je normalizacija najčešće korišćenih oblika relativnih kalendarskih datuma.



Slika 5.10: Transduktor *Datum relativni*, izlaz

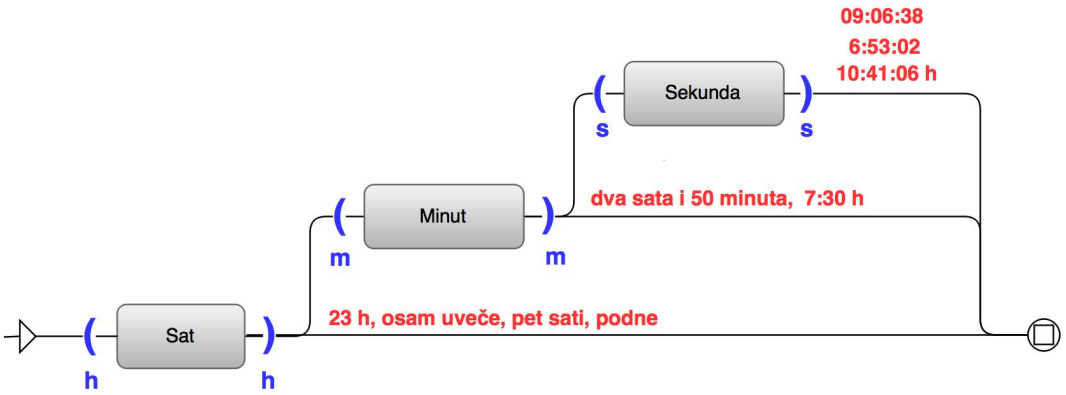
Početna i završna etiketa, dodeljene u fazi prepoznavanja, biće zamenjene `<TIMEX3>` etiketama, koje sadrže odgovarajući podatak o semantičkoj klasi izraza iskazanog atributima `type` i `temporalFunction`. S obzirom na to da se radi o relativnim kalendarskim datumima iz kojih podatak o godini može da bude izostavljen, već u prvom koraku postavljanja vrednosti atributa `value` neophodno je izvršiti testiranje promenljive `$g$`. U slučaju da je dat nedovoljno precizan podatak o godini, sekvencijom `$p.SET$ val="$p.CODE.ATTR=XVAL$` omogućava se korišćenje ranije uvedene promenljive `$p$`, odnosno vrednosti njenog atributa XVAL (`$p.CODE.ATTR=XVAL$`), koja će se naći na prvoj poziciji atributa `value`. Na taj način će za relativni izraz *prošle godine* normalizovana vrednost biti data u obliku `value="XXXX"`. S druge strane, ako se radi o izrazu iz koga je izostavljen bilo kakav podatak o godini, sekvencijom `$p.UNSET$ val="XXXX` biće omogućeno obavezno popunjavanje prve pozicije atributa `value`. Na drugoj poziciji atributa `value` moguće je pojavljivanje eksplicitno iskazane vrednosti meseca, koja je smeštena u izlaznu promenljivu `$m$`, što je definisano sekvencijom `$m.SET$-$m$`. Ako se radi o izrazu iz koga je podatak o mesecu potpuno izostavljen, popunjavanje i druge pozicije omogućeno je definisanom vrednošću (-XX) u okviru sekvencije `$m.UNSET$-XX`, pa će atribut `value` biti predstavljen oblikom `value="XXXX-XX`. U sledećem koraku se na isti način vrši i provera postojanja postavljene vrednosti promenljive `$d$`, odnosno podatka o danu, kao poslednjeg elementa atributa `value`. U slu-

čaju da je ovaj podatak eksplicitno naveden u tekstu, treća pozicija se popunjava vrednošću koja je definisana u okviru izlazne promenljive `$d$`, dok se u situacijama kada je ovaj podatak izostavljen predviđa samo zatvaranje vrednosti atributa `value` dodavanjem znaka navoda (`$d.UNSET$"`). S obzirom na to da je u slučaju relativnih vremenskih izraza potrebno obezbediti što više informacija neophodnih za računanje njihove krajnje vrednosti i pozicije na vremenskoj osi, u pretposlednjem koraku se vrši testiranje definisanih vrednosti koje će činiti sadržaj atributa `valueFromFunction`. Na primeru nedovoljno precizno iskazanog izraza *sledeće godine*, sekvencijom `$p.SET$ valueFromFunction="$pre$$p.CODE.ATTR=KOL$"` određena je vrednost ovog atributa u obliku `valueFromFunction="+1Y"`. U jednoj od putanja grafa ilustrovanog na slici 5.9 dat je primer smeštanja prideva *sledeći* u izlaznu promenljivu `$pre$`, čija će vrednost biti +1. Na drugoj poziciji atributa `valueFromFunction` će se naći ranije uvedena promenljiva `$p$`, odnosno vrednost njenog atributa KOL (`$p.CODE.ATTR=KOL$`). Kompletna reprezentacija ovog atributa (`valueFromFunction="+1Y"`) ukazuje na to da je potrebno dodati jednu godinu vremenskom izrazu koji će poslužiti kao referentna vrednost. Određivanje vremenskih izraza koji predstavljaju orijentir u računanju apsolutne vrednosti relativnog izraza, pa ni same računske operacije, neće biti predmet ovog rada. Na kraju, kao i u slučaju apsolutnih kalendarskih datuma, testiranjem promenljive `$mod$` je u izlazu transduktora *Datum relativni* omogućeno postavljanje atributa `mod` i unošenje vrednosti koja modifikuje prepoznat relativni vremenski izraz. Rezultat primene transduktora *Datum relativni*, koji je zadužen za normalizaciju relativnih vremenskih izraza u obliku kalendarskih datuma, ilustrovan je u okviru priloga C primerom C.2.

Transduktori za normalizaciju vrednosti vremenskih izraza datih u vidu vremena dana

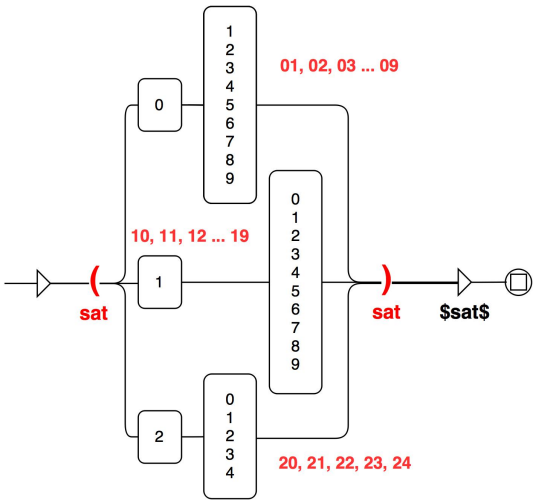
Graf *Vreme* dana namenjen je normalizaciji vrednosti vremenskih izraza iskazanih na nivou granularnosti uže od nivoa dana. Tokom normalizacije vrednosti ove klase izraza koriste se pojednostavljena pravila, već upotrebljena za njihovu identifikaciju u prethodnoj fazi, ali ovog puta uz korišćenje precizno određenog konteksta u vidu postavljenih početnih i završnih etiketa, koje ukazuju na tip prepoznatog vremenskog izraza (npr. `<time.hour.abs>8:45 h.</time.hour.abs>`). U okviru svakog od upotrebljenih transduktora definisane su vrednosti koje predstavljaju osnovne elemente vremena dana (npr. sat, minut, sekunda) i koje će se naći u normalizovanom obliku, kao elementi atributa `value`. Vrednosti prepoznate kao element sata smeštaju se u izlaznu promenljivu `h`, dok se vrednosti koje

se odnose na minut i sekundu smeštaju u izlazne promenljive `m` i `s`. Na slici 5.11 dat je primer jednog od transduktora koji je zadužen za izdvajanje prepoznatih elemenata vremena dana i njihovo smeštanje u odgovarajuće izlazne promenljive.

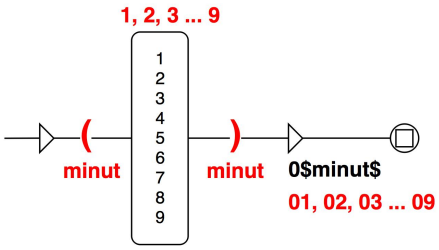


Slika 5.11: Primer nekih putanja zaduženih za izdvajanje prepoznatih elemenata vremena dana i njihovo smeštanje u odgovarajuće izlazne promenljive

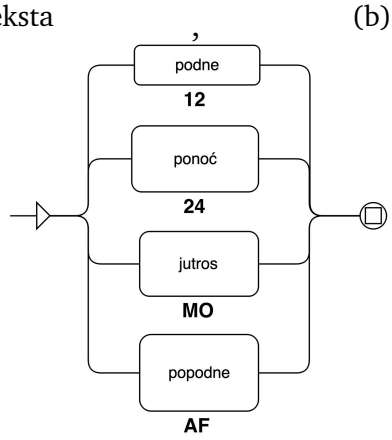
Vrednosti koje će biti upotrebljene za normalizovanu interpretaciju vremenskog izraza definisane su izlaznim promenljivima na isti način kao i u slučaju vremenskih izraza u obliku kalendarskih datuma. Izlazne vrednosti, pre svega, može da čini upravo ona niska koja predstavlja deo ulaznog teksta (slika 5.12a), kao što je to u primeru sata, minuta ili sekunde iskazanih dvocifrenim arapskim brojem. S druge strane, vrednosti osnovnih elemenata vremena dana, koje će biti upotrebljene u izlazu, mogu biti i rezultat primene gramatike koja u izlazu dodaje određene vrednosti prepoznatim sekvencijama (slika 5.12b) ili ih u potpunosti menja drugačijim niskama, a koje predstavljaju standardizovani oblik njihove interpretacije (slika 5.12c). Kodovi koji se koriste na poziciji sata za reprezentaciju relativnih vremenskih izraza sa značenjem vremena dana (npr. *jutros*, *noćas*, *popodne*) detaljno su predstavljeni u delu 5.2.1. Prilikom definisanja vrednosti koje se odnose na sat potrebno je obratiti posebnu pažnju na levi i desni kontekst pojavljivanja ovog elementa u tekstu. U slučaju eksplicitno iskazanog sata, u čijem se levom ili desnom kontekstu nalazi imenica vremenskog značenja ili vremenski prilog (npr. *šest sati* ***uveče***, ***večeras*** *oko 11 h*), neophodno je u izlazu uskladiti identifikovanu vrednost sata sa značenjem imenice ili priloga iz konteksta, što je ilustrovano slikom 5.13.



(a) Deo ulaznog teksta

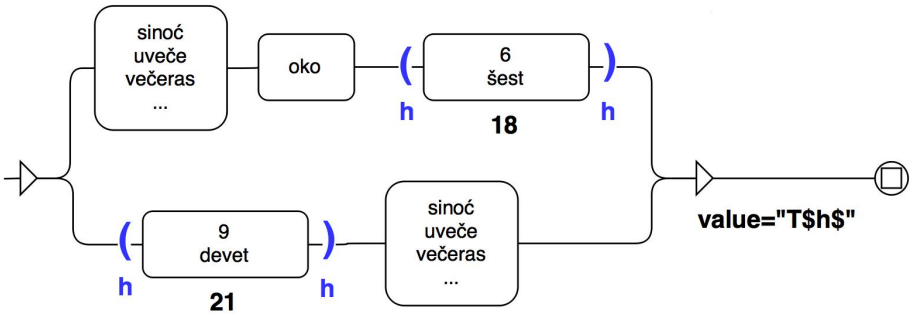


(b) Umetanje vrednosti



(c) Novi izlaz

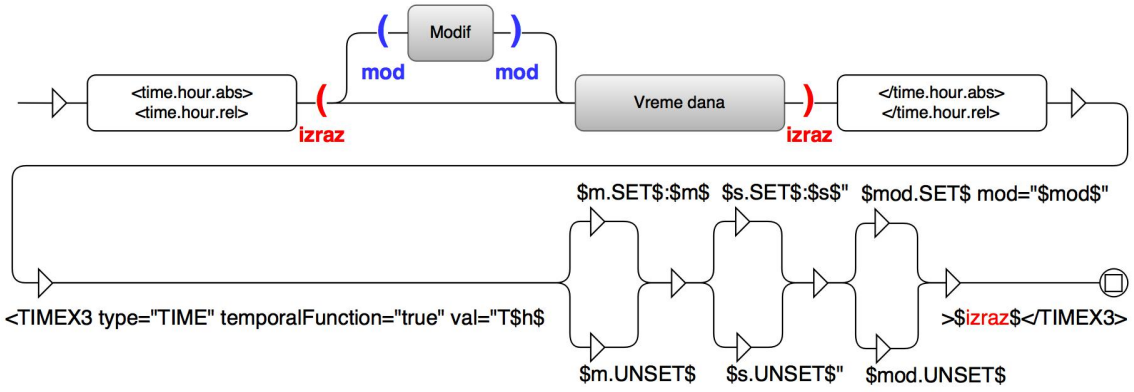
Slika 5.12: Neki primeri definisanja vrednosti elemenata vremena dana



Slika 5.13: Primer definisanja vrednosti elementa sat

Nakon izdvajanja osnovnih elemenata i njihovih vrednosti koje će biti upotrebljene za popunjavanje atributa `value`, glavni transduktor *Vreme* proizvodi izlaz, kao konačnu reprezentaciju vremenskog izraza sa značenjem vremena dana, obeleženog u skladu sa TimeML shemom (slika 5.14).

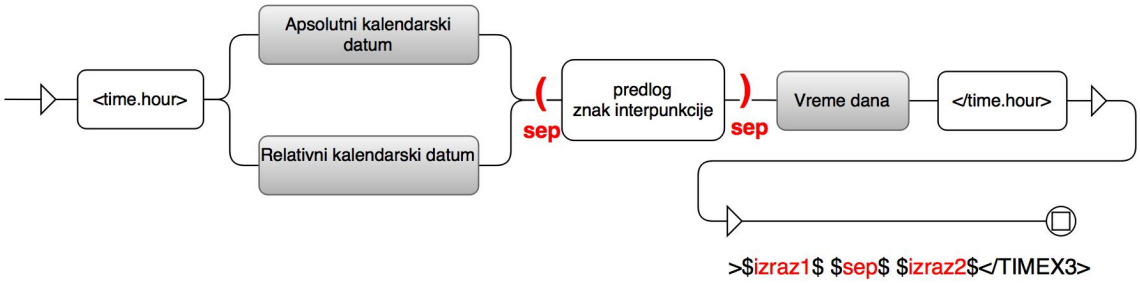
Podgraf *Vreme dana* poziva sve grafove namenjene prepoznavanju ove klase



Slika 5.14: Primer izlaza koji proizvodi transduktor *Vreme*

vremenskih izraza, u okviru kojih su svi postojeći elementi vremena dana definisani kao izlazne promenljive `h`, `m` i `s`. Osim ovih vrednosti, u slučaju modifikovanih izraza potrebno je izdvojiti i vrednosti modifikatora, koje se smeštaju u izlaznu promenljivu `mod`. U izlazu transduktora *Vreme* su na prvoj poziciji `<TIMEX3>` etikete definisani atributi `type` i `temporalFunction`, sa odgovarajućim informacijama o semantičkoj klasi prepoznatog izraza. Sledeći obavezan atribut `<TIMEX3>` etikete je atribut `value`, koji na prvom mestu sadrži karakter T iza koga sledi podatak o satu, definisan u izlazu kao izlazna promenljiva `$h$` (`val="T$h$"`). Budući da ova klasa izraza može, a ne mora da sadrži i elemente minuta i sekunde, u sledećem koraku se vrši testiranje promenljivih, čime se omogućava sravnjivanje ulaznog teksta i u slučajevima nepostojanja promenljivih `$m$` i `$s$`. Ako u okviru vremenskog izraza, osim sata, postoji i podatak o minutu, sekvencijom `$m.SET$` će se uneti vrednost promenljive `$m$` u obliku `:$m$`, odmah iza već definisane vrednosti `val="T$h$`. Tako će atribut `value` u ovom koraku imati sledeći oblik `val="T$h$:$m$`. U slučaju da vremenski izraz ne sadrži podatak o minutu, sekvencijom `$m.UNSET$` je omogućeno dalje sravnjivanje ulaznog teksta i testiranje promenljive `$s$`, kao sledećeg elementa izraza. Ako je ovaj podatak naveden u tekstu, pozicija sekunde se popunjava vrednošću koja je definisana u okviru izlazne promenljive `$s$`, dok se u situacijama kada je ovaj podatak izostavljen sekvencijom `$s.UNSET$"` predviđa samo zatvaranje vrednosti atributa `value` dodavanjem znaka navoda (npr. `val="T$h$:$m$"`). Nakon postavljanja vrednosti atributa `value`, testiranjem promenljive `$mod$` omogućeno je postavljanje atributa `mod` i unošenje vrednosti koja modifikuje vremenski izraz. S obzirom na to da vremenski izraz na nivou granularnosti sata, minuta ili sekunde može da sadrži i podatak o kalendarskom datumu, podgraf *Datum_Vreme* proizvodi kompletnu reprezentaciju ovog oblika, koji na osnovu TimeML sheme treba da bude obeležen jednom `<TIMEX3>` etiketom (slika 5.15). Podgrafovi *Apsolutni kalendarski datum* i *Relativni*

kalendarski datum proizvode prvi deo početne `<TIMEX3>` etikete, koji sadrži atribute `type`, `temporalFunction` i postojeće kalendarske elemente atributa `value`, što se vrši na prethodno opisan način testiranjem promenljivih u koje su ovi elementi datuma smešteni. Sâm kalendarski datum smešta se u ulaznu promenljivu `$izraz1$`, kako bi bio upotrebljen u izlazu transduktora *Datum_Vreme*. Za kreiranje onog dela izlaza koji čine elementi atributa `value` sa značenjem vremena dana, zadužen je podgraf *Vreme dana*, u okviru koga je vreme dana obeleženo ulaznom promenljivom `$izraz2$`. U samom izlazu transduktora *Datum_Vreme* definisan je samo završetak početne `<TIMEX3>` etikete, iza koga slede prepoznat vremenski izraz (`$izraz1$$izraz2$`) i završna `</TIMEX3>` etiketa.



Slika 5.15: Primer izlaza koji proizvodi transduktor *Datum_Vreme*

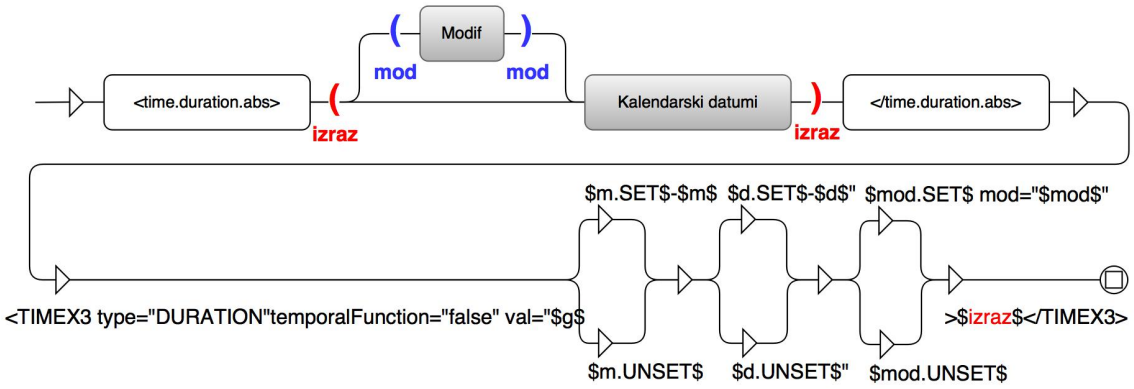
Primeri normalizovanih vremenskih izraza u vidu vremena dana dati su u okviru priloga C (primer C.3).

5.3.2 Transduktori za normalizaciju vrednosti vremenskih izraza koji impliciraju trajanje

Vremenski izrazi ovoga tipa jesu izrazi kojima se ispoljava temporalna kvantifikacija, odnosno odmerava dužina trajanja u vremenu. Kako značenje trajanja može biti implicirano i kalendarskim datumima, čiji se oblik normalizovane vrednosti razlikuje od oblika predviđenog za interpretaciju izraza datih u vidu kvantifikovanih jedinica mere vremena, glavni graf *Trajanje* poziva dva transduktora, koji odgovaraju navedenim oblicima ove klase izraza.

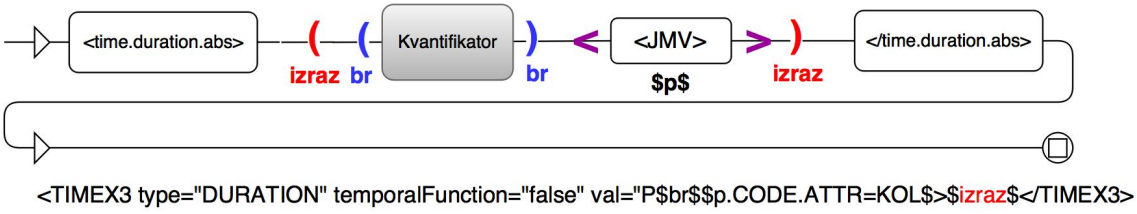
Za normalizaciju vrednosti vremenskih izraza koji ukazuju na trajanje iskazano nekim od oblika kalendarskog datuma zadužen je transduktor *Trajanje_datum*. Ovaj transduktor koristi ista pravila i metod definisanja vrednosti osnovnih elemenata datuma koji se primenjuju u normalizaciji vrednosti vremenskih izraza koji ukazuju na tačku u vremenu. Jedina razlika u odnosu na transduktor *Datum* je u izlazu koji transduktor *Trajanje_datum* proizvodi i u okviru koga je vrednost atributa `type` definisana kao DURATION (slika 5.16), što odgovara semantičkoj klasi

kojoj izraz u ovom slučaju pripada.



Slika 5.16: Izlaz koji proizvodi transduktor *Trajanje_datum*

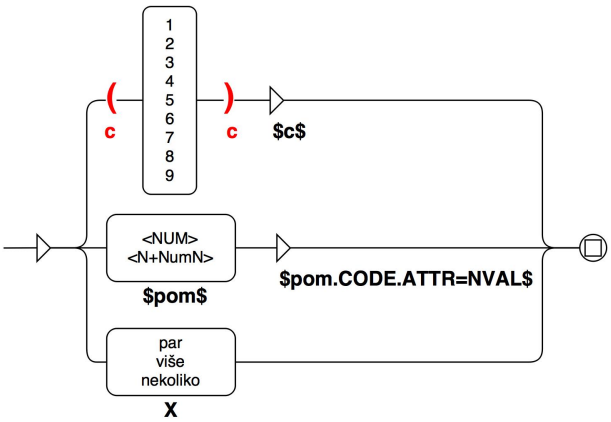
Kada je reč o normalizaciji vrednosti vremenskih izraza koji ukazuju na trajanje iskazano kvantifikovanim jedinicama mere vremena, potrebno je primeniti drugačiji pristup u definisanju izlazne standardizovane vrednosti, predviđene TimeML uputstvom za obeležavanje. U prvom koraku se vrši izdvajanje osnovnih elemenata ovog tipa izraza, neophodnih za normalizovanu interpretaciju. Osnovni elementi, čije će vrednosti biti upotrebljene u okviru atributa `value`, jesu numerički ili nenumerički kvantifikator i jedinica mere vremena. Delovi teksta koji odgovaraju kvantifikatoru, odnosno broju jedinica vremenske mere, smeštaju se u izlazne promenljive (npr. `br`), u transduktoru predstavljene plavim zagradama (slika 5.17). Za izdvajanje vrednosti drugog elementa atributa `value`, odnosno jedinica mere vremena, koje su, kao što je već opisano u delu 5.3.1, zastupljene u rečniku teksta, koriste se leksičke maske, kojima se bilo koji oblik prepoznate jedinice mere vremena definiše kao promenljiva (npr. promenljiva `$p$`). Na taj način je, u izlazu transduktora zaduženog za popunjavanje vrednosti atributa `value`, omogućeno korišćenje onih vrednosti koje se odnose na standardizovani oblik reprezentacije perioda trajanja određene jedinice mere vremena, definisan pridruženim semantičkim kodom KOL (`$p.CODE.ATTR=KOL$`).



Slika 5.17: Jedna od putanja transduktora koji proizvodi TIMEX3 izlaz za apsolutne izraze koji ukazuju na trajanje

Izrazi kojima je kvantifikovan period trajanja neke jedinice mere vremena mogu

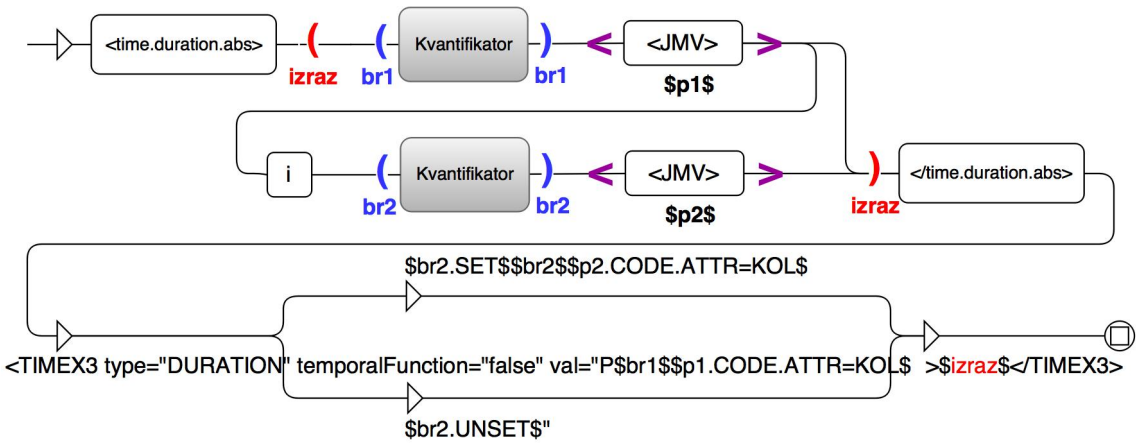
biti u obliku numeričkih vrednosti, iskazanih na više različitih načina (npr. *pet*, *23*, *desetak*, *30-ak*, *pola*, *hiljadu i 200*, *12,4*, *1.700*). Kako bi se omogućilo prepoznavanje svih mogućih oblika kvantifikatora, kao i upotreba njihove normalizovane vrednosti u standardizovanom numeričkom obliku, u procesu izdvajanja koriste se leksičke maske, koje u rečniku teksta pronalaze brojeve zapisane različitim oblicima numeričkih i alfabetskih karaktera, kao i njihovom kombinacijom (slika 5.18). Uvođenjem promenljive (u ovom slučaju promenljive `$pom$`) u izlazu je definisana vrednost koja će se koristiti za interpretaciju i koja je prepoznatom kvantifikatoru u rečniku pridružena u vidu atributa NVAL (`$pom.CODE.ATTR=NVAL$`) (Krstev and Vitas 2006; Krstev, Jaćimović, and Vitas 2012). S obzirom na to da prvih devet osnovnih brojeva, iskazanih ciframa, nije zastupljeno u rečnicima, njihova normalizacija se vrši pomoću ulaznih promenljivih, u koje se smeštaju prepoznate niske, koje će biti upotrebljene i u izlazu u istom obliku. U slučaju relativnih vremenskih izraza sa značenjem trajanja, koje je iskazano nekim od nenumeričkih kvantifikatora (npr. *nekoliko*, *najviše*, *par*), za interpretaciju se koristi izlazom predviđen karakter X.



Slika 5.18: Neki primeri izdvajanja normalizovanih vrednosti kvantifikatora jedinica mere vremena

Transduktor *Trajanje_izraz* poziva sve grafove namenjene prepoznavanju vremenskih izraza koji ukazuju na trajanje, u okviru kojih su svi postojeći elementi, neophodni za popunjavanje atributa `value`, definisani izlaznim promenljivim i promenljivim rečničkih odrednica. S obzirom na to da izrazi ovoga tipa mogu biti iskazani i uzastopnim navođenjem više kvantifikovanih jedinica mere vremena različite granularnosti, uz poštovanje hijerarhijskog redosleda (npr. *dve godine i 7 meseci* ili *dva sata, tri minuta i 9 sekundi*), izlazom transduktora je neophodno omogućiti ispravnu interpretaciju i ovih složenijih oblika. Na slici 5.19 su prikazane neke od putanja transduktora *Trajanje_izraz*, koje testiranjem promenljivih

omogućavaju korektan oblik normalizacije vrednosti ovih izraza.



Slika 5.19: Neke od putanja transduktora *Trajanje_izraz*

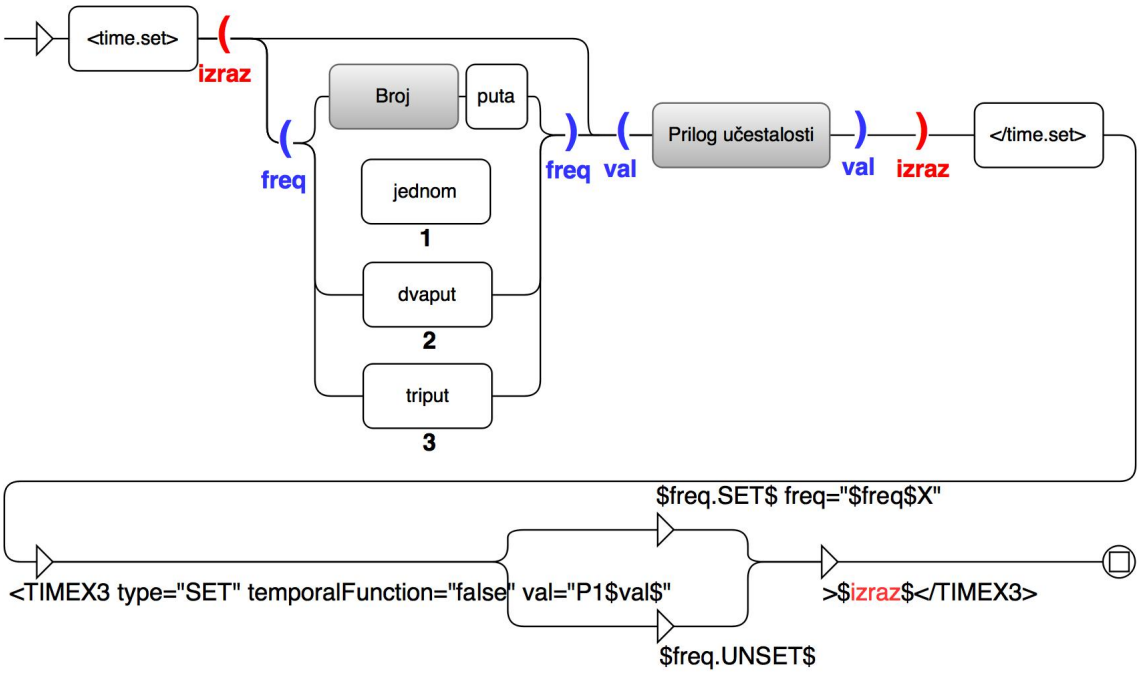
U izlazu transduktora *Trajanje_izraz* početna i završna etiketa, dodeljene u fazi prepoznavanja, menjaju se `<TIMEX3>` etiketama, koje sadrže odgovarajući podatak o semantičkoj klasi izraza iskazanog atributima `type` i `temporalFunction`. Kao i na primeru prethodno opisanih izraza koji ukazuju na tačku u vremenu, vrednosti modifikatora koje mogu da se nađu u okviru izraza se smeštaju u izlaznu promenljivu `mod`, kako bi bile upotrebljene za popunjavanje vrednosti atributa `mod`. U okviru priloga C dati su primeri (primer C.4 i C.5) koji ilustruju rezultat primene transduktora *Trajanje*, zaduženog za normalizaciju apsolutnih i relativnih vremenskih izraza sa značenjem trajanja.

5.3.3 Transduktori za normalizaciju izraza koji impliciraju učestalost

Vremenski izrazi sa značenjem učestalosti ukazuju na temporalnu frekvenciju, odnosno učestalost pojavljivanja u vremenu. S obzirom na to da predstavljaju poseban vid temporalnog kvantifikovanja, koje se može ostvariti kao povremeno ili regularno ponavljanje, radi normalizacije njihovih vrednosti kreirana su dva grafa koja odgovaraju navedenim frekvencijama pojavljivanja.

Graf *Povremeno* u okviru pojednostavljenih pravila, upotrebljenih za prepoznavanje izraza u prethodnoj fazi, vrši izdvajanje vrednosti koje će biti upotrebljene za popunjavanje atributa `value` i `freq` i proizvodi izlaz koji predstavlja vremenski izraz obeležen i normalizovan u skladu sa smernicama TimeML uputstva. Početna i završna etiketa, koje su dodeljene u fazi prepoznavanja vremenskih izraza, menjaju se `<TIMEX3>` etiketama, u okviru kojih su preneti atributi `type` i `temporalFunction`, zajedno sa dodeljenim vrednostima. Kako izrazi ovoga tipa

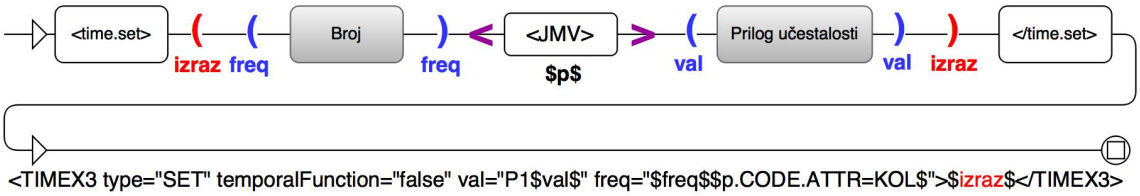
prenose značenje učestalosti ponavljanja u okviru nekog perioda, atribut `value` će i imati oblik vrednosti koja ukazuje na period vremena (PnX). Ovaj period je u izrazima najčešće dat u vidu priloga učestalosti (npr. *dnevno*, *nedeljno*, *godišnje*), kojima se u izlazu pridružuju određeni slovni kodovi, koji služe za normalizovanu interpretaciju tog perioda vremena u okviru kog se dešava ponavljanje (slika 5.20).



Slika 5.20: Pojednostavljen prikaz nekih putanja transduktora *Povremeno*

Identifikovani period smešta se u izlaznu promenljivu `val`, kako bi se njome definisana vrednost iskoristila u izlazu kao jedan od elemenata atributa `value`. U slučaju da frekvencija ponavljanja nije data izrazom, sekvencijom `$freq.UNSET$` se u izlazu vrši testiranje promenljive `freq` koja ukazuje na postojanje ovog podataka, čime je omogućeno dalje sravnjivanje teksta i kreiranje izlaza. Učestalost ponavljanja u okviru nekog perioda može biti iskazana prilozima *jednom*, *dvaput*, *triput* ili izrazom koji se sastoji od broja i priloga *puta* (npr. ***sedam puta*** *mesečno*). Normalizovana vrednost broja koji govori o frekventnosti ponavljanja dobija se na isti način kao i u slučaju vremenskih izraza koji ukazuju na trajanje (slika 5.18). Podatak o učestalosti ponavljanja se smešta u izlaznu promenljivu `freq` i koristi u izlazu za popunjavanje istoimenog atributa. Učestalost ponavljanja u okviru nekog perioda može biti iskazana i kvantifikovanim jedinicama mere vremena (npr. *dva sata dnevno*, *20 dana mesečno* itd.), pa je za ove primere predviđen nešto drugačiji izlaz u pogledu definisanja vrednosti atributa `freq` (slika 5.21). U tom slučaju, osim normalizovane vrednosti broja (smeštena u izlaznu promenljivu `$freq$`) ko-

jim je kvantifikovana jedinica mere vremena, atribut `freq` treba da sadrži i standardizovani oblik reprezentacije perioda trajanja jedinice mere vremena. Kao i na primeru izraza koji ukazuju na trajanje, korišćenjem rečničke promenljive `$p$` se u izlazu omogućava upotreba ovih informacija, u rečniku teksta pridruženih jedinici mere vremena u vidu semantičkog koda KOL (`$p.CODE.ATTR=KOL$`).

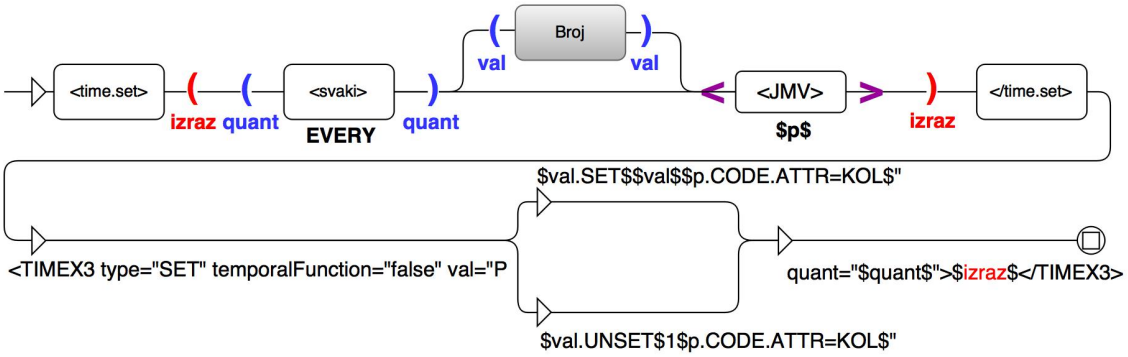


Slika 5.21: Jedna od putanja transduktora *Povremeno*

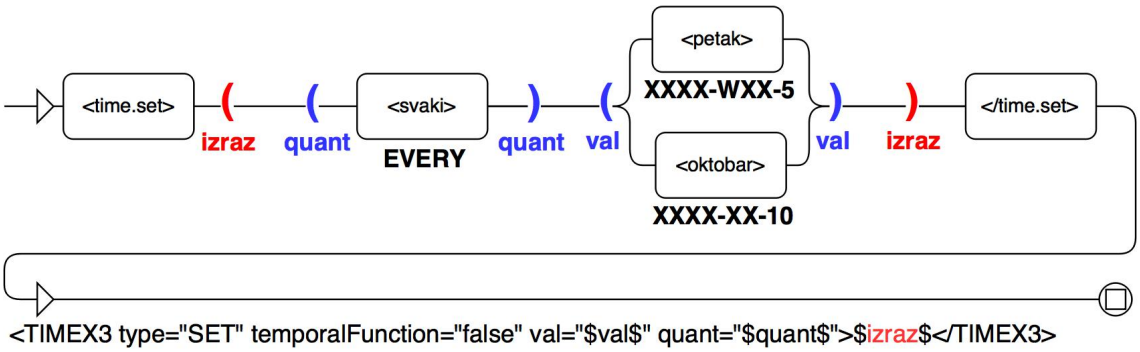
Za normalizaciju najčešćih oblika vremenskih izraza koji ukazuju na regularnu učestalost ponavljanja u vremenu predviđen je transduktor *Regularno*. Izraze ovoga tipa karakteriše, pre svega, upotreba pridevske zamenice *svaki* uz jedinicu mere vremena (npr. *svake godine*, *svakog dana*), koja može biti i numerički kvantifikovana (npr. *svakih jedanaest godina*, *svakih 13 dana*). Budući da identifikovana jedinica mere vremena predstavlja period učestalosti ponavljanja, njena vrednost, definisana kao atribut KOL na osnovu rečničke promenljive `$p$`, biće iskorišćena za popunjavanje atributa `value` (slika 5.22). Za izraze u okviru kojih je jedinica mere vremena kvantifikovana numeričkom vrednošću, u izlazu je sekvencijom `$val.SET$` predviđeno unošenje normalizovane vrednosti kvantifikatora (izlazna promenljiva `$val$`), iza koje sledi oznaka za standardizovani oblik perioda trajanja identifikovane jedinice mere vremena (`$p.CODE.ATTR=KOL$`). U slučaju da jedinica mere vremena nije kvantifikovana, u okviru atributa `value` se sekvencijom `$val.SET$` unosi podrazumevana vrednost 1, iza koje sledi standardizovana oznaka perioda trajanja jedinice mere vremena (`$p.CODE.ATTR=KOL$`). U sledećem koraku je izlazom transduktora *Regularno* predviđeno unošenje vrednosti atributa `quant`.

Značenje perioda ponavljanja, osim osnovnim jedinicama vremena, može biti iskazano i nazivima meseca u godini ili dana u nedelji. Budući da su ovo izrazi koji na osnovu svoje granularnosti pripadaju datumima, neophodno je sačuvati kalendarsku informaciju i vrednost atributa `value` reprezentovati odgovarajućim kalendarskim oblikom. Tako transduktor *Regularno* omogućava kreiranje izlaza koji vrši normalizaciju i ovog tipa izraza koji ukazuju na učestalost (slika 5.23).

Najčešći oblici vremenskiih izraza koji ukazuju na učestalost i koji su identifikovani u analiziranim novinskim tekstovima ilustrovani su u okviru priloga C (primer C.6).



Slika 5.22: Primeri nekih putanja transduktora *Regularno*



Slika 5.23: Primeri putanje transduktora *Regularno* za normalizaciju u kalendarskom obliku

5.3.4 Vremenski izrazi obeleženi kao periodi u fazi prepoznavanja

Tokom prethodne faze prepoznavanja vremenskih izraza vršeno je obeležavanje parova vremenskih izraza (npr. *14–16. januara 2008*), kao i nizova izraza koji su u određenom međusobnom odnosu (npr. *14, 15. i 16. feruara 2009*). Kako smernicama TimeML uputstva ovaj način obeležavanja nije predviđen, u okviru ovoga rada tokom procesa normalizacije ove postavljene etikete nisu uzete u obzir. Međutim, u okviru daljeg rada na razvoju sistema moguće je iskoristiti ih za definisanje vrednosti atributa `beginPoint` i `endPoint`, odnosno trajanja koje je implicirano ovim tačkama u vremenu. Osim toga, na osnovu ovako definisanih tačaka sa značenjem ingresivnosti i terminativnosti, biće omogućeno i precizno predstavljanje perioda na koji ukazuju, u obliku predviđenom TimeML uputstvom.

Glava 6

Evaluacija uspešnosti sistema za automatsku obradu vremenskih izraza srpskog jezika

Proces automatskog obeležavanja vremenskih izraza, koji, kao što je već rečeno ranije, uključuje dve faze – fazu prepoznavanja vremenskih izraza, kao i fazu normalizacije njihovih vrednosti – opisan je u okviru četvrtog i petog poglavlja. Postignuti rezultati i procena uspešnosti primenjene metodologije u rešavanju pomenutih zadataka detaljno su predstavljeni u ovom poglavlju. Vrednovanje učinka sistema u prepoznavanju i normalizaciji vremenskih izraza novinskih tekstova srpskog jezika vršena je na osnovu standardnih mera za procenu uspešnosti sistema za pronalaženje i ekstrakciju informacija. Na samom kraju poglavlja, izvršena je analiza grešaka, radi uočavanja propusta i ograničenja sistema, kako bi bili uzeti u obzir prilikom kasnijeg rada na unapređenju.

6.1 Primena sistema na novinske tekstove srpskog jezika

Sistem namenjen automatskoj obradi vremenskih izraza srpskog jezika primenjen je na novinske tekstove radi identifikacije i normalizacije vremenskih izraza sadržanih u ovoj vrsti dokumenata. Metodologija korišćena za potrebe prepoznavanja i normalizacije vremenskih izraza detaljno je opisana u poglavljima 4 i 5. Prepoznavanje vremenskih izraza podrazumeva identifikaciju i određivanje opsega lingvističkih izraza apsolutnog ili relativnog vremenskog značenja, koji su predstavljeni u novinskim tekstovima različitim formalnim jedinicama. Na osnovu

TimeML uputstva, osnovni semantički tipovi vremenskih izraza koje je potrebno prepoznati su oni koji ukazuju na temporalnu lokaciju u vidu kalendarskog vremena (DATE) ili vremena kao dela dana (TIME), kao i temporalnu kvantifikaciju, odnosno dužinu trajanja (DURATION) i učestalost pojavljivanja u vremenu (SET). Proces normalizacije vremenskih izraza odnosi se na interpretaciju vrednosti prepoznatih vremenskih izraza, u standardizovanom obliku koji je u skladu sa ISO 8601 standardom. Svi identifikovani vremenski izrazi obeležavaju se umetanjem `<TIMEX3>` etikete u okviru koje su definisani sledeći atributi, detaljno opisani u poglavljima 4 i 5:

- atribut `type`, kojim se specifikuje semantička klasa prepoznatog izraza;
- atribut `temporalFunction`, koji ukazuje na apsolutnu ili relativnu vrednost identifikovanog vremenskog izraza;
- atribut `value`, kojim se reprezentuju konačne vrednosti vremenskih izraza u standardizovanom obliku;
- atribut `mod`, koji se koristi za prikazivanje značenja kvantifikovanih ili modifikovanih vremenskih izraza;
- atribut `valueFromFunction`, koji se koristi u slučaju relativnih vremenskih izraza, ukazujući na računsku operaciju koju je potrebno izvršiti radi računanja apsolutne vrednosti identifikovanog izraza;
- atributi `quant` i `freq`, koji se koriste radi upotpunjavanja informacija o značenju vremenskih izraza koji ukazuju na učestalost ponavljanja u vremenu.

Nekoliko navedenih primera (primer 6.1) ilustruje tražene semantičke klase vremenskih izraza, kao i način obeležavanja rezultata, što je bogatije predstavljeno primerima u prilogu C.

Primer 6.1.
(a) *17. juna 2015*

```
<TIMEX3 type="DATE" temporalFunction="false" val="2015-06-17">
   17.  juna 2015.</TIMEX3>
```

aprila sledeće godine

```
<TIMEX3 type="DATE" temporalFunction="true" val="XXXX-04"
   valueFromFunction="+1Y">aprila sledeće godine</TIMEX3>
```

(b) *02.12.2010. godine oko 15 h*

```
<TIMEX3 type="TIME" temporalFunction="false" val="2010-12-02T15"
```

```
   mod="APPROX">02.12.2010.  godine oko 15 h</TIMEX3>
<TIMEX3 type="TIME" temporalFunction="true" val="TEV">uveče</TIMEX3>
```

(c) *manje od dve godine*

```
<TIMEX3 type="DURATION" temporalFunction="false" val="P2Y"
   mod="LESS_THAN">manje od dve godine</TIMEX3>
```

nekoliko meseci

```
<TIMEX3 type="DURATION" temporalFunction="true" val="PXM">
   nekoliko meseci</TIMEX3>
```

(d) *3 puta godišnje*

```
<TIMEX3 type="SET" temporalFunction="false" val="P1Y" freq="3X">
   3 puta godišnje</TIMEX3>
```

svake nedelje

```
<TIMEX3 type="SET" temporalFunction="false" val="P1W" quant="EVERY">
   svake nedelje</TIMEX3>
```

(e) *od 23.11.2007. do 25.11.2007.*

```
OD <TIMEX3 type="DATE.PERIOD">
   <TIMEX3 type="DATE" temporalFunction="false" val="2007-11-23">
   23.11.2007.</TIMEX3>
   DO
   <TIMEX3 type="DATE" temporalFunction="false" val="2007-11-25">
   25.11.2007.</TIMEX3>
   </TIMEX3>
<TIMEX3 type="DURATION.PERIOD">
```

9 - 13 dana

```
   <TIMEX3 type="DURATION" temporalFunction="false" val="P9D">
   9</TIMEX3>
   -
   <TIMEX3 type="DURATION" temporalFunction="false" val=„P13D">
   13 dana</TIMEX3>
   </TIMEX3>
```

6.2 Rezultati primene i evaluacija uspešnosti sistema

Za evaluaciju uspešnosti sistema u automatskoj obradi vremenskih izraza novinskih tekstova srpskog jezika korišćen je deo Korpusa savremenog srpskog je-

Tabela 6.1: Kvantitativni podaci o korpusu novinskih tekstova

Korpus	**Broj reči**	**Broj rečenica**
rts_101201	24.096	1.209
rts_101202	20.110	985
rts_101206	21.673	1.044
rts_101208	19.575	934
rts_101209	20.851	939
Ukupno	105.801	5.111

zika[1], odnosno kolekcija novinskih tekstova javnog servisa Radio televizija Srbije, prikupljenih tokom 2010. godine. Ovaj deo korpusa nije korišćen u fazi obuke sistema. Dimenzije korišćenog korpusa merene brojem reči i rečenica predstavljene su u tabeli 6.1. Podela teksta na rečenice izvršena je automatski (na način opisan u delu 4.5), te se, zbog određenog broja mogućih grešaka nastalih tokom segmentacije teksta na rečenice, dati brojevi posmatraju samo kao približne vrednosti. Nakon primene elektronskih rečnika, samo 3,8% reči teksta je ostalo nepoznato, tj. još uvek nije uključeno u elektronske rečnike standardnog srpskog jezika ili su u pitanju greške.

S obzirom na to da se među rezultatima primene bilo kog automatizovanog sistema za pronalaženje i ekstrakciju informacija javlja određeni broj grešaka, pre izvođenja zaključaka i sprovođenja daljeg istraživanja neophodno je utvrditi njihov stepen i vrstu. Kada je reč o mogućim vrstama grešaka automatizovanog sistema za obradu vremenskih izraza, one se mogu javiti i u odnosu na prepoznavanje i u odnosu na normalizaciju vrednosti prepoznatih vremenskih izraza. Tako se, pre svega, u fazi prepoznavanja javljaju greške u vidu *propuštenih* izraza (eng. *missing slots*), koji nisu, a trebalo je da budu prepoznati kao vremenski izraz; zatim, greške u vidu *uljeza* (eng. *incorrect slots*), koji predstavljaju izraze pogrešno identifikovane kao vremenske izraze; na kraju, greške u vidu *nepoklapanja* opsega izraza koji je trebalo prepoznati u odnosu na opseg prepoznat sistemom ili nepoklapanja u određivanju tipa vremenskog izraza (eng. *spurious slots*). Tokom normalizacije vrednosti vremenskih izraza, moguće su greške koje se odnose na nepoklapanja u vidu pogrešno definisanih ili izostavljenih vrednosti atributa koji se koriste za interpretaciju vrednosti prepoznatog vremenskog izraza. U pogledu izvora grešaka, to mogu biti greške nastale usled grešaka u samom tekstu ili izostavljenih ili ne-

[1]Dostupno na: http://korpus.matf.bg.ac.rs/index.html.

Tabela 6.2: Opšte vrednosti atributa korišćenih za proveru uspešnosti sistema

Vrsta greške	**Vrednost**	**Značenje**
–	OK	Korektno prepoznat i normalizovan vremenski izraz
nepoklapanja	UOK	Opseg ili neki od atributa nisu ispravno definisani ili su propušteni
	UOK/E	Opseg ili neki od atributa nisu ispravno definisani ili su propušteni zbog greške u tekstu
uljezi	NOK	Neispravno definisani svi atributi (NOKp – prepoznavanja, NOKn – normalizacije)
propušteni	MISS	Propušten vremenski izraz
	MISS/E	Propušteno zbog greške u tekstu

dovoljno precizno definisanih pravila korišćene gramatike.

Sva automatski izvršena prepoznavanja i dodeljene normalizovane vrednosti su proverene, a rezultat provere je zabeležen dodavanjem atributa proveraP i proveraN u okviru postavljene <TIMEX3> etikete, s ciljem definisanja vrste i izvora nastale greške (tabela 6.2), dok se neprepoznatim izrazima dodeljuju <TIMEX3> etikete s atributom proveraP, čija je vrednost MISS. Proveru je vršio jedan anotator (diplomirani bibliotekar-informatičar sa iskustvom u evaluaciji sistema za ekstrakciju imenovanih entiteta) na osnovu iscrpnog uputstva za rad, uz kontrolu autora primenjenog sistema. U slučaju ispravno izvršenog prepoznavanja i normalizacije atributima proveraP, odnosno proveraN, se pripisuje vrednost OK. U situacijama kada ni jedan od atributa, koji se dodeljuju u fazi prepoznavanja ili normalizacije, nije ispravno definisan, atributima korišćenim za proveru se pripisuje vrednost NOK. S obzirom na strukturu sistema, koja ne daje mogućnost normalizovanja vrednosti vremenskih izraza koji prethodno nisu prepoznati, greška u vidu propuštenih izraza zbog greške u tekstu može se javiti samo u vezi sa prepoznavanjem vremenskih izraza, kada se atributu proveraP dodeljuje vrednost MISS/E.

U slučaju grešaka nepoklapanja u vidu pogrešno definisanih ili izostavljenih vrednosti atributa <TIMEX3> etikete, atributu koji se koristi za proveru (proveraP ili proveraN) se pripisuje vrednost UOK, uz oznaku atributa koji je pogrešno određen. Detaljan prikaz oznaka <TIMEX3> atributa, korišćenih za potrebe evaluacije ilustrovan je u okviru tabele 6.3. Na primer, ako je pogrešno određen tip vremenskog izraza, atributu za proveru prepoznavanja proveraP biće dodeljena vrednost UOKt, dok će u slučaju pogrešno određenih atributa mod i valueFromFunction atri-

Tabela 6.3: Oznake <TIMEX3> atributa, korišćene u evaluaciji

Atribut	**Oznaka**
type	t
opseg	o
value	v
mod	m
valueFromFunction/funkcija	f
quant	q
freq/učestalost	u

butu proveraN biti dodeljena vrednost UOKmf. Ako su pogrešno definisan opseg ili neispravno određen tip prepoznatog izraza rezultat greške u tekstu, atributu proveraP biće dodeljena vrednost UOKo/E, odnosno UOKt/E.

Na osnovu svega ovoga, moguće vrednosti atributa proveraP su sledeće:

- OK – ispravno određen tip i precizno određen pun opseg vremenskog izraza;
- UOK – neki od atributa nije ispravan;

  UOKt (tip izraza nije ispravno određen, ali pun opseg jeste ispravno definisan);

  UOKt/E (tip izraza nije ispravno određen zbog greške u tekstu, ali pun opseg jeste ispravno definisan);

  UOKo (tip izraza je tačno određen, ali pun opseg nije ispravno definisan);

  UOKo/E (tip izraza je tačno određen, ali pun opseg nije ispravno definisan zbog greške u tekstu);
- NOKp – neispravno definisani i tip i opseg vremenskog izraza;
- NOK – izraz koji je pogrešno definisan i obeležan kao vremenski izraz, a nije trebalo (prepoznato je nešto što nije vremenski izraz);
- MISS – propušten izraz (nije prepoznat, a trebalo je da bude);
- MISS/E – izraz nije prepoznat zbog greške u tekstu.

Moguće vrednosti atributa proveraN su sledeće:

- OK – ispravno je izvršena normalizacija;

- `UOK` – neki od atributa nije ispravan

  `UOKv` (atribut `value` nije ispravan);

  `UOKm` (atribut `mod` nije ispravan);

  `UOKf` (atribut `valueFromFunction` nije ispravan);

  `UOKmf` (atributi `mod` i `valueFromFunction` nisu ispravni);

  `UOKvf` (atributi `value` i `valueFromFunction` nisu ispravni);

  `UOKvm` (atributi `value` i `mod` nisu ispravni);

  `UOKq` (atributi `quant` nije ispravan);

  `UOKu` (atributi `freq` nije ispravan);

  `UOKvq` (atributi `value` i `quant` nisu ispravni);

  `UOKvu` (atributi `value` i `freq` nisu ispravni);

- `NOKn` – neispravno definisani svi atributi normalizacije.

Primeri korektnog obeležavanja vremenskih izraza, kao i obeležavanja mogućih grešaka dati su u prilogu D (primeri D.1 i D.2).

Rezultati prebrojavanja pokazuju da se od ukupnog broja postojećih vremenskih izraza (2.050) u korišćenom korpusu novinskih narativnih tekstova javlja 1.967 (96%) vremenskih izraza koja ukazuju na datum, vreme, trajanje i učestalost, kao i 83 (4%) izraza kojima su označeni vremenski periodi (tabela 6.4). U tekstu je identifikovano 1.220 (59,5%) apsolutnih vremenskih izraza, dok 713 (34,8%) pripada grupi relativnih izraza koji nose značenje tačke u vremenu i trajanja (DATE, TIME i DURATION). Kada je reč o izrazima koji ukazuju na učestalost ponavljanja u vremenu (SET), u korpusu korišćenih tekstova ih je pronađeno svega 34 (1,7%).

Procena uspešnosti sistema u prepoznavanju i normalizaciji vremenskih izraza novinskih tekstova srpskog jezika vršena je na osnovu standardnih mera za procenu učinka sistema za pronalaženje i ekstrakciju informacija. Tokom druge u nizu MUC-konferencija (1989), o kojima je bilo više reči u delu 3.3.1, postignut je dogovor u pogledu metrike koja će se koristiti za potrebe evaluacije sistema namenjenih ekstrakciji informacija. Od tada se za procenu uspešnosti koriste dve vrednosti – preciznost (engl. *precision*) i odziv (engl. *recall*). Osnovnu jedinicu evaluacije

Tabela 6.4: Učešće tipova vremenskih izraza u korpusu novinskih tekstova

Korpus	DATE			TIME			DURATION			SET	Ukupno	DATE. PERIOD	DURATION. PERIOD	Ukupno
	abs	rel	Ukupno	abs	rel	Ukupno	abs	rel	Ukupno					
rts_101201	148	140	288	6	17	23	82	14	96	11	**418**	17	2	**437**
rts_101202	168	85	253	12	10	22	95	20	115	8	**398**	15	2	**418**
rts_101206	136	113	249	10	22	32	66	12	78	4	**363**	10	0	**373**
rts_101208	156	119	275	18	10	28	102	10	112	1	**416**	21	0	**437**
rts_101209	118	100	218	12	20	32	91	21	112	10	**372**	11	2	**385**
Ukupno	726	557	1.283	58	79	137	436	77	513	34	**1.967**	74	9	**2.050**

nad kojom se vrši procena uspešnosti ovog sistema predstavlja kompletan vremenski izraz. Preciznost (p) sistema izražava odnos ukupnog broja ispravno prepoznatih vremenskih izraza (OK) prema ukupnom broju prepoznatih izraza (M) (formula 6.1). Na taj način, preciznost ukazuje na tačnost s kojom sistem radi. S druge strane, odziv (r) predstavljen kao odnos ukupnog broja ispravno prepoznatih vremenskih izraza (OK) prema ukupnom broju postojećih izraza u tekstu (N) (formula 6.2), ukazuje na to koliko je sistem sveobuhvatan tokom ekstrakcije relevantnih informacija.

$$p = \frac{OK}{M} = \frac{OK}{OK + UOK + NOK} \tag{6.1}$$

$$r = \frac{OK}{N} = \frac{OK}{OK + MISS} \tag{6.2}$$

Bez obzira na značaj koji obe ove vrednosti – i preciznost i odziv – imaju tokom procene uspešnosti sistema, u velikom broju slučajeva nije moguće istovremeno ih optimizovati, budući da se nalaze u obrnuto proporcionalnom odnosu. Ako odgovor sistema teži da bude sveobuhvatan, to će uticati na smanjenje njegove preciznosti; i obrnuto, ako odgovor sistema teži da bude precizniji, to će negativno uticati na njegovu potpunost. Međutim, u zavisnosti od namene sistema, prilikom njegovog kreiranja moguće je dati veću važnost jednom od ova dva kriterijuma. Ako će se ekstrahovani podaci koristiti za neka dalja istraživanja, veoma je važno da te informacije budu tačne, pa čak i ako veći deo tih informacija ostane neprepoznat. Na primer, ukoliko bi se odgovor ovog sistema namenjenog automatskom prepoznavanju i normalizaciji vremenskih izraza srpskog jezika koristio za obuku sistema iste namene, zasnovanog na mašinskom učenju, utoliko bi postizanje izuzetne preciznosti bilo značajnije u odnosu na njegovu potpunost. S druge strane, ako bi cilj primene ovog sistema bio deidentifikacija medicinskih tekstova, tj. prepoznavanje i uklanjanje svih podataka na osnovu kojih se može identifikovati ličnost pacijenta, što, između ostalog, podrazumeva prepoznavanje i modifikovanje, odnosno uklanjanje i vremenskih izraza, onda je odziv daleko značajniji u odnosu na preciznost. Pogrešno određena semantička klasa kojoj prepoznati vremenski izraz pripada ne otkriva podatak o datumu nekog medicinski relevantnog događaja, ali doprinosi potpunosti sistema, bez koje je teško uspešno izvršiti deidentifikaciju.

Kako bi bilo moguće porediti učinak različitih sistema u odnosu na obe vrednosti, ustanovljena je još jedna mera – F mera, koja kao zajednička mera preciznosti i odziva, odražava njihovu geometrijsku sredinu i može se izraziti na sledeći način (formula 6.3):

$$F = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}, \tag{6.3}$$

pri čemu je *p* preciznost, *r* odziv, a β težinski koeficijent. Parametar β se koristi za održavanje ravnoteže između preciznosti i odziva. Ako je $\beta = 1$, preciznost i odziv imaju istu težinu, pa je *F* mera balansirana; kada je $\beta > 1$, značajnija je preciznost, dok za $\beta < 1$ veću težinu ima odziv. U evaluaciji opisanog sistema korišćena je balansirana F_1 mera, koja ne daje prednost ni odzivu ni preciznosti (formula 6.4).

$$F_1 = 2\frac{pr}{p + r} \tag{6.4}$$

U pogledu uspešnosti sistema u prepoznavanju vremenskih izraza, odnosno određivanju opsega i tipa, sistem je postigao veoma visoku preciznost (99%) u odnosu na odziv (80%), uz ukupnu F_1 meru od 88% (tabela 6.5). Sistem je postigao najveću preciznost (100%) u određivanju opsega i tipa izraza koji ukazuju na trajanje i period vremena trajanja, tj. predstavljaju nizanje više vremenskih izraza trajanja koji su u određenom odnosu, dok je istovremeno odziv prilikom identifikacije ovih klasa izraza najniži, uz postignutu F_1 meru od 72%, odnosno 78%. Takođe, dosta niži odziv (81%) u odnosu na postignutu preciznost (98%) javlja se i u slučaju izraza koji ukazuju na kalendarske datume, što imajući u vidu izvor grešaka može biti objašnjeno propustima prilikom definisanja pravila za identifikaciju vremenskih izraza (videti naredni odeljak). I kada je reč o izrazima sa značenjem tačke u vremenu, odnosno vremenima dana, ostvarena je bolja preciznost (98%) prema postignutom odzivu (89%). Jednaki odziv i preciznost u slučaju izraza koji ukazuju na učestalost (90%) nedvosmisleno ukazuju na potrebu za modifikovanjem postojećih pravila, s ciljem ispravnog obeležavanja opsega identifikovanih vremenskih izraza. Detaljni podaci o uspešnosti sistema u prepoznavanju vremenskih izraza dati su u tabeli 6.5, dok će podrobna analiza grešaka biti izvršena u daljem tekstu. Kolona N ukazuje na ukupan broj vremenskih izraza koji postoje u korpusu novinskih tekstova, dok kolona M sadrži ukupan broj vremenskih izraza koje je sistem identifikovao.

Postignuti rezultati u pogledu uspešnosti procesa normalizacije vremenskih izraza su, može se reći, izuzetno dobri (tabela 6.6). Vrednosti svih korektno prepoznatih vremenskih izraza su normalizovane (postignuti odziv je 100%), dodeljivanjem svih odgovarajućih atributa neophodnih za njihovu interpretaciju. S obzirom na to da su u slučaju izraza koji ukazuju na vreme dana i učestalost, zbog pogre-

šno određenog tipa, unete neodgovarajuće vrednosti atributa `value`, preciznost u vršenju normalizacije iznosi 99,7%. Kolona N ukazuje na ukupan broj vremenskih izraza koji su identifikovani sistemom u korpusu novinskih tekstova i za koje se očekuje sprovođenje procesa normalizacije, dok kolona M sadrži ukupan broj vremenskih izraza čije je vrednosti sistem normalizovao.

6.3 Analiza grešaka

Greške nastale prilikom primene sistema, kao i njegova uspešnost u prepoznavanju i normalizaciji vremenskih izraza u odnosu na postojeće semantičke klase vremenskih izraza prikazani su u okviru tabela 6.5 i 6.6. Uočene greške i postignuti preciznost, odziv i F_1 mera u pogledu određivanja opsega i dodeljivanja neophodnih atributa, bez uzimanja u obzir klase kojoj vremenski izraz pripada, ilustrovani su podacima u tabeli 6.7. Kolona N ukazuje na ukupan broj vremenskih izraza koji postoje u korpusu novinskih tekstova i za koje je očekivano određivanje opsega i dodeljivanje atributa, dok kolona M sadrži ukupan broj vremenskih izraza kojima je sistem odredio opseg i dodelio atribute.

Imajući u vidu postignutu uspešnost sistema, ilustrovanu F_1 merama, najveći problem u procesu automatske obrade vremenskih izraza, identifikovan je već u prvom koraku, prilikom prepoznavanja izraza. Izrazi koje sistem nije identifikovao i obeležio prilikom obrade čine 20% ukupnog broja postojećih vremenskih izraza u korišćenom korpusu (N). Većim delom (95,4%) su propušteni izrazi oni koji ukazuju na kalendarske datume (235) i trajanja (143), dok su propusti prevashodno rezultat primene pravila, koja zarad veće preciznosti sistema, nisu omogućavala identifikaciju numeričkih izraza koji u kontekstu novinskih tekstova mogu da budu izuzetno višeznačni (primer 6.2).

Primer 6.2.

```
{S}<TIMEX3 proveraP="MISS" type="DATE">1820.</TIMEX3>{S} - Knez Miloš
   Obrenović odobrio ...
Strelci su bili Spektor u <TIMEX3 proveraP="MISS" type="DURATION">
   22.</TIMEX3> i <TIMEX3 proveraP="MISS" type="DURATION">
   27.</TIMEX3> i Kol u <TIMEX3 proveraP="MISS" type="DURATION">
   56.</TIMEX3> i <TIMEX3 proveraP="MISS" type="DURATION">
   66.  minutu</TIMEX3>
```

Tabela 6.5: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u prepoznavanju vremenskih izraza novinskih tekstova na osnovu postojećih semantičkih klasa izraza

	N	**M**	**OK**	**UOKt**	**UOKo**	**UOKo/E**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
DATE	1.283	1.048	1.032	1	8	1	231	4	6	0,98	0,81	0,89
TIME	137	122	119	1	0	0	12	3	2	0,98	0,89	0,93
DURATION	513	370	369	0	1	0	141	2	0	1	0,72	0,84
SET	34	31	28	0	0	0	3	0	3	0,90	0,90	0,90
Ukupno	1.967	1.571	1.548	2	9	1	387	9	11	**0,99**	**0,80**	**0,88**
DATE.PERIOD	74	64	63	0	0	0	9	1	1	0,98	0,86	0,92
DURATION.PERIOD	9	7	7	0	0	0	2	0	0	1	0,78	0,88
Ukupno	2.050	1.642	1.618	2	9	1	398	10	12	**0,99**	**0,80**	**0,88**

Tabela 6.6: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u normalizaciji vremenskih izraza novinskih tekstova na osnovu postojećih semantičkih klasa izraza

	N	**M**	**OK**	**UOKv**	**UOKm**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
DATE	1.048	1.048	1.047	0	1	0	0	0	1	1	1
TIME	122	122	120	4	0	0	0	0	0,97	1	0,98
DURATION	370	370	369	0	1	0	0	0	1	1	1
SET	31	31	30	1	0	0	0	0	0,97	1	0,98
Ukupno	1.571	1.571	1.566	5	2	0	0	0	**0,997**	**1**	**1**

Tabela 6.7: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u određivanju opsega i atributa vremenskih izraza novinskih tekstova

	N	**M**	**OK**	**UOK**	**UOKo/E**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
opseg	2.050	1.642	1.620	9	1	398	10	12	0,99	0,80	0,88
`type`	1.642	1.642	1.628	2	0	0	0	12	0,99	1	1
`temporalFunction`	1.571	1.571	1.571	0	0	0	0	0	1	1	1
`value`	1.571	1.571	1.568	5	0	0	0	0	0,997	1	1
`mod`	42	42	41	1	0	0	0	0	0,98	1	0,99
`valueFromFunction`	120	120	119	1	0	0	0	0	0,99	1	1
`quant`	6	6	6	0	0	0	0	0	1	1	1
`freq`	0	0	0	0	0	0	0	0	N/A	N/A	N/A

Propusti u identifikaciji vremenskih izraza, nastalih usled postojećih tipografskih grešaka ili grešaka koje su rezultat pogrešne segmentacije teksta na rečenice, čine samo 2,3% ukupnog broja izostavljenih izraza. Neki slučajevi ilustrovani su primerom 6.3.

Primer 6.3.

```
<TIMEX3 proveraP="MISS/E" type="DATE">četvratak</TIMEX3>
Program:  {S}<TIMEX3 proveraP="MISS/E" type="TIME">21.{S}00</TIMEX3>
   Izmišljene ljubavi
```

Preciznost sistema u identifikovanju vremenskih izraza i određivanju njihovog opsega veoma je visoka (99%). Analizom grešaka koje ukazuju na neispravno određen opseg identifikovanog vremenskog izraza uočeno je da 90% slučajeva zaista jesu greške sistema (primer 6.4), dok samo jedan slučaj predstavlja grešku nastalu usled postojećih pravopisnih i tipografskih grešaka u tekstu.

Primer 6.4.

```
<TIMEX3 proveraP="UOKo" type="DATE" temporalFunction=""true"
   val="XXXX-WXX-5">Petak</TIMEX3> 03.12.
```

Greške u određivanju tipa prepoznatog vremenskog izraza javile su se u samo dva slučaja, i predstavljaju primere izraza koji ukazuju na trajanje, a koji su pogrešno prepoznatih kao izraz sa značenjem tačke u vremenu (primer 6.5).

Primer 6.5.

```
složio program od <TIMEX3 proveraP="UOKt" proveraN="UOKv" type="TIME"
   temporalFunction="false" val="T02">dva sata</TIMEX3>
```

Otkrivanje i rešavanje grešaka u vidu *uljeza*, odnosno izraza koji su pogrešno identifikovani kao vremenski izrazi, posebno je značajno u obradi narativnih tekstova, pogotovo ako je odgovor sistema namenjen za neko dalje istraživanje. Analizom grešaka nakon primene sistema uočeno je 12 pogrešnih prepoznavanja ovog tipa (primer 6.6), gde se prevashodno radi o pogrešnom prepoznavanju ličnih imena kao vremenskih izraza. Kako bi se izbegle greške ovoga tipa, pre primene sistema za prepoznavanje i normalizaciju vremenskih izraza, potrebno je primeniti graf koji prepoznaje lična imena, a koji je deo sistema za prepoznavanje

imenovanih entiteta srpskog jezika (Krstev et al. 2014).

Primer 6.6.

```
<TIMEX3 proveraP="NOK" type="DATE" temporalFunction="true"
   val="XXXX-03">Marte</TIMEX3> Dominges.
<TIMEX3 proveraP="NOK" type="DATE" temporalFunction="true"
   val="XXXX-05">Maja</TIMEX3> Nikolić
```

Kada je reč o ostalim atributima koji su definisani TimeML uputstvom, sistem je napravio samo pet grešaka prilikom određivanja vrednosti najvažnijeg atributa u procesu normalizacije tj. atributa `value`. Ove greške su prevashodno rezultat pogrešno određenog opsega ili tipa vremenskog izraza (primer 6.7), na osnovu kog je i izvršena normalizacija vrednosti prepoznatog izraza.

Primer 6.7.

```
domaćin 20-<TIMEX3 proveraP="NOK" proveraN="UOKv" type="SET"
   temporalFunction="false" val="P1Y">godišnje</TIMEX3> sednice
```

Na osnovu sračunate uspešnosti sistema u vidu tradicionalne F_1 mere, mera greške sistema računa se kao E=1-F_1. Međutim, kako je prilikom izračunavanja harmonijske sredine odziva i preciznosti važnost grešaka u vidu uljeza i propuštenih izraza umanjena (tako da ukupna stopa greške nikada ne može da bude veća od stope greške bilo kog tipa grešaka posebno), za potrebe evaluacije u okviru ovog rada sprovedeno je i utvrđivanje ukupne mere svih tipova grešaka na osnovu stope greške slota (eng. *slot error rate*, *SER*) (Makhoul et al. 1999). Kao jednostavna mera greške, SER je jednak količniku zbira sva tri tipa grešaka (propuštenih – `MISS`, uljeza – `NOK` i nepoklapanja – `UOK`) i ukupnog broja postojećih izraza u referentnom korpusu (N) (Formula 6.5).

$$SER = \frac{MISS + NOK + UOK}{N} = \frac{MISS + NOK + UOK}{OK + UOK + MISS} \tag{6.5}$$

Na osnovu podataka u okviru tabela 6.8, 6.9 i 6.10 jasno je da je stopa greške sračunata na osnovu *SER*-a za oko 50% viša u odnosu na stopu greške predstavljenu putem F_1 mere. Dobijene vrednosti *SER*-a u većoj meri ističu greške sistema u procesu prepoznavanja i normalizacije vremenskih izraza novinskih tekstova.

Tabela 6.8: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u prepoznavanju vremenskih izraza novinskih tekstova na osnovu postojećih semantičkih klasa izraza

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	1.032	10	235	6	0,89	0,11	0,20
TIME	119	1	15	2	0,93	0,07	0,13
DURATION	369	1	143	0	0,84	0,16	0,28
SET	28	0	3	3	0,90	0,10	0,19
Ukupno	1.548	12	396	11	0,88	0,12	**0,21**
DATE.PERIOD	63	0	10	1	0,92	0,08	0,15
DURATION.PERIOD	7	0	2	0,88	0,13	0	0,22
Ukupno	1.618	12	408	12	0,88	0,12	**0,21**

Tabela 6.9: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u normalizaciji vremenskih izraza novinskih tekstova na osnovu postojećih semantičkih klasa izraza

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	1.047	1	0	0	1	0	0
TIME	120	4	0	0	0,98	0,02	0,03
DURATION	369	1	0	0	1	0	0
SET	30	1	0	0	0,98	0,016	0,03
Ukupno	1.566	7	0	0	1	0,002	**0,004**

6.4 Zaključak

Rezultati sprovedene evaluacije ukazuju na činjenicu da kreiran sistem može da se koristi za potrebe automatskog prepoznavanja i normalizacije vremenskih izraza srpskog jezika, uz postizanje veoma visoke preciznosti (99%). Na osnovu analize grešaka otkriven je određeni broj propusta, koji utiču na smanjeni odziv (80%) a koji su mahom rezultat nepostojanja pravila ili, pak, primene pravila, koja zarad veće preciznosti sistema, nisu omogućavala identifikaciju određenih izraza, izuzetno višeznačnih u kontekstu novinskih tekstova. Modularnost sistema omogućava da se, u zavisnosti od namene, određena pravila uključuju u proces identifikacije po potrebi. Kada je reč o normalizaciji prepoznatih izraza, identifikovane greške ukazuju na to da ovaj proces u potpunosti zavisi od uspešnosti prepoznavanja, odnosno ispravnog određivanja opsega i tipa vremenskih izraza, budući da su vrednosti svih korektno identifikovanih i obeleženih izraza na ispravan način i

normalizovane.

Tabela 6.10: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u određivanju opsega i atributa vremenskih izraza

	OK	**UOK**	**UOK/E**	**MISS**	**MISS/E**	**NOK**	F_1	***E***	***SER***
opseg	1.620	9	1	398	10	12	0,88	0,12	**0,21**
`type`	1.628	2	0	0	0	12	1	0,004	**0,009**
`temporalFunction`	1.571	0	0	0	0	0	1	0	0
`value`	1.568	5	0	0	0	0	1	0,002	**0,003**
`mod`	41	2	0	0	0	0	0,98	0,02	0,05
`valueFromFunction`	119	0	0	0	0	0	1	0	0
`quant`	6	0	0	0	0	0	1	0	0
`freq`	0	0	0	0	0	0	N/A	N/A	N/A

Glava 7

Prepoznavanje i normalizacija vremenskih izraza medicinskih narativnih tekstova

Vreme, shvaćeno kao odraz promena fizičke stvarnosti u ljudskoj svesti, odnosno način na koji ljudski um opaža i tumači događaje, bitno određuje čovekovo poimanje sveta. Kao jedna od osnovnih, vremenska dimenzija predstavlja središnji aspekt ne samo našeg svakodnevnog života, već i sagledavanja i razumevanja promena koje se dešavaju i problema koji nastaju i u okviru različitih, specifično određenih domena. Tako se i u oblasti medicine vreme javlja kao jedan od suštinskih koncepata (Shahar 1999; Augusto 2005; Zhou and Hripcsak 2007; Reeves et al. 2013), pružajući osnovu za pravilnu interpretaciju i razumevanje medicinski relevantnih informacija. Redosled razvijanja određenih simptoma, pravo vreme primene različitih terapija, kao i trajanje i učestalost korišćenja lekova značajni su samo u određenom vremenskom kontekstu. Tako je, na primer, tokom postavljanja dijagnoze veoma bitno znati hronološki redosled pojavljivanja određenih simptoma ili dužinu njihovog trajanja. Gotovo da nije moguće adekvatno predstavljanje klinički relevantnih podataka, kao ni ispravno zaključivanje i donošenje odluka na osnovu njih, bez uzimanja u obzir vremenske dimenzije. Različite medicinske intervencije se dešavaju u jednoj ili više tačaka u vremenu (npr. *hirurška intervencija zakazana za **29.02.2000. g. u 9 h***), kao što i određeni podaci, poput rezultata laboratorijskih testova ili podataka o propisanim terapijama i postavljenim dijagnozama, važe u okviru određenog, eksplicitno ili implicitno iskazanog vremenskog perioda (npr. *po dobijanju gore navedenog rezultata ordinirana je infuzija Tygacil 50 mg **10 dana**, tokom **hospitalizacije** ordinirana konzervativna terapija*). Osim toga, pomenuti klinički relevantni podaci i predložene ili izvršene

medicinske intervencije, kao medicinski značajni koncepti, često se nalaze u određenim međusobnim vremenskim odnosima (npr. *potrebno nekoliko HD dijaliza* ***posle*** *implantacije grafta*), koje je potrebno otkriti i razrešiti radi ispravnog utvrđivanja hronologije događaja ili stanja, što omogućava, na primer, uspešno praćenje toka razvoja bolesti i efikasnosti primenjene terapije.

7.1 Značaj automatske obrade vremenskih informacija medicinskih narativnih tekstova

Posebno sa stanovišta izgradnje i korišćenja medicinskih informacionih sistema, vreme je izuzetno važno za reprezentovanje informacija sadržanih u njima. Takve informacije se odnose na pacijente i sadrže tačne i iscrpne podatke o medicinskim stanjima, postavljenim dijagozama, tokovima i ishodima lečenja. Savremeni napredak u oblasti medicinskih informacionih tehnologija omogućio je automatizaciju većine procesa koji se odnose na pružanje usluga zdravstvene zaštite, olakšavajući prikupljanje i čuvanje informacija o pacijentima. Stvarna potreba za čuvanjem ovih podataka, koji su pravilno orijentisani u okviru vremenske dimenzije, zaista postoji, pre svega, zbog mogućnosti njihove kasnije upotrebe u nekom drugom kontekstu, radi unapređenja kvaliteta usluga zdravstvene nege, dijagnostičkih i teorijskih procesa u medicini.

Elektronski zdravstveni kartoni sadrže određenu količinu informacija, koja je data u strukturiranom obliku, što podrazumeva unapred definisane tipove podataka i relacije koje postoje među njima (npr. lično ime pacijenta, lični broj zdravstvenog osiguranja pacijenta, datum posete lekaru itd.), čime je omogućena njihova mašinska obrada. Međutim, sa stanovišta potrebe za izgradnjom inteligentnijeg sistema, koji može obezbediti podršku u odlučivanju, ova vrsta podataka predstavlja oskudan izvor, budući da im nedostaju ključne informacije koje se odnose na tokove i ishode kliničkog lečenja. Ove ključne informacije, kao što su, na primer, stanje bolesti i tok njenog razvoja, zapravo su često zabeležene samo u vidu nestrukturiranih tekstova (npr. kliničkih beleški lekara), koji sadrže obilje klinički relevantnih informacija, predstavljenih raznorodnim, nestandardnim oblicima prirodnog jezika. S ciljem njihovog pronalaženja i ekstrakcije iz ovakvih nestrukturiranih medicinskih tekstova primenjuju se tehnologije obrade prirodnih jezika. Kako bi ekstrahovani medicinski koncepti bili i hronološki organizovani i u tom kontekstu upotrebljeni, neophodno je obezbediti i automatsku ekstrakciju i in-

terpretaciju vremenskih informacija iz medicinskih tekstova. Način predstavljanja i zaključivanje na osnovu vremenski orijentisanih kliničkih podataka, podjednako je važno, kako za ustanove koje pružaju usluge zdravstvene zaštite ili stručnjake iz oblasti medicine koji se bave naučno-istraživačkim radom, tako i za automatizovane sisteme namenjene opštem pretraživanju elektronskih zdravstvenih kartona pacijenata ili podršci u odlučivanju prilikom npr. postavljanja dijagnoze, određivanja terapije ili procene njene efikasnosti. Na primer, nakon primene „sublingvalne alergen specifične imunoterapije“, radi procene kliničke efikasnosti u lečenju alergijskog rinitisa i astme kod dece, bilo bi korisno znati da li su, na primer, simptomi nazalne kongestije i rinoreje statistički značajno smanjeni već nakon prvog merenja, u okviru prvih šest meseci od početka primene terapije.

7.2 Priroda medicinskih narativnih tekstova i izazovi u automatskoj obradi

Automatska obrada vremenskih izraza medicinskih tekstova podrazumeva proces identifikacije vremenskih izraza i njihovo formalizovanje na jezik razumljiv računarima, što kao i u slučaju vremenskih izraza opštih, novinskih tekstova nije jednostavan zadatak. Na složenost ovog problema, osim osnovnih pitanja koja se odnose na prirodu pojma vremena i način modelovanja njegove strukture (npr. linearni, granajući ili cirkularni), utiču i određene poteškoće koje proizilaze iz specifičnosti domena i korišćenog jezika, karakterističnog za ovu vrstu tekstova.

Kliničke beleške lekara, kao jedan od najznačajnijih oblika medicinskih narativnih tekstova, predstavljaju pisane izveštaje lekara o poseti, razgovoru i pregledu pacijenta, u cilju prikupljanja svih informacija neophodnih za otkrivanje bolesti i tačno postavljanje dijagnoze. Njihova forma često podrazumeva postojanje više jasno izdvojenih celina, koje se odnose na različite aspekte zdravstvene nege i prisutne bolesti kod pacijenta (npr. opšti administrativni podaci, kratak opis zdravstvenog problema, podaci o ličnoj prošlosti pacijenta tj. ranijim bolestima, porodična anamneza, podaci o sadašnjoj bolesti i sl.). Iako su pojedini termini i zaključci karakteristični samo za određeni tip kliničke prakse (kao što je npr. onkologija), nezavisno od institucije ili medicinske specijalnosti poštuje se određeni obrazac redosleda unošenja podataka. Tako, na primer, u okviru jednog od najznačajnijih delova anamneze, koji se odnosi na podatke o sadašnjoj bolesti (lat. *anamnesis morbi*), lekar prirodnim jezikom opisuje glavne tegobe pacijenta, kao i

sve intervencije i dijagnostičke procedure koju su do tog trenutka bile preduzete (primer 7.1).

Primer 7.1.

ANAMNESIS MORBI
navodi da je promenu u vidu ranice u predelu poda usta primetio aprila meseca 2012 godine. U ovoj Klinici operisan jula meseca 2012 godine kada je učinjena OP Dissectio colli supraomohyoidea bill. Excisio tu baseos oris lat. sin ressectio mandibulae segmentalis reg symphiseos. HP br 1413-17 Ca planocellulare. Supraomohioidni disektat levo 1/4. Tumor erodira ali ne infiltriše kost. Linije resekcije su negativne. Postoperativno sprovedena zračna terapja.Na jednoj od kontrola jugulodigastrično sa leve strane, primećena promena, učinjen EHO, opservacija., potom zakazan za prijem. Učinjena OP Extirpatio tu regio colli lat sin. Hp br 1813/13 Carcinoma planocellulare infiltrativum. Potom pacijent prikazan konzilijumu za MF regiju IORKCS., gde je odlučeno da se sprovede radikalne disekcija vrata sa leve strane.

Kliničke beleške lekara sadrže obilje vremenskih informacija i činjenica koje se navode hronološkim redosledom, od prethodnih terapija i simptoma, preko tekućeg stanja do budućih intervencija. Kako bi se na osnovu prikupljenih podataka i analize omogućilo izvođenje pouzdanog i pravog zaključka, neophodno je da ovi tekstovi sadrže što više podrobnih detalja, čije beleženje oduzima ogromno vreme, koje lekari retko imaju. Budući da ove tekstove pišu lekari za lekare i druge medicinske stručnjake, koji poseduju isto profesionalno znanje, izuzetno je česta upotreba raznih nestandardnih izraza i skraćenica, koje istovremeno obezbeđuju sažet, ali i dovoljno detaljan opis. Tako se i za iskazivanje vremena, osim već ranije opisanih raznorodnih oblika o kojima je bilo reči u 2. poglavlju, u medicinskim narativnim tekstovima često koriste određeni akronimi vremenskog značenja (npr. qid od lat. *quarter in die* - četiri puta dnevno, bid od lat. *bis in die* – dva puta dnevno, *post op* itd.). I u slučaju vremenskih izraza medicinskih narativnih tekstova, ista vremenska informacija može biti saopštena na razne načine (npr. *3. januara 2007*, *03.01.07*, *3.01.2007*). Opsežan pregled i klasifikacija vremenskih izraza pronađenih u otpusnim listama, kao jednoj od vrsta medicinskih narativnih tekstova, dati su u radu (Zhou et al. 2006). Ipak, imajući u vidu da se radi o vrsti tekstova pisanih u brzini posebnim, medicinskim jezikom, koji karakteriše često ignorisanje gramatičkih i pravopisinih pravila datog jezika, broj

oblika vremenskih izraza, koje je zbog značaja u interpretaciji klinički relevantnih podataka potrebno prepoznati, postaje značajno veći (npr. *03 01 07*, *03.01,2007*, *3.01..2007*, *0'3.01.2007*). Osim apsolutnih vremenskih izraza, česta je upotreba i relativnih vremenskih izraza, čija vrednost nije eksplicitno iskazana, već im je neophodan potpuno precizan vremenski izraz koji će poslužiti kao orijentir u odnosu na koji će se računati njihova krajnja vrednost i pozicija na vremenskoj osi. Na primer, vrednost vremenskog izraza *12. marta o.g.* je relativna u odnosu na datum kada je ovaj izraz napisan, dok će u slučaju izraza *trećeg p.op dana* kao orijentir poslužiti kalendarski datum pomenute operacije.

Sledeći izazov prilikom prepoznavanja i formalizacije vremenskih izraza medicinskih narativnih tekstova predstavlja i činjenica da vremenska informacija ne mora nužno da bude izrečena eksplicitno, već je često implicitna i zahteva tumačenja izvedena na osnovu opšteg znanja, neophodnog u procesu zaključivanja. U iskazu „ranije lečena od dijabetične flegmone" sistem za obradu vremenskih informacija bi trebalo da odredi da li se „ranije" odnosi na period od npr. nedelju ili godinu dana ranije, što zahteva više nivoe analize i razrešavanje koreferenci, što predstavlja nerešen problem o kome se još uvek nije raspravljalo na adekvatan način u kontekstu obrade medicinskih narativnih podataka (Sun, Rumshisky, and Uzuner 2013b).

S obzirom na prirodu jezika medicinskih narativnih tekstova, koje karakteriše nedosledna primena gramatičkih i pravopisnih pravila, česta upotreba skraćenica i delova tekstova koji su kopirani iz drugih tekstova (Meystre et al. 2008), potrebno je sakupiti i analizirati što veći broj raznorodnih tipova medicinskih narativnih tekstova, kako bi se omogućio razvoj sistema koji bi sa visokim odzivom i preciznošću vršio automatsku analizu ove vrste tekstova. Međutim, ni to nije tako jednostavan zadatak, imajući u vidu potrebu zaštite prava na poverljivost podataka o zdravstvenom stanju pacijenata. Osim velike količine korisnih, medicinski relevantnih informacija, narativni tekstovi ovog domena takođe sadrže i mnoge detalje koji se odnose na ličnost pacijenta, odnosno njegovo zdravstveno stanje. Stoga su svi zdravstveni radnici, odnosno zdravstveni saradnici, kao i druga lica zaposlena u zdravstvenim ustanovama, privatnoj praksi, organizacionoj jedinici visokoškolske ustanove zdravstvene struke koja obavlja zdravstvenu delatnost, druga pravna lica koja obavljaju određene poslove iz zdravstvene delatnosti u skladu sa zakonom, organizacija obaveznog zdravstvenog osiguranja, kao i pravno lice koje obavlja poslove dobrovoljnog zdravstvenog osiguranja, kod kojih je pacijent zdravstveno

osiguran, dužni da čuvaju podatke iz medicinske dokumetacije, bez mogućnosti njihovog iznošenja u javnost, osim u slučaju pristanka pacijenta ili zakonskog zastupnika.[1]

7.3 Deidentifikacija medicinskih narativnih tekstova

Kada se poverljivi klinički podaci dele i koriste u istraživačke svrhe, iz etičkih i pravnih razloga neophodno je zaštiti privatnost pacijenta i ukloniti sve one identifikatore koji se odnose na njegovu ličnost, i to najčešće putem procesa deidentifikacije. Deidentifikacija je usmerena na otkrivanje i uklanjanje, odnosno modifikovanje svih eksplicitno iskazanih ličnih podataka, koji se odnose na zdravstveno stanje pacijenta (eng. *Protected Health Information*), a koji se nalaze u tekstovima medicinske dokumentacije. S druge strane, prilikom ovog procesa neophodno je voditi računa o tome da sve medicinski relevantne informacije o nekom pacijentu ostanu u tekstu sačuvane.

Različiti standardi i regulative za zaštitu zdravstvenih podataka definišu više različitih pravaca u postizanju deidentifikacije, ali najčešće korišćena regulativa jeste Zakon o odgovornosti i prenosivosti zdravstvenih informacija Sjedinjenih Američkih Država (eng. *US Health Information Portability and Accountability Act*, *HIPAA*) (HIPAA 1996). Na osnovu HIPAA pristupa, kliničke beleške se smatraju deidentifikovanim kada je 18 kategorija (17 tekstualnih i jedna koja se odnosi na slike) zaštićenih zdravstvenih informacija uklonjeno iz teksta, i kada preostale informacije ne mogu poslužiti, samostalno ili u kombinaciji jedna sa drugom, za identifikovanje neke osobe. Ove kategorije zaštićenih zdravstvenih informacija uključuju imena, geografske lokacije, elemente datuma (osim godine), brojeve telefona i faksa, brojeve zdravstvenih kartona, zdravstvenog osiguranja ili bilo koje druge jedinstvene identifikacione brojeve. Budući da se ručno uklanjanje ove vrste informacija, koje bi sproveli zdravstveni radnici i saradnici, pokazalo kao izuzetno naporno i vremenski nedostižno, monotono, skupo i nepouzdano (Douglass et al. 2004; Neamatullah et al. 2008; Deleger et al. 2013), ekstrahovanje zaštićenih zdravstvenih podataka zahteva mnogo pouzdanije, brže i jeftinije automatske sisteme za deidentifikaciju, koji su zasnovani na metodama obrade prirodnih jezika (Meystre et al. 2010).

[1] Član 21, Zakon o pravima pacijenata (Sl. glasnik RS, broj 45/13).

Ektrakcija zaštićenih zdravstvenih informacija može se posmatrati kao problem prepoznavanja imenovanih entiteta, koji je primenjen u medicinskom domenu radi deidentifikacije (Nadeau and Sekine 2007). Ipak, iako i tradicionalni zadatak prepoznavanja imenovanih entiteta, kao i deidentifikacija, uključuju automatsko prepoznavanje određenih izraza u tekstu (osobe, lokacije, datumi itd.), deidentifikacija se značajno razlikuje od tradicionalnog zadatka prepoznavanja imenovanih entiteta (Wellner et al. 2007). Osim što se opšti i medicinski tekstovi prirodnog jezika stilski veoma razlikuju, način interpretacije identifikovanih podataka sadržanih u njima takođe zahteva drugačiji pristup, pogotovo kada je reč o vremenskim izrazima. Od svih ranije opisanih semantičkih klasa vremenskih izraza (poglavlje 4.2) jedino se kalendarski datumi i izrazi koji ukazuju na starost neke osobe posmatraju kao informacije koje je potrebno zaštititi procesom deidentifikacije. Prisutni elementi datuma koji se odnose na dan i mesec, osim što bi trebalo da budu automatski identifikovani i obeleženi u tekstu, u okviru postavljene etikete trebalo bi da sadrže normalizovanu vrednost kalendarskog datuma, ali ne onog koji je prepoznat u tekstu, već datuma koji je u odnosu na prepoznati datum pomeren za određeni nasumično izabran interval, kako bi se onemogućilo eventualno utvrđivanje identiteta pacijenta na osnovu ovog i drugih podataka iz teksta. S obzirom na to da podatak o godini igra važnu ulogu u kliničkom kontekstu, ovaj element kalendarskog datuma treba da ostane sačuvan. Na primer, nakon izvršene deidentifikacije kalendarski datum *5. mart 1998. godine* trebalo bi da bude predstavljen u tekstu nekom vrednošću kojom će elementi dana i meseca biti prikriveni, kao što je npr. *17. april 1998. godine*. U tom slučaju potrebno je obratiti posebnu pažnju na očuvanje postojećih intervala između događaja u tekstu, tako što će period pomeranja biti dosledan kroz ceo dokument. Kada je reč o izrazima koji ukazuju na starost neke osobe, u slučaju osoba starijih od 89 godina potrebno je identifikovati i obeležiti izraz, uz definisanje normalizovane vrednosti koja će samo dati uvid u to da se radi o osobama starosti od 90 do 120 godina.

S obzirom na to da je deidentifikacija prvi korak ka procesu prepoznavanja i ekstrakcije drugih relevantnih kliničkih podataka, veoma je važno rešiti ovaj problem i omogućiti širu upotrebu ovih tekstova, ali na takav način da se očuva korisnost i sveukupno značenje kliničkih beleški, pa sam tim i preciznost kasnijih automatskih procesa koji se mogu sprovoditi na deidentifikovanim medicinskim tekstovima. Primer medicinskog narativnog teksta dat je u prilogu E (primer E.1), dok je deidentifikovan medicinski narativni tekst ilustrovan primerom E.2.

Tokom poslednjih dvadeset godina razvijeni su različiti pristupi za automatsku deidentifikaciju medicinskih narativnih tekstova, ali je ipak relativno mali broj njih bio usmeren na nestrukturirane podatke. Opsežan pregled skorijih istraživanja sprovedenih u oblasti deidentifikacije medicinskih narativnih tekstova dat je u (Meystre et al. 2010). Ipak, većina njih je usmerena na deidentifikaciju samo određenih tipova dokumenata ili određenih tipova zaštićenih zdravstvenih informacija, odnosno identifikatora. Bez obzira na upotrebljen metod za deidentifikaciju, većina sistema je za evaluaciju koristila samo jedan ili dva tipa medicinskih dokumenata (Neamatullah et al. 2008; Uzuner et al. 2008; Gardner et al. 2010), dok je samo nekoliko njih koristilo raznovrsniji korpus (Taira, Bui, and Kangarloo 2002; Ferrández et al. 2013). Pristupi korišćeni za rešavanje ovog zadatka u medicinskom domenu prevashodno su klasifikovani na one zasnovane na pravilima i mašinskom učenju, dok neki hibridni pristupi (Ferrández et al. 2013) efikasno koriste prednosti oba prethodno pomenuta metoda. Sistemi koji koriste metodu zasnovanu na pravilima (Neamatullah et al. 2008; Gupta, Saul, and Gilbertson 2004; Beckwith et al. 2006; Friedlin and McDonald 2008; Morrison, Lai, and Hripcsak 2009) koriste rečnike i pravila za identifikovanje zaštićenih zdravstvenih informacija, bez prethodno obeleženih podataka koji će se koristiti za obuku sistema. Iako su ovi sistemi često okarakterisani kao sistemi čija opšta primena zavisi od kvaliteta obrazaca i pravila, oni mogu jednostavno i brzo da se modifikuju dodavanjem pravila, termina u rečnike ili regularnih izraza kako bi se poboljšala ukupna uspešnost sistema (Meystre et al. 2014). S druge strane, sistemi zasnovani na mašinskom učenju (Wellner et al. 2007; Gardner and Xiong 2008; Uzuner et al. 2008; Aramaki et al. 2006; Guo et al. 2006; Aberdeen et al. 2010) su se pokazali opštijim, i automatski uče iz primera za obuku da otkriju i predvide zaštićene zdravstvene informacije. Ipak, ove metode zahtevaju velike količine anotiranih podataka i adaptacija ovog sistema bi mogla da bude teška jer je teško predvideti efekte promena. Iscrpan pregled publikovanih strategija i tehnika posebno razvijenih za deidentifikaciju medicinskih narativnih tekstova dat je u (Kushida et al. 2012). Osim sistema koji su posebno dizajnirani za potrebe deidentifikacije, neki alati za prepoznavanje imenovanih entiteta obučeni za novinske tekstove su takođe postigli pristojne performanse u pogledu deidentifikacije određenih zaštićenih zdravstvenih informacija (Wellner et al. 2007; Benton et al. 2011).

Tako je za srpski jezik sistem koji automatski vrši deidentifikaciju kliničkih narativnih tekstova (Jaćimović, Krstev, and Jelovac 2014, 2015) razvijen upravo na

osnovu adaptacije već postojećeg sistema zasnovanog na pravilima, namenjenog prepoznavanju imenovanih entiteta u prevashodno novinskim tekstovima srpskog jezika. Cilj tog rada bio je evaluacija preciznosti u uklanjanju i zameni zaštićenih zdravstvenih podataka, uz očuvanje svih medicinski relevantnih podataka o pacijentu, čime se čuva i upotrebljivost proizvedenog deidentifikovanog dokumenta za kasnije procese ekstrakcije informacija. U pogledu vremenskih izraza posmatrani su samo oni tipovi koji se smatraju za zaštićene zdravstvene informacije, čije su vrednosti normalizovane pomeranjem identifikovanih vrednosti. Pomenuti sistem je postigao obećavajuće rezultate, ostvarivši visoku uspešnost u automatskoj deidentifikaciji kliničkih narativnih tekstova (postignuta F_1 mera je iznosila 94%). Stoga je važno naglasiti da način na koji će se automatski obrađivati vremenski izrazi medicinskih tekstova zavisi, pre svega, od postavljenog zadatka (npr. deidentifikacije ili nekog drugog zadatka).

U okviru priloga E dat je primer medicinskog narativnog teksta (primer E.1), kao i primer deidentifikovanog medicinskog narativnog teksta (primer E.2).

7.4 Pregled radova iz oblasti obrade vremenskih izraza medicinskih narativnih tekstova

Automatska obrada vremenskih informacija predstavlja jednu od dominantnih oblasti obrade prirodnih jezika, iako se ne bi moglo reći da je imala potpuno jednak tretman u svim domenima upotrebe jezika. Uprkos napretku koji je postignut u opštem domenu, odnosno novinskim tekstovima koje, kao bogatom izvoru primera vremenskih izraza i relacija, karakteriše narativni stil standardnog jezika, aktivniji rad na otkrivanju vremenskih informacija sadržanih u medicinskim narativnim tekstovima javlja se tek poslednjih pet-šest godina (Lin et al. 2015). Ipak, problem reprezentacije i zaključivanja pomoću nekog od aspekata okolnosti tipa vreme privlači pažnju istraživača biomedicinske oblasti već više decenija (Savova et al. 2009; Meystre et al. 2008; Augusto 2005). Još krajem osamdesetih godina XX veka sprovedeno je nekoliko studija usmerenih na obradu vremenskih informacija medicinskih tekstova (Johnson 1987; Hirschman 1981; Obermeier 1985). Prvi sistem kojim je vršena identifikacija onih reči i jezičkih formi koje nose vremensko značenje u medicinskim tekstovima bio je razvijen na Univerzitetu u Njujorku u okviru programa LSP (eng. *Linguistic String Project*) (Sager 1967). Ovaj sistem, namenjen obradi standardnog engleskog jezika, prilagođen je potrebama

primene na medicinske tekstove – LSP-MLP (eng. *Linguistic String Project- Medical Language Processor*), s ciljem prepoznavanja, analize i formalizacije različitih oblika vremenskih informacija u reprezentaciju koja će omogućiti lekarima da ekstrahuju i sumiraju informacije o simptomima, primeni i doziranju lekova, kao i reakcijama na njihovu primenu i mogućim neželjenim efektima (Hirschman 1981; Lyman et al. 1985). Takođe, jedan od prvih sistema namenjenih analizi vremenskih informacija medicinskih tekstova, poznat pod nazivom GROK (skr. od eng. Grammatical Representation of Objective Knowledge) (Obermeier 1985), razvijen je pomoću tekstova koji se odnose na oboljenja jetre. Rezultat primene ovog sistema bila je reprezentacija teksta zasnovana na poznavanju ključnih koncepata ove medicinske specijalnosti, u okviru koje su relevantni medicinski događaji bili ekstrahovani i hronološki poređani.

Tokom poslednje decenije XX veka sve veći broj istraživača se bavi ulogom vremena u medicini i temporalnošću medicinskih narativnih tekstova, objavljujući postignute rezultate, kako u medicinskim časopisima i časopisima iz oblasti računarstva, tako i u zbornicima radova izlaganih na konferencijama (Keravnou 1991; Goodwin and Hamilton 1996; Combi and Shahar 1997). Najznačajniji sistem razvijen tokom tog perioda na Univerzitetu Kolumbija u Njujorku (Friedman, Cimino, and Johnson 1993), poznat pod nazivom MedLEE (eng. *Medical Language Extraction and Encoding System*), od 1995. godine se koristi i za potrebe Prezbiterijanske bolnice Kolumbija u gradu Njujorku. Sistem MedLEE, metodom zasnovanom na znanju, uz korišćenje rečnika i pravila u vidu gramatika, identifikuje klinički značajne informacije i jednostavne reference na apsolutne datume u različitim tipovima medicinskih narativnih tekstova (npr. otpusne liste, radiološki izveštaji, patološki izveštaji itd.), transformišući ih u strukturirani oblik reprezentacije koji se može pohraniti u bazu podataka, na osnovu koje je omogućeno postavljanje upita i pružanje podrške lekarima u odlučivanju. Ipak, tokom ovog perioda veliki broj istraživača, zainteresovanih za temporalnu rezoluciju u kliničkom domenu, bavio se prevashodno strukturiranim podacima u vidu eksplicitno kodiranih događaja (npr. laboratorijskih testova, poseta lekaru itd.), koji su vremenski označeni i pohranjeni u baze podataka. Iscrpan metodološki pregled razvijenih pristupa i predloženih standarda za formalizaciju vremenskih informacija domena kliničke medicine, koji će omogućiti razmenu podataka među informacionim sistemima zdravstvene zaštite, dat je u (Augusto 2005).

U skorije vreme sve je veće interesovanje istraživača za efektivno korišćenje

vremenskih informacija medicinskih narativnih tekstova i njihovo ugrađivanje u sisteme namenjene ekstrakciji informacija. Rešavanje problema ekstrakcije vremenskih informacija, za razliku od opšteg domena, u oblasti medicine zahteva uzimanje u obzir postavljenih zadataka, odnosno namene sistema koji će se koristiti za npr. podršku u odlučivanju prilikom postavljanja dijagnoze, savetovanje oko određivanja terapije, sumarizaciju kliničkih podataka, sprovođenje epidemioloških studija itd. (Adlassnig et al. 2006). Na primer, u radu (Denny et al. 2010) autori su razvili sistem koji iz elektronskih zdravstvenih kartona ekstrahuje vreme i status skrining testova primenjenih s ciljem ranog otkrivanja kolorektalnog karcinoma. Sistem identifikuje kliničke koncepte iz ovih kartona, pronalazi i normalizuje informacije o datumu, koje pomoću heuristike pripisuje određenim kliničkim konceptima, štedeći vreme i trud koje bi bilo potrebno uložiti kada bi se ove informacije utvrđivale ručno. Slične ideje su bile primenjene i za utvrđivanje statusa korišćenih lekova. Kako bi proučili izloženost pacijenata Varfarinu[2] u vreme bolničkog prijema, Liu i saradnici (2011) su razvili sistem koji iz elektronskih zdravstvenih kartona pacijenata ekstrahuje podatke o korišćenju određenih lekova u dato vreme. Isto tako su Irvine i saradnici (2008) kreirali sistem koji iz beleški prilikom trijaže pacijenata na urgentnom odeljenju ekstrahuje i interpretira vremenske izraze. Kada je reč o automatskoj obradi vremenskih izraza nestrukturiranih podataka, kao što su medicinski narativni tekstovi, nekoliko radova daje sumiran prikaz publikovane literature (Meystre et al. 2008; Zhou and Hripcsak 2007).

Problem automatske obrade vremenskih informacija narativnih tekstova opšteg domena bio je u fokusu više međunarodnih izazova (detaljnije opisanih u delu 3), kojima se odazvao veliki broj istraživača, pre svega, zahvaljujući postojanju dovoljne količine javno dostupnih anotiranih korpusa neophodnih za razvoj ovih sistema. Međutim, imajući u vidu prirodu tekstova medicinskog domena, čije javno korišćenje nije moguće, aktivnije bavljenje automatskom obradom vremenskih informacija medicinskih tekstova omogućeno je tek 2012. godine, kada je za potrebe i2b2 (eng. *Informatics for Integrating Biology and the Bedside*) izazova u međunarodnim okvirima ustanovljen prvi klinički korpus engleskog jezika anotiran vremenskim informacijama (Sun, Rumshisky, and Uzuner 2013a). Prethodni zadaci ovog izazova, koji se u organizaciji Nacionalnog centra za biomedicinsku informatiku Sjedinjenih Američkih Država održava od 2007. godine, bili

[2]Varfarin (*warfarin*) je lek koji smanjuje koagulaciju krvi i sprečava stvaranje krvnih ugrušaka u krvnim sudovima.

su usmereni na one elemente koji su važni u procesu automatske obrade kliničkih tekstova i rešavanja problema, između ostalog, deidentifikacije zaštićenih zdravstvenih informacija, ekstrakcije medicinski relevantnih koncepata i relacija koje postoje među njima, kao i prepoznavanja koreferencija među osnovnim konceptima.

Tema šestog po redu i2b2 izazova, u kome je učestvovalo 18 timova iz celog sveta, bila je ekstrakcija vremenskih informacija iz kliničkih narativnih tekstova, odnosno ekstrakcija događaja (npr. problema pacijenata, testova, terapija), vremenskih izraza i vremenskih relacija. Korpus koji je obezbeđen za učesnike ovog izazova sastojao se od 310 medicinskih narativnih tekstova Medicinskog centra „Đakonica Bet Izrael“ u Bostonu, u vidu otpusnih listi anotiranih vremenskim informacijama. S obzirom na široku upotrebu TimeML uputstva za obeležavanje vremenskih informacija, koji je poslužio i kao osnova za razvoj standarda ISO-TimeML, u procesu anotiranja ovih medicinskih dokumenata korišćena je modifikovana verzija ovog uputstva, razvijenog posebno za klinički domen i trenutno poznatog pod nazivom THYME-TimeML (eng. *Temporal Histories of Your Medical Events*) (Styler IV et al. 2014a, 2014b). U skorije vreme je sprovedeno nekoliko istraživanja koja su se bavila mogućnošću prilagođavanja TimeML uputstva za obeležavanje vremenskih informacija u medicinskim narativnim tekstovima (Savova et al. 2009; Galescu and Blaylock 2012) i vodičima dobre kliničke prakse (Wenzina and Kaiser 2014a, 2014b), koja su postigla ohrabrujuće rezultate. Za razliku od zadataka identifikacije događaja i vremenskih relacija, ekstrakcija i normalizacija vremenskih izraza medicinskog domena se ne razlikuje u velikoj meri od ovog zadatka u opštem domenu, osim u pogledu često korišćenih izraza u vidu skraćenica sa značenjem učestalosti korišćenja medikamentne terapije. Osim prilagođavanja TimeML uputstva za potrebe obeležavanja vremenskih informacija medicinskih tekstova, pojedini autori su se bavili i istraživanjem drugih mogućnosti (Zhou et al. 2006; Tao, Solbrig, and Chute 2011).

Najuspešniji sistemi i2b2 izazova 2012. godine su za rešavanje zadatka prepoznavanja i normalizacije vremenskih izraza koristili već postojeće sisteme namenjene prepoznavanju i normalizaciji novinskih tekstova, poput HeidelTime, SUTIME, GUTIME, NorMA sistema (detaljno opisanih u delu 3), uz primenu metoda zasnovanih na pravilima ili hibridnog pristupa, kombinujući pravila i metode mašinskog učenja (Sun, Rumshisky, and Uzuner 2013a). Bez obzira na uspešnost primene postojećih sistema, većina timova je razvila sopstvena pravila za prepo-

znavanje i normalizaciju izraza koji ukazuju na učestalost. Analiza postignutih rezultata je pokazala da je za većinu sistema zadatak normalizacije relativnih vremenskih izraza složeniji problem, koji zahteva dalja istraživanja u pravcu obuhvatanja šireg konteksta vremenskih izraza.

Nakon organizovanog i2b2 izazova, interesovanje naučne zajednice za ekstrakciju vremenskih informacija kliničkog domena se nastavlja i u okviru poslednjeg SemEval takmičenja, održanog 2015. godine. Kao jedan od zadataka izazova SemEval-2015, Clinical TempEval (Bethard et al. 2015) je bio zadatak usmeren na identifikaciju opsega i osnovnih odlika vremenskih izraza u kliničkim beleškama i patološkim izveštajima Klinike Mejo, čije je sedište u Ročesteru u američkoj državi Minesota. Proces obeležavanja vremenskih izraza je podrazumevao primenu kompletnog THYME uputstva za obeležavanje (Styler IV et al. 2014b), kojim su, osim vremenskih izraza tipa DATE, TIME, DURATION i SET (opisani u delu 4.2), definisane još dve semantičke klase karakteristične za ovu vrstu tekstova, i to: QUANTIFIER i PREPOSTEXP. Na osnovu uputstva THYME, izrazi koji ukazuju na broj pojavljivanja nekog događaja takođe spadaju u grupu vremenskih izraza (npr. *pacijent je **dva puta** povratio pre operacije*), čija će vrednost atributa `type` biti QUANTIFIER. Izrazi, poput *predoperativni* ili *postoperativno*, koji ukazuju na određeni vremenski period koji je u vezi sa nekim događajem (npr. pre operacije ili posle operacije) trebalo bi da budu obeleženi kao posebna klasa vremenskih izraza – PREPOSTEXP. U okviru postavljenog zadatka od učesnika se očekivalo određivanje opsega i tipa vremenskih izraza, ali ne i normalizacija vrednosti prepoznatih izraza. Evaluacija uspešnosti sistema vršena je pomoću standardnih mera preciznosti, odziva i F_1 mere, opisanih u delu 6. Sistemi koji su bili najuspešniji u identifikovanju i klasifikaciji vremenskih izraza (najuspešniji sistem je postigao *F* meru od 0,709%) jesu sistemi zasnovani na mašinskom učenju, dok su u identifikaciji vremenskih izraza sistemi zasnovani na pravilima postigli najbolji odziv (0,795%). Ovakav rezultat je iznenađujuć, imajući u vidu uobičajenu pretpostavku da su sistemi zasnovani na pravilima daleko precizniji od sistema zasnovanih na mašinskom učenju, koji na uštrb preciznosti postižu dobar odziv. Interesantno je i da su u ovom izazovu, od ukupno 15 prijavljenih, učestvovala samo tri tima sa ukupno 13 verzija sistema namenjenih prepoznavanju vremenskih izraza. Organizatori smatraju da je slab odziv učesnika rezultat složene i vremenski zahtevne procedure potpisivanja sporazuma o korišćenju pripremljenog kliničkog korpusa, koja bi trebalo značajno da se pojednostavi za buduće korisnike zainteresovane za razvoj ovih sistema.

Osim pomenutih izazova organizovanih u međunarodnim okvirima, razvijani su i drugi sistemi namenjeni ekstrakciji vremenskih informacija iz medicinskih narativnih tekstova, prevashodno zasnovanih na već postojećim sistemima opšte namene. Na primer, modifikovanjem postojećih i kreiranjem novih pravila sistema TARSQI razvijen je sistem Med-TTK namenjen upotrebi u medicinskom domenu (Reeves et al. 2013). Na osnovu 200 kliničkih narativnih tekstova, ovaj sistem vrši identifikaciju vremenskih izraza koji ukazuju na datum, vreme dana, trajanje i učestalost, uz ukupnu meru uspešnosti od 85%. Zadatak normalizacije prepoznatih vremenskih izraza nije obuhvaćen ovim sistemom. Kada je reč o normalizaciji, aktuelan problem predstavlja normalizacija relativnih vremenskih izraza kliničkih tekstova (Sun, Rumshisky, and Uzuner 2015).

Većim delom se sprovedena istraživanja u oblasti automatskog obeležavanja vremenskih izraza medicinskih narativnih tekstova odnose na dokumente pisane engleskim jezikom. Ipak, zabeleženi su i određeni napori istraživača koji se bave prilagođavanjem i razvijanjem sistema namenjenih medicinskim narativnim tekstovima drugih jezika, poput francuskog (Hamon and Grabar 2014), kineskog (Xiaojia et al. 2011), švedskog (Velupillai 2014) itd.

Kada je reč o medicinskim narativnim tekstovima srpskog jezika, automatska obrada vremenskih informacija karakterističnih za ovu vrstu tekstova do sad nije bila predmet istraživanja. Imajući u vidu značaj vremena u interpretaciji medicinski relevantnih informacija, bilo bi korisno razviti sistem za automatsku obradu vremenskih izraza medicinskih narativnih tekstova, što bi uticalo na razvoj drugih aplikacija automatske obrade tekstova medicinskog domena, pa samim tim i unapređenje kvaliteta usluga zdravstvene nege, dijagnostičkih i teorijskih procesa u medicini. S obzirom na to da se narativni tekstovi medicinskog domena stilski veoma razlikuju od opštih tekstova pisanih prirodnim jezikom, tehnologije obrade prirodnog jezika razvijene za druge domene ne mogu se jednostavno primeniti na tekstove kliničkog domena, već su potrebna određena modifikovanja i prilagođavanja. Stoga je jedan od ciljeva ovog rada primena kreiranog sistema za automatsku obradu vremenskih izraza srpskog jezika u novinskim tekstovima na medicinske narativne tekstove, radi procene efikasnosti sistema u prepoznavanju i normalizaciji vremenskih izraza i određivanja stepena potrebnog prilagođavanja karakteristikama i zahtevima medicinskog domena. Kako automatsko prepoznavanje i normalizacija vremenskih izraza medicinskih narativnih tekstova predsta-

vlja dovoljno složen problem koji zahteva sistematičniji pristup rešavanja u okviru zasebnog istraživanja, u okviru ovog rada biće primenjena neka *ad-hoc* rešenja i preliminarna evaluacija uspešnosti sistema, koja će poslužiti kao osnova daljim istraživanjima na ovom polju automatske obrade srpskog jezika.

7.5 Primena sistema za prepoznavanje i normalizaciju vremenskih izraza na medicinske narativne tekstove srpskog jezika

Sistem namenjen automatskoj obradi vremenskih izraza novinskih tekstova srpskog jezika primenjen je na medicinske narativne tekstove radi identifikacije i normalizacije vremenskih izraza sadržanih u ovoj vrsti dokumenata. Metodologija korišćena za potrebe prepoznavanja i normalizacije vremenskih izraza detaljno je opisana u poglavljima 4 i 5. Kao i u slučaju vremenskih izraza standardnog srpskog jezika, prepoznavanje vremenskih izraza medicinskih narativnih tekstova podrazumeva identifikaciju i određivanje opsega lingvističkih izraza apsolutnog ili relativnog vremenskog značenja, koji su predstavljeni u medicinskim tekstovima različitim formalnim jedinicama. Na osnovu TimeML uputstva, osnovni semantički tipovi vremenskih izraza koje je potrebno prepoznati su oni koji ukazuju na temporalnu lokaciju u vidu kalendarskog vremena (DATE) ili vremena kao dela dana (TIME), trajanje (DURATION) i temporalnu frekvenciju, odnosno učestalost pojavljivanja u vremenu (SET). Semantičke klase vremenskih izraza QUANTIFIER i PREPOSTEXP, predložene THYME uputstvom za obeležavanje, u ovom trenutku neće biti uzete u obzir prilikom identifikacije vremenskih izraza jer se radi o primeni postojećeg sistema, a ne njegovoj nadogradnji.

Proces normalizacije vremenskih izraza odnosi se na interpretaciju vrednosti prepoznatih vremenskih izraza, u standardizovanom obliku koji je u skladu sa ISO 8601 standardom. Svi identifikovani vremenski izrazi obeležavaju se umetanjem `<TIMEX3>` etikete u okviru koje su definisani sledeći atributi, detaljno opisani u poglavljima 4 i 5:

- atribut `type`, kojim se specifikuje semantička klasa prepoznatog izraza;
- atribut `temporalFunction`, koji ukazuje na apsolutnu ili relativnu vrednost identifikovanog vremenskog izraza;

- atribut `value`, kojim se reprezentuju konačne vrednosti vremenskih izraza u standardizovanom obliku;
- atribut `mod`, koji se koristi za prikazivanje značenja kvantifikovanih ili modifikovanih vremenskih izraza;
- atribut `valueFromFunction`, koji se koristi u slučaju relativnih vremenskih izraza, ukazujući na računsku operaciju koju je potrebno izvršiti radi računanja apsolutne vrednosti identifikovanog izraza;
- atributi `quant` i `freq`, koji se koriste radi upotpunjavanja informacija o značenju vremenskih izraza koji ukazuju na učestalost ponavljanja u vremenu.

Nekoliko navedenih primera (primer 7.2) ilustruje tražene semantičke klase vremenskih izraza, kao i način obeležavanja rezultata.

Primer 7.2.
(a)
```
<TIMEX3 type="DATE" temporalFunction="false" val="2011-12-23">
23.12.2011.</TIMEX3>
<TIMEX3 type="DATE" temporalFunction="true" val="XXXX-03"
valueFromFunction="-1Y">martu prošle godine</TIMEX3>
```
(b)
```
<TIMEX3 type="TIME" temporalFunction="false" val="2012-10-07T16"
mod="APPROX">07.10.2012.  godine oko 16 h</TIMEX3>
<TIMEX3 type="TIME" temporalFunction="true" val="TMO">ujutru</TIMEX3>
```
(c)
```
<TIMEX3 type="DURATION" temporalFunction="false" val="P1Y"
mod="MORE_THAN">više od godinu dana</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true" val="PXY">
više godina</TIMEX3>
```
(d)
```
<TIMEX3 type="SET" temporalFunction="false" val="P1W" freq="6X">
6 puta nedeljno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1D" quant="EVERY">
svaki dan</TIMEX3>
```
(e)
```
OD <TIMEX3 type="DATE.PERIOD">
<TIMEX3 type="DATE" temporalFunction="false" val="2009-09-29">
```

```
29.09.2009.
</TIMEX3>
DO <TIMEX3 type="DATE" temporalFunction="false" val="2009-10-03">
03.10.2009.</TIMEX3>
</TIMEX3>
<TIMEX3 type="DURATION.PERIOD">
<TIMEX3 type="DURATION" temporalFunction="false" val="P10D">10</TIMEX3>
-
<TIMEX3 type="DURATION" temporalFunction="false" val=„P14D">
14 dana</TIMEX3>
</TIMEX3>
```

7.6 Rezultati primene i evaluacija uspešnosti sistema

Za evaluaciju uspešnosti sistema u automatskoj obradi vremenskih izraza medicinskih narativnih tekstova korišćen je korpus koji se sastoji od 150 nasumično izabranih dokumenata i koje čine otpusne liste (100) i izveštaji lekara (50) dve nastavne jedinice Stomatološkog fakulteta Univerziteta u Beogradu. Tekstovi korišćeni za evaluaciju su prethodno automatski deidentifikovani (Jaćimović, Krstev, and Jelovac 2015), zamenom ličnih imena, naziva lokacija i drugih zdravstvenih zaštićenih informacija fiktivnim podacima, dok vremenski izrazi nisu prepravljani.[3] Primer deidentifikovanog teksta koji je korišćen za evaluaciju sistema dat je u prilogu E (primer E.3). Izabrani tekstovi nisu korišćeni prilikom razvoja sistema i predstavljaju potpuno nov materijal, pogodan za sprovođenje preliminarne evaluacije.

Otpusne liste i izveštaji lekara predstavljaju nestrukturirane tekstove pisane prirodnim jezikom, koje su kucali lekari prilikom zaključenja bolničkog lečenja ili nakon više poseta pacijenata, te sadrže istorije bolesti pacijenata, podatke o trenutnom fizičkom stanju, propisanoj medikamentnoj terapiji, rezultate laboratorijskih testova, dijagnostičke nalaze, preporuke prilikom otpusta pacijenata i druge informacije koje se odnose na zdravstveno stanje pacijenta. Dimenzije korišćenog korpusa merene brojem reči i rečenica predstavljene su u tabeli 7.1. Podela teksta

[3] Budući da ovi tekstovi predstavljaju materijal koji nije u potpunosti deidentifikovan, medicinska dokumentacija je korišćena uz dobijenu saglasnost Stomatološkog fakulteta.

Tabela 7.1: Dimenzije korišćenog korpusa medicinskih narativnih tekstova

Korpus	**Broj reči**	**Broj rečenica**
Otpusne liste	23.175	2.167
Izveštaji lekara	13.373	1.219
Ukupno	36.529	3.386

na rečenice izvršena je automatski (na način opisan u delu 4.5), te se dati brojevi posmatraju samo kao približne vrednosti, zbog određenog broja mogućih grešaka nastalih tokom segmentacije teksta na rečenice. Korpus medicinskih narativnih tekstova koji je korišćen za evaluaciju sistema, kao i sve ostale medicinske narativne tekstove, odlikuju nedovršene rečenice, nedostatak znakova interpunkcije, kao i neuobičajeno veliki broj pravopisnih i tipografskih grešaka, mnogo veći nego, na primer, u slučaju novinskih tekstova. Imajući u vidu okolnosti u kojima ovi tekstovi nastaju, sasvim je očekivano pojavljivanje većeg broja grešaka, kao jedne od osnovnih karakteristika medicinskih narativnih tekstova. Iz tog razloga su tekstovi obrađivani u izvornom obliku, bez ispravljanja grešaka. Za razliku od novinskih tekstova pisanih standardnim srpskim jezikom, u čijem slučaju je nakon primene elektronskih rečnika samo 3-4% reči teksta nepoznato, medicinski narativni tekstovi sadrže nešto više od 20% neprepoznatih reči, što je razumljivo, s obzirom na to da se radi o tekstovima specifičnog domena i terminologiji koja još uvek nije uključena u elektronske rečnike standardnog srpskog jezika.

Evaluacija uspešnosti sistema za automatsko obeležavanje vremenskih izraza medicinskih narativnih tekstova sprovedena je na isti način kao i evaluacija sistema primenjenog na novinske tekstove (opisano u delu 6). S obzirom na to da se među rezultatima primene bilo kog automatizovanog sistema za pronalaženje i ekstrakciju informacija javlja određeni broj grešaka, pre izvođenja zaključaka i sprovođenja daljeg istraživanja neophodno je utvrditi njihov stepen i vrstu. Kada je reč o mogućim vrstama grešaka automatizovanog sistema za obradu vremenskih izraza, one se mogu javiti i u odnosu na prepoznavanje i u odnosu na normalizaciju vrednosti prepoznatih vremenskih izraza. Tako se, pre svega, u fazi prepoznavanja javljaju greške u vidu *propuštenih* izraza (eng. *missing slots*), koji nisu, a trebalo je da budu prepoznati kao vremenski izraz; zatim, greške u vidu *uljeza* (eng. *incorrect slots*), koji predstavljaju izraze pogrešno identifikovane kao vremenski izraz; na kraju, greške u vidu *nepoklapanja* opsega izraza koji je trebalo prepoznati u odnosu na opseg prepoznat sistemom ili nepoklapanja u određivanju tipa vremenskog izraza (eng. *spurious slots*). Tokom normalizacije vrednosti vre-

Tabela 7.2: Opšte vrednosti atributa korišćenih za proveru (proveraP i proveraN)

Vrednost	**Značenje**
OK	Ispravno prepoznato/normalizovano
UOK	Opseg ili neki od atributa nisu ispravno definisani ili su propušteni
UOK/E	Opseg ili neki od atributa nisu ispravno definisani ili su propušteni zbog greške u tekstu
NOK	Neispravno definisani svi atributi (NOKp – prepoznavanja, NOKn – normalizacije)
MISS	Propušten vremenski izraz
MISS/E	Propušteno zbog greške u tekstu

menskih izraza, moguće su greške koje se odnose na nepoklapanja u vidu pogrešno definisanih ili izostavljenih vrednosti atributa koji se koriste za interpretaciju vrednosti prepoznatog vremenskog izraza. U pogledu izvora grešaka, to mogu biti greške nastale usled grešaka u samom tekstu ili izostavljenih ili nedovoljno precizno definisanih pravila korišćene gramatike.

Sva automatski izvršena prepoznavanja i dodeljene normalizovane vrednosti su proverene dodavanjem atributa proveraP i proveraN u okviru postavljene <TIMEX3> etikete, s ciljem definisanja vrste i izvora nastale greške (tabela 7.2), dok se neprepoznatim izrazima dodeljuju <TIMEX3> etikete i atribut proveraP, čija je vrednost MISS. Proveru je vršio jedan anotator (diplomirani bibliotekar-informatičar sa iskustvom u evaluaciji sistema za ekstrakciju imenovanih entiteta) na osnovu iscrpnog uputstva za rad, uz kontrolu autora primenjenog sistema. U slučaju ispravno izvršenog prepoznavanja ili normalizacije atributima proveraP i proveraN se pripisuje vrednost OK. U situacijama kada ni jedan od atributa, koji se dodeljuju u fazi prepoznavanja ili normalizacije, nije ispravno definisan, atributima korišćenim za proveru se pripisuje vrednost NOK. S obzirom na strukturu sistema, koja ne daje mogućnost normalizovanja vrednosti vremenskih izraza koji prethodno nisu prepoznati, greška u vidu propuštenih izraza zbog greške u tekstu može se javiti samo u vezi sa prepoznavanjem vremenskih izraza, kada se atributu proveraP dodeljuje vrednost MISS/E.

U slučaju grešaka nepoklapanja u vidu pogrešno definisanih ili izostavljenih vrednosti atributa <TIMEX3> etikete, atributu koji se koristi za proveru (proveraP ili proveraN) se pripisuje vrednost UOK, uz oznaku atributa koji je pogrešno određen. Detaljan prikaz oznaka <TIMEX3> atributa, korišćenih za potrebe evaluacije

Tabela 7.3: Oznake <TIMEX3> atributa

Atribut	**Oznaka**
type	t
Opseg	o
value	v
mod	m
valueFromFunction/funkcija	f
quant	q
freq/učestalost	u

ilustrovan je u okviru tabele 7.3. Na primer, ako je pogrešno određen tip vremenskog izraza, atributu za proveru prepoznavanja proveraP biće dodeljena vrednost UOKt, dok će u slučaju pogrešno određenih atributa mod i valueFromFunction atributu proveraN biti dodeljena vrednost UOKmf. Ako su pogrešno definisan opseg ili neispravno određen tip prepoznatog izraza rezultat greške u tekstu, atributu proveraP biće dodeljena vrednost UOKo/E, odnosno UOKt/E.

Moguće vrednosti atributa proveraP su sledeće:

- OK – ispravno određen tip i precizno određen pun opseg vremenskog izraza;
- UOK – neki od atributa nije ispravan;

  UOKt (tip izraza nije ispravno određen, ali pun opseg jeste ispravno definisan);

  UOKt/E (tip izraza nije ispravno određen zbog greške u tekstu, ali pun opseg jeste ispravno definisan);

  UOKo (tip izraza je tačno određen, ali pun opseg nije ispravno definisan);

  UOKo/E (tip izraza je tačno određen, ali pun opseg nije ispravno definisan zbog greške u tekstu);
- NOKp – neispravno definisani i tip i opseg vremenskog izraza;
- NOK – izraz koji je pogrešno definisan i obeležan kao vremenski izraz, a nije trebalo (prepoznato je nešto što nije vremenski izraz);
- MISS – propušten izraz (nije prepoznat, a trebalo je da bude);
- MISS/E – izraz nije prepoznat zbog greške u tekstu.

Moguće vrednosti atributa `proveraN` su sledeće:

- `OK` – ispravno je izvršena normalizacija;
- `UOK` – neki od atributa nije ispravan

  `UOKv` (atribut `value` nije ispravan);

  `UOKm` (atribut `mod` nije ispravan);

  `UOKf` (atribut `valueFromFunction` nije ispravan);

  `UOKmf` (atributi `mod` i `valueFromFunction` nisu ispravni);

  `UOKvf` (atributi `value` i `valueFromFunction` nisu ispravni);

  `UOKvm` (atributi `value` i `mod` nisu ispravni);

  `UOKq` (atributi `quant` nije ispravan);

  `UOKu` (atributi `freq` nije ispravan);

  `UOKvq` (atributi `value` i `quant` nisu ispravni);

  `UOKvu` (atributi `value` i `freq` nisu ispravni);
- `NOKn` – neispravno definisani svi atributi normalizacije.

Primer medicinskog narativnog teksta obeleženog vremenskim izrazima nakon primene sistema dat je u okviru priloga E (primer E.4). Još neke ilustracije korektnog obeležavanja vremenskih izraza, kao i obeležavanja mogućih grešaka predstavljene su primerima E.5 i E.6 u prilogu E.

Rezultati prebrojavanja nakon provere pokazuju da se od ukupnog broja postojećih vremenskih izraza (884) u korišćenom korpusu medicinskih narativnih tekstova javlja 772 (87,3%) vremenska izraza koja ukazuju na datum, vreme, trajanje i učestalost, kao i 112 (12,7%) izraza kojima su označeni vremenski periodi (tabela 7.4). U ovom korpusu je pronađeno 613 (69,3%) apsolutnih vremenskih izraza koji nose značenje tačke u vremenu i trajanja (DATE, TIME i DURATION), dok 99 (11,2%) pripada grupi relativnih izraza. U slučaju izraza koji ukazuju na učestalost ponavljanja u vremenu identifikovano ih je ukupno 60, odnosno 6,8%, što je i očekivano kada je reč o narativnim tekstovima medicinskog domena.

Tokom provere obeležavanja vremenskih izraza definisanih TimeML uputstvom, za potrebe kasnijih istraživanja ručno su izvršeni identifikacija i obeležavanje opsega izraza, koji na osnovu THYME uputstva, predstavljaju klasu izraza karakterističnu za medicinske narativne tekstove i koji ukazuju na određeni vremenski

Tabela 7.4: Učešće tipova vremenskih izraza u korpusu medicinskih narativnih tekstova

TIMEX3	**abs**	**rel**	**Ukupno**
DATE	402	40	442
TIME	43	37	80
DURATION	168	22	190
SET			60
Ukupno	613	99	772
DATE.PERIOD			109
DURATION.PERIOD			3
Ukupno	613	99	884

period vezan za neki događaj. Tom prilikom je u korišćenom korpusu pronađeno 26 izraza koji pripadaju novodefinisanoj klasi PREPOSTEXP (primer 7.3).

Primer 7.3.
(a)
```
<TIMEX3 type="PREPOSTEXP">Postoperativno</TIMEX3> sprovedena RT.
```
(b)
```
Rezultat trombektomije dobar,ali <TIMEX3 type="PREPOSTEXP" val="POD8">
trećeg p.op dana</TIMEX3> dolazi do ponovne tromboze.
```
(c)
```
...otpušten je iz bolnice <TIMEX3 type="PREPOSTEXP" val="HD10">
10.  dana</TIMEX3> kao oporavljen ...
```
(d)
```
Međutim <TIMEX3 type="PREPOSTEXP">intraoperativno</TIMEX3> su nađeni
nepovoljni uslovi ...
```
(e)
```
<TIMEX3 type="PREPOSTEXP">Preoperativno</TIMEX3> korigovana
kardiološka terapija.
```

Procena uspešnosti sistema u prepoznavanju i normalizaciji vremenskih izraza medicinskih narativnih tekstova srpskog jezika vršena je na osnovu standardnih mera za procenu učinka sistema za pronalaženje i ekstrakciju informacija, detaljnije opisanih u poglavlju 6. Osnovnu jedinicu evaluacije nad kojom se vrši procena predstavlja kompletan vremenski izraz. Posmatrani su preciznost (p) si-

stema, koja predstavlja odnos ukupnog broja ispravno prepoznatih vremenskih izraza (OK) prema ukupnom broju prepoznatih izraza (M) (formula 7.1); zatim, odziv (*r*), kao odnos ukupnog broja ispravno prepoznatih vremenskih izraza (OK) prema ukupnom broju postojećih izraza u tekstu (N) (formula 7.2), i F_1 mera, kao zajednička mera koja odražava harmonijsku sredinu odziva i preciznosti i uvek ima vrednost koja teži manjoj od vrednosti postignutog odziva ili preciznosti (formula 7.3). Kao i u slučaju vrednovanja uspešnosti sistema primenjenog na novinske tekstove, u evaluaciji je korišćena balansirana F_1 mera, koja ne daje prednost ni odzivu ni preciznosti.

$$p = \frac{OK}{M} = \frac{OK}{OK + UOK + NOK} \tag{7.1}$$

$$r = \frac{OK}{N} = \frac{OK}{OK + MISS} \tag{7.2}$$

$$F_1 = 2\frac{pr}{p + r} \tag{7.3}$$

U pogledu uspešnosti sistema u prepoznavanju vremenskih izraza, odnosno određivanju opsega i tipa, sistem je postigao nešto veću preciznost (94%) u odnosu na odziv (90%), uz ukupnu F_1 meru od 92% (tabela 7.5). Najbolji rezultati su ostvareni prilikom identifikacije i određivanja opsega i tipa izraza koji ukazuju na period vremena ili predstavljaju nizanje više vremenskih izraza koji su u određenom odnosu. Greške nastale u ovom procesu isključivo su rezultat postojanja grešaka u tekstu, nastalih prilikom kucanja teksta ili usled nedosledne primene pravopisnih pravila.

Sistem je postigao najveću preciznost u određivanju opsega i tipa izraza koji ukazuju na učestalost, dok je istovremeno odziv prilikom identifikacije ove klase izraza najniži, uz postignutu F_1 meru od 67%. Takođe, dosta niži odziv (86%) u odnosu na postignutu preciznost (96%) javlja se i u slučaju izraza koji ukazuju na trajanje, što imajući u vidu izvor grešaka može biti objašnjeno propustima prilikom definisanja pravila za identifikaciju vremenskih izraza. Kada je reč o izrazima sa značenjem tačke u vremenu, odnosno kalendarskim datumima i vremenima dana, ostvaren bolji odziv (95% i 97%) prema postignutoj preciznosti (94% i 84%) nedvosmisleno ukazuje na potrebu za modifikovanjem postojećih pravila, s ciljem ispravnog obeležavanja opsega identifikovanih vremenskih izraza. Detaljni podaci

o uspešnosti sistema u prepoznavanju vremenskih izraza dati su u tabeli 7.5, dok će podrobna analiza grešaka biti izvršena u daljem tekstu. Kolona N ukazuje na ukupan broj vremenskih izraza koji postoje u korpusu medicinskih tekstova, dok kolona M sadrži ukupan broj vremenskih izraza koje je sistem identifikovao.

Postignuti rezultati u pogledu uspešnosti procesa normalizacije vremenskih izraza su, može se reći, izuzetno dobri (tabela 7.6). Vrednosti svih korektno prepoznatih vremenskih izraza su normalizovane (postignuti odziv je 100%), dodeljivanjem svih odgovarajućih atributa neophodnih za njihovu interpretaciju. S obzirom na to da su u slučaju izraza koji ukazuju na vreme dana, zbog pogrešno određenog opsega, unete neodgovarajuće vrednosti atributa `value`, preciznost u vršenju normalizacije iznosi 99,7%. Kolona N ukazuje na ukupan broj vremenskih izraza koji su identifikovani sistemom u korpusu medicinskih tekstova i za koje je očekivano sprovođenje procesa normalizacije, dok kolona M sadrži ukupan broj vremenskih izraza čije je vrednosti sistem normalizovao.

7.6.1 Analiza grešaka

Greške nastale prilikom primene sistema, kao i njegova uspešnost u prepoznavanju i normalizaciji vremenskih izraza u odnosu na postojeće semantičke klase vremenskih izraza prikazani su u okviru tabela 7.5 i 7.6. Uočene greške i postignuti preciznost, odziv i F_1 mera u pogledu određivanja opsega i dodeljivanja neophodnih atributa, bez uzimanja u obzir klase kojoj vremenski izraz pripada, ilustrovani su podacima u tabeli 7.7. Kolona N ukazuje na ukupan broj vremenskih izraza koji postoje u korpusu medicinskih tekstova i za koje je očekivano određivanje opsega i dodeljivanje atributa, dok kolona M sadrži ukupan broj vremenskih izraza kojima je sistem odredio opseg i dodelio atribute. Imajući u vidu postignute F_1 mere sistema, najveći problem u procesu automatske obrade vremenskih izraza, ako ne i jedini, identifikovan je već u prvom koraku, prilikom prepoznavanja i određivanja opsega izraza.

Izrazi koje sistem nije identifikovao i obeležio prilikom obrade čine 9% ukupnog broja postojećih vremenskih izraza u korišćenom korpusu (N). Gotovo trećinu propuštenih izraza (n=28) čine izrazi koji ukazuju na učestalost primene terapija (npr. *na **drugi dan**, infuzijom 25000 j / **24 h**, 0,8 ml na **12 h*** i sl.), dok su u slučaju izraza koji ukazuju na kalendarske datume i trajanje propusti prevashodno rezultat primene pravila, koja zarad veće preciznosti sistema, nisu omogućavala identifikaciju numeričkih izraza koji u kontekstu novinskih tekstova mogu da budu

Tabela 7.5: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u prepoznavanju vremenskih izraza medicinskih narativnih tekstova na osnovu postojećih semantičkih klasa izraza

	N	**M**	**OK**	**UOKt**	**UOKo**	**UOKo/E**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
DATE	442	419	393	0	17	9	14	7	0	0,94	0,95	0,94
TIME	80	80	67	1	5	5	2	0	2	0,84	0,97	0,90
DURATION	190	165	158	3	2	2	21	4	0	0,96	0,86	0,91
SET	60	30	30	0	0	0	28	2	0	1	0,5	0,67
Ukupno	772	694	648	4	24	16	65	13	2	**0,93**	**0,89**	**0,91**
DATE.PERIOD	109	105	104	0	0	1	0	4	0	0,99	0,96	0,98
DURATION.PERIOD	3	3	3	0	0	0	0	0	0	1	1	1
Ukupno	884	802	755	4	24	17	65	17	2	**0,94**	**0,90**	**0,92**

Tabela 7.6: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u normalizaciji vremenskih izraza medicinskih narativnih tekstova na osnovu postojećih semantičkih klasa izraza

	N	**M**	**OK**	**UOKv**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
DATE	419	419	419	0	0	0	0	1	1	1
TIME	80	80	78	2	0	0	0	0,98	1	0,99
DURATION	165	165	165	0	0	0	0	1	1	1
SET	30	30	30	0	0	0	0	1	1	1
Ukupno	694	694	692	2	0	0	0	**0,997**	**1**	**1**

Tabela 7.7: Podaci o rezultatima evaluacije i postignutoj uspešnosti sistema u određivanju opsega i atributa vremenskih izraza

	N	**M**	**OK**	**UOK**	**UOKo/E**	**MISS**	**MISS/E**	**NOK**	***p***	***r***	F_1
opseg	884	802	761	24	17	65	17	0	0,95	0,90	0,93
`type`	802	802	796	4	0	0	0	2	0,99	1	1
`temporalFunction`	694	694	694	0	0	0	0	0	1	1	1
`value`	694	694	692	2	0	0	0	0	0,997	1	1
`mod`	46	46	46	0	0	0	0	0	1	1	1
`valueFromFunction`	11	11	11	0	0	0	0	0	1	1	1
`quant`	12	12	12	0	0	0	0	0	1	1	1
`freq`	4	4	4	0	0	0	0	0	1	1	1

izuzetno dvosmisleni ili im nisu svojstveni (primer 7.4).

Primer 7.4.

```
Navodi da je <TIMEX3 proveraP="MISS" type="DATE">1995</TIMEX3> a potom
<TIMEX3 proveraP="MISS" type="DATE">96</TIMEX3> operisala karcinom
u <TIMEX3 proveraP="MISS" type="DURATION">7 god.</TIMEX3>operisao
Pušač unazad <TIMEX3 proveraP="MISS" type="DURATION">oko 20 g.</TIMEX3>
```

Propusti u identifikaciji vremenskih izraza, nastalih usled postojećih tipografskih grešaka, čine skoro 17% ukupnog broja izostavljenih izraza. Neki slučajevi ilustrovani su primerom 7.5.

Primer 7.5.

```
<TIMEX3 proveraP="MISS/E" type="DURATION">emesec dana</TIMEX3>
<TIMEX3 proveraP="MISS/E" type="DATE">18.o6.2009.</TIMEX3>
<TIMEX2 proveraP="MISS/E" type="DATE">26 07 2013</TIMEX3>
<TIMEX3 proveraP="MISS/E" type="DATE">sptembru o.g.</TIMEX3>
<TIMEX3 proveraP="MISS/E" type="SET">svaakodnevno</TIMEX3>
<TIMEX3 proveraP="MISS/E" type="SET">( sati</TIMEX3>
```

Na osnovu analize grešaka koje ukazuju na neispravno određen opseg identifikovanog vremenskog izraza uočeno je da 58,5% slučajeva zaista jesu greške sistema (primer 7.6), dok preostalih 41,5% čine greške nastale usled postojećih pravopisnih i tipografskih grešaka u tekstu (primer 7.7).

Primer 7.6.

```
<TIMEX3 proveraP="UOKo" proveraN="OK" type="DATE"
   temporalFunction="true" val="XXXX-09" mod="START">početkom
   septembra</TIMEX3> o.g.
povređen <TIMEX3 proveraP="UOKo" proveraN="OK" type="TIME"
   temporalFunction="false" val="2011-12-22T01" mod="APPROX">22.12.
   2011.  godine oko 01h </TIMEX3>30 min.
Pre <TIMEX3 proveraP="UOKo" proveraN="OK" type="DURATION"
   temporalFunction="false" val="P1.5Y">godinu i po</TIMEX3> dana
```

Primer 7.7.

```
po podnevnim časovima <TIMEX3 proveraP="UOKo" proveraN="OK"
   type="DATE" temporalFunction="true" val="XXXX-XX-XX"
   valueFromFunction="=1D">istog dana</TIMEX3>
0'<TIMEX3 proveraP="UOKo" proveraN="OK" type="DATE"
   temporalFunction="false" val="2012-12-06">6.12.2012.</TIMEX3>
```

S obzirom na to da vremenski izrazi u kontekstu medicinskih narativnih tekstova predstavljaju posebnu kategoriju zaštićenih zdravstvenih informacija koje podležu tajni, izuzetno je važno postići što bolji odziv prilikom njihove identifikacije i visoku preciznost u određivanju opsega, kako neki deo ovih informacija ne bi ostao otkriven. Imajući to u vidu, rezultati analize grešaka u određivanju opsega identifikovanih vremenskih izraza pokazuju da u većini slučajeva načinjeni propusti ne utiču negativno na određivanje tipa izraza i sprovođenje normalizacije njegove vrednosti, što je izuzetno značajno za kasniji proces prikrivanja stvarnih datuma i njihovo pomeranje za nasumično odabrani interval. Na primer, u određivanju opsega kalendarskih datuma najčešće je izostavljan deo koji se odnosi na element godine iskazan skraćenim oblikom relativnog izraza *o.g.* (*ove godine*) (prvi izraz u okviru primera 7.6). Identifikovani deo izraza koji je potrebno prikriti *početkom septembra* ispravno je definisan kao relativni kalendarski datum, kome su dodeljene korektne vrednosti atributa `value` i `mod`.

Greške u određivanju tipa prepoznatog vremenskog izraza čine samo 0,7% od ukupnog broja grešaka. U većini slučajeva radi se o izrazima koji ukazuju na učestalost primene terapije, a prepoznati su kao izrazi sa značenjem trajanja (primer 7.8). Bez obzira na pogrešno određen tip izraza, oblik i vrednost atributa `value` odgovaraju propisanom obliku izraza koji ukazuje na učestalost (opisano u delu 5.2.1). S obzirom na to da je izrazom implicirana frekvencija ponavljanja na svakih 8 sati, podatak koji nedostaje jeste atribut `freq`, čija bi vrednost bila EVERY. Budući da TimeML uputstvo predviđa definisanje `freq` atributa samo na osnovu eksplicitno iskazane frekvencije ponavljanja, u ovom trenutku izostavljeni atribut nije tretiran kao greška.

Primer 7.8.

```
cefalosporinski preparat na <TIMEX3 proveraP="UOKt" proveraN="OK"
   type="DURATION" temporalFunction="false" val="P8H">8 sati</TIMEX3>
```

Otkrivanje i rešavanje grešaka u vidu *uljeza*, odnosno izraza koji su pogrešno identifikovani kao vremenski izrazi, posebno je značajno u obradi medicinskih narativnih tekstova, pogotovo ako je pogrešno obeležen izraz zapravo izraz koji ukazuje na neki medicinski relevantan podatak. Analizom grešaka nakon primene sistema uočena su samo dva pogrešna prepoznavanja ovog tipa (primer 7.9). Na primer, u oblasti maksilofacijalne hirurgije se prilikom dijagnostikovanja preloma zigomatične kosti koristi izraz koji, pomoću analogije sa pozicijom kazaljke na satu, vrši opis i lokalizaciju pojave koštanog stepenika na ramu orbite (npr. *na oko 5 sati*).

Primer 7.9.

```
koštani stepenik u predelu rama leve orbite na <TIMEX3 proveraP="NOK"
   proveraN="OK" type="TIME" temporalFunction="false" val="T07"
   mod="APPROX">oko 7 sati</TIMEX3>
```

Kada je reč o ostalim atributima koji su definisani TimeML uputstvom, sistem je napravio samo dve greške prilikom određivanja vrednosti najvažnijeg atributa u procesu normalizacije tj. atributa `value`. Obe greške su rezultat pogrešno određenog opsega vremenskog izraza (primer 7.10), na osnovu kog je i izvršena normalizacija vrednosti prepoznatog izraza. Broj koji je greškom identifikovan kao apsolutno vreme dana i uključen u opseg izraza zapravo predstavlja podatak o količini propisane terapije, koji bi na osnovu uputstva THYME trebalo da bude obeležen kao izraz tipa QUANTIFIER.

Primer 7.10.

```
Detralex tbl <TIMEX3 proveraP="UOKo" proveraN="UOKv" type="TIME"
   temporalFunction="false" val="T02">2 ujutro</TIMEX3>
```

Na osnovu sračunate uspešnosti sistema u vidu tradicionalne F_1 mere, mera greške sistema računa se kao E=1-F. Međutim, kako je prilikom izračunavanja harmonijske sredine odziva i preciznosti važnost grešaka u vidu uljeza i propuštenih izraza umanjena (tako da ukupna stopa greške nikada ne može da bude veća od stope greške bilo kog tipa grešaka posebno), za potrebe evaluacije u okviru ovog rada sprovedeno je i utvrđivanje ukupne mere svih tipova grešaka na osnovu stope greške slota (eng. *slot error rate*, *SER*) (Makhoul et al. 1999). Kao jednostavna mera greške, SER je jednak količniku zbira sva tri tipa grešaka (propuštenih – `MISS`, uljeza – `NOK` i nepoklapanja – `UOK`) i ukupnog broja izraza postojećih u re-

ferentnom korpusu (N) (Formula 7.4).

$$SER = \frac{MISS + NOK + UOK}{N} = \frac{MISS + NOK + UOK}{OK + UOK + MISS} \tag{7.4}$$

Na osnovu podataka u okviru tabela 7.8, 7.9 i 7.10 jasno je da je stopa greške sračunata na osnovu *SER*-a za oko 50% viša u odnosu na stopu greške predstavljenu putem F_1 mere. Dobijene vrednosti *SER*-a u većoj meri ističu greške sistema u procesu prepoznavanja i normalizacije vremenskih izraza medicinskih narativnih tekstova.

Tabela 7.8: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u prepoznavanju vremenskih izraza medicinskih narativnih tekstova na osnovu postojećih semantičkih klasa izraza

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	393	26	21	0	0,94	0,06	0,11
TIME	67	11	2	2	0,90	0,10	0,19
DURATION	158	7	25	0	0,91	0,09	0,17
SET	30	0	30	0	0,67	0,33	0,5
Ukupno	648	44	78	2	0,91	0,09	**0,16**
DATE.PERIOD	104	1	4	0	0,98	0,02	0,05
DURATION.PERIOD	3	0	0	0	1	0	0
Ukupno	755	45	82	2	0,92	0,08	**0,15**

Tabela 7.9: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u normalizaciji vremenskih izraza medicinskih narativnih tekstova na osnovu postojećih semantičkih klasa izraza

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	419	0	0	0	1	0	0
TIME	78	2	0	0	0,99	0,01	0,03
DURATION	165	0	0	0	1	0	0
SET	30	0	0	0	1	0	0
Ukupno	692	2	0	0	1	0,001	**0,003**

S obzirom na to da se medicinski narativni tekstovi stilski razlikuju od tekstova opšteg domena (Sun, Rumshisky, and Uzuner 2013b) i da trećinu ovih tekstova

Tabela 7.10: Vrednosti postignute F_1 mere i mere grešaka *E* i *SER* u određivanju opsega i atributa vremenskih izraza

	OK	**UOK**	**UOK/E**	**MISS**	**MISS/E**	**NOK**	F_1	***E***	***SER***
opseg	761	24	17	65	17	0	0,93	0,07	**0,14**
`type`	796	4	0	0	0	2	1	0,004	**0,008**
`temporalFunction`	694	0	0	0	0	0	1	0	0
`value`	692	2	0	0	0	0	1	0,001	**0,003**
`mod`	46	0	0	0	0	0	1	0	0
`valueFromFunction`	11	0	0	0	0	0	1	0	0
`quant`	12	0	0	0	0	0	1	0	0
`freq`	4	0	0	0	0	0	1	0	0

čine nekompletne rečenice koje teško mogu da budu identifikovane konvencionalnim jezičkim parserima (Savova et al. 2009), automatska obrada medicinskih narativnih tekstova zahteva nešto drugačiji pristup. Imajući u vidu činjenicu da se u korišćenom korpusu medicinskih tekstova 26,4% ukupnog broja grešaka javlja usled nedosledne primene gramatičkih i pravopisnih pravila, kao i postojanja većeg broja tipografskih grešaka, korišćena pravila za prepoznavanje vremenskih izraza je potrebno modifikovati, kako bi se poboljšali i preciznost i odziv procesa automatske obrade.

Semantičke klase vremenskih izraza koje se javljaju u novinskim tekstovima, postoje i u medicinskim narativnim tekstovima, od kojih su najzastupljeniji kalendarski datumi i izrazi koji ukazuju na trajanja. Međutim, za razliku od opšteg domena, u okviru medicinskih tekstova mnogo je češća upotreba izraza koji ukazuju na učestalost, i to prevashodno učestalost primene medikamentnih terapija. U korišćenom korpusu je identifikovano čak 7,7% izraza ovoga tipa, što je neuobičajeno visok procenat u odnosu na njihovu zastupljenost u novinskim tekstovima. Iako veliki broj izraza za koje su kreirana pravila prepoznavanja nije pronađen u korpusu medicinskih tekstova, identifikovana je posebna klasa vremenskih izraza (PREPOSTEXP), svojstvenih samo tekstovima medicinskog domena.

7.7 Zaključak

Sistem za prepoznavanje i normalizaciju vremenskih izraza srpskog jezika, razvijen na primerima iz novinskih članaka, primenjen je na tekstove medicinskog domena, s ciljem utvrđivanja uspešnosti njegove primene i u ovoj oblasti. Rezultati sprovedene evaluacije ukazuju na činjenicu da kreiran sistem, bez ikakve prethodne pripreme i prilagođavanja pravila ovom specifičnom domenu, vrši automatsko prepoznavanje i normalizaciju vremenskih izraza medicinskih narativnih tekstova srpskog jezika, uz postizanje visoke preciznosti (94%). Iako je postignuta preciznost u automatskoj obradi vremenskih izraza niža u odnosu na preciznost ostvarenu u opštem domenu (99%), ona ukazuje na to da se sistem može uspešno primeniti i u ovom domenu, uz prilagođavanje pravila prepoznavanja na osnovu analize grešaka identifikovanih evaluacijom. Kada je reč o postignutom odzivu, sistem primenjen na medicinske narativne tekstove postiže bolje rezultate u odnosu na rezultate postignute u opštem domenu (90%, odnosno 80%), što je i razumljivo s obzirom na prirodu medicinskih tekstova koji, u odnosu na novinske tekstove, sadrže ograničen broj oblika iskazivanja vremena.

Analizom grešaka otkriven je određeni broj propusta, koji su mahom rezultat nepostojanja pravila ili, pak, primene pravila, koja zarad veće preciznosti sistema, nisu omogućavala identifikaciju određenih izraza. Zahvaljujući modularnosti sistema moguće je na jednostavan način doraditi postojeća i uključiti nova pravila za prepoznavanje onih oblika koji su karakteristični za medicinske tekstove, što će uticati na poboljšanje i preciznosti i odziva. Kada je reč o normalizaciji prepoznatih izraza, kao i u slučaju novinskih tekstova, identifikovane greške ukazuju na to da ovaj proces u potpunosti zavisi od uspešnosti prepoznavanja, odnosno ispravnog određivanja opsega i tipa vremenskih izraza, s obzirom na to da su vrednosti svih korektno identifikovanih i obeleženih izraza na ispravan način i normalizovane. S obzirom na to medicinski narativni tekstovi spadaju u grupu tekstova koje karakteriše postojanje velikog broja pravopisnih i tipografskih grešaka, bilo bi korisno uključiti u primenu i postojeći sistem za korekciju tekstova (Stanković et al. 2011).

Glava 8

Zaključak

Ovaj rad se bavi istraživanjem i razumevanjem različitih načina iskazivanja vremenskih izraza srpskog jezika, a zatim i kreiranjem sistema koji će omogućiti njihovu što precizniju automatsku obradu, uz postizanje visokog odziva. Iako je predmet rada, između ostalog, sagledavanje i opis jezičkih struktura kojima se iskazuje vreme u srpskom jeziku, ono nije deo lingvističkih istraživanja usmerenih na sintaksičko-semantičku analizu određivanja temporalnosti, već predstavlja doprinos u oblasti obrade prirodnih jezika, odnosno doprinos razvoju jezičkih tehnologija srpskog jezika. Osnovni cilj prilikom kreiranja ovog sistema bio je, ne samo postizanje dobrog učinka u automatskoj obradi vremenskih izraza, već i kreiranje alata koji je robustan i pouzdan, ali i dovoljno modularan, kako bi se omogućilo njegovo uključivanje i primena i u okviru većih sistema namenjenih automatskoj obradi srpskog jezika.

Osnovni rezultat ovog istraživanja jeste alat koji vrši automatsko obeležavanje vremenskih izraza u nestrukturiranim tekstovima srpskog jezika sa postizanjem visokog nivoa odziva (80%) i preciznosti (99%), uz ukupnu F_1 meru od 88%. Na osnovu analize primera identifikovanih u korpusu novinskih tekstova, izvršen je opis elemenata formalne strukture tri osnovna značenjska tipa vremenskih izraza srpskog jezika. Metodom automata konačnih stanja precizno su definisana jezička pravila za prepoznavanje, odnosno identifikaciju niski reči koje čine strukturu i opseg vremenskih izraza. Svaki identifikovani vremenski izraz je u fazi normalizacije obeležen u skladu sa TimeML shemom, pripisivanjem niza atributa i njihovih vrednosti. Na ovom nivou, pomoću složenih gramatika plitkog parsiranja, apsolutnim vremenskim izrazima su određene njihove standardizovane vrednosti, dok su nedovoljno preciznim izrazima dodeljeni svi oni atributi na osnovu kojih se mogu izračunati njihove konačne vrednosti. Postignuti rezultati u pogledu uspe-

šnosti procesa normalizacije vremenskih izraza su izuzetno dobri (ostvarena F_1 mera je 99,7%), imajući u vidu činjenicu da su vrednosti svih korektno prepoznatih vremenskih izraza ispravno normalizovane, dodeljivanjem svih odgovarajućih atributa neophodnih za njihovu interpretaciju.

Prilikom metodološkog utvrđivanja osnove sistema i definisanja pravila za prepoznavanje vremenskih izraza, jedan od osnovnih zadataka je bio da se ona učine što opštijim, kako bi se sa većom sigurnošću moglo očekivati njihovo uspešno prilagođavanje i široka primena. Procena efikasnosti kreiranog sistema u medicinskom domenu izvršena je primenom već definisanih pravila za automatsku obradu vremenskih izraza novinskih tekstova na medicinske nestrukturirane tekstove srpskog jezika. Na osnovu analize dobijenih rezultata primene, utvrđeno je da kreirani sistem, razvijen na primerima iz novinskih članaka, bez ikakve prethodne pripreme i prilagođavanja pravila, vrši automatsko prepoznavanje vremenskih izraza medicinskih narativnih tekstova, uz postizanje visokog nivoa preciznosti (94%), odziva (90%) i F_1 mere (92%). Kada je reč o normalizaciji prepoznatih izraza, budući da su vrednosti svih korektno prepoznatih izraza ispravno normalizovane, postignuta je apsolutna uspešnost. Iako je metoda kojom se u okviru ovog sistema vrši obeležavanje vremenskih izraza razvijena na primerima iz novinskih članaka, nakon testiranja i evaluacije uspešnosti u obeležavanju vremenskih izraza medicinskih narativnih tekstova, utvrđeno je da se u osnovi sistema nalaze dovoljno opšta pravila koja mogu uspešno da se primene i na druge tipove tekstova.

Postignuti rezultati ukazuju na to da su postavljeni ciljevi ostvareni. Kreiran je alat, koji bez prethodne pripreme i sa visokom preciznošću i odzivom, vrši automatsko prepoznavanje i normalizaciju vremenskih izraza u tekstovima srpskog jezika. Isto tako je pokazano da se ovaj alat može koristiti i jednostavno prilagoditi primeni u domenu medicine, na narativnim medicinskim tekstovima srpskog jezika. Zahvaljujući postignutoj visokoj preciznosti, omogućeno je i obeležavanje proizvoljne količine tekstova, koji se mogu koristiti za obuku i testiranje mašinski zasnovanih metoda koje se bave automatskom obradom vremenskih informacija. Mogućnost preciznog obeležavanja proizvoljne količine tekstova i anotacije tih korpusa mogu da budu podjednako korisne i za lingvističku analizu određenih vremenskih fenomena, što predstavlja još jedan od doprinosa istraživačkoj zajednici.

8.1 Budući rad

Buduća istraživanja koja proizilaze iz ovog rada uključuju određene zadatke koji će uticati na poboljšanje funkcionalnosti sistema opisanog u ovoj tezi, ili će omogućiti njegovu širu primenu, s ciljem rešavanja i nekih složenijih problema u oblasti obrade prirodnih jezika.

Iako je postignuta preciznost u radu ovog sistema izuzetno visoka, problemi identifikovani u ovom trenutku, u smislu određivanja opsega i tipa vremenskih izraza, ostavljaju prostor za dalji rad, koji bi podrazumevao modifikaciju postojećih pravila. Budući da je veoma važno povećati odziv sistema, uz posebnu pažnju usmerenu na održavanje visoke preciznosti, buduća istraživanja bi mogla da budu usmerena na kreiranje novih pravila za prepoznavanje formalnih jedinica koje u ovoj fazi istraživanja nisu uključene u sistem. Radi razrešavanja vrednosti nedovoljno preciznih izraza i koreferenci, bilo bi veoma korisno primeniti program koji će omogućiti dodeljivanje identifikacionih brojeva prepoznatim izrazima. Korišćene sheme za obeležavanje vremenskih izraza i strukture anotacije, predstavljaju polje istraživanja u okviru koga je moguće detaljnije se pozabaviti i mogućnostima proširivanja anotacija određenim informacijama koje bi bile korisne za sprovođenje lingvističkih istraživanja.

Jedan od zadataka, koji bi mogao da bude veoma značajan u pogledu kreiranja kompletnog sistema namenjenog obradi vremenskih informacija, jeste razvoj metodologije koja će se upotrebiti za prepoznavanje vremenskih entiteta, koji zajedno sa vremenskim izrazima, čine osnovu temporalne determinacije – događaja i vremenskih relacija koje postoje među njima, a koji su predviđeni TimeML shemom.

Osim u opštem domenu, budući pravci istraživanja bi mogli da budu usmereni na uključivanje pravila koja će omogućiti uspešnu primenu kreiranog sistema i u medicinskom domenu. Nakon automatske obrade vremenskih izraza, a radi potpunosti sistema, bilo bi korisno pozabaviti se i prepoznavanjem medicinski relevantnih događaja i uspostavljanjem njihove hronologije.

Potencijal primene kreiranog sistema u okviru nekih drugih aplikacija jezičkih tehnologija, koje mogu imati koristi od informacija dobijenih ovim sistemom, stvara nove perspektive i otvara široko područje istraživanja, čije su mogućnosti beskrajne.

Literatura

Aberdeen, John, et al. 2010. “The MITRE Identification Scrubber Toolkit: Design, training, and assessment”. *International Journal of Medical Informatics* 79 (12): 849–859. doi:10.1016/j.ijmedinf.2010.09.007.

Abney, Steven. 1996. “Partial parsing via finite-state cascades”. *Natural Language Engineering* 2 (04): 337–344.

Abot, P. 2009. *Uvod u teoriju proze*. Beograd: Službeni glasnik.

ACE. 1999. *Automatic Content Extraction Evaluation*. Web Page. http://www.itl.nist.gov/iad/mig//tests/ace/.

Adlassnig, Klaus-Peter, et al. 2006. “Temporal representation and reasoning in medicine: research directions and challenges”. *Artificial intelligence in medicine* 38 (2): 101–113.

Ahn, D, J Rantwijk, and M Rijke. 2007. “A Cascaded Machine Learning Approach to Interpreting Temporal Expressions”, 420–427. Association for Computational Linguistics.

Ahn, David, Sisay Fissaha Adafre, and Maarten de Rijke. 2005. “Extracting Temporal Information from Open Domain Text: A Comparative Exploration”. In *Fifth Dutch-Belgian Information Retrieval Workshop*, 3–10.

Ahn, David, Sisay Fissaha Adafre, and Maarten De Rijke. 2005. “Towards task-based temporal extraction and recognition”. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Ašić, Tijana. 2005. “Predlozi po, na i u u srpskom jeziku i njihova fizička i temporalna interpretacija”. *Zbornik Matice srpske za slavistiku* 68:161–178.

Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. 2007. “On the value of temporal information in information retrieval”. In *ACM SIGIR Forum*, 41:35–41. 2. ACM.

Antonić, Ivana. 2001. *Vremenska rečenica*. Novi Sad: Izdavačka Knjižarnica Zorana Stojanovića.

Aramaki, Eiji, et al. 2006. “Automatic deidentification by using sentence features and label consistency”. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 10–11.

Augusto, Juan Carlos. 2005. “Temporal Reasoning for Decision Support in Medicine”. *Artificial intelligence in medicine* 33 (1): 1–24.

Badurina, Lada. 2013. “Vremenski odnosi na razini složene rečenice i teksta”. *Vrijeme u jeziku/Nulti stupanj pisma, Zbornik radova 41. seminara Zagrebačke slavističke škole*: 75–97.

Beckwith, Bruce A, et al. 2006. “Development and evaluation of an open source software tool for deidentification of pathology reports”. *BMC medical informatics and decision making* 6 (1): 1.

Benton, Adrian, et al. 2011. “A system for de-identifying medical message board text”. *BMC bioinformatics* 12 (3): 1.

Bethard, Steven. 2013. “ClearTK-TimeML: A minimalist approach to TempEval 2013”. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2:10–14.

Bethard, Steven, et al. 2015. “Semeval-2015 task 6: Clinical tempeval”. *Proc. SemEval*.

Bittar, Andrè. 2009. “Annotation of events and temporal expressions in French texts”. In *Computational linguistics in the Netherlands 2009 : selected papers from the nineteenth CLIN meeting (CLIN 2009), January 2009, Groningen, The Netherlands*, ed. by Barbara Plank, Erik Tjong Kim Sang, and Tim Van de Cruys. Utrecht : LOT.

Black, William J, Fabio Rinaldi, and David Mowatt. 1998. “FACILE: Description of the NE System Used for MUC-7”. In *Proceedings of the 7th Message Understanding Conference*.

Blekburn, Sajmon. 2013. *Filozofski rečnik*. Novi Sad: Adresa.

Boguraev, Branimir, and Rie Kubota Ando. 2006. “Analysis of TimeBank as a Resource for TimeML Parsing”. In *Proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation, May 22-28*, 71–76. Genova, Italy: ELRA.

Boguraev, Branimir, and Rie Kubota Ando. 2005. “TimeBank-Driven TimeML Analysis”. In *Annotating, extracting and reasoning about time and events, number 05151*, ed. by Graham Katz, James Pustejovsky, and Frank Schilder. Dagstuhl seminar proceedings. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum fur Informatik (IBFI). http://drops.dagstuhl.de/opus/volltexte/2005/335.

Carpenter, Bob. 2004. “Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval.” In *TREC*.

Caselli, Tommaso, Felice Dell’Orletta, and Irina Prodanof. 2009. “TETI: A TimeML compliant TimEx tagger for Italian”. In *Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT ’09, October 12-14, 2009. Mrągowo, Poland*, ed. by Maria Ganzha and Marcin Paprzycki, 4:185–192. Piscataway, NJ, USA: IEEE.

Cassel, David M, et al. 2006. “Automated capture and representation of date/time to support intelligence analysis”. In *Intelligence Tools Workshop*, 12.

Chambers, Nathanael. 2013. *Navytime: Event and time ordering from raw text*. Tech. rep. DTIC Document.

Chang, Angel X, and Christopher D Manning. 2012. “SUTime: A library for recognizing and normalizing time expressions.” In *LREC*, 3735–3740.

Chang, Angel X, and Christopher D Manning. 2013. “SUTIME: Evaluation in TempEval-3”. *Atlanta, Georgia, USA*: 78.

Chinchor, Nancy. 1998. “MUC-7 Named Entity Task Definition (version 3.5)”. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

Ciravegna, Fabio, et al. 1999. “Facile: Classifying texts integrating pattern matching and information extraction”. In *IJCAI*, 890–897.

Collobert, Ronan, and J Weston. 2011. “SENNA”. *NEC Laboratories America, Inc* 6.

Combi, Carlo, and Yuval Shahar. 1997. “Temporal reasoning and temporal data maintenance in medicine: issues and challenges”. *Computers in biology and medicine* 27 (5): 353–368.

Costa, Francisco, and Antonio Branco. 2010. “Temporal information processing of a new language: fast porting with minimal resources”. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16*, 671–677. Uppsala, Sweden: Association for Computational Linguistics.

Cowie, Jim, and Yorick Wilks. 2000. *Handbook of natural language processing. chapter Information Extraction*.

Cunningham, Hamish, et al. 1997. "GATE–a TIPSTER-based general architecture for text engineering". In *Proceedings of the TIPSTER Text Program (Phase III)*, vol. 6.

Daniel, Naomi, Dragomir Radev, and Timothy Allison. 2003. "Sub-event based multi-document summarization". In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, 9–16. Association for Computational Linguistics.

Day, David, et al. 2004. "Callisto: A Configurable Annotation Workbench". In *Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, Centro Cultural de Belem, Lisbon, Portugal, 26th, 27th 28th May*, 2073–2076. Lisbon, Portugal: European Language Resources Association (ELRA).

Day, David, et al. 1997. "Mixed-initiative development of language processing systems". In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, March 31-April 3*, ed. by Ralf Grishman, 348–355. Washington, DC, USA: Association for Computational Linguistics. doi:10.3115/974557.974608.

Deleger, Louise, et al. 2013. "Large-scale Evaluation of Automated Clinical Note De-identification and its Impact on Information Extraction". *Journal of the American Medical Informatics Association* 20 (1): 84–94. doi:10.1136/amiajnl-2012-001012.

Denny, Joshua C, et al. 2010. "Extracting timing and status descriptors for colonoscopy testing from electronic medical records". *Journal of the American Medical Informatics Association* 17 (4): 383–388.

Derczynski, Leon, and Robert Gaizauskas. 2010. "Analysing Temporally Annotated Corpora with CAVaT". In *Proceedings of LREC 2010: Seventh conference on International Language Resources and Evaluation*, ed. by Nicoletta Calzolari et al. Valletta, Malta.

Derczynski, Leon, and Robert Gaizauskas. 2010. "USFD2: Annotating temporal expresions and TLINKs for TempEval-2". In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 337–340. Association for Computational Linguistics.

Deschacht, Koen, and Marie-Francine Moens. 2009. “Semi-supervised semantic role labeling using the latent words language model”. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 21–29. Association for Computational Linguistics.

Douglass, M., et al. 2004. “Computer-assisted De-identification of Free Text in the MIMIC II Database”. In *Computers in Cardiology 2004, Vol 31*, 31:341–344. Computers in Cardiology.

Fekete, Egon. 2009. “Atributske odredbe uz temporalne imenice”. *Slavistika* (Beograd) Knjiga XIII:296–303.

Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

Ferrández, Oscar, et al. 2013. “BoB, a best-of-breed automated text de-identification system for VHA clinical documents”. *Journal of the American Medical Informatics Association* 20 (1): 77–83.

Ferro, Lisa. 2004. *The TERN 2004 Evaluation Plan (DRAFT): Time Expression Recognition and Normalization*. Report.

Ferro, Lisa, et al. 2004. *ACE Time Normalization (TERN) 2004 English Training Data v 1.3. Linguistic Data Consortium LDC2004E23*. Report.

Ferro, Lisa, et al. 2003. *TIDES 2003 Standard for the Annotation of Temporal Expressions. Technical report*. Report.

Ferro, Lisa, et al. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report*. Report.

Ferro, Lisa, et al. 2000. *TIDES Temporal Annotation Guidelines, Draft v.1.0. Mitre technical report MTR00W0000094*. Report.

Ferro, Lisa, et al. 2001. *TIDES Temporal Annotation Guidelines. Version 1.0.2. Technical report*. Report.

Filannino, Michele. 2012. “Temporal expression normalisation in natural language texts”. *arXiv preprint arXiv:1206.2010*.

Filannino, Michele, Gavin Brown, and Goran Nenadic. 2013. “ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge”. *Atlanta, Georgia, USA*: 53.

Filatova, Elena, and Eduard Hovy. 2001. “Assigning time-stamps to event-clauses”. In *Proceedings of the workshop on Temporal and spatial information processing*, 1–8. Morristown, NJ, USA: ACL.

Forascu, Corina. 2008. “GMT to+2 or How Can TimeML Be Used in Romanian”. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, 3238–3242. ELRA.

Friburger, N., and D. Maurel. 2004. “Finite-state transducer cascades to extract named entities in texts”. *Theoretical Computer Science* 313 (1): 93–104. ISSN: 0304-3975. doi:10.1016/j.tcs.2003.10.007.

Friburger, Nathalie. 2002. “Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques”. PhD thesis, Tours.

Friedlin, F. Jeff, and Clement J. McDonald. 2008. “A software tool for removing patient identifying information from clinical documents”. *Journal of the American Medical Informatics Association* 15 (5): 601–610. doi:10.1197/jamia.M2702.

Friedman, Carol, James J Cimino, and Stephen B Johnson. 1993. “A conceptual model for clinical radiology reports.” In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 829. American Medical Informatics Association.

Gaizauskas, Robert, et al. 1995. “University of Sheffield: description of the LaSIE system as used for MUC-6”. In *Proceedings of the 6th conference on Message understanding*, 207–220. Association for Computational Linguistics.

Galescu, Lucian, and Nate Blaylock. 2012. “A corpus of clinical narratives annotated with temporal information”. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 715–720. ACM.

Gardner, James, and Li Xiong. 2008. “HIDE: an integrated system for health information DE-identification”. In *Computer-Based Medical Systems, 2008. CBMS’08. 21st IEEE International Symposium on*, 254–259. IEEE.

Gardner, James, et al. 2010. “An evaluation of feature sets and sampling techniques for de-identification of medical records”. In *Proceedings of the 1st ACM International Health Informatics Symposium*, 183–190. ACM.

Goodwin, Scott D, and Howard J Hamilton. 1996. “It’s about time: an introduction to the special issue on temporal representation and reasoning”. *Computational Intelligence* 12 (3): 357–358.

Graff, David. 2002. *The AQUAINT Corpus of English News Text*. Catalog number LDC2002T3. Philadelphia: Linguistic Data Consortium. ISBN: 1-58563-240-6.

Grishman, Ralph, and Beth Sundheim. 1996. “Message understanding conference-6: A brief history”. In *COLING*, 466–471.

Gross, Maurice. 1993. “Local grammars and their representation by finite automata”. *Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair*: 26–38.

Grover, Claire, et al. 2010. “Edinburgh-LTG: TempEval-2 system description”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 333–336. Association for Computational Linguistics.

Gucul-Milojević, Sandra, Vanja Radulović, and Cvetana Krstev. 2008. “Usage of NooJ Graphs and Annotation for Information Extraction”. In *Proceedings of the 2007 International Nooj Conference*, ed. by Xavier Blanco and Max Silberztein, 103–120. Cambridge Scholars Publishing.

Guo, Yikun, et al. 2006. “Identifying personal health information using support vector machines”. In *i2b2 workshop on challenges in natural language processing for clinical data*, 10–11.

Gupta, Dilip, Melissa Saul, and John Gilbertson. 2004. “Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research”. *American journal of clinical pathology* 121 (2): 176–186.

Hacioglu, Kadri, Ying Chen, and Benjamin Douglas. 2005. “Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings”. Chap. Automatic Time Expression Labeling for English and Chinese Text, ed. by Alexander Gelbukh, 548–559. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-30586-6_59.

Hamon, Thierry, and Natalia Grabar. 2014. “Tuning HeidelTime for identifying time expressions in clinical texts in English and French”. *EACL 2014*: 101–105.

Han, N, Martin Chodorow, and C Leacock. 2006. “Detecting errors in English article usage by non-native speakers”. 12.

Hepple, Mark, Andrea Setzer, and Rob Gaizauskas. 2007. “USFD: preliminary exploration of features and classifiers for the TempEval-2007 tasks”. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 438–441. Association for Computational Linguistics.

HIPAA. 1996. *Health Insurance Portability and Accountability Act. Pub. L. 104-191*.

Hirschman, L. 1981. “Retrieving time information from natural language texts”. In *Information retrieval research*, ed. by RN Oddy et al., 154–171. Butterworths, London.

Hobbs, Jerry R, et al. 1997. “13 FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text”. *Finite-state language processing*: 383.

Humphreys, Kevin, et al. 1999. “University of sheffield trec-8 q & a system”. *NIST Special Publication 500-246*: 707–716.

Im, S., et al. 2009. “KTimeML: Specification of temporal and event expressions in Korean text”. In *Proceedings of the 7th Workshop on Asian Language Resources*, 115–122.

Irvine, Ann K, Stephanie W Haas, and Tessa Sullivan. 2008. “Tn-ties: A system for extracting temporal information from emergency department triage notes”. In *AMIA Annual Symposium proceedings*, 2008:328–332. American Medical Informatics Association.

ISO. 2004. *ISO 8601 Data Elements and Interchange formats - Information interchange - Representation of Dates and Times*. Standard. Geneva, Switzerland: International Organization for Standardization.

ISO. 1986. *ISO 8879 Information processing - Text and office systems - Standard Generalized Markup Language (SGML)*. Standard. Geneva, Switzerland: International Organization for Standardization.

ISO. 2009. *ISO/DIS 24617-1 Language Resources Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events (SemAF-Time, ISO-TimeML)*. Standard. Geneva, Switzerland: International Organization for Standardization.

ISO. 2007. *ISO/TC 37/SC 4/WG 2. Language Resource Management – Semantic Annotation Framework (SemAF) - Part 1: Time and events. Tech. rep.* Standard. Geneva, Switzerland: International Organization for Standardization.

Ittycheriah, Abraham, et al. 2003. “Identifying and tracking entity mentions in a maximum entropy framework”. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, 40–42. Association for Computational Linguistics.

Ivić, Milka. 1955–1956. “Iz problematike padežnih vremenskih konstrukcija”. *Južnoslovenski filolog* (Beograd) XXI:165–217.

Ivić, Milka. 1983. *Lingvistički ogledi*. Beograd: Prosveta.

Jaćimović, Jelena, Cvetana Krstev, and Drago Jelovac. 2015. “A rule-based system for automatic de-identification of medical narrative texts”. *Informatica* 39 (1): 43–51.

Jaćimović, Jelena, Cvetana Krstev, and Drago Jelovac. 2014. “Automatic de-identification of protected health information”. In *9th Language Technologies Conference, Information Society - IS 2014*, 73–78.

Jijkoun, Valentin, et al. 2008. “Named entity normalization in user generated content”. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, 23–30. ACM.

Johnson, S. 1987. “Temporal information in medical narrative”. In *Medical Language Processing: Computer Management of Narrative Data*, ed. by N Sager, C Friedman, and MS Lyman, 175–94. Reading, MA: Addison-Wesley.

Jung, Hyuckchul, and Amanda Stent. 2013. “ATT1: Temporal annotation using big windows and rich syntactic and semantic features”. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2:20–24.

Katz, G., and F. Arosio. 2001. “The annotation of temporal information in natural language sentences”. In *Proceedings of the workshop on Temporal and spatial information processing*, vol. Vol. 13, 151–158. Morristown, NJ, USA.

Keravnou, Elpida. 1991. “Medical temporal reasoning”. *Artificial Intelligence in Medicine* 3 (6): 289–290.

Kolomiyets, Oleksandr, and Marie-Francine Moens. 2013. “KUL: A data-driven approach to temporal parsing of documents”. In *Proceedings of the second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, 83–87. ACL.

Kolomiyets, Oleksandr, and Marie-Francine Moens. 2010. “KUL: recognition and normalization of temporal expressions”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 325–328. Association for Computational Linguistics.

Kolya, Anup Kumar, Asif Ekbal, and Sivaji Bandyopadhyay. 2010. “JU_CSE_TEMP: a first step towards evaluating events, time expressions and temporal relations”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 345–350. Association for Computational Linguistics.

Kolya, Anup Kumar, et al. 2013. “JU_CSE: A CRF Based Approach to Annotation of Temporal Expres-sion, Event and Temporal Relations”. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2:64–72.

Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Belgrade: University of Belgrade, Faculty of Philology.

Krstev, Cvetana, Jelena Jaćimović, and Duško Vitas. 2012. “Recognition and normalization of some classes of named entities in Serbian”. In *Proceedings of the Fifth Balkan Conference in Informatics*, 52–57. BCI ’12. Novi Sad, Serbia: ACM. ISBN: 978-1-4503-1240-0. doi:10.1145/2371316.2371327.

Krstev, Cvetana, and Duško Vitas. 2006. “Finate State Transducers for Recognition and Generation of Compound Words”. In *Proceedings of the 5th Slovenian and 1st International Conference Language Technologies, IS-LTC 2006*, ed. by Tomaž Erjavec and Jerneja Žganec Gros, 192–197. Ljubljana, Slovenia: Institut "Jožef Stefan".

Krstev, Cvetana, Duško Vitas, and Sandra Gucul. 2005. “Recognition of Personal Names in Serbian Texts”. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, ed. by G Angelova, 288–292. Borovets, Bulgaria.

Krstev, Cvetana, Duško Vitas, and Agata Savary. 2006. “Prerequisites for a Comprehensive Dictionary of Serbian Compounds”. In *FinTAL*, ed. by Tapio Salakoski et al., 4139:552–563. Lecture Notes in Computer Science. Springer. ISBN: 3-540-37334-9. doi:http://dx.doi.org/10.1007/11816508_55.

Krstev, Cvetana, et al. 2014. “A system for Named Entity Recognition Based on Local Grammars”. *Journal of Logic and Computation* 24 (2): 473–489. doi:10.1093/logcom/exs079.

Krstev, Cvetana, et al. 2011. “E-Dictionaries and Finite-State Automata for the Recognition of Named Entities”. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, 48–56. Blois, France: Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-4407.

Krupka, George R. 1995. “SRA: Description of the SRA system as used for MUC-6”. In *Proceedings of the 6th conference on Message understanding*, 221–235. Association for Computational Linguistics.

Krupka, George R, and Kevin Hausman. 1998. “Isoquest, Inc: Description of the NetOw(TM) extractor system as used for MUC-7”. In *Proceedings of The Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, 29 April-1 May, 1998*. MUC-7.

Kushida, Clete A., et al. 2012. “Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies”. *Medical Care* 50 (7): S82–S101. doi:10.1097/MLR.0b013e3182585355.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. 2001. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”: 282–289.

Lecuit, Émeline, et al. 2009. “Temporal Expressions: Comparisons in a Multilingual Corpus”. In *Proceedings of the 4th Language & Technology Conference*, ed. by Zygmunt Vetulani, 531–535. Poznań, Poland: IMPRESJA Widawnictwa Elektroniczne S.A.

Li, Wenjie, Kam-Fai Wong, and Chunfa Yuan. 2001. “A model for processing temporal references in Chinese”. In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, 5. Association for Computational Linguistics.

Li, Wenjie, et al. 2004. “Applying machine learning to Chinese temporal relation resolution”. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 582. Association for Computational Linguistics.

Lin, Ching-Heng, et al. 2015. “Comparison of a semi-automatic annotation tool and a natural language processing application for the generation of clinical statement entries”. *Journal of the American Medical Informatics Association* 22 (1): 132–142.

Liu, Mei, et al. 2011. “Modeling drug exposure data in electronic medical records: an application to warfarin”. In *AMIA annual symposium proceedings*, 2011:815–823. American Medical Informatics Association.

Llorens, H., E. Saquete, and B. Navarro-Colorado. 2012. “Automatic system for identifying and categorizing temporal relations in natural language”. *International Journal of Intelligent Systems* 27 (7): 680–703.

Llorens, Hector, Estela Saquete, and Borja Navarro. 2010. “TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 284–291. SemEval ’10. Los Angeles, California: Association for Computational Linguistics.

Llorens, Hector, Estela Saquete, and Borja Navarro-Colorado. 2013. "Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language". *Information Processing Management* 49 (1): 179–197. doi:10.1016/j.ipm.2012.05.005.

Llorens, Hector, et al. 2012. "TIMEN: An Open Temporal Expression Normalisation Resource." In *LREC*, 3044–3051.

Lyman, M, et al. 1985. "Computer-structured narrative in ambulatory care: its use in longitudinal review of clinical data". In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 82–86. American Medical Informatics Association.

Makhoul, John, et al. 1999. "Performance measures for information extraction". In *Proceedings of DARPA broadcast news workshop*, 249–252.

Mani, Inderjeet, James Pustejovsky, and Robert Gaizauskas, eds. 2005. *The language of time: a reader*. Vol. 126. Oxford University Press Oxford.

Mani, Inderjeet, and George Wilson. 2000. "Robust temporal processing of news". In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 69–76. Association for Computational Linguistics.

Mani, Inderjeet, et al. 2006. "Machine learning of temporal relations". In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 753–760. Association for Computational Linguistics.

Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*. Vol. 1. 1. Cambridge university press Cambridge.

Marşic, Georgiana. 2011. "Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations". PhD thesis, University of Wolverhampton.

Mazur, Paweł, and Robert Dale. 2007. "The DANTE temporal expression tagger". In *Human language technology. Challenges of the information society*, 245–257. Springer.

McTaggart, J. Ellis. 1908. "The Unreality of Time". *Mind* 17 (68): 457–474.

Meystre, Stephane M., et al. 2010. "Automatic de-identification of textual documents in the electronic health record: a review of recent research". *Bmc Medical Research Methodology* 10. doi:10.1186/1471-2288-10-70.

Meystre, Stéphane M, et al. 2008. "Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research". *Yearb Med Inform* 35:128–44.

Meystre, Stéphane M, et al. 2014. "Text de-identification for privacy protection: A study of its impact on clinical text information content". *Journal of biomedical informatics* 50:142–150.

Mikheev, Andrei, Claire Grover, and Marc Moens. 1998. "Description of the LTG system used for MUC-7". In *Proceedings of 7th Message Understanding Conference (MUC-7)*, 1–12. Fairfax, VA.

Miljković, Vanja. 2013. "Tip glagolske situacije i tip temporalnog konstituenta kao faktori iterativnog značenja u srpskom jeziku". *Anali Filološkog fakulteta* Knj. 25 (sv. 1): 195–226.

Miller, George A. 1995. "WordNet: a lexical database for English". *Communications of the ACM* 38 (11): 39–41.

Miller, Scott, et al. 1998. "Algorithms that learn to extract information: BBN: TIPSTER phase III". In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, 75–89. Association for Computational Linguistics.

Morrison, Frances P., Albert M. Lai, and George Hripcsak. 2009. "Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes?" *Journal of the American Medical Informatics Association* 16 (1): 37–39. doi:10.1197/jamia.M2862.

Nadeau, David, and Satoshi Sekine. 2007. "A survey of named entity recognition and classification". *Lingvisticae Investigationes* 30 (1): 3–26.

Neamatullah, Ishna, et al. 2008. "Automated De-identification of Free-text Medical Records". *Bmc Medical Informatics and Decision Making* 8. doi:10.1186/1472-6947-8-32.

Negri, Matteo, and Luca Marseglia. 2005. *Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004*. Report.

Obermeier, Klaus K. 1985. "Temporal inferences in medical texts". In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, 9–17. Association for Computational Linguistics.

Oklander, Natan. 1998. "Metafizika vremena i temporalnosti". *Theoria* 41 (3): 13–41.

Paumier, Sébastien. 2016. *Unitex 3.1 User manual*. http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf?.

Piper, Predrag. 1997. *Jezik i prostor*. Zemun: Biblioteka XX vek.

Piper, Predrag, et al. 2005. *Sintaksa savremenoga srpskog jezika: prosta rečenica*. Beograd: Institut za srpski jezik SANU / Beogradska knjiga.

Poveda, Jordi, Mihai Surdeanu, and Jordi Turmo. 2009. "An analysis of bootstrapping for the recognition of temporal expressions". In *Proceedings of the NAACL HLT 2009 workshop on semi-supervised learning for natural language processing*, 49–57. Association for Computational Linguistics.

Pustejovsky, James, et al. 2002. *Annotation Guideline to TimeML 1.0*. Report. http://timeml.org.

Pustejovsky, James, et al. 2006. *TimeBank 1.2. Linguistic Data Consortium LDC2006T08*. Philadelphia: Linguistic Data Consortium.

Pustejovsky, James, et al. 2003. "TimeML: Robust Specification of Event and Temporal Expressions in Text". In *Fifth International Workshop on Computational Semantics (IWCS-5)*.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1-55860-238-0.

Reeves, Ruth M, et al. 2013. "Detecting temporal expressions in medical narratives". *International journal of medical informatics* 82 (2): 118–127. doi:10.1016/j.ijmedinf.2012.04.006.

Robaldo, Livio, et al. 2011. "From italian text to timeml document via dependency parsing". In *Computational Linguistics and Intelligent Text Processing*, 177–187. Springer.

Roche, Emmanuel, and Yves Schabes. 1997. *Finite-state language processing*. MIT press.

Sager, Naomi. 1967. "Syntactic analysis of natural language". *Advances in computers* 8 (153–188): 35.

Saquete, E., R. Muñoz, and P. Martínez-Barco. 2006. "Event ordering using TERSEO system". *Data Knowledge Engineering* 58 (1): 70–89. doi:http://dx.doi.org/10.1016/j.datak.2005.05.011.

Saquete, Estela. 2010. "ID 392: TERSEO+ T2T3 Transducer: a systems for recognizing and normalizing TIMEX3". In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 317–320. Association for Computational Linguistics.

Saquete, Estela, et al. 2009. “Enhancing QA systems with complex temporal question processing capabilities”. *Journal of Artificial Intelligence Research*: 775–811.

Saurı, Roser, Estela Saquete, and James Pustejovsky. 2009. *Annotating time expressions in Spanish. TimeML annotation guidelines. Tech. rep.* Report.

Sauri, Roser, et al. 2006. *TimeML Annotation Guidelines Version 1.2.1*. Report. http://www.timeml.org/site/publications/timeMLdocs/annguide1.2.1.pdf.

Savova, G, et al. 2009. “Towards temporal relation discovery from the clinical narrative”. In *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*, 568–572.

Schilder, Frank, and Christopher Habel. 2001. “From temporal expressions to temporal information: Semantic tagging of news messages”. In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, 9. Association for Computational Linguistics.

Schilder, Frank, and Christopher Habel. 2001. “From temporal expressions to temporal information: semantic tagging of news messages”. In *Proceedings of the workshop on Temporal and spatial information processing*, vol. Vol. 13, 91–98. Morristown, NJ, USA: ACL.

Schmid, Helmut. 1994. “Treetagger”. *TC project at the Institute for Computational Linguistics of the University of Stuttgart*.

Seker, Sadi Evren, and Banu Diri. 2010. “TimeML and Turkish Temporal Logic”. In *Proceedings of the 2010 International Conference on Artificial Intelligence, ICAI 2010, July 12-15*, 881–887. Las Vegas Nevada, USA.

Setzer, Andrea. 2001. “Temporal information in newswire articles: An annotation scheme and corpus study. Ph.D. thesis”. Thesis.

Setzer, Andrea, and Robert Gaizauskas. 2000. “Annotating Events and Temporal Information in Newswire Texts”. In *Proceedings of LREC 2000 - 2nd International Conference on Language Resources Evaluation, 31 May - 2 June*, 1287–1294. Athens, Greece.

Shahar, Yuval. 1999. “Timing Is Everything: Temporal Reasoning and Temporal Data Maintenance in Medicine”. In *Artificial Intelligence in Medicine*, ed. by Werner Horn et al., 1620:30–46. Lecture Notes in Computer Science. Springer Berlin Heidelberg. doi:10.1007/3-540-48720-4_3.

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

Stanković, Ranka, et al. 2011. "Production of Morphological Dictionaries of Multi-Word Units Using a Multipurpose Tool". In *Proceedings of the Computational Linguistics-Applications Conference*, ed. by K Jassem et al., 77–84. Jachranka, Poland: Polish Information Processing Society.

Stanojčić, Živojin, and Ljubomir Popović. 2011. *Gramatika srpskog jezika: za gimnazije i srednje škole*. Beograd: Zavod za udžbenike.

Stevanović, Mihailo. 1967. *Funkcije i značenja glagolskih vremena*. Beograd: Naučno delo.

Stevanović, Mihailo. 1979. *Savremeni srpskohrvatski jezik: gramatički sistemi i književnojezička norma. 2, Sintaksa*. Beograd: Naučna knjiga.

Strötgen, Jannik, and Michael Gertz. 2010. "HeidelTime: High quality rule-based extraction and normalization of temporal expressions". In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 321–324. Association for Computational Linguistics.

Strötgen, Jannik, and Michael Gertz. 2012. "Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards." In *LREC*, 12:3746–3753.

Strötgen, Jannik, and Michael Gertz. 2011. "WikiWarsDE: A German corpus of narratives annotated with temporal expressions". In *Proceedings of the conference of the German society for computational linguistics and language technology (GSCL 2011)*, 129–134.

Strötgen, Jannik, Julian Zell, and Michael Gertz. 2013. "HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3". *Atlanta, Georgia, USA*: 15.

Styler IV, William F, et al. 2014. "Temporal annotation in the clinical domain". *Transactions of the Association for Computational Linguistics* 2:143–154.

Styler IV, William F, et al. 2014. *THYME annotation guidelines*.

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. 2013. "Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge." *Journal of the American Medical Informatics Association* 20 (5): 806–813.

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. 2015. "Normalization of relative and incomplete temporal expressions in clinical narratives". *Journal of the American Medical Informatics Association* 22 (5): 1001–1008.

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. 2013. "Temporal Reasoning over Clinical Text: the State of the Art". *Journal of the American Medical Informatics Association* 20 (5): 814–819.

Taira, Ricky K, Alex AT Bui, and Hooshang Kangarloo. 2002. "Identification of patient name references within medical documents using semantic selectional restrictions." In *Proceedings of the AMIA Symposium*, 757–761. American Medical Informatics Association.

Tanev, Hristo, Jakub Piskorski, and Martin Atkinson. 2008. "Real-time news event extraction for global crisis monitoring". In *Natural Language and Information Systems*, 207–218. Springer.

Tao, Cui, Harold R Solbrig, and Christopher G Chute. 2011. "CNTRO 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives". *AMIA Summits Transl Sci Proc* 2011:64–8.

Tobin, Richard, et al. 2010. "Evaluation of georeferencing". In *proceedings of the 6th workshop on geographic information retrieval*, 7. ACM.

Utvić, Miloš. 2014. "Izgradnja referentnog korpusa savremenog srpskog jezika". Doktorska disertacija, Univerzitet u Beogradu, Filološki fakultet.

Uzuner, Oezlem, et al. 2008. "A de-identifier for medical discharge summaries". *Artificial Intelligence in Medicine* 42 (1): 13–35. doi:10.1016/j.artmed.2007.10.001.

UzZaman, Naushad, and James Allen. 2010. "TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text". In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 276–283. Association for Computational Linguistics.

UzZaman, Naushad, et al. 2013. "SemEval-2013 Task 1: TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations". In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*.

Vazov, Nikolai. 2001. "A system for extraction of temporal expressions from French texts based on syntactic and semantic constraints". In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, 14. Association for Computational Linguistics.

Velupillai, Sumithra. 2014. "Temporal expressions in swedish medical text–a pilot study". In *Proceedings of BioNLP*, 88–92.

Vendler, Z. 1957. “Verbs and Times”. *Philosophical Review* 56:143–160.

Verhagen, Marc. 2010. “The Brandeis Annotation Tool”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), May 19-21*, ed. by Nicoletta Calzolari et al. Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

Verhagen, Marc, et al. 2005. “Automating temporal annotation with TARSQI”. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, 81–84. Association for Computational Linguistics.

Verhagen, Marc, et al. 2007. “Semeval-2007 task 15: Tempeval temporal relation identification”. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 75–80. Prague: ACL.

Verhagen, Marc, et al. 2010. “Semeval-2010 task 13: Tempeval-2”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 57–62. Uppsala, Sweden.

Vicente-Díez, Maía Teresa, Julián Moreno Schneider, and Paloma Martínez. 2010. “UC3M system: determining the extent, type and value of time expressions in TempEval-2”. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 329–332. Association for Computational Linguistics.

Vitas, Duško. 2006. *Prevodioci i interpretatori: uvod u teoriju i metode kompilacije programskih jezika*. Beograd: Matematički fakultet.

Vitas, Duško, et al. 2003. “A Processing Serbian Written Texts: An Overview of Resources and Basic Tools”. In *Workshop on Balkan Language Resources and Tools*, ed. by S Piperidis and V Karkaletsis, 97–104. Thessaloniki, Greece.

Vitas, Duško, et al. 2012. *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Dostupno na http://www.meta-net.eu/whitepapers. Springer. ISBN: 978-3-642-30754-6.

Wellner, Ben, et al. 2007. “Rapidly retargetable approaches to de-identification in medical records”. *Journal of the American Medical Informatics Association* 14 (5): 564–573. doi:10.1197/jamia.M2435.

Wenzina, Reinhardt, and Katharina Kaiser. 2014. “Towards the Application of TimeML in Clinical Guidelines”. In *Modellierung im Gesundheitswesen. Tagungsband des Workshops im Rahmen der Modellierung 2014*, 37–48. ICB-Research Report.

Wenzina, Reinhardt, and Katharina Kaiser. 2014. “Using TimeML to support the modeling of computerized clinical guidelines.” In *MIE*, 8–12.

Xiaojia, Zhou, et al. 2011. “Temporal expression recognition and temporal relationship extraction from Chinese narrative medical records”. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, 1–4. IEEE.

Xue, Nianwen, and Yuping Zhou. 2010. “Applying Syntactic, Semantic and Discourse Constraints in Chinese Temporal Annotation”. In *Coling 2010: Posters*, 1363–1372. Beijing, China.

Zavarella, Vanni, and Hristo Tanev. 2013. “FSS-TimEx for TempEval-3: Extracting temporal information from text”. *Atlanta, Georgia, USA*: 58.

Zhang, Chunxia, et al. 2008. “A transformation-based error-driven learning approach for Chinese temporal information extraction”. In *Information Retrieval Technology*, 663–669. Springer.

Zhou, Li, and George Hripcsak. 2007. “Temporal reasoning with medical data—a review with emphasis on medical natural language processing”. *Journal of biomedical informatics* 40 (2): 183–202.

Zhou, Li, et al. 2006. “A Temporal Constraint Structure for Extracting Temporal Information from Clinical Narrative”. *Journal of biomedical informatics* 39 (4): 424–439.

Znika, Marija. 1979. “O sustavu jedinica vremenske mjere”. *Rasprave Zavoda za jezik* (Zagreb) 4–5:69–80.

Prilozi

Prilog A

Primeri obeležavanja prepoznatih vremenskih izraza leksičkim etiketama

Primer A.1. Konkordance koje predstavljaju rezultat primene transduktora *Kalendarski datum*, odnosno izraze obeležene leksičkim etiketama

kanala Dunav-Tisa-Dunav.	{01. 02. 2010.,.NE+time+date+abs} - 00:02h Radnici Zastave od danas prelaze
Bariju.{S} Istraga je u toku.	{01.02.2010. godine,.NE+time+date+abs} Fijat preuzima prvih 1.000 radnika.
ako rešenja budu dostavljena u	{ponedeljak, 01. 02. 2010.,.NE+time+date+abs}, realno je da RRK proglasi
Vulović je ispričao kako se	{dana 01.02.2010. godine,.NE+time+date+abs} našao sa optuženim Ratiborom
nije ispunilo obavezu da do	{31.12.2008. god.,.NE+time+date+abs} , zatvori finansijsku konstrukciju kako bi
kompanija FCC i Alpina	{30. 3. 2007. god.,.NE+time+date+abs} potpisala ugovor o davanju koncesije
bivšu Jugoslaviju u Hagu je	{24. 07. 1995. g.,.NE+time+date+abs} podigao optužnicu protiv Karadžića i
Evrope. Tanjug Torino,	{07-03-2009.,.NE+time+date+abs} - 17:48. Reprezentativac Srbije u atletici
Bogunović (eventualno Ilić).	{01.03.2009. godine,.NE+time+date+abs} 19:57, Los Anđeles Prema pisanju
(Beta) PRENETO 23 Press	{2.15.2007.,.NE+time+date+abs} Vest 9 POKUŠAO DA PODMITI POLICAJCA Krivična
država. NN Večernje novosti	{3.28.2006.,.NE+time+date+abs} Vest 18 NOVA DONACIJA OSEČINA - Savremena
kako bi se ublažila kriza.	{01/03/2009,.NE+time+date+abs} 20:11, Beograd Danas sednica Monetarnog odbora
Ubijen policajac. Tanjug	{10/03/2009,.NE+time+date+abs} LONDON - U Severnoj Irskoj je ubijen policajac
stalnog dopisnika [objavljeno:	{04/02/2010,.NE+time+date+abs}] Zagreb - Uz izuzetno stroge mere bezbednosti
maraton, 42,195 km.{S}	{13. septembar 490. p.n.e.,.NE+time+date+abs} .{S} Na istočnoj obali Atike
Isus je preminuo u	{petak 3. aprila, 33. godine nove ere,.NE+time+date+abs} prema istraživanju
prognao ih iz Rima.{S}	{18. jula 64. godine naše ere,.NE+time+date+abs} Rim je u požaru, koji
izvojevala je pobedu	{9. 8. 480. pre Hrista,.NE+time+date+abs} posle trodnevne bitke u

Primer A.2. Konkordance koje predstavljaju rezultat primene transduktora *Datum prošireni*, odnosno izraze obeležene leksičkim etiketama

Izrazi koji ukazuju na tačku u vremenu

Tokom suđenja, koje je počelo	{10. jula 2006,.NE+time+date+abs}, a završeno je u avgustu prošle godine
Gudurić je bio u bekstvu od	{23. okt. 2008. god.,.NE+time+date+abs}, kada su u Zagrebu ubijeni Pukanić
više od 18 godina pauze, od	{1. III 2012. godine,.NE+time+date+abs} ponovo leti na liniji Ljubljana -
tri vojna objekta u Srbiji do	{subote, prvog juna 2002.,.NE+time+date+abs}, poslednji put na Povlenu kod
okončaju proces reformi do	{2007. godine, 1. januara,.NE+time+date+abs} i predviđenog roka za integraciju
prebačen iz Argentine,	{ove godine 21. februara,.NE+time+date+abs}, gde je uhapšen u avgustu 2005. godine
službi Policijske uprave	{24. XI prošle godine,.NE+time+date+abs} prijavio da je u gimnaziji podmetnuta
Srbije i Crne Gore će od	{1. januara sledeće godine,.NE+time+date+abs} biti potrebne vize.{S} Jedina šansa
Andreas Cobel potpisali su	{juče, 12. aprila 2006,.NE+time+date+abs} sporazum o vojnoj saradnji, a našem
ali to nisu uspeli.{S}	{21. aprila 753. godine p.n.e.,.NE+time+date+abs} Romul je, prema istoričaru
hrama - Zid plača.{S}	{8. VII 70. godine nove ere,.NE+time+date+abs} Rimljani su u Judejskom ratu
Sveti Marin je	{3. septembra 301. godine naše ere,.NE+time+date+abs} osnovao Republiku San Marino
diktature.{S} Julije Cezar je	{10. 01. 49. pre Hrista,.NE+time+date+abs} izgovorio čuvenu rečenicu *Alea iacta*

Izrazi koji ukazuju na trajanje

Vrhovnog saveta odbrane tokom {meseca aprila 2011. godine,.NE+time+duration+abs}, na kojoj je bilo reči
Kandić zbog izjave date u toku {juna 2007.,.NE+time+duration+abs} da je on 1991. u hrvatskom selu Antin
pitanja, pošto će Belgija tokom {novembra iduće godine,.NE+time+duration+abs} preuzeti predsedavanje Oebsu
očigledan način vladanja SPS tokom {1990-ih,.NE+time+duration+abs}.{S} Ministar spoljnih poslova Belgije Karol
svoje bogatstvo.{S} Tokom {osamdesetih prošlog veka,.NE+time+duration+abs} vratio se u Srbiju i ponovo
Ruski milijarderi su tokom {zime 2005.,.NE+time+duration+abs}, iako kriza još nije prošla, uspeli da
stvore preduslovi za to da se tokom {leta ove godine,.NE+time+duration+abs} pokrene postupak ratifikacije
proizvodnju autodelova Jura tokom {petog marta,.NE+time+duration+abs} treba da se sastane sa predstavnicima
spisak sporednih troškova u toku {meseca aprila,.NE+time+duration+abs} iznosi dodatnih 10.000 do 15.000 dinara
pregleda, koji je počeo u toku {10-og,.NE+time+duration+abs}, a na njihov zahtev obustavljen nakon redovnog
zdravstveno stanje stabilizuje tokom {ove godine,.NE+time+duration+abs} izašao pred sud.{S} Priprema planova
da je dosad jedino u EU, tokom {prošlog meseca,.NE+time+duration+abs} usvojena makroregionalna strategija za
dođe vrlo brzo.{S} Tokom {predstojećeg vikenda,.NE+time+duration+abs} i državnog praznika, primenjivaće
gori trebalo bi da počne u toku {godine,.NE+time+duration+abs}, a Vlada Srbije će finansirati izgradnju samo
taj predmet.{S} Vlada Srbije će tokom {sedmice,.NE+time+duration+abs} potpisati prve ugovore sa bankama za
s kraja prošle godine, evro će tokom {2013.,.NE+time+duration+abs} dostići barem 102 dinara, a ako se slabljenje

Primer A.3. Konkordance koje predstavljaju rezultat primene transduktora *Datum skraćeni*, odnosno izraze obeležene leksičkim etiketama

Hrvatske, koji je rođen u Zagrebu	{aprila 1965.,.NE+time+date+abs}.{S} Jedan je od retkih kandidata za
avionska nesreća tu desila	{1998. godine juna meseca,.NE+time+date+abs}.{S} Avion je krenuo
Dobrovoljačkoj ulici u Sarajevu	{početkom maja meseca 1992. godine,.NE+time+date+abs}.{S} Predmet je
o napretku Srbije ka EU, iz	{novembra prošle godine,.NE+time+date+rel}, bilo dosta negativnih ocena
primenjivali postojeće propise.{S} U	{drugoj polovini aprila ove godine,.NE+time+date+rel} navršiće se tačno
policijskoj akciji Memfis, u kojoj je	{sredinom aprila 2007. g.,.NE+time+date+abs} na beogradskom aerodromu
u Seriju A iz španske Maljorke	{2007. godine,.NE+time+date+abs}.{S} Međutim, posle postignutog dogovora
precizirani su na sastanku u Minhenu	{krajem 2007. god.,.NE+time+date+abs} i poslednji detalji ovog plana
I taj posao mora da se uradi	{početkom zime 2014. godine,.NE+time+date+abs}, jer posle para neće biti.
susretu.{S} Dominga sam upoznala	{2002,.NE+time+date+abs} u Vašingtonu na takmičenju Operalija.{S} Reprezent
kriminalnoj grupi osumnjičenoj da je	{drugog juna,.NE+time+date+rel} u Zemunu, pokušala ubistvo Nikole
Premijera Milionera sa ulice	{13. aprila,.NE+time+date+rel} će u Srbiji biti prikazan (distributer je
Najpoznatija japanska reprezentativka u	{julu,.NE+time+date+rel} će napuniti 34 godine.{S} Za veliki trud
mu biti uručena nagrada za MVP-ja	{oktobra meseca,.NE+time+date+rel}.{S} Partizanovi navijači imaće
Meč Partizan - Panatinaikos igra se	{5-og,.NE+time+date+rel}, a prodaja karata za utakmicu počinje
pretkrivičnog postupka.{S} Mi smo	{krajem prošle godine,.NE+time+date+rel} podneli zahtev za proširenje
zasedanja borda MMF, koji se održava	{sledećeg meseca,.NE+time+date+rel} i tada će se znati njihov stav i
Ninas fitnes klub je oformljen u	{prošlom milenijumu,.NE+time+date+rel}, a vlasnik je optužen za
u Benidormu.{S} Đoković će	{narednog petka,.NE+time+date+rel} predvoditi Dejvis kup reperezentaciju
koja je s najmanje primljenih golova	{jesenas,.NE+time+date+rel} bila najsigurnija u Superligi, nego
odredio pritvor.{S} Dragićević je	{početkom decenije,.NE+time+date+rel} uhapšen u okviru policijske akcije
Broj stanovnika u svetu će	{sredinom veka,.NE+time+date+rel} porasti na 9,7 milijardi, sa sadašnjih

Primer A.4. Konkordance koje predstavljaju rezultat primene transduktora *Period datuma*, odnosno izraze obeležene leksičkim etiketama

{{10. maja 1995.,.NE+time+date+abs} do {14. decembra 1996.,.NE+time+date+abs},.NE+time+date+period}

{{13. maja 1995.,.NE+time+date+abs} i {14. jula 1998.,.NE+time+date+abs},.NE+time+date+period}

{{početka marta 20015.,.NE+time+date+abs} do {kraja aprila ove godine,.NE+time+date+rel},.NE+time+date+period}

{{septembru,.NE+time+date+rel} i {oktobru 2015.,.NE+time+date+abs},.NE+time+date+period}

{{11,.NE+time+date+rel} ili {12. februara,.NE+time+date+rel},.NE+time+date+period}

{{12,.NE+time+date+rel} - {16. januara,.NE+time+date+rel},.NE+time+date+period}

{{1995,.NE+time+date+abs} - {1997.,.NE+time+date+abs},.NE+time+date+period}

{{2007,.NE+time+date+abs}/{08,.NE+time+date+abs},.NE+time+date+period}

{{2008,.NE+time+date+abs}, {2009.,.NE+time+date+abs} i {2011.,.NE+time+date+abs},.NE+time+date+period}

{{februaru,.NE+time+date+rel} ili {početkom marta,.NE+time+date+rel},.NE+time+date+period}

{{decembru,.NE+time+date+rel}, {januaru,.NE+time+date+rel} i {polovinom februara,.NE+time+date+rel}
,.NE+time+date+period}

{{drugoj polovini ove godine,.NE+time+date+rel} ili u {prvoj polovini 2011.,.NE+time+date+abs},.NE+time+date+period}

{{nedelju 1.,.NE+time+date+rel} i {ponedeljak 2. januara,.NE+time+date+rel},.NE+time+date+period}

Primer A.5. Konkordance koje predstavljaju rezultat primene transduktora *Učestalost*, odnosno izraze obeležene leksičkim etiketama

dovodi i u Srbiju.{S} Domingo	{svakodnevno,.NE+time+set} vežba sa njima i vodi ih na turneju
klijentima objave i dostave NBS	{jednom dnevno,.NE+time+set} i da sve transakcije obavljaju po
Studenti uglavnom peru svoj veš	{dva puta mesečno,.NE+time+set} i za tu uslugu im je potrebno
Letovi će se obavljati	{triput nedeljno,.NE+time+set}, ponedeljkom, utorkom i sredom.
po potrebi, i da održe sastanke	{5 puta mesečno,.NE+time+set}, naveo je norveški diplomata u
budućim supružnicima razgovaraće	{sedmično,.NE+time+set} tim od šest terapeuta, a radionice
Tako se mesecima spremao	{7 sati dnevno,.NE+time+set} za plivački maraton u Bejrutu, gde je
ugroziti zdravlje.{S} Vežba se	{dva dana nedeljno,.NE+time+set} ako želite da ostanete u dobroj
vreme.{S} U Beogradu je	{pet meseci godišnje,.NE+time+set} temperatura iznad 20 stepeni
ne smanjuju broj zaposlenih i da	{svakog meseca,.NE+time+set} dostavljaju Poreskoj upravi OD
postojanje uredbe, prema kojoj se	{svakih 15 dana,.NE+time+set} vrši korekcija propisa
od 30 radnih dana.{S} Na	{svakih pet godina,.NE+time+set} odmor će biti produžavan za jedan
radu parlamenta predviđeno je da	{svakog poslednjeg četvrtka,.NE+time+set} premijer i ministri
dužinu rada Skupštine za	{svaku sedmicu,.NE+time+set}.{S} Na sastanku je predloženo i da
Trajanove table i planine Miroč, a	{svake večeri,.NE+time+set} biće održavane žurke na plaži
Upravni odbor EBRD-a zaseda	{svakog drugog utorka,.NE+time+set}.{S} Poslednja sednica
godina, a ako se menja na	{svakih jedanaest godina,.NE+time+set} onda se radi o mnogo
na levoj obali Dunava radiće	{četiri sata svakog dana,.NE+time+set}.{S} Do sada su ove ustanove

Primer A.6. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje modifikovano*, odnosno izraze obeležene leksičkim etiketama

dosadašnju privatizaciju od	{skoro 20 godina,.NE+time+duration+abs}, prihod od privatizacije od oko
prošle godine analitičari su	{gotovo godinu i po,.NE+time+duration+abs} samo posmatrali berzu, dok
To je njihov drugi susret za	{manje od nedelju dana,.NE+time+duration+abs}.{S} Sastanak je počeo
oko pola devet ujutru i nakon	{više od tri sata i 15 minuta,.NE+time+duration+abs} leta trebalo je da sleti
poslanička pitanja može trajati	{najduže sat i dvadeset pet minuta,.NE+time+duration+abs}, a obavlja se
da Turska usvoji zakon za	{manje od desetak dana,.NE+time+duration+abs}.{S} Vlada Srbije prihvatila
za koja preti zatvorska kazna	{duža od 10 meseci,.NE+time+duration+abs}.{S} Produženje trajanja
srpskih pripovedača već	{skoro tri decenije,.NE+time+duration+abs}.{S} Albahari je jedan od
cd-ova i rokovnika starih	{preko 15 godina,.NE+time+duration+abs}, kao i određenu sumu novca.
strategiju velikih razmera, jer će	{idućih godinu i po dana,.NE+time+duration+abs} moći da računa
dok lekari upozoravaju da će u	{naredna dva dana,.NE+time+duration+abs} umreti još ljudi ukoliko
pa su i optužbe koje su	{poslednjih petnaestak dana,.NE+time+duration+abs} znak nezadovoljstva
takvu pretragu nije obavio u	{proteklih šest godina,.NE+time+duration+abs}, proveravajući iskaz svedoka
poslove precizirao da je za	{ta tri meseca,.NE+time+duration+abs} prijavljeno za 620 krivičnih dela
otmičari kriju u gradu Sabratu,	{sledećih sat i 45 minuta,.NE+time+duration+abs} ključno u rešavanju ove
ove pojave.{S} Širom Srbije je u	{prethodna 3 sata,.NE+time+duration+abs} zabeleženo čak 17 slučajeva
to je najjači potres u ovoj zemlji u	{poslednjih sto godina,.NE+time+duration+abs}.{S} Epicentar zemljotresa

Primer A.7. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje u kontekstu*, odnosno izraze obeležene leksičkim etiketama

novosti.{S} Ovo pravilo od pre	{tri nedelje,.NE+time+duration+abs} kada je doneto na preporuku Republičke
poslova u izveštaju navodi da je za	{18 meseci,.NE+time+duration+abs} smanjen broj zaposlenih za oko 200
prvog poena.{S} Meč je trajao	{sat i 32 minuta,.NE+time+duration+abs}, a selektor Srbije Dejan Vraneš
Skupštini Srbije na usvajanje već za	{mesec i po,.NE+time+duration+abs} ako bi se na tome intenzivno radilo
iskoristio brejk priliku, pa je posle	{jednog sata i 36 minuta,.NE+time+duration+abs} igre Đoković iskoristio
biti održan u Moskvi.{S} Posle	{dve večeri,.NE+time+duration+abs} festivala Beovizija konačno smo dobili
zvani Bajone, koji je bio u bekstvu	{10 i po meseci,.NE+time+duration+abs} i za kojim je hrvatsko pravosuđe
komercijalnih letova, prvi put posle	{pet decenija,.NE+time+duration+abs}.{S} Američki sekretar za
određena mera zadržavanja u trajanju od	{48 sati,.NE+time+duration+abs}.{S} Nakon toga, on će biti pušten
isporuka uglja termoelektrani na	{2,5 sata,.NE+time+duration+abs} i da će štrajk i biti nastavljen
razmatrali pitanje obelezxavanja	{1.700 godina,.NE+time+duration+abs} Milanskog edikta, koje će se 2013.
u Bulevaru biti posađen do jeseni, za	{180 dana,.NE+time+duration+abs}.{S} Takva promena će koštati oko 40
saopštila je juče da će u periodu od	{petnaestak dana,.NE+time+duration+abs} posle referenduma u Crnoj Gori
nakon izdržane zatvorske kazne od	{jedne i po godine,.NE+time+duration+abs} pušteno još 13 osuđenih
Krokobabića s gostima iz Vašingtona	{oko sat i po,.NE+time+duration+abs} i, mimo svih očekivanja, nije ni
rasta.{S} Materijal se resorbuje za	{oko šest meseci,.NE+time+duration+abs}, a ceo proces obnavljanja
da bi sednica bila prekinuta posle	{približno 2 sata i 15 minuta,.NE+time+duration+abs} rasprave odbora za
obaviti.{S} Bivši američki vojnik će za	{oko dve sedmice,.NE+time+duration+abs} postati prvi građanin SAD-a

Primer A.8. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana*, odnosno izraze obeležene leksičkim etiketama

Izrazi koji ukazuju na tačku u vremenu

B92 četvrtak,	{01:43:45,.NE+time+hour+abs} {S} Statutarno pravna
kod Kraljeva	{rano jutros oko 6 h,.NE+time+hour+abs} u sudaru vozila
kolena u	{15 časova,.NE+time+hour+abs} nije mogao da igra u meču
zakazan za	{13:35 h,.NE+time+hour+abs}, pa traži produženje za još
na programu	{oko 17.30,.NE+time+hour+abs}.{S} Nagradni fond
je počeo u	{3 časa poslepodne,.NE+time+hour+abs}.{S} Dragan Ilić
je pozvao u	{7 sati ujutro,.NE+time+hour+abs} kazavši mu da odmah
odnosno	{četiri sata jutros po lokalnom vremenu,.NE+time+hour+abs}
stigne do	{deset časova,.NE+time+hour+abs} saopštili su u centru
zakazan za	{dva sata i 50 minuta,.NE+time+hour+abs} ipak pomeren
u terminu od	{pola devet uveče,.NE+time+hour+abs}.{S} Svi kupci
Međutim,	{dva sata posle ponoći,.NE+time+hour+abs} njihovi
odjeknule su	{noćas oko dva,.NE+time+hour+abs} u centru grada
dogodila se	{oko ponoći,.NE+time+hour+abs} kada je iz za sada
dozvolama u	{ponoć,.NE+time+hour+abs} i da otad nijedan srpski
Grujića 11,	{sinoć oko 22 časa,.NE+time+hour+abs} začuo se pucanj

Izrazi koji ukazuju na trajanje

misiju MMF-a tokom	{poslepodneva,.NE+time+duration+abs} poveo je u
uhapšena su tokom	{noći,.NE+time+duration+abs} u Austriji, u blizini
navodi da bi tokom	{prepodneva,.NE+time+duration+abs} stigla isporuka
jedan od njih tokom	{prethodne noći,.NE+time+duration+abs} operisan i da
dogovorio u toku	{prethodne večeri,.NE+time+duration+abs}, kada su ga

Primer A.9. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana skraćeno*, odnosno izraze obeležene leksičkim etiketama

pa je autobus u {7.30 sati,.NE+time+hour+abs} krenuo iz Beograda ka
oglašavaju se od {08:00 časova,.NE+time+hour+abs}, što je označilo
dogovoreno za {10 časova,.NE+time+hour+abs} u zgradi novosadskog
će se sastati do {11 sati,.NE+time+hour+abs}, što nije učinjeno zbog
da ističe posle {14 časova,.NE+time+hour+abs}, saopštilo je
neće stići pre {2 sata i 15 minuta,.NE+time+hour+abs}, kada počinje
primljen je u {17 h,.NE+time+hour+abs} u Klinički centar Srbije
objavila je {oko 15.35,.NE+time+hour+abs} RTV B92.{S} Dve
sreću, upalile {oko 12 i 15,.NE+time+hour+abs} i dalji rad Vlade
pomereno za {7 uveče,.NE+time+hour+abs}, nakon prenosa
organi su od {osam,.NE+time+hour+abs}, preko medija obaveštavali
pravde, izvedena {oko pola dva,.NE+time+hour+abs}, osumnjičenom
zadatak da u {18 sati,.NE+time+hour+abs}, po nalogu Šarića
poručio je u {20 časova,.NE+time+hour+abs}, državni sekretar
pregovorima do {pola tri,.NE+time+hour+abs}, finale razgovora
dogodila se u {23 časa i 59 minuta,.NE+time+hour+abs} i još se ne
institucijama, za {dva popodne,.NE+time+hour+abs} ugovorenu proveru
predstavio do {9.30,.NE+time+hour+abs}, kada je očekivana
neće stići pre {18.30 časova,.NE+time+hour+abs}, osim ako
zabrana do {21 čas,.NE+time+hour+abs} za sve članove
sa početkom u {15.15 časova,.NE+time+hour+abs} na Fakultetu
nepoznatih su {oko 7,.NE+time+hour+abs} sedeli na terasi
opljačkali su {oko dva,.NE+time+hour+abs}, saopštila je policija
igrati u petak od {13 časova i 15 minuta,.NE+time+hour+abs} samo
čekali su do {dva i 15,.NE+time+hour+abs}, da bi nakon toga

Primer A.10. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana u kontekstu*, odnosno izraze obeležene leksičkim etiketama

a narodnjaci polaze u {cik zore,.NE+time+hour+rel}.{S} J. Cerovina Beograd, 17. februara
otežano se odvijao posebno u {jutarnjim časovima,.NE+time+hour+rel} i na seoskom području.{S} Prema
kompaniji, telefoni usijali od {ranog jutra,.NE+time+hour+rel}.{S} Građani su se raspitivali o subve
prolazim ovuda i tvrdim da se ovde {noću,.NE+time+hour+rel} svašta dešava, ali od ovog prizora su mi zaklecale
Mađarska policija saopštila je {jutros,.NE+time+hour+rel} da su dvojica od trojice osumnjičenih za ubistvo
Vuk Tufegdžić uhapšen je juče {kasno popodne,.NE+time+hour+rel}.{S} Agenciji Beta nije potvrđeno da li je
na sednici Saveta bezbednosti {kasno sinoć,.NE+time+hour+rel} ocenili su da vlasti Srbije ne sarađuju u
radikalne stranke (SRS) izvučen je {noćas,.NE+time+hour+rel} iz kanala Dunav-Tisa-Dunav.{S} Predrag Marić,
policija pokušava da otkrije od juče {po podne,.NE+time+hour+rel}.{S} Na osnovu tragova guma i probijene ograde
međubankarskom deviznom tržištu {popodne,.NE+time+hour+rel} bio samo 4,8 miliona evra.{S} On je ipak dodao
Vladi Republike Srpske, Tadić je u {popodnevnim časovima,.NE+time+hour+rel} trebalo da se sastane s najvišim
Saznajemo da je Albahari juče {posle podne,.NE+time+hour+rel} iz Kanade potvrdio da dolazi u Beograd
štimungu.{S} Ali, u nedeljno {praskozorje,.NE+time+hour+rel}, odjednom su sevnuli noževi.{S} Na
koji su ubrzo i smenjene, a {preksinoć,.NE+time+hour+rel} beogradski parlament je po kratkom postupku
Priština je blokiran u ponedeljak u {prepodnevnim satima,.NE+time+hour+rel}.{S} Pošto ova deonica magistrale
spasavanje MUP-a Srbije, rekao je {rano jutros,.NE+time+hour+rel} agenciji Beta da se u vozilu nalazila tri
Međutim, i ova pretpostavka je {sinoć,.NE+time+hour+rel} uzeta s rezervom.{S} Nastradalima se, naime,
{8. marta,.NE+time+date+rel} u {večernjim časovima,.NE+time+hour+rel} u blizini "Brodoremonta".{S} U
se, naime, trag gubi od {prethodne noći, {sat posle ponoći,.NE+time+hour+abs},.NE+time+hour+abs}
demokrate.{S} Beograd - {Večeras u {ponoć,.NE+time+hour+abs},.NE+time+hour+abs} ističe

Primer A.11. Konkordance koje predstavljaju rezultat primene transduktora *Period vremena dana*, odnosno izraze obeležene leksičkim etiketama

{{10,.NE+time+hour+abs} do {11 časova,.NE+time+hour+abs},.NE+time+hour+period}

{{12,.NE+time+hour+abs} i {17 sati,.NE+time+hour+abs},.NE+time+hour+period}

{{8,.NE+time+hour+abs} do {podneva,.NE+time+hour+abs},.NE+time+hour+period}

{{20 časova,.NE+time+hour+abs} i {22 časa,.NE+time+hour+abs},.NE+time+hour+period}

{{11,.NE+time+hour+abs} ili {12 h,.NE+time+hour+abs},.NE+time+hour+period}

{{17,.NE+time+hour+abs} - {18 sati,.NE+time+hour+abs},.NE+time+hour+period}

{{9,.NE+time+hour+abs}, {11,.NE+time+hour+abs} i {13 časova,.NE+time+hour+abs},.NE+time+hour+period}

{{sinoć,.NE+time+hour+rel} ili {jutros,.NE+time+hour+rel},.NE+time+hour+period}

{{15,.NE+time+hour+abs}, {17,.NE+time+hour+abs} i {19 h,.NE+time+hour+abs},.NE+time+hour+period}

{{sedam,.NE+time+hour+abs} ili u {osam časova,.NE+time+hour+abs},.NE+time+hour+period}

{{pet popodne,.NE+time+hour+abs} i {11 ujutru,.NE+time+hour+abs},.NE+time+hour+period}

Primer A.12. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje osnovno*, odnosno izraze obeležene leksičkim etiketama

radije bio {1000 godina,.NE+time+duration+abs} u Hagu nego prihvatio
raspisan za {11 i po meseci,.NE+time+duration+abs}, a održan u
traje samo {12,4 godine,.NE+time+duration+abs}.{S} Analize
građani su {18 sati,.NE+time+duration+abs} čekali odgovor iz
potrebna {24 časa,.NE+time+duration+abs}.{S} Sve se obavlja
vratio se pre {tri nedelje i dva dana,.NE+time+duration+abs} i odlučio
možda dugo {25 godina i 220 dana,.NE+time+duration+abs}, što nije
što je pre {dve nedelje,.NE+time+duration+abs} izraelskoj teniserki
Kosovu.{S} {Vek i po,.NE+time+duration+abs} je bilo ovde, a
već u prvih {desetak minuta,.NE+time+duration+abs} imali po tri
kaznu na {godinu i šest meseci,.NE+time+duration+abs}.{S} Golubov
da odigram {jednu godinu,.NE+time+duration+abs} u Americi, ali ne
imenovana pre {oko dva meseca,.NE+time+duration+abs}.{S} Odmah
Troicki i za {sat i 14 minuta,.NE+time+duration+abs} preslišao
loptu za {sat,.NE+time+duration+abs} igre i istim rezultatom
i da živi {sto godina,.NE+time+duration+abs}, te da ima oko
crkve je {tri i po decenije,.NE+time+duration+abs} bio episkop
poskupelo za {mesec dana,.NE+time+duration+abs} i struja će
gubi od pre {dve noći,.NE+time+duration+abs}, kada su im

Primer A.13. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje relativno*, odnosno izraze obeležene leksičkim etiketama

neefikasnosti parlamenta.{S}	{Danima,.NE+time+duration+rel} se spekuliše koliko je koštao njen
sagovornika, oseća dušu narodnu,	{decenijama,.NE+time+duration+rel} je posvećen svojoj veri i svom
strategiju velikih razmera, jer će	{idućih godina,.NE+time+duration+rel} moći da računa i na priliv
struje i uglja ne samo više nego u	{mesecima,.NE+time+duration+rel} pre poplava, već su potavljeni
nastavljena i ova akcija tako da se	{narednih dana,.NE+time+duration+rel} očekuju nova hapšenja.{S}
ustava može napisatiu roku od	{narednih nekoliko nedelja,.NE+time+duration+rel}.{S} Nacrt ustava
vlade Kosova, koja je trajala	{prošlih par minuta,.NE+time+duration+rel} da će vladini prioriteti biti
i odrediti put kojim će ići u	{narednim godinama,.NE+time+duration+rel}.{S} Ako bi se opredelili na
crnogorske policije, uhapšen je	{nekoliko časova,.NE+time+duration+rel} pošto se njegov saradnik
što je skrivila lokalna uprava, jer	{više decenija,.NE+time+duration+rel} nisu isplaćivale doprinose na
otpremninu napuste prosvetu.{S}	{Ovih dana,.NE+time+duration+rel} ćemo imati sastanak i sa
za Mladićem pripadnici policije su	{ovih nekoliko sati,.NE+time+duration+rel} pretresali porodičnu
Vaša uloga biće veoma važna u	{ovim godinama,.NE+time+duration+rel} približavanja Evropskoj
Srbiji, verovatno odlučiti da za	{par meseci,.NE+time+duration+rel} odloži podnošenje predloga
državi, pa su i optužbe koje su	{poslednjih dana,.NE+time+duration+rel} znak nezadovoljstva
sirove nafte.{S} Iako je u	{poslednjih nekoliko sedmica,.NE+time+duration+rel} zabeležen pad
Trojkom Marinovića u	{poslednjim sekundama,.NE+time+duration+rel} Zvezda na kratku
za plasman u dalji krug	{prethodnih dana,.NE+time+duration+rel} kružila glasina da će Željko
veliki problem, kojim se država	{proteklih godina,.NE+time+duration+rel} nije dovoljno bavila
dresu Milana i da, za razliku od	{ranijih godina,.NE+time+duration+rel} kada je takođe
uticala je i jaka kiša koja je	{više časova,.NE+time+duration+rel} padala u okolini Knjaževca
s obzirom na to da je DEA u	{zadnjih nekoliko meseci,.NE+time+duration+rel} u više navrata

Primer A.14. Konkordance koje predstavljaju rezultat primene transduktora *Datum relativni*, odnosno izraze obeležene leksičkim etiketama

Na sednici u	{četvrtak,.NE+time+date+rel} Vlada je usvojila
troškovi za penzije	{danas,.NE+time+date+rel} su oko 12 odsto BDP
afera koja ni do	{današnjeg dana,.NE+time+date+rel} nije potpuno
pošto će Belgija	{iduće godine,.NE+time+date+rel} preuzeti presed
tom cilju, Đelić će	{idućeg utorka,.NE+time+date+rel} na samitu koji
troškovi života za	{isti mesec,.NE+time+date+rel} povećani su 10,4
najavili da će od	{istog dana,.NE+time+date+rel} na svakoj gradskoj
države ni o ovome	{juče,.NE+time+date+rel} nije imalo jedinstven
više u odnosu na	{jučerašnji dan,.NE+time+date+rel}.{S} Narodna
potvrdio je da se	{minule sedmice,.NE+time+date+rel} sastao u gradu
u završnoj fazi,	{minulog leta,.NE+time+date+rel} tvrdili su da su
budućnost Balkana u	{narednoj deceniji,.NE+time+date+rel}.{S} Što se
Đoković će	{narednog vikenda,.NE+time+date+rel} predvoditi
plus ocenio da je	{ova godina,.NE+time+date+rel} bila godina
najavio je da će	{ove nedelje,.NE+time+date+rel} predložiti
neto zarada u	{ovom mesecu,.NE+time+date+rel} nominalno je
politika za	{ovu godinu,.NE+time+date+rel} bazirana na više
ambiciozno ušla u	{predstojeću godinu,.NE+time+date+rel} i kada je
dogodila se	{prekjuče,.NE+time+date+rel} na Uvačkom jezeru
otpadnih voda	{pretprošle godine,.NE+time+date+rel} smo uložili
stanicu Palilula	{prošle srede,.NE+time+date+rel}, kada je
Srbije trebalo bi	{sutra,.NE+time+date+rel}, između ostalog,
dela.{S} Ona je	{toga dana,.NE+time+date+rel}, kako nezvanično

Primer A.15. Konkordance koje predstavljaju rezultat primene transduktora *Period trajanja*, odnosno izraze obeležene leksičkim etiketama

{{godinu,.NE+time+duration+abs} - {dve,.NE+time+duration+abs},.NE+time+duration+period}

{{dva,.NE+time+duration+abs} - {tri dana,.NE+time+duration+abs},.NE+time+duration+period}

{{2 meseca,.NE+time+duration+abs} do {pola godine,.NE+time+duration+abs},.NE+time+duration+period}

{{2 dana,.NE+time+duration+abs} i {10 sati,.NE+time+duration+abs},.NE+time+duration+period}

{{dve nedelje,.NE+time+duration+abs} ili {mesec dana,.NE+time+duration+abs},.NE+time+duration+period}

{{17,.NE+time+duration+abs} - {18 godina,.NE+time+duration+abs},.NE+time+duration+period}

{{6,.NE+time+duration+abs} na {8 meseci,.NE+time+duration+abs},.NE+time+duration+period}

{{juče,.NE+time+date+rel} do {petka,.NE+time+date+rel},.NE+time+date+period}

{{1968. godine,.NE+time+date+abs} do {danas,.NE+time+date+rel},.NE+time+date+period}

{{četvrtak,.NE+time+date+rel} i {petak,.NE+time+date+rel},.NE+time+date+period}

{{decembar prošle godine,.NE+time+date+rel} i {januar 2006.,.NE+time+date+abs},.NE+time+date+period}

Primer A.16. Konkordance koje predstavljaju rezultat primene transduktora *Vreme*, odnosno izraze obeležene leksičkim etiketama

{{01.02.2010.,.NE+time+date+abs} {20:41:59,.NE+time+hour+abs},.NE+time+hour}

{{24. avgusta 2008.,.NE+time+date+abs}, {oko tri sata ujutru,.NE+time+hour+abs},.NE+time+hour}

{{25 februara,.NE+time+date+rel} u {12 časova,.NE+time+hour+abs},.NE+time+hour}

{{30. januara,.NE+time+date+rel} {ujutru,.NE+time+hour+rel},.NE+time+hour}

{{8. marta,.NE+time+date+rel} u {večernjim časovima,.NE+time+hour+rel},.NE+time+hour}

{{oko ponoći,.NE+time+hour+abs} između {{četvrtka,.NE+time+date+rel} i {petka,.NE+time+date+rel}},.NE+time+date+period}
,.NE+time+hour}

{{nedelju, 21. maja,.NE+time+date+rel}, {ujutru u 8,.NE+time+hour+abs},.NE+time+hour}

{{ponedeljak,.NE+time+date+rel} {uveče oko jedan,.NE+time+hour+abs},.NE+time+hour}

{{subotu, 28, maja,.NE+time+date+rel}, od {{11,.NE+time+hour+abs} do {13 sati,.NE+time+hour+abs}},.NE+time+hour+period}
,.NE+time+hour}

{{noći,.NE+time+hour+rel} između {{19,.NE+time+date+rel} i {20. januara ove godine,.NE+time+date+rel}}
,.NE+time+date+period},.NE+time+hour}

{{juče,.NE+time+date+rel} u {podne,.NE+time+hour+abs},.NE+time+hour}

{{narednog dana,.NE+time+date+rel} {oko pet sati popodne,.NE+time+hour+abs},.NE+time+hour}

{{danas,.NE+time+date+rel} u {9:45 sati,.NE+time+hour+abs},.NE+time+hour}

Prilog B

Primeri obeležavanja prepoznatih vremenskih izraza XML etiketama

Primer B.1. Konkordance koje predstavljaju rezultat primene transduktora *Kalendarski datum*, odnosno izraze obeležene XML etiketama

kanala Dunav-Tisa-Dunav.	<time.date.abs>01. 02. 2010.</time.date.abs> - 00:02h Radnici Zastave od danas
Bariju.{S} Istraga je u toku.	<time.date.abs>01.02.2010. godine</time.date.abs> Fijat preuzima prvih 1.000
ako rešenja budu dostavljena u	<time.date.abs>ponedeljak, 01. 02. 2010.</time.date.abs>, realno je da RRK
Vulović je ispričao kako se	<time.date.abs>dana 01.02.2010. godine</time.date.abs> našao sa optuženim
nije ispunilo obavezu da do	<time.date.abs>31.12.2008. god.</time.date.abs> , zatvori finansijsku konstrukciju
kompanija FCC i Alpina	<time.date.abs>30. 3. 2007. god.</time.date.abs> potpisala ugovor o davanju
bivšu Jugoslaviju u Hagu je	<time.date.abs>24. 07. 1995. g.</time.date.abs> podigao optužnicu protiv
Evrope. Tanjug Torino,	<time.date.abs>07-03-2009.</time.date.abs> - 17:48. Reprezentativac Srbije u atletici
Bogunović (eventualno Ilić).	<time.date.abs>01.03.2009. godine</time.date.abs> 19:57, Los Anđeles Prema pisanju
(Beta) PRENETO 23 Press	<time.date.abs>2.15.2007.</time.date.abs> Vest 9 POKUŠAO DA PODMITI POLICAJCA
država. NN Večernje novosti	<time.date.abs>3.28.2006.</time.date.abs> Vest 18 NOVA DONACIJA OSEČINA - Savremena
kako bi se ublažila kriza.	<time.date.abs>01/03/2009</time.date.abs> 20:11, Beograd Danas sednica Monetarnog
Ubijen policajac. Tanjug	<time.date.abs>10/03/2009</time.date.abs> LONDON - U Severnoj Irskoj je ubijen
stalnog dopisnika [objavljeno:	<time.date.abs>04/02/2010</time.date.abs>] Zagreb - Uz izuzetno stroge mere
maraton, 42,195 km.{S}	<time.date.abs>13. septembar 490. p.n.e.</time.date.abs> .{S} Na istočnoj obali
Isus je preminuo u	<time.date.abs>petak 3. aprila, 33. godine nove ere</time.date.abs> prema
prognao ih iz Rima.{S}	<time.date.abs>8. jula 64. godine naše ere</time.date.abs> Rim je u požaru,
izvojevala je pobedu	<time.date.abs>9. 8. 480. pre Hrista</time.date.abs>posle trodnevne bitke u

Primer B.2. Konkordance koje predstavljaju rezultat primene transduktora *Datum prošireni*, odnosno izraze obeležene XML etiketama

Izrazi koji ukazuju na tačku u vremenu

suđenja, koje je počelo	<time.date.abs>10. jula 2006</time.date.abs>, a završeno je u avgustu prošle godine
je bio u bekstvu od	<time.date.abs>23. okt. 2008. god.</time.date.abs>, kada su u Zagrebu ubijeni Pukanić
od 18 godina pauze, od	<time.date.abs>1. III 2012. godine</time.date.abs> ponovo leti na liniji Ljubljana -
vojna objekta u Srbiji do	<time.date.abs>subote, prvog juna 2010.</time.date.abs>, poslednji put na Povlenu kod
proces reformi do	<time.date.abs>2007. godine, 1. januara</time.date.abs> i predviđenog roka za integraciju
prebačen iz Argentine,	<time.date.abs>ove godine 21. februara</time.date.abs>, gde je uhapšen u avgustu 2005.
službi Policijske uprave	<time.date.abs>24. XI prošle godine</time.date.abs> prijavio da je u gimnaziji podmetnuta
i Crne Gore će od	<time.date.abs>1. januara sledeće godine</time.date.abs> biti potrebne vize.{S} Jedina
Cobel potpisali su	<time.date.abs>juče, 12. aprila 2006</time.date.abs> sporazum o vojnoj saradnji,
ali to nisu uspeli.{S}	<time.date.abs>21. aprila 753. godine p.n.e.</time.date.abs> Romul je, prema istoričaru
hrama - Zid plača.{S}	<time.date.abs>8. VII 70. godine nove ere</time.date.abs> Rimljani su u Judejskom ratu
Sveti Marin je	<time.date.abs>3. septembra 301. godine naše ere</time.date.abs> osnovao Republiku San
Julije Cezar je	<time.date.abs>10. 01. 49. pre Hrista</time.date.abs> izgovorio čuvenu rečenicu *Alea*

Izrazi koji ukazuju na trajanje

odbrane tokom <time.duration.abs>meseca aprila 2011. godine</time.duration.abs>, na kojoj je bilo reči
date u toku <time.duration.abs>juna 2007.</time.duration.abs> da je on 1991. u hrvatskom selu Antin
Belgija tokom <time.duration.abs>novembra iduće godine</time.duration.abs> preuzeti predsedavanje Oebsu
vladanja SPS tokom <time.duration.abs>1990-ih</time.duration.abs>.{S} Ministar spoljnih poslova Belgije
bogatstvo.{S} Tokom <time.duration.abs>osamdesetih prošlog veka</time.duration.abs> vratio se u Srbiju
milijarderi su tokom <time.duration.abs>zime 2005.</time.duration.abs>, iako kriza još nije prošla, uspeli da
za to da se tokom <time.duration.abs>leta ove godine</time.duration.abs> pokrene postupak ratifikacije
autodelova Jura tokom <time.duration.abs>petog marta</time.duration.abs> treba da se sastane sa
troškova u toku <time.duration.abs>meseca aprila</time.duration.abs> iznosi dodatnih 10.000 do 15.000
je počeo u toku <time.duration.abs>10-og</time.duration.abs>, a na njihov zahtev obustavljen nakon redovnog
stabilizuje tokom <time.duration.abs>ove godine</time.duration.abs> izašao pred sud.{S} Priprema planova
jedino u EU, tokom <time.duration.abs>prošlog meseca</time.duration.abs> usvojena makroregionalna
vrlo brzo.{S} Tokom <time.duration.abs>predstojećeg vikenda</time.duration.abs> i državnog praznika,
bi da počne u toku <time.duration.abs>godine</time.duration.abs>, a Vlada Srbije će finansirati izgradnju
Vlada Srbije će tokom <time.duration.abs>sedmice</time.duration.abs> potpisati prve ugovore sa bankama za
godine, evro će tokom <time.duration.abs>2013.</time.duration.abs> dostići barem 102 dinara, a ako se slabljenje

Primer B.3. Konkordance koje predstavljaju rezultat primene transduktora *Datum skraćeni*, odnosno izraze obeležene XML etiketama

rođen u Zagrebu	<time.date.abs>aprila 1965.</time.date.abs>.{S} Jedan je od retkih kandidata za
nesreća tu desila	<time.date.abs>1998. godine juna meseca</time.date.abs>.{S} Avion je krenuo
ulici u Sarajevu	<time.date.abs>početkom maja meseca 1992. godine</time.date.abs>.{S} Predmet je
Srbije ka EU, iz	<time.date.rel>novembra prošle godine</time.date.rel>, bilo dosta negativnih ocena
ove propise.{S} U	<time.date.rel>drugoj polovini aprila ove godine</time.date.rel> navršiće se tačno
akciji Memfis, u kojoj je	<time.date.abs>sredinom aprila 2007. g.</time.date.abs> na beogradskom aerodromu
A iz španske Maljorke	<time.date.abs>2007. godine</time.date.abs>.{S} Međutim, posle postignutog dogovora
sastanku u Minhenu	<time.date.abs>krajem 2007. god.</time.date.abs> i poslednji detalji ovog plana
mora da se uradi	<time.date.abs>početkom zime 2014. godine</time.date.abs>, jer posle para neće biti.
Dominga sam upoznala	<time.date.abs>2002</time.date.abs> u Vašingtonu na takmičenju Operalija.{S} Reprezent
grupi osumnjičenoj da je	<time.date.rel>drugog juna</time.date.rel> u Zemunu, pokušala ubistvo Nikole
Milionera sa ulice	<time.date.rel>13. aprila</time.date.rel> će u Srbiji biti prikazan (distributer je
japanska reprezentativka u	<time.date.rel>julu</time.date.rel> će napuniti 34 godine.{S} Za veliki trud
uručena nagrada za MVP-ja	<time.date.rel>oktobra meseca</time.date.rel>.{S} Partizanovi navijači imaće
Panatinaikos igra se	<time.date.rel>5-og</time.date.rel>, a prodaja karata za utakmicu počinje
postupka.{S} Mi smo	<time.date.rel>krajem prošle godine</time.date.rel> podneli zahtev za proširenje
MMF, koji se održava	<time.date.rel>sledećeg meseca</time.date.rel> i tada će se znati njihov stav i
fitnes klub je oformljen u	<time.date.rel>prošlom milenijumu</time.date.rel>, a vlasnik je optužen za
Benidormu.{S} Đoković će	<time.date.rel>narednog petka</time.date.rel> predvoditi Dejvis kup reperezentaciju
najmanje primljenih golova	<time.date.rel>jesenas</time.date.rel> bila najsigurnija u Superligi, nego
pritvor.{S} Dragićević je	<time.date.rel>početkom decenije</time.date.rel> uhapšen u okviru policijske akcije
stanovnika u svetu će	<time.date.rel>sredinom veka</time.date.rel> porasti na 9,7 milijardi, sa sadašnjih

Primer B.4. Konkordance koje predstavljaju rezultat primene transduktora *Period datuma*, odnosno izraze obeležene XML etiketama

<time.date.period><time.date.abs>10. maja 1995.</time.date.abs> do <time.date.abs>14. decembra 1996.</time.date.abs> </time.date.period>

<time.date.period><time.date.abs>13. maja 1995.</time.date.abs> i <time.date.abs>14. jula 1998.</time.date.abs> </time.date.period>

<time.date.period><time.date.abs>početka marta 2015.</time.date.abs> do <time.date.rel>kraja aprila ove godine </time.date.rel></time.date.period>

<time.date.period><time.date.rel>septembru</time.date.rel> i <time.date.abs>oktobru 2015.</time.date.abs> </time.date.period>

<time.date.period><time.date.rel>11</time.date.rel> ili <time.date.rel>12. februara</time.date.rel></time.date.period>

<time.date.period><time.date.rel>12</time.date.rel>- <time.date.rel>16. januara</time.date.rel></time.date.period>

<time.date.period><time.date.abs>1995</time.date.rel> - <time.date.abs>1997.</time.date.abs></time.date.period>

<time.date.period><time.date.abs>2007</time.date.rel>/<time.date.abs>08</time.date.abs></time.date.period>

<time.date.period><time.date.abs>2008</time.date.rel>, <time.date.abs>2009.</time.date.abs> i <time.date.abs>2011. </time.date.abs></time.date.period>

<time.date.period><time.date.rel>februaru</time.date.rel> ili <time.date.rel>početkom marta</time.date.rel> </time.date.period>

<time.date.period><time.date.rel>decembru</time.date.rel>, <time.date.rel>januaru</time.date.rel> i <time.date.rel> polovinom februara</time.date.rel></time.date.period>

<time.date.period><time.date.rel>drugoj polovini ove godine</time.date.rel> ili u <time.date.abs>prvoj polovini 2011. </time.date.abs></time.date.period>

<time.date.period><time.date.rel>nedelju 1.</time.date.rel> i <time.date.rel>ponedeljak 2. januara</time.date.rel> </time.date.period>

Primer B.5. Konkordance koje predstavljaju rezultat primene transduktora *Učestalost*, odnosno izraze obeležene XML etiketama

dovodi i u Srbiju.{S} Domingo	<time.set>svakodnevno</time.set> vežba sa njima i vodi ih na turneju
klijentima objave i dostave NBS	<time.set>jednom dnevno</time.set> i da sve transakcije obavljaju po
Studenti uglavnom peru svoj veš	<time.set>dva puta mesečno</time.set> i za tu uslugu im je potrebno
Letovi će se obavljati	<time.set>triput nedeljno</time.set>, ponedeljkom, utorkom i sredom.
po potrebi, i da održe sastanke	<time.set>5 puta mesečno</time.set>, naveo je norveški diplomata u
budućim supružnicima razgovaraće	<time.set>sedmično</time.set> tim od šest terapeuta, a radionice
Tako se mesecima spremao	<time.set>7 sati dnevno</time.set> za plivački maraton u Bejrutu, gde je
ugroziti zdravlje.{S} Vežba se	<time.set>dva dana nedeljno</time.set> ako želite da ostanete u dobroj
vreme.{S} U Beogradu je	<time.set>pet meseci godišnje</time.set> temperatura iznad 20 stepeni
ne smanjuju broj zaposlenih i da	<time.set>svakog meseca</time.set> dostavljaju Poreskoj upravi OD
postojanje uredbe, prema kojoj se	<time.set>svakih 15 dana</time.set> vrši korekcija propisa
od 30 radnih dana.{S} Na	<time.set>svakih pet godina</time.set> odmor će biti produžavan za jedan
radu parlamenta predviđeno je da	<time.set>svakog četvrtka</time.set> premijer i ministri
dužinu rada Skupštine za	<time.set>svaku sedmicu</time.set>.{S} Na sastanku je predloženo i da
Trajanove table i planine Miroč, a	<time.set>svake večeri</time.set> biće održavane žurke na plaži
Upravni odbor EBRD-a zaseda	<time.set>svakog drugog utorka</time.set>.{S} Poslednja sednica
godina, a ako se menja na	<time.set>svakih jedanaest godina</time.set> onda se radi o mnogo
na levoj obali Dunava radiće	<time.set>četiri sata svakog dana</time.set>.{S} Do sada su ove ustanove

Primer B.6. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje modifikovano*, odnosno izraze obeležene XML etiketama

privatizaciju od	<time.duration.abs>skoro 20 godina</time.duration.abs>, prihod od privatizacije od oko
godine analitičari su	<time.duration.abs>gotovo godinu i po</time.duration.abs> samo posmatrali berzu, dok
drugi susret za	<time.duration.abs>manje od nedelju dana</time.duration.abs>.{S} Sastanak je počeo
devet ujutru i nakon	<time.duration.abs>više od tri sata i 15 minuta</time.duration.abs> leta trebalo je da
pitanja može trajati	<time.duration.abs>najduže sat i dvadeset pet minuta</time.duration.abs>, a obavlja se
usvoji zakon za	<time.duration.abs>manje od desetak dana</time.duration.abs>.{S} Vlada Srbije prihvatila
preti zatvorska kazna	<time.duration.abs>duža od 10 meseci</time.duration.abs>.{S} Produženje trajanja
pripovedača već	<time.duration.abs>skoro tri decenije</time.duration.abs>.{S} Albahari je jedan od
i rokovnika starih	<time.duration.abs>preko 15 godina</time.duration.abs>, kao i određenu sumu novca.
velikih razmera, jer će	<time.duration.abs>idućih godinu i po dana</time.duration.abs> moći da računa
upozoravaju da će u	<time.duration.abs>naredna dva dana</time.duration.abs> umreti još ljudi ukoliko
optužbe koje su	<time.duration.abs>poslednjih petnaestak dana</time.duration.abs> znak nezadovoljstva
pretragu nije obavio u	<time.duration.abs>proteklih šest godina</time.duration.abs>, proveravajući iskaz svedoka
precizirao da je za	<time.duration.abs>ta tri meseca</time.duration.abs> prijavljeno za 620 krivičnih dela
kriju u gradu Sabratu,	<time.duration.abs>sledećih sat i 45 minuta</time.duration.abs> ključno u rešavanju ove
Širom Srbije je u	<time.duration.abs>prethodna 3 sata</time.duration.abs> zabeleženo čak 17 slučajeva
potres u ovoj zemlji u	<time.duration.abs>poslednjih sto godina</time.duration.abs>.{S} Epicentar zemljotresa

Primer B.7. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje u kontekstu*, odnosno izraze obeležene XML etiketama

Ovo pravilo od pre	<time.duration.abs>tri nedelje</time.duration.abs> kada je doneto na preporuku
izveštaju navodi da je za	<time.duration.abs>18 meseci</time.duration.abs> smanjen broj zaposlenih za oko
poena.{S} Meč je trajao	<time.duration.abs>sat i 32 minuta</time.duration.abs>, a selektor Srbije Dejan
Srbije na usvajanje već za	<time.duration.abs>mesec i po</time.duration.abs> ako bi se na tome intenzivno
brejk priliku, pa je posle	<time.duration.abs>jednog sata i 36 minuta</time.duration.abs> igre Đoković
održan u Moskvi.{S} Posle	<time.duration.abs>dve večeri</time.duration.abs> festivala Beovizija konačno smo
Bajone, koji je bio u bekstvu	<time.duration.abs>10 i po meseci</time.duration.abs> i za kojim je hrvatsko
letova, prvi put posle	<time.duration.abs>pet decenija</time.duration.abs>.{S} Američki sekretar za
zadržavanja u trajanju od	<time.duration.abs>48 sati</time.duration.abs>.{S} Nakon toga, on će biti pušten
uglja termoelektrani na	<time.duration.abs>2,5 sata</time.duration.abs> i da će štrajk i biti nastavljen
rpitanje obelezxavanja	<time.duration.abs>1.700 godina</time.duration.abs> Milanskog edikta, koje će se
biti posađen do jeseni, za	<time.duration.abs>180 dana</time.duration.abs>.{S} Takva promena će koštati oko
juče da će u periodu od	<time.duration.abs>petnaestak dana</time.duration.abs> posle referenduma u Crnoj
izdržane zatvorske kazne od	<time.duration.abs>jedne i po godine</time.duration.abs> pušteno još 13 osuđenih
s gostima iz Vašingtona	<time.duration.abs>oko sat i po</time.duration.abs> i, mimo svih očekivanja, nije
Materijal se resorbuje za	<time.duration.abs>oko šest meseci</time.duration.abs>, a ceo proces obnavljanja
sednica bila prekinuta posle	<time.duration.abs>oko 2 sata i 15 minuta</time.duration.abs> rasprave odbora za
Bivši američki vojnik će za	<time.duration.abs>oko dve sedmice</time.duration.abs> postati prvi građanin SAD-a

Primer B.8. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana*, odnosno izraze obeležene XML etiketama

Izrazi koji ukazuju na tačku u vremenu

četvrtak, <time.hour.abs>01:43:45</time.hour.abs> {S} Statutarno
Kraljeva <time.hour.abs>rano jutros oko 6 h</time.hour.abs> u sudaru
kolena u <time.hour.abs>15 časova</time.hour.abs> nije mogao da igra
zakazan za <time.hour.abs>13:35 h</time.hour.abs>, pa traži produženje
programu <time.hour.abs>oko 17.30</time.hour.abs>.{S} Nagradni fond
počeo u <time.hour.abs>3 časa poslepodne</time.hour.abs>.{S} Dragan
pozvao u <time.hour.abs>7 sati ujutro</time.hour.abs> kazavši mu
odnosno <time.hour.abs>četiri sata jutros po lokalnom vremenu</time.hour.abs>
stigne do <time.hour.abs>deset časova</time.hour.abs> saopštili su u
zakazan za <time.hour.abs>dva sata i 50 minuta</time.hour.abs> ipak
terminu od <time.hour.abs>pola devet uveče</time.hour.abs>.{S} Svi kupci
Međutim, <time.hour.abs>dva sata posle ponoći</time.hour.abs> njihovi
odjeknule su <time.hour.abs>noćas oko dva</time.hour.abs> u centru grada
dogodila se <time.hour.abs>oko ponoći</time.hour.abs> kada je iz za sada
dozvolama u <time.hour.abs>ponoć</time.hour.abs> i da otad nijedan srpski
Grujića 11, <time.hour.abs>sinoć oko 22 časa</time.hour.abs> začuo se

Izrazi koji ukazuju na trajanje

MMF-a tokom <time.duration.abs>poslepodneva</time.duration.abs> poveo
su tokom <time.duration.abs>noći</time.duration.abs> u Austriji, u
da bi tokom <time.duration.abs>prepodneva</time.duration.abs> stigla
njih tokom <time.duration.abs>prethodne noći</time.duration.abs> operisan
u toku <time.duration.abs>prethodne večeri</time.duration.abs>, kada

Primer B.9. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana skraćeno*, odnosno izraze obeležene XML etiketama

je autobus u	<time.hour.abs>7.30 sati</time.hour.abs> krenuo iz
oglašavaju se od	<time.hour.abs>08:00 časova</time.hour.abs>, što je
dogovoreno za	<time.hour.abs>10 časova</time.hour.abs>u zgradi
se sastati do	<time.hour.abs>11 sati</time.hour.abs>, što nije
da ističe posle	<time.hour.abs>14 časova</time.hour.abs>, saopštilo
neće stići pre	<time.hour.abs>2 sata i 15 minuta</time.hour.abs> ,
primljen je u	<time.hour.abs>17 h</time.hour.abs> u Klinički centar
objavila je	<time.hour.abs>oko 15.35</time.hour.abs> RTV B92
sreću, upalile	<time.hour.abs>oko 12 i 15</time.hour.abs>i dalji rad
pomereno za	<time.hour.abs>7 uveče</time.hour.abs>, nakon prenosa
organi su od	<time.hour.abs>osam</time.hour.abs>, preko medija
pravde, izvedena	<time.hour.abs>oko pola dva</time.hour.abs>, osum
zadatak da u	<time.hour.abs>18 sati</time.hour.abs>, po nalogu
poručio je u	<time.hour.abs>20 časova</time.hour.abs>, državni
pregovorima do	<time.hour.abs>pola tri</time.hour.abs>, finale
dogodila se u	<time.hour.abs>23 časa i 59 minuta</time.hour.abs> i
institucijama, za	<time.hour.abs>dva popodne</time.hour.abs>ugovorenu
predstavio do	<time.hour.abs>9.30</time.hour.abs>, kada je očekivana
neće stići pre	<time.hour.abs>18.30 časova</time.hour.abs> , osim ako
zabrana do	<time.hour.abs>21 čas</time.hour.abs> za sve članove
sa početkom u	<time.hour.abs>15.15 časova</time.hour.abs> na Fakultetu
nepoznatih su	<time.hour.abs>oko 7</time.hour.abs>sedeli na terasi
opljačkali su	<time.hour.abs>oko dva</time.hour.abs> , saopštila je
u petak od	<time.hour.abs>13 časova i 15 minuta</time.hour.abs> ,
čekali su do	<time.hour.abs>dva i 15</time.hour.abs>, da bi nakon

Primer B.10. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana u kontekstu*, odnosno izraze obeležene XML etiketama

polaze u	<time.hour.rel>cik zore</time.hour.rel>.{S} J. Cerovina Beograd, 17. februara
posebno u	<time.hour.rel>jutarnjim časovima</time.hour.rel> i na seoskom području.{S} Prema
usijali od	<time.hour.rel>ranog jutra</time.hour.rel>.{S} Građani su se raspitivali o subve
da se ovde	<time.hour.rel>noću</time.hour.rel> svašta dešava, ali od ovog prizora su mi zaklecale
saopštila je	<time.hour.rel>jutros</time.hour.rel> da su dvojica od trojice osumnjičenih za ubistvo
uhapšen je juče	<time.hour.rel>kasno popodne</time.hour.rel>.{S} Agenciji Beta nije potvrđeno da li je
Saveta bezbednosti	<time.hour.rel>kasno sinoć</time.hour.rel> ocenili su da vlasti Srbije ne sarađuju u
izvučen je	<time.hour.rel>noćas</time.hour.rel> iz kanala Dunav-Tisa-Dunav.{S} Predrag Marić,
otkrije od juče	<time.hour.rel>po podne</time.hour.rel>.{S} Na osnovu tragova guma i probijene ograde mosta,
deviznom tržištu	<time.hour.rel>popodne</time.hour.rel> bio samo 4,8 miliona evra.{S} On je ipak dodao da je
Tadić je u	<time.hour.rel>popodnevnim časovima</time.hour.rel> trebalo da se sastane s najvišim
Albahari juče	<time.hour.rel>posle podne</time.hour.rel> iz Kanade potvrdio da dolazi u Beograd na uručenje
u nedeljno	<time.hour.rel>praskozorje</time.hour.rel>, odjednom su sevnuli noževi.{S} Na rukometaše je
smenjene, a	<time.hour.rel>preksinoć</time.hour.rel> beogradski parlament je po kratkom postupku razrešio
ponedeljak u	<time.hour.rel>prepodnevnim satima</time.hour.rel>.{S} Pošto ova deonica magistrale nema
rekao je	<time.hour.rel>rano jutros</time.hour.rel> agenciji Beta da se u vozilu nalazila tri tela, a
pretpostavka je	<time.hour.rel>sinoć</time.hour.rel> uzeta s rezervom.{S} Nastradalima se, naime, trag gubi
desilo u	<time.hour.rel>večernjim časovima</time.hour.rel> u blizini "Brodoremonta".{S} U policiji je
gubi od	<time.hour.abs>prethodne noći, <time.hour.abs>sat posle ponoći</time.hour.abs></time.hour.abs>
Beograd -	<time.hour.abs>Večeras u <time.hour.abs>ponoć</time.hour.abs></time.hour.abs> ističe rok za

Primer B.11. Konkordance koje predstavljaju rezultat primene transduktora *Period vremena dana*, odnosno izraze obeležene XML etiketama

<time.hour.period><time.hour.abs>10</time.hour.abs> do <time.hour.abs>11 časova</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>12</time.hour.abs> i <time.hour.abs>17 sati</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>8</time.hour.abs> do <time.hour.abs>podneva</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>20 časova</time.hour.abs> i <time.hour.abs>22 časa</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>11</time.hour.abs> ili <time.hour.abs>12 h</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>17</time.hour.abs> - <time.hour.abs>18 sati</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>9</time.hour.abs>, <time.hour.abs>11</time.hour.abs> i <time.hour.abs>13 časova
</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.rel>sinoć</time.hour.rel> ili <time.hour.rel>jutros</time.hour.rel></time.hour.period>

<time.hour.period><time.hour.abs>15</time.hour.abs>, <time.hour.abs>17</time.hour.abs> i <time.hour.abs>19 h
</time.hour.abs></time.hour.period>

<time.hour.period><time.hour.abs>sedam</time.hour.abs> ili u <time.hour.abs>osam časova</time.hour.abs>
</time.hour.period>

<time.hour.period><time.hour.abs>pet popodne</time.hour.abs> i <time.hour.abs>11 ujutru</time.hour.abs>
</time.hour.period>

Primer B.12. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje osnovno*, odnosno izraze obeležene XML etiketama

radije bio	<time.duration.abs>1000 godina</time.duration.abs> u Hagu nego prihvatio
raspisan za	<time.duration.abs>11 i po meseci</time.duration.abs>, a održan u maju
traje samo	<time.duration.abs>12,4 godine</time.duration.abs>.{S} Analize pokazuju
građani su	<time.duration.abs>18 sati</time.duration.abs> čekali odgovor iz uprave
potrebna	<time.duration.abs>24 časa</time.duration.abs>.{S} Sve se obavlja nakon
vratio se pre	<time.duration.abs>tri nedelje i dva dana</time.duration.abs> i odlučio
možda dugo	<time.duration.abs>25 godina i 220 dana</time.duration.abs>, što nije
što je pre	<time.duration.abs>dve nedelje</time.duration.abs> izraelskoj teniserki
Kosovu.{S}	<time.duration.abs>Vek i po</time.duration.abs> je bilo ovde, a sada
već u prvih	<time.duration.abs>desetak minuta</time.duration.abs> imali po tri poena
kaznu na	<time.duration.abs>godinu i šest meseci</time.duration.abs>.{S} Golubov
da odigram	<time.duration.abs>jednu godinu</time.duration.abs> u Americi, ali ne
imenovana pre	<time.duration.abs>oko dva meseca</time.duration.abs>.{S} Odmah zatim
Troicki i za	<time.duration.abs>sat i 14 minuta</time.duration.abs> preslišao protivnika
loptu za	<time.duration.abs>sat</time.duration.abs> igre i istim rezultatom završio
i da živi	<time.duration.abs>sto godina</time.duration.abs>, te da ima oko milion
crkve je	<time.duration.abs>tri i po decenije</time.duration.abs> bio episkop naše
poskupelo za	<time.duration.abs>mesec dana</time.duration.abs> i struja će nakon
gubi od pre	<time.duration.abs>dve noći</time.duration.abs>, kada su im javili da se

Primer B.13. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje relativno*, odnosno izraze obeležene XML etiketama

parlamenta.{S}	<time.duration.rel>Danima</time.duration.rel> se spekuliše koliko je koštao njen
oseća dušu narodnu,	<time.duration.rel>decenijama</time.duration.rel> je posvećen svojoj veri i svom
velikih razmera, jer će	<time.duration.rel>idućih godina</time.duration.rel> moći da računa i na priliv
uglja ne samo više nego u	<time.duration.rel>mesecima</time.duration.rel> pre poplava, već su potavljeni
i ova akcija tako da se	<time.duration.rel>narednih dana</time.duration.rel> očekuju nova hapšenja.{S}
može napisatiu roku od	<time.duration.rel>narednih nekoliko nedelja</time.duration.rel>.{S} Nacrt ustava
Kosova, koja je trajala	<time.duration.rel>prošlih par minuta</time.duration.rel> da će vladini prioriteti biti
put kojim će ići u	<time.duration.rel>narednim godinama</time.duration.rel>.{S} Ako bi se opredelili na
policije, uhapšen je	<time.duration.rel>nekoliko časova</time.duration.rel> pošto se njegov saradnik
okalna uprava, jer	<time.duration.rel>više decenija</time.duration.rel> nisu isplaćivale doprinose na
napuste prosvetu.{S}	<time.duration.rel>Ovih dana</time.duration.rel> ćemo imati sastanak i sa
pripadnici policije su	<time.duration.rel>ovih nekoliko sati</time.duration.rel> pretresali porodičnu
uloga biće veoma važna u	<time.duration.rel>ovim godinama</time.duration.rel> približavanja Evropskoj
verovatno odlučiti da za	<time.duration.rel>par meseci</time.duration.rel> odloži podnošenje predloga
pa su i optužbe koje su	<time.duration.rel>poslednjih dana</time.duration.rel> znak nezadovoljstva
sirove nafte.{S} Iako je u	<time.duration.rel>poslednjih nekoliko sedmica</time.duration.rel> zabeležen pad
Trojkom Marinovića u	<time.duration.rel>poslednjim sekundama</time.duration.rel> Zvezda na kratku
za plasman u dalji krug	<time.duration.rel>prethodnih dana</time.duration.rel> kružila glasina da će Željko
problem, kojim se država	<time.duration.rel>proteklih godina</time.duration.rel> nije dovoljno bavila
Milana i da, za razliku od	<time.duration.rel>ranijih godina</time.duration.rel> kada je takođe
i jaka kiša koja je	<time.duration.rel>više časova</time.duration.rel> padala u okolini Knjaževca
obzirom na to da je DEA u	<time.duration.rel>zadnjih nekoliko meseci</time.duration.rel> u više navrata

Primer B.14. Konkordance koje predstavljaju rezultat primene transduktora *Datum relativni*, odnosno izraze obeležene XML etiketama

sednici u	<time.date.rel>četvrtak</time.date.rel> Vlada je
za penzije	<time.date.rel>danas</time.date.rel> su oko 12 odsto
koja ni do	<time.date.rel>današnjeg dana</time.date.rel> nije
će Belgija	<time.date.rel>iduće godine</time.date.rel> preuzeti
Đelić će	<time.date.rel>idućeg utorka</time.date.rel> na samitu
života za	<time.date.rel>isti mesec</time.date.rel> povećani su
da će od	<time.date.rel>istog dana</time.date.rel> na svakoj
ni o ovome	<time.date.rel>juče</time.date.rel> nije imalo
odnosu na	<time.date.rel>jučerašnji dan</time.date.rel>.{S} Narod
je da se	<time.date.rel>minule sedmice</time.date.rel> sastao u
završnoj fazi,	<time.date.rel>minulog leta</time.date.rel> tvrdili su
Balkana u	<time.date.rel>narednoj deceniji</time.date.rel>.{S} Što
Đoković će	<time.date.rel>narednog vikenda</time.date.rel> predvod
ocenio da je	<time.date.rel>ova godina</time.date.rel> bila godina
je da će	<time.date.rel>ove nedelje</time.date.rel> predložiti
neto zarada u	<time.date.rel>ovom mesecu</time.date.rel> nominalno je
politika za	<time.date.rel>ovu godinu</time.date.rel> bazirana na
ušla u	<time.date.rel>predstojeću godinu</time.date.rel> i kada
dogodila se	<time.date.rel>prekjuče</time.date.rel> na Uvačkom jezeru
otpadnih voda	<time.date.rel>pretprošle godine</time.date.rel> smo
stanicu Palilula	<time.date.rel>prošle srede</time.date.rel>, kada je
trebalo bi	<time.date.rel>sutra</time.date.rel>, između ostalog,
Ona je	<time.date.rel>toga dana</time.date.rel>, kako nezvani

Primer B.15. Konkordance koje predstavljaju rezultat primene transduktora *Period trajanja*, odnosno izraze obeležene XML etiketama

<time.duration.period><time.duration.abs>godinu</time.duration.abs> - <time.duration.abs>dve</time.duration.abs>
</time.duration.period>
<time.duration.period><time.duration.abs>dva</time.duration.abs> - <time.duration.abs>tri dana</time.duration.abs>
</time.duration.period>
<time.duration.period><time.duration.abs>2 meseca</time.duration.abs> do <time.duration.abs>pola godine
</time.duration.abs></time.duration.period>
<time.duration.period><time.duration.abs>2 dana</time.duration.abs> i <time.duration.abs>10 sati</time.duration.abs>
</time.duration.period>
<time.duration.period><time.duration.abs>dve nedelje</time.duration.abs> ili <time.duration.abs>mesec dana
</time.duration.abs></time.duration.period>
<time.duration.period><time.duration.abs>17</time.duration.abs> - <time.duration.abs>18 godina</time.duration.abs>
</time.duration.period>
<time.duration.period><time.duration.abs>6</time.duration.abs> na <time.duration.abs>8 meseci</time.duration.abs>
</time.duration.period>
<time.date.period><time.date.rel>juče</time.date.rel> do <time.date.rel>petka</time.date.rel></time.date.period>
<time.date.period><time.date.abs>1968. godine</time.date.abs> do <time.date.rel>danas</time.date.rel>
</time.date.period>
<time.date.period><time.date.rel>četvrtak</time.date.rel> i <time.date.rel>petak</time.date.rel></time.date.period>
<time.date.period><time.date.rel>decembar prošle godine</time.date.rel> i <time.date.abs>januar 2006.</time.date.abs>
</time.date.period>

Primer B.16. Konkordance koje predstavljaju rezultat primene transduktora *Vreme*, odnosno izraze obeležene XML etiketama

<time.hour><time.date.abs>01.02.2010.</time.date.abs><time.hour.abs>20:41:59</time.hour.abs></time.hour>

<time.hour><time.date.abs>24. avgusta 2008.</time.date.abs>, <time.hour.abs>oko tri sata ujutru</time.hour.abs></time.hour>

<time.hour><time.date.rel>25 februara</time.date.rel> u <time.hour.abs>12 časova</time.hour.abs></time.hour>

<time.hour><time.date.rel>30. januara</time.date.rel> <time.hour.rel>ujutru</time.hour.rel></time.hour>

<time.hour><time.date.rel>8. marta</time.date.rel> u <time.hour.rel>večernjim časovima</time.hour.rel></time.hour>

<time.hour><time.hour.abs>oko ponoći</time.hour.abs> između <time.date.period><time.date.rel>četvrtka</time.date.rel> i <time.date.rel>petka</time.date.rel></time.date.period></time.hour>

<time.hour><time.date.rel>nedelju, 21. maja</time.date.rel>, <time.hour.abs>ujutru u 8</time.hour.abs></time.hour>

<time.hour><time.date.rel>ponedeljak</time.date.rel> <time.hour.abs>uveče oko jedan</time.hour.abs></time.hour>

<time.hour><time.date.rel>subotu, 28, maja</time.date.rel>, od <time.hour.period><time.hour.abs>11</time.hour.abs> do <time.hour.abs>13 sati</time.hour.abs></time.hour.period></time.hour>

<time.hour><time.hour.rel>noći</time.hour.rel> između <time.date.period><time.date.rel>19</time.date.rel> i <time.date.rel>20. januara ove godine</time.date.rel></time.date.period></time.hour>

<time.hour><time.date.rel>juče</time.date.rel> u <time.hour.abs>podne</time.hour.abs></time.hour>

<time.hour><time.date.rel>narednog dana</time.date.rel> <time.hour.abs>oko pet sati popodne</time.hour.abs></time.hour>

<time.hour><time.date.rel>danas</time.date.rel> u <time.hour.abs>9:45 sati</time.hour.abs></time.hour>

Prilog C

Primeri obeležavanja prepoznatih vremenskih izraza <TIMEX3> etiketama

Primer C.1. Konkordance koje predstavljaju rezultat primene transduktora *Datum apsolutni*, odnosno izraze obeležene <TIMEX3> etiketama

efekte u <TIMEX3 type="DATE" temporalFunction="false" val="2">2. milenijumu</TIMEX3>, kako je i očekivano
kada je u <TIMEX3 type="DATE" temporalFunction="false" val="15">XV veku</TIMEX3> postala pravo
kada se <TIMEX3 type="DATE" temporalFunction="false" val="198" mod="END">krajem osamdesetih</TIMEX3> vratio
iskustva <TIMEX3 type="DATE" temporalFunction="false" val="199">1990-ih</TIMEX3>, te je zato
Politika <TIMEX3 type="DATE" temporalFunction="false" val="2005-09-16">9.16.2005</TIMEX3> Vest dana
Politika <TIMEX3 type="DATE" temporalFunction="false" val="2009-02-05">četvrtak, 5. februar 2009.</TIMEX3>
koji se <TIMEX3 type="DATE" temporalFunction="false" val="2004-04-04">4. aprila 2004.</TIMEX3> prilikom
došli <TIMEX3 type="DATE" temporalFunction="false" val="1990-08-15">15.VIII 1990. godine</TIMEX3>
objavljeno <TIMEX3 type="DATE" temporalFunction="false" val="2010-01-20">20/01/2010</TIMEX3> Drvengrad
dodao je <TIMEX3 type="DATE" temporalFunction="false" val="2015-07-01">2015-07-01</TIMEX3>. Nadalova pobeda
da su u <TIMEX3 type="DATE" temporalFunction="false" val="2004-12">decembru 2004.</TIMEX3> objavili
tamo su od <TIMEX3 type="DATE" temporalFunction="false" val="1991-08">1991. godine avgusta meseca</TIMEX3>
Sarajevu <TIMEX3 type="DATE" temporalFunction="false" val="1992-05" mod="START">početkom maja 1992. godine</TIMEX3>.
letelice u <TIMEX3 type="DATE" temporalFunction="false" val="1998-SP">proleće 1998.</TIMEX3> zabeležene
dogodi <TIMEX3 type="DATE" temporalFunction="false" val="2013">2013. godine</TIMEX3> u Nišu, povodom
letelice u <TIMEX3 type="DATE" temporalFunction="false" val="1999" mod="START">prvoj polovini 1999.</TIMEX3>
depresije, <TIMEX3 type="DATE" temporalFunction="false" val="193" mod="MID">sredinom tridesetih godina minulog veka</TIMEX3>
sve do <TIMEX3 type="DATE" temporalFunction="false" val="2011-SU" mod="END">kraja leta 2011</TIMEX3>, kaže

Primer C.2. Konkordance koje predstavljaju rezultat primene transduktora *Datum relativni*, odnosno izraze obeležene <TIMEX3> etiketama

da su <TIMEX3 type="DATE" temporalFunction="true" val="XX" mod="MID">polovinom veka</TIMEX3> dostigli

Balkana u <TIMEX3 type="DATE" temporalFunction="true" val="XXX" valueFromFunction="+1E">narednoj deceniji</TIMEX3>

tek u <TIMEX3 type="DATE" temporalFunction="true" val="XXXX" mod="END">drugoj polovini godine</TIMEX3>, ako rast

uzeti u <TIMEX3 type="DATE" temporalFunction="true" val="XXXX" valueFromFunction="-1Y" mod="END">poslednjem kvartalu prošle godine</TIMEX3>, pri tome će biti dužni

državi do <TIMEX3 type="DATE" temporalFunction="true" val="XXXX" valueFromFunction="=1Y" mod="MID">sredine ove godine</TIMEX3> i tek tada

poštar, <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-01" valueFromFunction="=1Y" mod="END">krajem januara ove godine</TIMEX3> oštetio do 145

da je <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-01-01">prvog januara</TIMEX3> devizni kurs

početi u <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-02" mod="END">drugoj polovini februara</TIMEX3>.

je u <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-10" mod="START">prvih nekoliko dana oktobra</TIMEX3> značajno smanjena.

akcija <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-12-31">poslednjeg dana u godini</TIMEX3>

da će do <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-WXX" mod="END">kraja nedelje</TIMEX3>

održana u <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-WXX-3">sredu</TIMEX3>, a Savet ministara

frakcija <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-XX" valueFromFunction="+1M" mod="MID">sredinom sledećeg meseca</TIMEX3>, a Savet ministara

devojke, <TIMEX3 type="DATE" temporalFunction="true" val="XXXX-XX-XX" valueFromFunction="-1D">juče</TIMEX3> je bila

Primer C.3. Konkordance koje predstavljaju rezultat primene transduktora *Vreme dana*, odnosno izraze obeležene <TIMEX3> etiketama

tačno u <TIMEX3 type="TIME" temporalFunction="false" val="T18:27:05">18:27:05</TIMEX3> stigao je

viđeni u <TIMEX3 type="TIME" temporalFunction="false" val="T20:30">20.30 sati</TIMEX3>, trebalo je da

poslato u <TIMEX3 type="TIME" temporalFunction="false" val="T23:59">23 časa i 59 minuta</TIMEX3>, trebalo

zakazano za <TIMEX3 type="TIME" temporalFunction="false" val="T14:15">dva sata i 15 minuta popodne</TIMEX3>

važi do <TIMEX3 type="TIME" temporalFunction="false" val="T08">8 h</TIMEX3> narednog radnog dana

početi <TIMEX3 type="TIME" temporalFunction="false" val="T10" mod="APPROX">oko 10 sati</TIMEX3> za sve

desio <TIMEX3 type="TIME" temporalFunction="false" val="T21" mod="APPROX">oko devet uveče</TIMEX3>.

poskupelo od <TIMEX3 type="TIME" temporalFunction="false" val="T24">ponoći</TIMEX3> u proseku 85 para po litru

stižu do <TIMEX3 type="TIME" temporalFunction="false" val="T12">podneva</TIMEX3> u našu ustanovu

pozvala je <TIMEX3 type="TIME" temporalFunction="false" val="T23:50">deset minuta pre ponoći</TIMEX3>, kada

saopštila je <TIMEX3 type="TIME" temporalFunction="true" val="TMO">jutros</TIMEX3> da su dvojica od trojice

evocirane su <TIMEX3 type="TIME" temporalFunction="true" val="TNI">sinoć</TIMEX3> u prostorijama opštinskog

saopštio je <TIMEX3 type="TIME" temporalFunction="true" val="TNI">kasno sinoć</TIMEX3> da je od zemljotresa

vest od <TIMEX3 type="TIME" temporalFunction="false" val="2010-02-22T12:22:38">22.02.2010. 12:22:38 </TIMEX3>

promenjeno <TIMEX3 type="TIME" temporalFunction="true" val="XXXX-XX-XXT14" valueFromFunction="-1D" mod="APPROX">juče oko dva popodne</TIMEX3> bez najave

dolaze <TIMEX3 type="TIME" temporalFunction="true" val="XXXX-XX-XXTEV" valueFromFunction="+1D">sutra uveče</TIMEX3> na dogovor

Primer C.4. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje*, odnosno izraze obeležene <TIMEX3> etiketama

Izrazi koji ukazuju na trajanje, a dati su u obliku kalendarskog datuma

tokom <TIMEX3 type="DURATION" temporalFunction="false" val="2011-04">meseca aprila 2011. godine</TIMEX3>
u toku <TIMEX3 type="DURATION" temporalFunction="false" val="2007-06">juna 2007.</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-11">novembra iduće godine</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="true" val="199">1990-ih</TIMEX3>
u toku <TIMEX3 type="DURATION" temporalFunction="false" val="198">osamdesetih prošlog veka</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="2005-WI">zime 2005.</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-SU">leta ove godine</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-03-05">petog marta</TIMEX3>
u toku <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-04">meseca aprila</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-XX-10">10-og</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX">ove godine</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-XX">prošlog meseca</TIMEX3>
u toku <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-WXX-WE">predstojećeg vikenda</TIMEX3>
u toku <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX">godine</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-WXX">sedmice</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="2013">2013.</TIMEX3>
tokom <TIMEX3 type="DURATION" temporalFunction="false" val="XXXX-XX-XXTAF">poslepodneva</TIMEX3>

Primer C.5. Konkordance koje predstavljaju rezultat primene transduktora *Trajanje*, odnosno izraze obeležene <TIMEX3> etiketama

Izrazi koji ukazuju na trajanje

<TIMEX3 type="DURATION" temporalFunction="false" val="P10Y" mod="LESS_THAN">skoro 10 godina</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P1W" mod="APPROX">oko nedelju dana</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P3H15MIN" mod="MORE_THAN">više od tri sata i 15 minuta</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P10D" mod="LESS_THAN">manje od desetak dana</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P3DE" mod="LESS_THAN">skoro tri decenije</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P100Y">poslednjih sto godina</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P3W">tri nedelje</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P1.5M">mesec i po</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P2EV">dve večeri</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P180D">180 dana</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="false" val="P1H">sat</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true" val="PXD">danima</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true" val="PXW">narednih nekoliko nedelja</TIMEX3>
<TIMEX3 type="DURATION" temporalFunction="true" val="PXS">poslednjim sekundama</TIMEX3>
<time.duration.period><TIMEX3 type="DURATION" temporalFunction="false" val="P17Y">17</TIMEX3> ili <TIMEX3 type="DURATION" temporalFunction="false" val="P18Y">18 godina</TIMEX3></time.duration.period>

Primer C.6. Konkordance koje predstavljaju rezultat primene transduktora *Učestalost*, odnosno izraze obeležene <TIMEX3> etiketama

<TIMEX3 type="SET" temporalFunction="false" val="P1D" quant="EVERY">svakodnevno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1D" freq="1X">jednom dnevno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1M" freq="2X">dva puta mesečno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1W" freq="3X">triput nedeljno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1M" freq="5X">5 puta mesečno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1W">sedmično</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1D" freq="11H">11 sati dnevno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1W" freq="2D">dva dana nedeljno</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1Y" freq="5M">pet meseci godišnje</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1M" quant="EVERY">svakog meseca</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P15D" quant="EVERY">svakih 15 dana</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P5Y" quant="EVERY">svakih pet godina</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="XXXX-WXX-04" quant="EVERY">svakog četvrtka</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1W" quant="EVERY">svaku sedmicu</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1EV" quant="EVERY">svake večeri</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="XXXX-WXX-02" freq="2" quant="EVERY">svakog drugog utorka</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P11Y" quant="EVERY">svakih jedanaest godina</TIMEX3>
<TIMEX3 type="SET" temporalFunction="false" val="P1D" freq="3H" quant="EVERY">tri sata svakog dana</TIMEX3>

Prilog D

Primeri obeležavanja prepoznatih vremenskih izraza prilikom procene uspešnosti sistema

Primer D.1. Primeri ispravno obeleženih konkordanci

```
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false" val="2011-02-27">27. februar 2011. godine
   </TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false" val="2005-01">januara 2005. godine</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false" val="2007-SU">leto 2007. godine</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-XX" valueFromFunction="-1M">prošlog
   meseca</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX" valueFromFunction="+1Y">sledeće
   godine</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-10">oktobra</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false" val="2009" mod="END">kraja 2009. godine
   </TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-11" mod="START">početkom novembra
   </TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-02" mod="END">kraj februara</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="false" val="T16:15">16 časova i 15 minuta</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="false" val="T05">noćas u pet</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="true" val="XXXX-WXX-2TEV">utorak uveče</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DURATION" temporalFunction="false" val="P3D">tri dana</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="DURATION" temporalFunction="false" val="P3E" mod="MORE_THAN">više od tri
   decenije</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="SET" temporalFunction="false" val="P1D">dnevno</TIMEX3>
<TIMEX3 proveraP="OK" proveraN="OK" type="SET" temporalFunction="false" val="P1Y" quant="EVERY">svake godine</TIMEX3>
```

Primer D.2. Primeri pogrešno obeleženih ili propuštenih konkordanci

<TIMEX3 proveraP="UOKt" proveraN="UOKv" type="TIME" temporalFunction="false" val="T02">dva sata</TIMEX3>

mlađe od <TIMEX3 proveraP="UOKo" proveraN="UOKm" type="DURATION" temporalFunction="false" val="P15Y">15 godina</TIMEX3>

sledeći <TIMEX3 proveraP="UOKo" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-12-06">ponedeljak, 6. decembar</TIMEX3>

prvoj deceniji <TIMEX3 proveraP="UOKo" proveraN="UOKm" type="DATE" temporalFunction="true" val="XX" valueFromFunction="=1C">ovog veka</TIMEX3>

<TIMEX3 proveraP="OK" proveraN="UOKv" type="TIME" temporalFunction="true" val="XXXX-WXXTEV">nedelju uveče</TIMEX3>

<TIMEX3 proveraP="OK" proveraN="UOKv" type="TIME" temporalFunction="true" val="XXXX-WXXT13">nedelju, u 13 časova</TIMEX3>

20-<TIMEX3 proveraP="NOK" proveraN="UOKv" type="SET" temporalFunction="false" val="P1Y">godišnje</TIMEX3>

Rođena Madam Tiso <time.date.period proveraP="NOK"><TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false" val="1760">1760. godine</TIMEX3> na <TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="true" val="XXXX-XX-XX" valueFromFunction="=1D">današnji dan</TIMEX3></time.date.period>

<TIMEX3 proveraP="NOK" type="DATE" temporalFunction="true" val="XXXX-05">Maja</TIMEX3> Gojković

<TIMEX3 proveraP="MISS" type="DATE">1820</TIMEX3>

<TIMEX3 proveraP="MISS" type="DURATION">67. minutu</TIMEX3>

<TIMEX3 proveraP="MISS/E" type="DURATION">petodnenoj</TIMEX3>

<TIMEX3 proveraP="MISS/E" type="DATE">nredne godine</TIMEX3>

Prilog E

Primeri obeležavanja prepoznatih vremenskih izraza medicinskih narativnih tekstova

Primer E.1. Primer medicinskog narativnog teksta

Pacijentkinja Ivana Petrović, broj istorije 1234, iz Beograda, rođena 1984 godine.{S}

Primljena na Kliniku za maksilofacijalnu hirurgiju Stomatološkog fakulteta radi nastavka hirurškog lečenja planocelularnog karcinoma jezika sa desne strane.{S}

ANEMNESIS FAMILIAE:{S} Negira anemnestički značajna oboljenja za hereditet.{S}

ANAMNESIS VITTAE:{S} Negira ostala anamnestički značajna oboljenja.{S} Ne pije, ne puši.{S} Negira predhodne povrede.{S} Negira alergije.S} Od lekova koristi tegretorl , diklofen i.m. pp.S} I ranisan.{S}

ANAMNESIS MORBI Navodi da je prvi put promenu primetila u novembru 2010. kod nadležnog stomatologa, koji je tretirao promenu, nije došlo do regresije pa je upućena na Kliniku za Oralnu medicinu.{S} U julu u8činjena je biopsija na ovoj Klinici za maksilofacijalnu hirurgiju koja nije dala Dg, pa je ista ponovljena i u aprilu 2011. goidne HP br DG Displasio epithelii squamosi teškog stepena.{S} Lečena je potom na Klinici do avgusta meseca 2011., ali i dalje kod pacijentkinje prisutni bolovi i upućena u ovu Kliniku.{S} Dana 23.09.2012. u uslovima OET učinjena Op Excisio tu linguae lat dex in toto HP Ca Planocellulare infiltrativum HG 2 NG 2.{S} Širina 12 mm, Dubina 8 mm.{S} Na linijama resekcija nema tumorskog tkiva.{S} Shodno HP nalazu kod pacijenkinje je 12.10.2011. u uslovima OET učinjena OP Disectio collli radicalis lat dex type III.{S} HP 6789 Ukupno 42 limfna čvora (iz svih pet nivoa) – nisu nađeni znakovi maligniteta.{S} Kod pacijentkinje je potom sprovedena postoperativna zračna terapija odlukom konzilijuma za MF regiju IORKCS , a ista je završena januara meseca 2012..{S} Nakon toga učinejna dva CT snimka operativnog predela, koja su bez pouzdanih znakova recidiva bolesti (u prilogu, nalaz stacionaran) Međutim kod pacijenkinje prisutni bolovi koji zrače u uho obostrano.{S} Ukupno je izgubila oko 6 kg TT do sada.{S} Oteržano guta i otežano se hrani.{S}

STATUS LOCALIS Ekstraoralnom Inspekcioom spolja se ne konsatuju znakovi patoloških promena.{S}Otvaranje usta i funkcija oko 35 mm na nivou sečivnih ivica sekutića.,Iontraoralno se konstatuje ožIljak u predelu jezika sa desne strane.{S} Isti je palpatorno u predelu korena sa desne strane bolan, tvrdo elastičan.{S}Na vratu sa desne strane prisutan je ožiljak karakterističan za predhodni OP zahvat., formiran, linijski.{S}

Status presens universalis Pacijentkinja je svesna, samostalno aktivno pokretna, koža i vidljive sluznice nešto bleđe od uobičajenog nalaza.{S} Abdomen u ravni grudnog koša, bezbolan.{S} Ekstremiteti bez edema i varikoziteta.{S}

DG Ca planocellulare linguae lat. dex status post op A.A.I Status post IRR A. L. VII

dr Jovanović

Primer E.2. Primer deidentifikovanog medicinskog narativnog teksta

Pacijentkinja <persName.full><persName.full PHI="yes">Vilma Kremenko </persName.full></persName.full>, broj istorije <number PHI="yes">XXXX </number>, iz <top.gr PHI="yes">Kamengrada</top.gr>, rođena 1984 godine.{S}

Primljena na <org PHI="yes">Kliniku</org> radi nastavka hirurškog lečenja planocelularnog karcinoma jezika sa desne strane.{S}

ANEMNESIS FAMILIAE:{S} Negira anemnestički značajna oboljenja za hereditet.{S}

ANAMNESIS VITTAE:{S} Negira ostala anamnestički značajna oboljenja. {S} Ne pije, ne puši.{S} Negira predhodne povrede.{S} Negira alergije. {S} Od lekova koristi tegretorl , diklofen i.m. pp.{S} I ranisan.{S}

ANAMNESIS MORBI Navodi da je prvi put promenu primetila u <date PHI="yes">novembru 2010.</date> kod nadležnog stomatologa, koji je tretirao promenu, nije došlo do regresije pa je upućena na <org PHI="yes"> Kliniku</org> za Oralnu medicinu.{S} U <date PHI="yes">julu</date> u8činjena je biopsija na ovoj <org PHI="yes">Klinici</org> koja nije dala Dg, pa je ista ponovljena i u <date PHI="yes">aprilu 2011.</date> goidne HP br DG Displasio epithelii squamosi teškog stepena.{S} Lečena je potom na <org PHI="yes">Klinici</org> do <date PHI="yes">avgusta meseca 2011. </date>, ali i dalje kod pacijentkinje prisutni bolovi i upućena u ovu <org PHI="yes">Kliniku</org>.{S} Dana <date PHI="yes">23.09.2012.</date> u uslovima OET učinjena Op Excisio tu linguae lat dex in toto HP Ca Planocellulare infiltrativum HG 2 NG 2.{S} Širina 12 mm, Dubina 8 mm.{S} Na linijama resekcija nema tumorskog tkiva.{S} Shodno HP nalazu kod pacijenkinje je <date PHI="yes">12.10.2011.</date> u uslovima OET učinjena OP Disectio collli radicalis lat dex type III.{S} HP <number PHI="yes"> XXXX</number> Ukupno 42 limfna čvora (iz svih pet nivoa) - nisu nađeni znakovi maligniteta.{S} Kod pacijentkinje je potom sprovedena postoperativna zračna terapija odlukom konzilijuma za MF regiju IORKCS , a ista je završena <date PHI="yes">januara meseca 2012.</date>.{S} Nakon toga učinejna dva CT snimka operativnog predela, koja su bez pouzdanih znakova recidiva bolesti (u prilogu, nalaz stacionaran) Međutim kod pacijenkinje prisutni bolovi koji zrače u uho obostrano.{S} Ukupno je izgubila oko 6 kg TT do sada.{S} Oteržano guta i otežano se hrani.{S}

STATUS LOCALIS Ekstraoralnom Inspekcioom spolja se ne konsatuju

znakovi patoloških promena.{S}Otvaranje usta i funkcija oko 35 mm na nivou sečivnih ivica sekutića.,Iontraoralno se konstatuje ožIljak u predelu jezika sa desne strane.{S} Isti je palpatorno u predelu korena sa desne strane bolan, tvrdo elastičan.{S}Na vratu sa desne strane prisutan je ožiljak karakterističan za predhodni OP zahvat., formiran, linijski.{S}

Status presens universalis Pacijentkinja je svesna, samostalno aktivno pokretna, koža i vidljive sluznice nešto bleđe od uobičajenog nalaza.{S} Abdomen u ravni grudnog koša, bezbolan.{S} Ekstremiteti bez edema i varikoziteta.{S}

DG Ca planocellulare linguae lat. dex status post op A.A.I Status post IRR A. L. VII

<pers PHI="yes"><persName.last>Kremenko</persName.last></pers>

Primer E.3. Primer deidentifikovanog teksta koji je korišćen za evaluaciju sistema

Pacijentkinja <persName.full><persName.full PHI="yes">Vilma Kremenko
</persName.full></persName.full>, broj istorije <number PHI="yes">XXXX
</number>, iz <top.gr PHI="yes">Kamengrada</top.gr>, rođena 1984 godi-
ne.{S}

Primljena na <org PHI="yes">Kliniku</org> radi nastavka hirurškog
lečenja planocelularnog karcinoma jezika sa desne strane.{S}

ANEMNESIS FAMILIAE:{S} Negira anemnestički značajna oboljenja za
hereditet.{S}

ANAMNESIS VITTAE:{S} Negira ostala anamnestički značajna oboljenja.
{S} Ne pije, ne puši.{S} Negira predhodne povrede.{S} Negira alergije.
{S} Od lekova koristi tegretorl , diklofen i.m. pp.{S} I ranisan.{S}

ANAMNESIS MORBI Navodi da je prvi put promenu primetila u novembru
2010. kod nadležnog stomatologa, koji je tretirao promenu, nije došlo
do regresije pa je upućena na <org PHI="yes">Kliniku</org> za Oralnu
medicinu.{S} U julu u8činjena je biopsija na ovoj <org PHI="yes">Kli-
nici</org> koja nije dala Dg, pa je ista ponovljena i u aprilu 2011.
goidne HP br DG Displasio epithelii squamosi teškog stepena.{S} Lečena
je potom na <org PHI="yes">Klinici</org> do avgusta meseca 2011., ali i
dalje kod pacijentkinje prisutni bolovi i upućena u ovu <org PHI="yes">
Kliniku</org>.{S} Dana 23.09.2012. u uslovima OET učinjena Op Excisio
tu linguae lat dex in toto HP Ca Planocellulare infiltrativum HG 2 NG
2.{S} Širina 12 mm, Dubina 8 mm.{S} Na linijama resekcija nema tumorskog
tkiva.{S} Shodno HP nalazu kod pacijenkinje je 12.10.2011. u uslovima
OET učinjena OP Disectio collli radicalis lat dex type III.{S} HP <number
PHI="yes">XXXX</number> Ukupno 42 limfna čvora (iz svih pet nivoa)
- nisu nađeni znakovi maligniteta.{S} Kod pacijentkinje je potom sprove-
dena postoperativna zračna terapija odlukom konzilijuma za MF regiju IOR-
KCS , a ista je završena januara meseca 2012..{S} Nakon toga učinejna dva
CT snimka operativnog predela, koja su bez pouzdanih znakova recidiva
bolesti (u prilogu, nalaz stacionaran) Međutim kod pacijenkinje prisutni
bolovi koji zrače u uho obostrano.{S} Ukupno je izgubila oko 6 kg TT do
sada.{S} Oteržano guta i otežano se hrani.{S}

STATUS LOCALIS Ekstraoralnom Inspekcioom spolja se ne konsatuju
znakovi patoloških promena.{S}Otvaranje usta i funkcija oko 35 mm na
nivou sečivnih ivica sekutića.,Iontraoralno se konstatuje ožIljak u

predelu jezika sa desne strane.{S} Isti je palpatorno u predelu korena sa desne strane bolan, tvrdo elastičan.{S}Na vratu sa desne strane prisutan je ožiljak karakterističan za predhodni OP zahvat., formiran, linijski.{S}

Status presens universalis Pacijentkinja je svesna, samostalno aktivno pokretna, koža i vidljive sluznice nešto bleđe od uobičajenog nalaza.{S} Abdomen u ravni grudnog koša, bezbolan.{S} Ekstremiteti bez edema i varikoziteta.{S}

DG Ca planocellulare linguae lat. dex status post op A.A.I Status post IRR A. L. VII

<pers PHI="yes"><persName.last>Kremenko</persName.last></pers>

Primer E.4. Primer medicinskog narativnog teksta obeleženog vremenskim izrazima nakon primene sistema

Pacijent <persName.full><persName.full PHI="yes">Barni Kamenko
</persName.full></persName.full>, primljen na <org PHI="yes">Kliniku
</org> radi hirurškog lečenja preloma zigomatikomaksilarnog kompleksa sa
leve strane.{S}
ANEMNESIS FAMILIAE: negira anamnstički značajna oboljenja u porodici.{S}
ANAMNESIS Navodi da je alergičan na baktrim i andol Navodi da povremeno
konzumira alkohol.{S} Ne puši.{S} Ne koristi narkotike.{S} Negira
anamsntički značajna oboljenja.{S} Navodi operaciju devijacije nosa
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false"
val="1978">1978 GODINE</TIMEX3> U <top.gr PHI="yes">Kamengradu</top.gr>.
{S} U detinjstvu operacija krajnika, a <TIMEX2 proveraP="MISS/E"
type="DATE">199’ godine</TIMEX3> prelom leve noge.{S}
ANAMNESIS TRAUMAE Navodi da je povređen <TIMEX3 proveraP="UOKo"
proveraN="OK" type="DATE" temporalFunction="false" val="2009-07-12">
12.07.2009.</TIMEX3> gdoine <TIMEX3 proveraP="OK" proveraN="OK"
type="TIME" temporalFunction="false" val="T18" mod="APPROX">oko 18 h
</TIMEX3> prilikom slučajnog pada .{S} Rekonstruiše okolnosti povređivanja,
negira povraćanje.{S} Krvario iz rane čeonog predela sa leve strane.{S}
Inicijalno zbrinut u RMC <top.gr PHI="yes">Kamengrad</top.gr>, gde je
načinjena sutura razderine čeono levo i CT dijagnostika.{S} Potom je
<TIMEX3 proveraP="OK" proveraN="OK" type="DATE" temporalFunction="false"
val="2009-07-13">13.07.2009.godine</TIMEX3> transportovan u UC <org
PHI="yes">Klinike</org> gde je pregledan od strane neurohirurga i potom
upućen u <org PHI="yes">Kliniku</org>.{S} Žali se na utrnulost zuba sa
leve strane, kao i kože obraza i gornje usne sa te strane.{S}
STATUS PRESENS UNIVERSALIS:{S} Pacijent je svestan, orjentisan u sva tri
pravca, komunikativan, , zenice jednake, reaktivne, bez znakova grubog
neurološkog deficita.{S}
STATUS LOCALIS Inspekcioom spolja se konstatuje rana u predelu čela koja
je dužiune oko 2 cm, u KK pravcu, suturirana pojedinačnim šavovima.
{S}Lateroorbitalno, zigomatično kao i prisutan je prekidn kontinuiteta
kože, koža lišena natkožice, prekrivena prljavo sivim krustama.{S}
Prisutan je i defekt dela kože jagodično slepoočnog predela sa te strane.
{S}

```
Palpacijom u predelu donje očne ivice je prisutan koštani stepenik u
predelu rama leve orbite na <TIMEX3 proveraP="NOK" proveraN="OK"
type="TIME" temporalFunction="false" val="T07" mod="APPROX">oko 7
sati</TIMEX3>, kao i laterorobitalno sa iste strane.{S}
Otvaranje usta je ograničeno, i praćeno bolom.{S}
Intraoralno prilikom palpacije prisutna je osetljivost u predelu ZA levo
Rtg snimak u prilogu.{S}Vrat je simetričan, pokretn u svim pravcima.{S}
DG FRACTURA COMPLEXUS ZYGOMATICOMAXILLARIS LAT SIN
<pers PHI="yes"><persName.last>Kremenko</persName.last></pers>
```

Primer E.5. Primeri ispravno obeleženih konkordanci

<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="true" val="XXXX-WXX-3TAF">sredu posle podne</TIMEX3>

<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="false" val="2013-11-18T12" mod="APPROX">18.11.2013. godine oko 12 h</TIMEX3>

<TIMEX3 proveraP="OK" proveraN="OK" type="TIME" temporalFunction="false" val="2013-12-05T00" mod="APPROX">05.12.2013 godine oko 00 h</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="TIME" temporalFunction="true val="TMO>ujutru</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="DATE" temporalFunction="false val="2013-02-26>26.2.2013.</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="DATE" temporalFunction="false val="2005>2005 godine</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="DURATION temporalFunction="false val="P20Y mod="APPROX>oko 20 godina</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="DURATION temporalFunction="false val="P76Y>76 godina</TIMEX3>

<TIMEX3 proveraP="OK proveraN="OK type="DURATION temporalFunction="false val="P3Y>poslednje tri godine</TIMEX3>

<TIMEX3 proveraP="OK type="DURATION.PERIOD><TIMEX3 proveraP="OK proveraN="OK type="DURATION temporalFunction="false val="P10Y>10</TIMEX3 >-<TIMEX3 proveraP="OK proveraN="OK type="DURATION temporalFunction="false val="P15Y>15 godina</TIMEX3></TIMEX3>

Primer E.6. Primeri pogrešno obeleženih ili propuštenih konkordanci

<TIMEX3 proveraP="UOKo/E" proveraN="OK" type="DATE" temporalFunction="false" val="2008-06-20">20.06.2008</TIMEX3> goine

<TIMEX3 proveraP="UOKo" proveraN="OK" type="TIME" temporalFunction="false" val="2013-11-30T01">subotu 30.11.2013. godine u 1h</TIMEX3> posle ponoći

<TIMEX3 proveraP="UOKo/E" proveraN="OK" type="DATE" temporalFunction="false" val="2011-07-12">12.07.2011.</TIMEX3> gdoine

Detralex tbl <TIMEX3 proveraP="UOKo" proveraN="UOKv" type="TIME" temporalFunction="false" val="T02">2 ujutro</TIMEX3>

leve orbite na <TIMEX3 proveraP="NOK" proveraN="OK" type="TIME" temporalFunction="false" val="T07" mod="APPROX">oko 7 sati</TIMEX3>.{S} Pokreti bulbusa

Dinuorm 1x1 na <TIMEX3 proveraP="MISS" type="SET">2 dan</TIMEX3>

kontrole zavoja na <TIMEX3 proveraP="MISS" type="SET">drugi dan</TIMEX3>

operaciju leve ruke sa <TIMEX3 proveraP="MISS" type="DURATION">12 god.</TIMEX3>

u <TIMEX3 proveraP="MISS" type="DURATION">7 god.</TIMEX3> operisao slepo crevo

ili <TIMEX3 proveraP="MISS" type="DATE">98</TIMEX3>, ne zna tačno

Spisak tabela

Spisak slika

Biografija autora

Jelena Jaćimović je rođena 1. januara 1976. godine u Požarevcu, gde je završila osnovnu školu i gimnaziju. Diplomirala je na Katedri za bibliotekarstvo i informatiku Filološkog fakulteta Univerziteta u Beogradu februara 2002. godine, gde je 2009. godine završila i diplomske akademske studije – master.

Od 2003. godine zaposlena je kao bibliotekar na Stomatološkom fakultetu Univerziteta u Beogradu. Od 2009. godine angažovana je i kao saradnik u nastavi doktorskih akademskih studija na predmetu Biomedicinska naučna informatika.

U periodu 2008–2010. godine bila je član Komisije za unapređenje korisničkih servisa Bibliotekarskog društva Srbije. Od 2012. godine član je Sekcije za visokoškolske biblioteke Bibliotekarskog društva Srbije.

Tokom maja 2011. godine učestvovala je u studijskom putovanju bibliotekara Srbije u Nemačku, pod nazivom „Organizacija u naučnim bibliotekama – najbolji primeri iz prakse“, sprovedenog u organizaciji Gete instituta u Beogradu, Narodne biblioteke Srbije i *Bibliothek & Information International* (*BI-International*).

Dobitnik je počasnog zvanja „Arhont otvorenog pristupa Univerziteta u Beogradu“ za 2012. godinu, namenjenog istraživačima i bibliotekarima koji su tokom svog rada demonstrirali najviši stepen profesionalnosti u promociji principa otvorenog pristrupa naučnim informacijama.

Organizovala je i vodila kurseve o naučnim informacijama u Srbiji, njihovoj dostupnosti i načinima pronalaženja, namenjene studentima i nastavnicima Stomatološkog, Medicinskog i Veterinarskog fakulteta Univerziteta u Beogradu.

Bavi se istraživanjem u oblasti bibliotečke informatike i automatske obrade prirodnih jezika, posebno razvojem alata namenjenih ekstrakciji informacija iz tekstova srpskog jezika. Objavila je više naučnih radova iz ovih oblasti i učestvovala je na međunarodnim i nacionalnim skupovima.

Прилог 1.

Изјава о ауторству

Потписани-а Јелена Јаћимовић

број индекса 08129Д

Изјављујем

да је докторска дисертација под насловом

Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, _____

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Јелена Јаћимовић

Број индекса 08129Д

Студијски програм

Наслов рада Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику

Ментор Проф. др Цветана Крстев

Потписани/а Јелена Јаћимовић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, _____

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, _____

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.