T A R T U Ü L I K O O L
MATEMAATIKA-INFORMAATIKATEADUSKOND

Arvutiteaduse instituut
Infotehnoloogia eriala

Siim Orasmaa

Ajaväljendite tuvastamine eestikeelses tekstis

Magistritöö (30 EAP)

Juhendaja: Margus Treumuth

Autor: …… "….." mai 2010

Juhendaja: ……. "….." mai 2010

Lubada kaitsmisele:
Professor: .....…… "….." …….. 2010

TARTU 2010

Sisukord

Sissejuhatus

Käesoleva töö eesmärgiks on luua eesti keele ajaväljendite tuvastaja: programm, mis leiab loomuliku keele tekstist üles ajaväljendid ning esitab leitud väljendite semantika formaliseeritud kujul. Töö on jätkuks autori bakalaureusetööle, kus tehti esimesed sammud siinkirjeldatava programmi loomisel.

Ajaväljendite tuvastamine on oluline alamosa mitmetes loomuliku keele automaattöötluse ülesannetes, näiteks dialoogisüsteemides ning automaatsel sisukokkuvõtete tegemisel. Loomuliku keele korpus, kus on märgendatud ajaväljendid, avab uusi võimalusi infootsingute sooritamisel: dokumente ja artikleid on võimalik nendes sisalduvate ajaliste viidete järgi grupeerida ning infopäringut sooritav kasutaja saab täpsustada, millistele ajaperioodidele viitavad dokumendid teda kõige rohkem huvitavad.

Ajaväljendite tuvastamisel käesolevas töös keskendutakse eelkõige konkreetsetele ajaväljenditele – s.o ajaväljenditele, mida on võimalik kalendris positsioneerida (st seada vastavusse ajapunkt või ajavahemik), või mis avalduvad kalendriühikutes mõõdetavate ajalõikudena (lõikude otspunktid ei pea tingimata olema kalendris määratud).

Loodud rakendus kasutab ajaväljendite tuvastamisel reeglipõhist lähenemist. Tuvastamisreeglid toetuvad automaatse morfoloogilise analüüsi ja ühestamise tulemustele. Iga tuvastamisreegel kirjeldab ühelt poolt ajaväljendile vastavat fraasi tekstis ning teiselt poolt annab edasi instruktsioonide jada, mis tuleb ajaväljendi semantika leidmiseks läbi viia. Kõrvuti eraldatud väljenditest pikemate fraaside moodustamiseks kasutab süsteem täiendavat kihti reegleid ning süsteemisiseste protseduuridena realiseeritud heuristikuid.

Süsteemi arendamisel ja testimisel keskenduti eelkõige ajakirjandustekstidele. Seetõttu on tõenäoline, et loodud süsteemi rakendamisel mõnes teises valdkonnas (nt ilukirjanduses) saadakse siinmõõdetust madalamad tulemused.

Antud töö koosneb neljast osast. Esimeses osas selgitatakse mõisteid, mida ajaväljendite tuvastamise ülesande kirjeldamisel kasutatakse, ja tutvustatakse ajaväljendite liigitusaluseid ning märgendusformaate. Teises osas vaadeldakse erinevaid lähenemisviise, mida on kasutatud ajaväljendite automaatsel tuvastamisel. Kolmandas osas käsitletakse meetodit, mis võimaldab ajaväljendeid automaatselt tekstikorpusest kaevandada, ning antakse ülevaade töö käigus loodud reeglipõhisest ajaväljendite tuvastajast. Neljandas osas võrreldakse eksperimentaalsel teel erinevaid semantika lahendamise heuristikuid ning

analüüsitakse loodud süsteemi poolt tehtavaid vigu ajakirjanduskorpustel.

Tööle on kaasa pandud viis lisa: 1) näide ajaväljendite kaevandamise tulemustest, 2) ajaväljendite tuvastaja märgendusformaadi kirjeldus, 3) ülevaade ajaväljendite semantika leidmisel kasutatavatest instruktsioonidest, 4) testkorpuses märgendatud ajaväljendite sagedusprofiilid ning 5) materjalide (korpused, programmid) CD.

Kuna ajaväljendite tuvastaja kirjeldus sisaldab ka tehnilisi detaile, eeldatakse, et käesoleva töö lugeja omab baasteadmisi programmeerimisest keeles Java ning regulaaravaldiste koostamisest.

1 Ajaväljendite liigitusalused ja märgendusstandardid

Käesoleva peatüki eesmärgiks on anda töös lahendatava ülesande mõistmist hõlbustavad taustateadmised. Peatüki alguses selgitame mõisteid, mida kasutame nii selles kui ka järgnevates peatükkides ajaväljendite tuvastamise ülesande kirjeldamisel. Seejärel vaatleme praktilistes rakendustes kasutatud ajaväljendite liigitusi. Peatüki lõpus tutvustame lühidalt märgendusstandardeid TIMEX2 ja TimeML ning kirjeldame ajaväljendite liigitust antud töös.

1.1 Kasutatavad mõisted

Ajaväljend on keeles aja väljendamiseks kasutatav keeleüksus või mitmest üksusest koosnev fraas.

Ajaväljendite eraldamiseks (*temporal expression recognition / extraction / identification*) nimetatakse loomuliku keele ajaväljendite leidmist ja esiletõstmist tekstis.

Ajaväljendite tuvastamiseks (*temporal expression resolution / normalization / interpretation*) nimetatakse loomuliku keele ajaväljendile vastava (konkreetse või hägusa) kuupäeva, ajaintervalli või sageduse leidmist ning selle esitamist formaalsel kujul (**normaliseerimist**), mis oleks üheselt mõistetav ning kasutatav teksti edasisel automaattöötlusel [1].

Ajaväljendi granulaarsus (*granularity*) näitab, millise detailsusega ajalist informatsiooni sisaldab ajaväljend ilmutatud kujul. Näiteks ajaväljend *„aastal 2005“* sisaldab ainult *aasta*-granulaarsusega informatsiooni ning ajaväljend *„homme kell 13“* sisaldab *päev*- ja *tund*-granulaarsusega ajalist informatsiooni.

Käesoleva töö praktilises osas käsitletakse ajaväljendeid, mis võivad sisaldada 6 erineva granulaarsusega informatsiooni (*aasta*-, *kuu*-, *nädal*-, *päev*-, *tund*- ja *minut*-granulaarsus). Lahenduse tehnilisel kirjeldamisel kasutatakse terminiga *granulaarsus* samatähenduslikult ka terminit *kalendriväli*.

Granulaarsuste järjestus defineeritakse käesolevas töös kui vastavate ajaühikute järjestus pikkuse alusel: *aasta* > *kuu* > *nädal* > *päev* > *tund* > *minut*. Seega loetakse nt *aasta*-granulaarsust *suuremaks* kui *kuu*-granulaarsus ning *tund*-granulaarsust *väiksemaks* kui *päev*- või *nädal*-granulaarsus.

Referentsajaks (*reference time*) nimetatakse ajalist nullpunkti, mille teadmine on eelduseks mõningate ajaväljendite semantika lahendamisel. Referentsajaks võib olla kõnehetk (nt ajaväljendite *homme, tänavu, mullu* puhul), teise ajaväljendi semantika lahendus (nt lauses „*1996. aastal oli puudujääk 5% väiksem kui aasta varem.*“) või sündmuse toimumise aeg (nt ajaväljendis „*päev enne festivali algust*“). Kirjanduses nimetatakse referentsaja leidmist ka ajaväljendite **ankurdamiseks** (*anchoring*). Käesoleva töö praktilises osas keskendume kõnehetkest või teise ajaväljendi semantika lahendusest sõltuvatele ajaväljenditele.

1.2 Ajaväljendite liigitusaluseid

Ajaväljendeid tuvastava süsteemi loomisel on oluline saada ülevaade ajaväljendite võimalikest liigitustest. Järgnevalt vaatame mõningaid ajaväljendite liigitusi, mida on kasutatud praktilistes lähenemistes tuvastamise ülesandele.

E. Saquete, R. Muñoz ja P. Martínez-Barco [2] toovad välja kaks ajaväljendite liigitusviisi: 1) klassifitseerimine ajalise viite liigi põhjal ning 2) klassifitseerimine ajaväljendi semantika esituskuju põhjal.

1) Klassifitseerimisel ajalise viite liigi alusel eristatakse ilmutatud semantikaga väljendeid (*explicit temporal expressions,* nt „*June 1999*“) ning varjatud semantikaga väljendeid (*implicit temporal expressions*). Varjatud semantikaga väljendite puhul toovad autorid välja täiendava jaotuse: kõnehetkele[1] toetuvad ajaväljendid (nt *„yesterday“, „next month“*) ning mõnele teisele (tekstis mainitud) ajahetkele toetuvad väljendid (nt *„after next Christmas“, „a month later“*).

2) Ajaväljendi semantika esituskuju põhjal eristavad autorid kolme liiki väljendeid: konkreetse kalendrikuupäeva ja/või kellaajana esitatavad väljendid (nt *„yesterday“* esitus kujul `dd/mm/yyyy`[2]), konkreetse ajavahemikuna esitatavad väljendid (nt *„during the five following days*“ esitus intervallina `[dd/mm/yyyy - dd/mm/yyyy]`) ning hägusad ajaväljendid (nt *„a day of the last week“*). Autorite käsitluses on ka hägusate ajaväljendite esituskujuks intervall, mis võimaldab anda ajaväljendi semantikale orienteeruvad piirid.

Liigitusviisile 1) sarnaneb ka töös [3] toodud liigitus: autorid eristavad töö praktilises osas

1 Autorite praktilises käsitluses on „kõnehetkeks“ dokumendi loomise kuupäev.
2 `dd` – kuupäev, `mm` – kuu, `yyyy`– aastaarv.

ilmutatud ajaväljendeid (*explicit expressions*), kõnehetkest sõltuvaid deiktilisi väljendeid (*deictic expressions*) ning diskursuse kesksest ajast sõltuvaid relatiivseid ajaväljendeid (*relative expressions*). Lisaks sellele vaatlevad nad neljanda ajaväljendiliigina ajalist kestvust edasiandvaid väljendeid (*duration expressions*, nt *„less than 20 minutes“*).

M. T. Vicente-Díez, D. Samy ja P. Martínez [4] kasutavad kahte ajaväljendite liigitusviisi: a) liigitamine ajaväljendifraasi struktuuri alusel ning b) liigitamine ajaväljendi lahenduse järgi.

a) Ajaväljendifraasi struktuuris eristavad autorid kaht tüüpi põhimoodustajaid: aja üksused (*time units*) ning aja täpsustajad (*time modifiers*). Aja üksuste alla kuuluvad mõõtühikud (nt *„hour“, „week“*), deiktilised üksused (nt *„today“, „yesterday“*) ning nimelised üksused (nt *„Winter“, „January“, „2009“*). Aja täpsustajatena käsitlevad autorid aja üksuste semantikat täiendavaid sõnu (nt *„last“, „each“, „after“*).

b) Ajaväljendite liigitus lahenduse järgi sisaldab alamosana liigitust 1); lisaks sellele eristavad autorid ajaintervallina lahenduvaid ajaväljendeid, ajahulkadena lahenduvaid väljendeid (*time sets*, nt *„Mondays“, „every day“*) ja kestvusena lahenduvaid väljendeid (nt *„during two months“*). Lahendamise seisukohast vaatavad autorid eraldi liigina ka tähtpäevade nimesid (nt *„Christmas Day“*), sest nende lahendamine seisneb väljendile (maailmateadmuse alusel) konkreetse kuupäeva vastavusse seadmises.

1.3 Märgendusstandardid

Järgnevalt anname lühiülevaate ajaväljendite semantika esitamisest märgendusstandardites TIMEX2 ja TimeML.

1.3.1 TIMEX2

TIMEX2 [5] on XML-formaadis märgendusskeem, mis võimaldab tuua tekstis esile aja-väljendid (ümbritsedes need TIMEX2 märgenditega) ning normaliseerida need kujule, mis põhineb ISO-8601:1997 kalendriaegade esitamise standardil [6]. Märgendusskeem esitab konkreetsele ajapunktile viitavate ajaväljendite tähenduse kas kuupõhises formaadis (nt „*February 15th in 2010“* normaliseeritakse kujule „2010-02-15“) või nädalapõhises formaadis (nt „*in the last week*“ normaliseeritakse kujule „2010-W06“, kui referentsaeg viitab 2010. aasta seitsmendale nädalale). Samuti võimaldab formaat anda edasi ajalisi kestvuseid (nt *„half an hour long“* normaliseeritakse kujule „PT30M“) ning ajalisi

korduvusi (nt *„each year*“ normaliseerides tuuakse välja korduvate aegade vaheline periood (*periodicity*) – F1Y ning korduva ajalõigu granulaarsus (*granularity*) – G1Y[3]).
Mittekonkreetse semantika esitamiseks on kasutusel kaks strateegiat: päevaosade, aastaaegade, kvartalite ja poolaastate märkimiseks on eraldi tähised (nt *„summer of 1990“* normaliseeritakse kujule „1990-SU“) ning kui ajaväljendit on kasutatud üldises tähenduses, märgitakse puuduolevad ajalised granulaarsused X-sümbolitega (nt lauses *„April is usually wet*“ normaliseeritakse ajaväljend *„April“* kujule „XXXX-04“, kus XXXX tähistab puuduolevat aastaarvu).
Näide 1 toob TIMEX2 formaadi alusel märgendatud inglisekeelse lause. Normaliseeritud semantika tuuakse välja märgendi atribuudis `VAL`. Lisaks näeb standard ette semantika täpsustuste esitamist atribuudis `MOD` (nt väljendi *„about three years ago“* semantika ligikaudsust märgib `MOD="APPROX"`) ning ajaväljendite ankurdamise edasiandmist kasutades märgendeid `ANCHOR_VAL` (referentsaeg) ja `ANCHOR_DIR` (paiknemine ajateljel referentsaja suhtes).

Näide 1. TIMEX2 formaadis märgendatud lause.

```
I returned to work at <TIMEX2 VAL="1984-01-03T12:00">twelve o'clock
January 3, 1984</TIMEX2>.
```

1.3.2 TimeML

TimeML märgendusformaat [7] on TIMEX2 formaadi edasiarendus, mis võimaldab lisaks ajaväljenditele märgendada sündmuseid ning tuua välja seosed ajaväljendite ning sündmuste vahel.
Ajaväljendite semantika normaliseerimise osas kattub TimeML suures osas formaadiga TIMEX2: atribuudid `VAL` ja `MOD` võetakse üle terviklikul kujul. Lisaks eeltoodud atribuutidele nõuab TimeML ka ajaväljendi tüübi väljatoomist, lubades eristada nelja tüüpi ajaväljendeid: `DATE`, `TIME`, `DURATION` ja `SET`. Tüübi `DATE` alla loetakse kuuluvat kõik ajaväljendid, mille saab esitada ajapunktina kalendris ning mis ei sisalda kellaajalist informatsiooni (nt *the second of December*, *this year's summer*, *last week*). Tüüp `TIME` katab ajaväljendeid, mis sisaldavad lisaks kalendrilisele informatsioonile ka kellaajalist informatsiooni, olgu see siis avaldatud kas konkreetse kellaaja (nt *9 a.m. October 1, 1999*) või päevaosale viitamise kaudu (nt *late last night*). Tüübi `DURATION` alla kuuluvad ajalised

3 TIMEX2 viimases versioonis (v1.1) ei nõuta siiski korduvustel perioodi ja granulaarsuse täpsustamist, kuna nende määramine tekitas käsitsi märgendamisel segadust. Autorid möönavad, et korduvuste esitamise formaat ei ole lõplikult paika pandud ning vajab täiendavat uurimist.

kestvused ning tüübi `SET` alla ajalised korduvused.

TIMEX2 formaadis kasutatud ajaliste korduvuste esitamist lihtsustatakse: korduva perioodi pikkus tuuakse atribuudis `value`, ning selle täpsustamiseks võetakse täiendavalt kasutusele atribuudid `quant` (annab edasi korduvuse kvantori) ning `freq` (täisarvuline kordumissagedus, millele võib olla määratud granulaarsus). Näiteks ajaväljendi *„three days every month“* semantika esitatakse kujul `value=P1M;quant=EVERY;freq=P3D`.

1.4 Liigitamine käesolevas töös

Sarnaselt TIMEX2 ja TimeML standarditele, läheneme käesoleva töö praktilises osas ajaväljendite liigitamisele semantika normaliseerimise lähtepunktist, eristades järgmisi ajaväljendiliike:

- **Ajapunkt** on ajateljel punktina paigutuv ajaväljend. Kuna ajaväljendite granulaarsused on erinevad, saab „punkti“ kujul esitada nii ajaväljendi *„2009. aastal“* semantika kui ka ajaväljendi *„järgmisel hommikul kell 13.00“* semantika.
- **Ajavahemik** on ajaväljend, millele saab vastavuse seada konkreetsete otspunktidega lõigu ajateljel. Ajavahemiku otspunktid võivad olla ilmutatud kujul, nt väljendites *„esmaspäevast reedeni“* ning *„1.-3. märtsil“* või referentsajast (kõnehetkest) tuletatavad, nt väljendis *„viimase viie aasta jooksul.“*
- **Ajaline kestvus** on ajalõiku väljendav ajaväljend, mis ei täpsusta (ilmutatud kujul) ajalõigu alguspunkti ning lõpp-punkti. Kestvust väljendavateks loeme näiteks ajaväljendid *„pool tundi“* ja *„22-aastane.“*

  Kui ajaväljend annab edasi nii kestvuse kui ka ajapunkti tähendust, liigitame selle ajapunktide alla kuuluvaks (nt lauses „*Mari jäi nädalavahetuseks linna.“*).
- **Ajaline korduvus** on ajaväljend, mis annab edasi mingi sündmuse ajalist korduvust. Kordumist edasiandvateks loeme näiteks ajaväljendid *„esmaspäeviti“* ning *„kaks korda aastas.“*

Selline liigitusviis ühildub osaliselt ka M.Erelt jt poolt „Eesti keele grammatikas“ [8] kasutatud ajamääruste liigitusviisiga. Viidatud töös jagati ajamäärused nelja rühma: 1) toimumisaega väljendavad ajamäärused (käesolevas töös ajapunktid), 2) ajapiiri väljendavad ajamäärused (käesolevas töös eraldiseisvatena ajapunktid, koosesinevatena moodustavad ajavahemiku), 3) kestust näitavad ajamäärused ja 4) korduvust väljendavad ajamäärused.

2 Ajaväljendite tuvastamise probleemi lahendusi

Käesoleva peatüki eesmärgiks on anda ülevaade erinevatest lähenemisviisidest ja strateegiatest, mida on ajaväljendite tuvastamisel kasutatud. Ajaväljendeid tuvastavatest süsteemidest vaatleme eraldi kahte reeglipõhist süsteemi ning tutvustame masinõppel põhinevaid lähenemisi. Kuna käesoleva töö praktilises osas kasutatakse reeglipõhist lähenemist, on ka siin peatükis toodud reeglipõhiste süsteemide käsitlused detailsemad ning masinõppepõhiseid lähenemisi tutvustame ainult lühidalt. Peatüki lõpus anname lühiülevaate sellest, milliseid meetodeid on seni katsetatud ajaväljendite tuvastamisel eesti keeles.

2.1 Reeglipõhised lähenemised

Järgnevalt kirjeldame mõningaid lähenemisi, mida on kasutatud ajaväljendite reeglipõhisel tuvastamisel teistes keeltes. Kuigi vaadeldavate süsteemide juures on toodud välja ka testimise tulemused (saagis ja täpsus), ei ole need otseselt võrreldavad, kuna süsteeme on testitud erinevate korpuste peal.

2.1.1 TempEx märgendaja

TempEx märgendaja [9] on ajalooliselt üks esimesi inglise keelele loodud ajaväljendite märgendajatest. Süsteemi ajaväljendeid eraldav osa on üles ehitatud automaatse sõnaliikide märgendaja töö tulemustele ning leiab tekstist ajaväljendid üles käsitsi koostatud reeglite alusel. Autorid ei täpsusta, millisel kujul on ajaväljendite eraldamisel kasutatavad reeglid, küll aga annavad põhjaliku ülevaate ajaväljendite semantika leidmise strateegiatest. Nad keskenduvad oma töös ajapunktina lahenduvatele ajaväljenditele, eristades iseseisvaid ajaväljendeid (*self-contained temporal expressions*, näiteks „*June 1999*“), mille lahendamine taandub lihtsalt ajaväljendi ümberkirjutamisele normaliseeritud kujule, ning kontekstist sõltuvaid ajaväljendeid (*context-dependent expressions*, näiteks „*next Tuesday*“, „*2 weeks ago*“), mille lahendamiseks kasutavad autorid erinevaid heuristikuid.
Lahendamisstrateegia järgi eristavad autorid kolme liiki kontekstist sõltuvaid ajaväljendeid: (a) ainult referentsajast sõltuvad väljendid, (b) referentsajast ja fraasi leksikaalsetest tunnustest (*lexical markers*) sõltuvad väljendid ning (c) referentsajast ja verbi grammatilisest ajast sõltuvad väljendid.

Ainult referentsajast sõltuvateks (a) loetakse nt väljendid „*tomorrow*“, „*yesterday*“, „*this afternoon*“; selliste väljendite lahendamisel nähakse olulise probleemina üldise ja konkreetse tähenduse eristamist (nt „*today*“ tähendused *tänapäeval* ja *kõnehetkega samal päeval*) ning lahendusena katsetatakse masinõppe meetodeid kontekstide eristamisel.

Referentsajast ja fraasi leksikaalsetest tunnustest sõltuvaiks (b) loetakse näiteks väljendid „*next month*“ ja „*this coming Thursday*“, mille puhul otsustatakse ajapunkti paiknemissuund referentsaja suhtes ajaväljendifraasis sisalduvate leksikaalsete tunnuste põhjal (eeltoodud näidetes allajoonitud).

Referentsajast ja verbi grammatilisest ajast sõltuvaiks (c) loetakse üksikult esinevad kuu- ja nädalapäevanimed (nt „*Thursday“* lauses „*The Iraqi news agency said the first shipment of 600,000 barrels was loaded Thursday by the oil tanker Edinburgh“*). Selliste ajaväljendite puhul tuginetakse ajaväljendile lähima verbi grammatilisele ajale, et määrata ajapunkti paiknemissuund referentsaja suhtes (minevikus, olevikus või tulevikus). Lähima verbi leidmisel vaadatakse kõigepealt ajaväljendile eelnevaid sõnu kuni eelmise ajaväljendini (või lause alguseni), seejärel ajaväljendile järgnevaid sõnu kuni lauselõpuni, ning lõpuks kõiki ajaväljendi ja lausealguse vahele jäävaid sõnu. Kui eeltoodud heuristik siiski ebaõnnestub, otsustatakse ajaväljendi paiknemissuund referentsaja suhtes mõningate lisatunnuste (nt suunale vihjavad sõnad „*since“* või „*until“*) esinemise järgi lauses või valitakse lahenduseks lähim sobiv kuupäev, mis ei ole referentsajast kaugemal kui üks kuu.

Autorid katsetavad süsteemi ajalehekorpuste peal ning mõõdavad ajaväljendite eraldamisel süsteemi saagiseks 95,6% ja täpsuseks 96,8%. Ajaväljendite semantika määramisel[4] saadakse tulemuseks vastavalt 82,7% ja 83,7%. Tulemust analüüsides leiavad autorid, et ajaväljendite eraldamisel põhjustasid kõige rohkem vigu implementeerimata jäänud reeglid ning semantika lahendamisel suutmatus eristada ajaväljendi kasutust üldise ja konkreetse tähenduse edasi andmisel.

2.1.2 Süsteem Chronos

M.Negri ja L.Marseglia annavad raportis [10] ülevaate süsteemist Chronos, mis märgendab inglisekeelseid ajaväljendid TIMEX2 formaadi alusel. Etteantud formaadist juhindudes käsitlevad autorid eraldi absoluutseid ajaväljendeid (nt „*July 17, 1999“*) ning relatiivseid ajaväljendeid (nt *„three years ago“*), lisaks sellele märgendatakse ajalisi kestvuseid (nt

4 Normaliseerimine seisneb ainult ühe atribuudi täitmises, mille formaat on sarnane TIMEX2 atribuudi VAL formaadile.

„three weeks“) ning ajalisi korduvusi (nt *„every week“*).

Süsteemi arhitektuuris on kaks peakomponenti: eraldamise ja sulustamise komponent (*detection and bracketing component*), mille ülesandeks on ajaväljendite leidmine ja esialgse märgendusega ümbritsemine, ning normaliseerimise komponent (*normalization component*), mis formaliseerib ajaväljendi semantika tuginedes esialgse märgenduse tulemusele.

Eraldamise ja sulustamise etapis eristatakse omakorda kolme alamosa: a) lingvistiline eeltöötlus, mille käigus märgendatakse tekstis sõnaliigid ning leitakse üles püsiühendid; b) põhireeglite rakendamine, mille käigus leitakse kõik potentsiaalsed ajaväljendid ning määratakse nende ulatus; ning c) kompositsioonireeglite rakendamine, mille eesmärgiks on lahendada märgenduste ülekattuvusest[5] tulenevaid probleeme.

Eraldamise käigus otsitakse esmalt nn leksikaalseid võtmesõnu (*lexical triggers*, nt *„day“*, *„weekly“, „Christmas“*), et määrata kindlaks ajaväljendi olemasolu lauses, ning seejärel püütakse ajaväljendi ulatust laiendada, otsides kontekstist täiendavaid ajaväljendifraasi koostisosi (nt sõnad *„following“, „every“, „during“*). Esialgsete märgendite lisamisel kasutatakse ajutisi atribuute, et anda edasi instruktsioone relatiivsete ajaväljendite semantika leidmiseks. Näiteks ajaväljendi *„three years later“* semantika lahenduskäigu annavad edasi atribuudid `OP="+" T-CAT="year" QUANT="3"` (suurendada referentsaja aastaarvu 3 võrra).

Ajaväljendite lõplikule normaliseerimisele eelneb ankurdamise etapp, mille käigus leitakse igale relatiivsele ajaväljendile referentsaeg, millest lähtuvalt selle väärtuse saab arvutada. Autorid kasutavad ankurdamisel kahte heuristikut: esimese järgi tuleb relatiivne ajaväljend ankurdada dokumendi loomise kuupäeva külge, teise alusel aga ankurdatakse relatiivne väljend tekstis eelneva ning antud väljendile lähima absoluutse ajaväljendi külge. Selle, millist heuristikut parajasti kasutada, määrab ankurdatava ajaväljendi võtmesõna ja seda ümbritsevad leksikaalsed tunnused. Näiteks deiktilised võtmesõnad *„today“, „tonight“,* üksikud nädalapäeva- või kuunimed nõuavad dokumendi kuupäeva külge ankurdamist. Samuti nõuab seda sõnade *„this“, „last“, „next“* olemasolu ajaväljendifraasis. Aga kui ajaväljendifraasi koosseisu kuuluvad sõnad *„following“, „previous“* või *„same“*, tuleb ajaväljend ankurdada tekstis vahetult eelneva ajaväljendi külge.

Ankurdamist eelneva ajaväljendi külge kontrollib kitsendus: valitud ankur peab sisaldama

5 Näiteks fraasist *„the whole Monday night“* tehakse kolm ülekattuvat eraldust: *„the whole Monday“*, *„Monday night“* ja *„the whole Monday night“*.

ankurdatava väljendiga sama või väiksema granulaarsusega ajalist informatsiooni. Näiteks kui relatiivne väljend sisaldab *kuu*-granulaarsusega informatsiooni, siis võib ankruks sobiv absoluutne väljend sisaldada *kuu-, nädal-* ja *päev*-granulaarsusega informatsiooni, aga mitte *aasta*-granulaarsusega ajalist informatsiooni.

Absoluutsete ajaväljendite lõplik normaliseerimine on triviaalne ning seisneb lihtsalt väärtuste ümberkirjutamises TIMEX2 formaati. Relatiivsete väljendite puhul kasutatakse esialgsetes märgendites peituvat informatsiooni (ajutised atribuudid `OP`, `T-CAT`, `QUANT`), et viia valitud ankur-ajahetke peal läbi arvutuskäik. Lihtsamatel juhtudel piirdubki kogu arvutus vaid ühe aritmeetilise tehtega kalendris. Keerulisemad on juhud, kus relatiivne väljend ei sisalda ilmutatud kujul liidetavat või lahutatavat ajakvantiteeti ning normaliseerimise komponent peab selle ise kindlaks määrama (nt lause „*He started studying on March 30 2004, and passed the exam the following Friday"* puhul tuleb süsteemil leida, mitu päeva jääb „*March 30 2004"* ja „*the following Friday"* vahele).

Süsteemi Chronos autorid osalesid TERN 2004 hindamisel,[6] kus mõõdeti süsteemi tulemuseks ajaväljendite eraldamisel saagis 88,0% ning täpsus 97,6%. Semantika normaliseerimisel (täpsemalt – TIMEX2 atribuudi `VAL` täitmisel) saadi saagiseks 87% ning täpsuseks 87,5%. TERN hindamisel osalevate süsteemide seas oli Chronos nende tulemustega parim.

2.2 Statistilised meetodid

Vältimaks aeganõudvat ja vigadealdist reeglite koostamise protsessi, on proovitud mitmeid lähenemisi masinõppe kaasamiseks ajaväljendite tuvastamisel. Käesoleva töö autorile teadaolevalt on tuvastamisel edukalt rakendust leidnud vaid juhendatud masinõppe (*supervised learning*) meetodid, mis eeldavad käsitsi märgendatud treeningkorpuse olemasolu. Ajaväljendite eraldamisel on lootustandvaid tulemusi saadud ka pooljuhendatud masinõppe (*semi-supervised learning*) meetoditega, mis lubavad väikese arvu positiivsete näidete abil õppida märgendamata korpusest uute ajaväljendite kuju ja fraasistruktuuri.

Järgnevalt vaatleme lähemalt kahte juhendatud masinõppe lähenemist ning ühte pooljuhendatud masinõppe meetodit.

6 TERN (*Temporal Expression Recognition and Normalization evaluation*) võistlus TIMEX2 formaadis ajaväljendeid märgendavate süsteemide hindamiseks; vt http://fofoca.mitre.org/tern.html (viimati vaadatud 23.02.2010)

2.2.1 Juhendatud masinõpe

J. A. Baldwin katsetab TIMEX2 märgendamisreeglite automaatset õppimist käsitsi märgendatud treeningkorpuselt [11]. Autori lähenemise kandvaks ideeks on grupeerida sarnaselt normaliseeritud TIMEX2 märgendid (nt kõik kalendrikuuks normaliseeritud ajaväljendid) ning õppida automaatselt, millised märgendite vahele jäävad sõnad väljendavad mingit osa normaliseeritud semantikast. Õppimist alustatakse ühesõnalistest ajaväljenditest, mille puhul lihtsalt memoreeritakse sõna „tähendus“, ning minnakse järkjärgult edasi pikemate fraaside juurde, mille puhul püütakse leida tundmatute fraasi liikmete tähendusi, välistades juba teadaolevaid tähendus-sõna seoseid. Autor katsetab loodud süsteemi inglise- ja prantsusekeelsete uudistekstide peal ning mõõdab inglisekeelsetel tekstidel semantika normaliseerimise saagiseks 84,0% ja täpsuseks 83,7% ning prantsusekeelsetel tekstidel vastavalt 84,4% ja 83,6%. Problemaatilisem on ajaväljendite eraldamine, mille mõõtetulemused prantsusekeelsetel tekstidel on 53,9% (täpsus) ja 68,8% (saagis) ning inglisekeelsetel tekstidel vastavalt 50,0% ja 64,5%. Tulemusi analüüsides näeb autor ühe olulise probleemina vähest keeleteadmuse kaasamist: süsteem ei kasuta automaatseid keeletöötluse vahendeid (nt sõnaliikide märgendajat või lausepiiride määrajat).

D.Ahn, J. van Rantwijk ja M. de Rijke loovad TIMEX2 standardile toetuva märgendaja, mis kasutab masinõpitud klassifikaatoreid[7] ajaväljendite eraldamisel ning osaliselt ka semantika normaliseerimisel [12].

Toetudes automaatse süntaksianalüüsi tulemustele, käsitlevad autorid ajaväljendite eraldamist kui süntaktiliste fraaside binaarset klassifitseerimist (kas fraas on ajaväljend või mitte), mille puhul klassifitseerimisotsus langetatakse fraasi sisu ja lähikonteksti tunnuste põhjal. Fraasi sisu tunnustena (*features*) arvestatakse näiteks nädalapäeva nime või numbrilise aastaarvu esinemist fraasis, fraasi tüüpi ning fraasi peasõna koos sõnaliigiga; lähikonteksti tunnustena arvestatakse kahte fraasile eelnevat sõna. Samade tunnuste põhjal treenitakse ka klassifikaator, mille ülesandeks on määrata ajaväljendi liik (toetudes TIMEX2 normaliseerimiskujudele, eristavad autorid 6 liiki ajaväljendeid).

Kui ajaväljendi liik on määratud, rakendatakse käsitsi koostatud reegleid, et seada ajaväljendiga vastavusse selle esmane semantiline esituskuju ning viia läbi ankurdamine (kui

7 Autorid kasutavad klassifikaatoritena tugivektormasinaid (*Support Vector Machines*). Meetodit tutvustab näiteks http://www.statsoft.com/textbook/support-vector-machines/ (Viimati vaadatud: 01.03.2010)

tegu on relatiivse ajaväljendiga). Pärast ankurdamist kasutatakse masinõpitud klassifikaatorit, et määrata relatiivse ajaväljendi paiknemissuund referentsaja suhtes (minevikus, olevikus või tulevikus). Suuna üle otsustamisel võetakse aluseks samad tunnused, mis ajaväljendite eraldamiselgi; täiendavate tunnustena tuuakse sisse ka ajaväljendile lähima verbi grammatiline aeg ning ajaväljendi esmase semantilise esituskuju võrdlus dokumendi loomise kuupäevaga.[8]

Autorid treenivad süsteemi TERN 2004 inglisekeelsel treeningkorpusel ning mõõdavad TERN testkorpusel ajaväljendite eraldamise saagiseks 81,3% ja täpsuseks 92,9%. Semantika normaliseerimise parimaks tulemuseks saadakse saagis 88,7% ja täpsus 91,0%.

2.2.2 Pooljuhendatud masinõpe

O.Craveiro, J.Macedo ja H.Madeira katsetavad ajaväljendite eraldamisel portugali keeles tehnikat, mis näeb ette eraldamisreeglite automaatset „kaevandamist“ korpusest [13]. Meetod on kaheosaline: esimeses osas kaevandatakse korpusest ajaväljendifraasid ning luuakse nende põhjal ajaväljendimustrid (ajaväljendeid kirjeldavad regulaaravaldised), teises osas kasutatakse leitud mustreid ajaväljendite automaatsel eraldamisel ning klassifitseerimisel.

Ajaväljendite kaevandamisel kasutatakse „seemnetena“ leksikaalseid tunnuseid (kuunimed, aastaajanimed, nädalapäevanimed ning ajaühikute nimetused): korpusest kogutakse iga leksikaalse tunnuse kõige sagedamini esinevad lähikontekstid (konteksti ulatus on piiratud: *n* sõna tunnusest vasakul ja *n* sõna paremal) ning moodustatakse nende põhjal nimekiri fraasikandidaatidest. Loodud nimekirjast filtreeritakse stoppsõnade abil välja mitte-ajaväljendifraasid ning viiakse läbi mustrite agregeerimine, asendades ajalised võtmesõnad eritähistega (nt fraasist *in May* saadakse muster *in tag_MONTH*) ja ühildades üksikutel fraasipositsioonidel erinevusi sisaldavad fraasid (nt fraasidest *In the last year* ning *In the following year* saadakse muster *In the* (*last*|*following*) *year*). Viimases kaevandamise etapis sorteeritakse allesjäänud kandidaadid esinemissageduse järgi ning jäetakse alles fraasimustrid, mis ületavad süsteemile etteantud sageduslävendi.

Ajaväljendite märgendamist muudetakse efektiivsemaks kahe sammuga. Esiteks jäetakse sisendtekstist välja laused, mis ei sisalda ajalisi võtmesõnu (nt eemaldatakse lause *Lisbon is the capital of Portugal*, aga jääb alles lause *Today is a beautiful day*). Teiseks, allesjäänud lausetes asendatakse ajalised võtmesõnad eritähistega (nt *tag_DATE, tag_MONTH, tag_WEEK*), seega ei tule ajaväljendite leidmisel sobitada mitte kõiki

8 Näiteks võrreldakse ajaväljendis sisalduvat kuunime dokumendi loomise aja kuuga ning leitakse, kas aasta-pikkuse tsükli raamides on tegemist samade kuudega või eelneb/järgneb üks teisele.

korpusest leitud mustreid, vaid võib keskenduda lausega samasid eritähiseid sisaldavatele mustritele.

Süsteemi testimisel mõõtsid autorid ajaväljendite eraldamise saagiseks 64,10% ning täpsuseks 84,27%. Autorid võrdlesid tulemusi ühe reeglipõhise portugali keele ajaväljendeid eraldava süsteemiga ning leidsid, et nende süsteemi saagis oli ~13% võrra väiksem, samas täpsus ligi 9% võrra suurem.

2.3 Ajaväljendite eraldamine ja tuvastamine eesti keeles

Järgnevalt anname ülevaate senitehtud töödest ajaväljendite automaatsel eraldamisel ja tuvastamisel eesti keeles.

M.Treumuth vaatleb ajaväljendite tuvastamist rakendusespetsiifilisest vaatepunktist, käsitledes seda alamülesandena kasutajasisendi parsimisel dialoogisüsteemides [14]. Ta loob reeglipõhise tuvastaja, mis formaliseerib ajaväljendite semantika SQL-päringu kitsenduste (*constraints*) kujule. Selliselt formaliseeritud semantikat rakendab autor infopäringute tegemisel sündmuste andmebaasis.

Süsteemis kasutatakse automaatset morfoloogilist analüüsi, mis võimaldab ajaväljendite eraldamise üles ehitada sõnalemmade sobitamisele. Lisaks sõnalemmadele kirjeldavad eraldamisreeglid ka semantilist osa, täpsustades näiteks ajaväljendi granulaarsuse ning esitades poolformaalse semantikadefinitsiooni (nt väljendi *„emadepäev“* semantika esitatakse kujul `PÄEV=(08.05 kuni 14.05) JA PÄEV=pühapäev`). Reeglid on koostatud eesmärgiga eraldada võimalikult lühikesi ajaväljendifraase (valdavalt ühest sõnast koosnevaid ajaväljendeid, nt *„aprillis“*, *„nädalalõppudel“*), pikemate fraaside semantika pannakse kokku pärast eraldamist, arvestades ajaväljendite granulaarsuseid. Näiteks kui sisendist eraldatakse väljendid *„esmaspäeval“* ja *„kell 17“,* esitatakse nende semantika konjunktsioonina `PÄEV=esmaspäev JA KELL=17`, kuna eraldatud väljendid ei sisalda kattuvaid granulaarsusi.

E.Saue keskendub ajaväljendite eraldamise ülesandele ning kasutab reeglipõhist lähenemist ajaväljendite märgendaja loomisel [1]. Ta võtab ajaväljendite kirjeldamiseks kasutusele grammatika, mille alusel genereerib eraldamist teostavad regulaaravaldised. Autor mõõdab süsteemi täpsuseks ajakirjandustekstidel 92% ja ilukirjandustekstidel 83%, saagiseks vastavalt 92% ja 77%.

3 Eesti keele ajaväljendite tuvastaja

Käesolevas peatükis antakse ülevaade eesti keele jaoks loodud reeglipõhisest ajaväljendite tuvastajast. Valminud programm on autori bakalaureusetöös[15] loodud süsteemi (edaspidi: vana süsteem) edasiarendus.

Süsteem on realiseeritud programmeerimiskeeles Java (versiooniga J2SE 5.0 ühilduvalt). Valdav osa realisatsioonist toetub Java standardteegi vahenditele, erandiks on kalendriaritmeetika teostamine, milleks kasutatakse vabavaralise teegi Joda Time[9] abi. Süsteem kasutab oma töös ka morfoloogilist analüüsi. Morfoloogilise analüsaatori ja ühestajana on kasutusel programm `t3mesta` [16].

Järgnevas alampeatükis vaadeldakse meetodit, mis hõlbustab keeles sagedasti kasutatavate ajaväljendite kaardistamist ning pakub seega teatavat pidepunkti reeglite koostamisel. Seejärel tuuakse alampeatükis 3.2 välja kõige olulisemad uuendused, mis on süsteemi lisatud, ning alampeatükis 3.3 antakse ülevaade süsteemi ülesehitusest. Alampeatükis 3.4 kirjeldatakse detailsemalt süsteemi toimeloogikat ajaväljendite eraldamisel ning alampeatükis 3.5 vaadeldakse süsteemi tööd ajaväljendite normaliseerimisel.

3.1 Ajaväljendite automaatne kaevandamine korpusest

Ajaväljendite tuvastamise süsteemi ülesehitamisel on oluline saada ülevaade sellest, millised ajaväljendifraasid on loomulikus keeles kasutusel ning milline on nende kasutuskontekst. Käesolevas töös kasutame ajaväljendite otsimiseks Eesti kirjakeele korpuse alamkorpust – tasakaalustatud ajalehekorpust[10] aastate 1990-2001 väljaannetega, mis koondab endas ca 5 miljonit sõna.

Meid huvitas, millistes fraasides[11] (aga ka – milliste liitsõnade koosseisus ning millistes vormides) esinevad keeles enimkasutatavad aega väljendavad sõnad (näiteks *aasta*, *kuu, päev*). Sellise info leidmiseks tekstikorpusest võib kasutada näiteks UNIX käsureatööriista `grep`, mis võimaldab leida kõiki etteantud sõne või regulaaravaldise esinemisi korpusest. Tööriist `grep` tagastab meile küll kõik read, kus otsitud sõne esines, ent ei ütle midagi konkreetse sõnavormi esinemissageduse või selle erinevate kontekstide kohta. Näiteks

9 vt http://joda-time.sourceforge.net/ (05.12.2009)
10 Tasakaalus korpus: http://www.cl.ut.ee/korpused/grammatikakorpus/ (Viimati vaadatud: 09.02.2010)
11 Siin ja edaspidi mõeldakse *fraasi* all eeskätt ühest või mitmest sõnast koosnevat sõnade järjendit, esitamata sellele mingeid täpsemaid süntaktilisi nõudeid.

sõna *aasta* esinemisi otsides tagastab `grep` ~29000 rida, mille käsitsi läbivaatamine ning sõnavormide ja erinevate esinemiskontekstide loetlemine on töömahukas ettevõtmine.
Selle ajakuluka ettevõtmise kiiremaks teostamiseks kasutasime artiklis [13] kirjeldatule sarnast fraasikaevandamise tehnikat. Ajaväljendifraaside automaatseks kaevandamiseks koostasime programmi, mis leiab regulaaravaldise kujul etteantud sõna (nn. võtmesõna) kõik esinemised korpuses ning toob välja sagedasemad esinemiskontekstid sõnavormide kaupa. Algoritmi sisendiks on võtmesõna kirjeldav regulaaravaldis *reg*, maksimaalne sõnakonteksti raadius *r* ning fraasi esinemissageduse lävend *l*. Lühidalt kirjeldades on algoritmi tegevusloogika järgmine:

1. Leitakse korpusest kõik erinevad sõnavormid (sh ka liitsõnad), mis vastavad regulaaravaldisele *reg*. Leitud sõnavormidest valitakse välja need, mille esinemissagedus on suurem-võrdne kui lävend *l*, ning jäädvustatakse koos esinemissagedusega.
2. Iga punktis 1. jäädvustatud sõnavormi (võtmesõna) korral: leitakse korpusest kõik fraasid (sõnajärjendid), mis sisaldavad antud võtmesõna ning mille pikkus (sõnades) on väiksem-võrdne kui 2**r* + 1. Jäädvustatakse leitutest fraasid, mille esinemissagedus on suurem-võrdne kui *l*.

   Fraaside leidmisel ja jäädvustamisel asendatakse teatud liiki alamsõned üldistavate tähistega:

   - Number asendatakse tähisega [d]
   - Nädalapäevanimi asendatakse tähisega NADALAPAEV
   - Kuunimi asendatakse tähisega KUU
   - Arvsõnad (*üks, kaks*, ... , *kümme*)[12] asendatakse tähisega ARV
   - Järgarvsõnad (*esimene, teine*, ... , *kümne*) asendatakse tähisega JARG_ARV

   Lisaks kasutatakse fraasiulatuse piiramisel ka stoppsõnade filtrit; näiteks katkestatakse fraasi laiendamine, kui järgmine lisatav sõna on üks järgnevaist: *on, oli, olnud*, *kuid, kui, siis, et, ta, mil, millal*.
3. Tulemuste väljastamisel:
   - Grupeeritakse leitud sagedased fraasid sõnavormide kaupa.
   - Sorteeritakse fraasid grupisiseselt vastavalt võtmesõna paiknemisjärjekorrale fraasis: kõige esimesena tulevad fraasid, kus võtmesõna on lõpus ning

12 Arvestatakse ka võimalikke tüvemuutuseid, nt tähisega asendatakse nii sõna *„üks“* kui ka sõna *„ühe“*

viimasena tulevad fraasid, mille alguses on võtmesõna.

Lisa 1 toob näite programmi töö tulemustest: otsitud on võtmesõna *kevad* sisaldavaid sagedasi fraase. Lisaks aastaaegade nimedele kasutasime sagedaste fraaside leidmisel võtmesõnadena veel ajaühikute nimesid (*minut, tund, päev, nädal, kuu, aasta*), kalendrikuude nimesid, nädalapäevade nimesid, päevaosade nimetuseid (*hommik, lõuna, õhtu, öö*), deiktilisi ajaväljendeid (*eile, täna, homme, mullu, tänavu*) ning numbrilisi kellaaja- ja aastaarvumustreid. Lisa 5 (materjalide CD) sisaldab fraaside kaevandamise programmi ning tulemusi.

Käesolevas töös oli fraaside kaevandamise programmi peamine eesmärk kaardistada keeles sagedasti kasutatavaid ajaväljendifraase, et leitud fraaside põhjal saaks koostada reeglipõhisele ajaväljendite tuvastajale uusi tuvastamismustreid. Kuigi potentsiaalselt võib sellest programmist välja arendada ka automaatse ajaväljendite eraldamise meetodi (nagu on tehtud töös [13]), on problemaatiline just ajaväljendite normaliseerimise komponendi lisamine meetodi koosseisu ning seetõttu käesolevas töös meetodit edasi ei arendata.

3.2 Uuendused võrreldes vana süsteemiga

Järgnevalt tuuakse välja kõige olulisemad uuendused võrreldes vana süsteemiga:

1. Süsteemi poolt kasutatavat ajaväljendite märgendamise formaati on muudetud. Kui vana süsteemi märgendusformaat lubas ainult ajapunktide ja ajavahemike semantika väljatoomist, siis uus märgendusformaat võimaldab kirjeldada ka ajalisi kestvuseid ning korduvusi. Mitmete konventsionaalsete väljendite (aastaajad, päevaosad) ning semantika täpsustuste esitamine on uues formaadis paindlikum.
2. Ajaväljendite eraldamisloogika kirjeldamisel saab uues süsteemis toetuda taaskasutatavatele mustriosadele (nn sõnaklassidele). Lisaks sellele võivad mustriosad uues süsteemis olla valikulised ning ärajäetavad ja vale-eralduste vähendamiseks on kasutusele võetud nn negatiivsed mustrid.
3. Osa pikemate ajaväljendifraaside moodustamise loogikast on uues süsteemis viidud heuristiliselt aluselt (liitmine granulaarsusi arvestades) reeglipõhistele alustele: kasutusele on võetud liitumisreeglid. Samuti on uus sisseehitatud heuristikute kasutamine ajavahemike moodustamisel.
4. Ajaväljendi semantika defineerimist saab uues süsteemis teostada paindlikumalt:

semantikakirjeldus võib oleneda ajaväljendifraasi morfoloogilistest tunnustest ning arvestada ajaväljendi lähikonteksti. Ajaväljendi semantika kirjelduses on võimalik kasutada ankurdamist, st võtta semantika lahendamisel aluseks mõne eelneva ajaväljendi semantika lahendus.

5. Semantika lahendamisel kasutatav kalendrimudel on välja vahetatud: Java standardteegi kalendriaritmeetika vahendite asemel kasutatakse paindlikumat teeki Joda Time. Kuigi kalendriväljade muutmise põhioperatsioonid on jäänud suures osas samaks, on lisatud uusi võimalusi semantika lahendamiseks (nt lähima verbi grammatilise aja arvestamine semantika lahendamisel) ning märgendusformaadispetsiifilisi operatsioone (märgenduse atribuutide muutmine).
6. Reeglite kirjeldamise formaati on muudetud.

Eelmainitud uuendusi (sõnaklassid, valikulised mustriosad jms) kirjeldatakse detailsemalt järgnevates alampeatükkides.

3.3 Süsteemi ülesehitus

Käesolevas töös on ajaväljendeid tuvastav süsteem üles ehitatud kahte liiki reeglitele. Primaarsed on **tuvastamisreeglid**, mis kirjeldavad ühelt poolt ajaväljenditele vastavaid fraase tekstis nn **fraasimustrite** abil ning teiselt poolt annavad edasi operatsioonide jada, mis tuleb ajaväljendi semantika leidmiseks läbi viia. Sekundaarsed on **liitumisreeglid**, mis täpsustavad, kuidas kõrvutipaiknevad ajaväljendid ühendatakse.[13]

Süsteemi tegevuse ajaväljendite tuvastamisel võib üldises plaanis jagada järgmisteks alametappideks:

1. *Eeltöötlus,* mille käigus viiakse läbi sisendteksti morfoloogiline analüüs ning tükeldamine sõnadeks. Samaaegselt sõnadeks tükeldamisega teostatakse ka mitmed heuristilised tekstianalüüsi sammud, näiteks tuvastatakse arvsõnafraasid (st leitakse, millised sõnad kuuluvad arvsõnafraaside koosseisu ja nende semantika) ning tehakse kindlaks verbide grammatilised ajad.
2. *Ajaväljendikandidaatide eraldamine tekstist,* mille käigus leitakse üles kõik fraasimustritele vastavad sõnajärjendid ning eraldatakse need kui potentsiaalsed ajaväljendikandidaadid. Iga ajaväljendikandidaadiga seotakse ka jada arvutus-

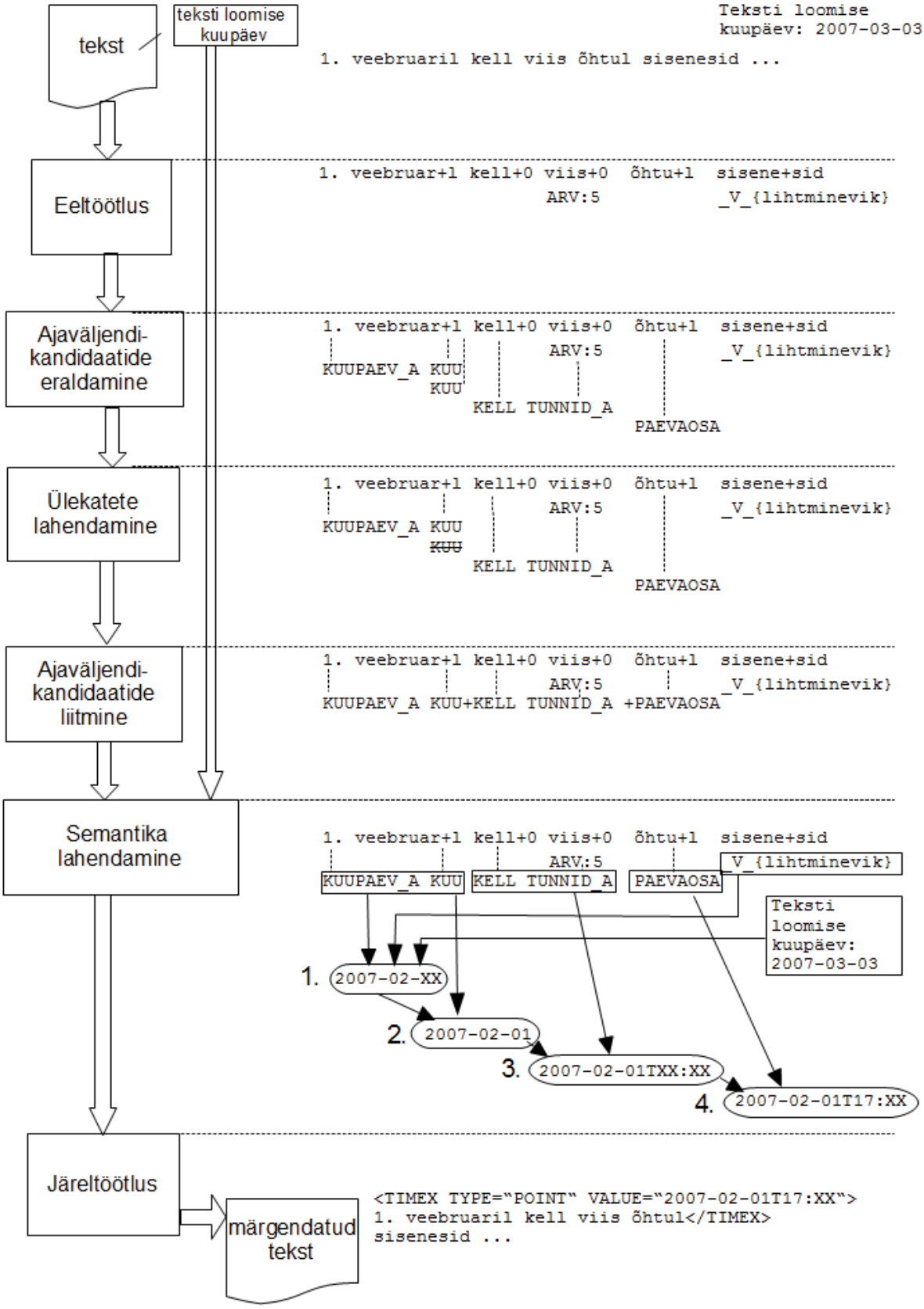
13 Potentsiaalselt saab iga ajaväljendi anda edasi tuvastamisreegli abil. Ainult sellist kirjeldamisviisi kasutades loodaks aga suur hulk reegleid, mille poolt kirjeldatud ajaväljendid erineksid üksteisest vaid sõnajärje poolest. Sõnajärje-probleemi lahendusena kirjeldavad tuvastamisreeglid enamasti kinnistunud sõnajärjega fraasiosi, liitumisreeglid aga näitavad, kuidas tuvastamisreeglite poolt eraldatud ajaväljendid võivad liituda pikemateks fraasideks.

operatsioone (nn semantikareegleid), mille rakendamine peaks viima kandidaadi semantika lahenduseni.

3. *Ülekatete lahendamine*, mille käigus kustutatakse teiste ajaväljendikandidaatide poolt täielikult ülekaetud kandidaadid. Nt kuupäevana eraldatud ajaväljendikandidaadiga „*30. jaanuar*“ paralleelselt eraldatakse ka ainult kuunimest koosnev kandidaat „*jaanuar*“. Ülekatete lahendamisel kustutatakse just viimane, kuna see on täielikult üle kaetud pikema kandidaadi poolt.
4. *Ajaväljendikandidaatide liitmine*, mille käigus moodustatakse tekstis kõrvutipaiknevatest ajaväljendikandidaatidest uusi ajaväljendifraase. Uute fraaside moodustamisel juhindutakse liitumisreeglitest; erandiks on teatud kujul ajavahemike moodustamine, mida kontrollivad reeglitevälised heuristikud.
5. *Ajaväljendite semantika lahendamine*, mis seisneb iga ajaväljendikandidaadiga seotud arvutusoperatsioonide järjestikuses rakendamises. Kui operatsioonide tulemusena jõutakse normaliseeritud semantikani (nt konkreetse kuupäevani), leiab ajaväljend ka lõppväljundis märgendamist; vastasel juhul kandidaat hüljatakse.
6. *Järeltöötlus*, mille käigus teostatakse morfoloogiliselt analüüsitud teksti tagasijoondamine esialgse sisendtekstiga[14] ning seejärel märgendatakse tekstis leitud ja lahendatud ajaväljendikandidaadid.

Joonis 1 kujutab ajaväljendite tuvastamisel läbitavaid etappe (vasakul) ning toob lihtsustatud näite tuvastamisprotsessi läbimise kohta (paremal). Näites on igal etapil toodud analüüsi tulemused kihtide kaupa: esimene kiht on morfoloogilise analüüsi tulemus, teine kiht eeltöötluse tulemus ning alates kolmandast kihist märgib iga kiht eraldiseisvat ajaväljendikandidaati (`KUUPAEV_A KUU`, `KELL TUNNID_A` jt). Kuigi näites ei ole kujutatud ajaväljendikandidaatide semantilist osa, on semantika lahendamise etapis toodud välja lahendamise vahetulemused (ümardatud nurkadega kastides) ning märgitud vahetulemusteni jõudmise järjekord (kastidele eelnevad numbrid).

14 morfoloogilise analüüsi käigus lähevad tekstist kaduma tühikud ja reavahetused; märgendite korrektseks paigutamiseks joondatakse morfoloogilise analüüsi väljund esialgse sisendtekstiga.



Joonis 1. Ajaväljendite tuvastamisel läbitavad sammud. Vasakul on kujutatud protsessi alametappe, paremal on toodud lihtsustatud näide protsessi läbimise kohta.

3.4 Ajaväljendite eraldamine

3.4.1 Eeltöötlus

Eeltöötluse etapi eesmärgiks on valmistada tekst ette järgnevaks ajaväljendite eraldamiseks ning semantika normaliseerimiseks. Eeltöötluse esimeses pooles viiakse läbi teksti morfoloogiline analüüs ja ühestamine; teises pooles tükeldatakse tekst sõnadeks ning kasutatakse mitmesuguseid heuristikuid, et leida järgmistel analüüsietappidel olulised tunnused (nt lauselõpud ning arvsõnafraaside asukohad ja semantika).

Morfoloogiline analüüs ja ühestamine teostatakse programmi `t3mesta` abil. Morfoloogilise analüüsi käigus leitakse iga sõnavormi puhul selle sõna algvorm, struktuur (formatiivid) ja morfoloogiline informatsioon (nt sõnaliik, kääne või pööre, arv jms). Kuna morfoloogiline analüüs on sageli mitmene (nt sõna „*kuus“* võib tähistada nii arvsõna nimetavas kui ka nimisõna seesütlevas), rakendatakse selle lõpuleviimiseks ühestamist, st valitakse välja üks antud kontekstis sobivaim analüüs.

Morfoloogilise analüsaatori väljundi põhjal toimub teksti tükeldamine sõnadeks. Tükeldamisel pööratakse täiendavat tähelepanu ainult numbritest ja (mittetäht-)sümbolitest koosnevatele üksustele (nt „*29.10.2009*“) – need tükeldatakse sümboli ja sellele järgneva numbri vahelistelt positsioonidelt (nt „*29.10.2009*“ jagatakse üksusteks „*29.*“, „*10.*“ ja „*2009*“). Selline tükeldamine hõlbustab hilisemat töötlust, võimaldades neid tükke käsitleda eraldiseisvate üksustena.

Üheaegselt teksti tükeldamisega toimub viis tegevust: 1) tuvastatakse arvsõnafraasid, 2) leitakse potentsiaalsed lauselõpud, 3) leitakse potentsiaalsed ajavahemike piirid, 4) määratakse verbide grammatilised ajad ning 5) kapseldatakse iga sõna koos selle morfoloogiliste tunnuste (algvorm, sõnalõpp, sõnaliik, vormi nimetused) ning punktides 1) – 4) leitud tunnustega klassi `AjavtSona`.

Arvsõnafraaside tuvastamise käigus leitakse tekstis üles arvulist semantikat väljendavad fraasid[15] ning seotakse nendega nende „tähendus“ täisarvu või murdarvu kujul. Süsteem peaks suutma leida ja normaliseerida fraase, mille semantika jääb täisarvulisse vahemikku

15 Arvsõnafraasina käsitletakse ühest või mitmest järjestikusest põhiarvsõnast või järgarvsõnast koosnevat fraasi (nt *viiesaja kahekümne neljas*). Numbritega edasi antud arvud siia alla ei kuulu.

0-999999[16]. Murdarvudest leitakse ja normaliseeritakse ainult sõnad *pool, veerand*, *kolm-veerand* ja *poolteist*.

Potentsiaalsete lauselõppude kindlaks tegemisel järgitakse lihtsat heuristikut: kui sõna lõpus on lauselõpumärk (punkt, hüüumärk või küsimärk) ning järgmine sõna algab suure tähega, eeldatakse, et on tegemist lauselõpuga. Potentsiaalsete ajavahemiku algustena käsitletakse seestütlevas käändes sõnu ning sidekriipsuga lõppevaid numbreid (nt „*15*-“). Potentsiaalsete ajavahemiku lõpp-punktidena käsitletakse rajavas käändes sõnu.

Verbi grammatilise aja määramisel püütakse iga verbiga siduda üks viiest morfoloogiliselt avalduvast ajakategooriast: olevik, lihtminevik, täisminevik, enneminevik ja üldminevik. Olevik ja lihtminevik määratakse kindlaks vastavalt morfoloogilise analüsaatori väljundi selgitusele.[17] Täismineviku, ennemineviku ja üldmineviku moodustamisel lähtutakse „Eesti keele käsiraamatus“[17] toodud vastavate ajakategooriate kirjeldustest. Eraldiseisva ajalise tähisega seotakse veel kesksõna minevikuvorm, mis ei kuulu täismineviku või enne-mineviku alla.

Eeltöötluse etapi lõpuks moodustatakse sisendtekstist sõnade nimestik (täpsemalt – nimestik klassi `AjavtSona` isenditest), millest rakenduse töö järgmises etapis hakatakse eraldama ajaväljendeid.

3.4.2 Fraasimustrid ja sõnamallid

Ajaväljendite eraldamiseks tekstist kasutatakse tuvastamisreeglite alla kuuluvaid *fraasi-mustreid*, mille toimeloogika sarnaneb lõpliku automaadi toimeloogikale.

Fraasimuster sisaldab üldistavat fraasikirjeldust ning säilitab sobitamise olekut. Ajaväl-jendite eraldamisel antakse fraasimustrile ükshaaval sisendteksti sõnu ette ning iga kord, kui järjestikku etteantud sõnad vastavad mingile mustri poolt kirjeldatud fraasile (terves fraasi ulatuses), teostatakse uue ajaväljendikandidaadi eraldamine.

Fraasimuster kirjeldab fraase *sõnamallide* abil. Sõnamalliks loetakse sõna kirjakuju või morfoloogiliste omaduste kirjeldust. Loodud rakendus lubab fraaside kirjeldamisel kasu-tada järgmisi sõnamalle:

- Tavatekstiga määratud sõnamall – kirjeldab mallile vastavat üksiksõna, tuues välja selle (tõstutundetu) kirjakuju. Tähistatakse edaspidi ka `'sõna'`.

16 Erand: süsteem ei toeta sõnaga *tuhat* kokkukirjutatavate arvsõnade tuvastamist (nt *kahetuhandes*, *neljasaja-tuhandene*), lahkukirjutatud variandid leitakse üles (nt *kaks tuhat, nelisada tuhat*).

17 Morfoloogilise analüsaatori väljundi selgitus: http://www.filosoft.ee/html_morf_et/morfoutinfo.html (Viimati vaadatud: 24.04.2010)

- Regulaaravaldisega määratud sõnamall – kirjeldab mallile vastavaid sõnu regulaaravaldise[18] abil. Tähistatakse edaspidi ka `/regulaaravaldis/`.

  Näide 2. Sõnamallile `/(sündinud|s|snd|sünd)/` vastavad sõna *sündinud* ning lühendid *s*, *snd* ja *sünd*.

- Algvormiga määratud sõnamall – kirjeldab mallile vastavaid sõnu algvormi kaudu. Seda liiki sõnamallis saab täpsustada ka sõnaliigi, aga see ei ole kohustuslik.

  Tähistatakse `|algvorm|` ja `|algvorm(sõnaliik)|`.

  Näide 3. Sõnamallile `|esmaspäev|` vastavad sõna *esmaspäev* erinevates käänetes vormid: *esmaspäeval, esmaspäevaks, esmaspäevani* jms.

Tavatekstiga sõnamalle kasutatakse juhtudel, kui ei ole tarvis morfoloogiat arvestada ning sõna kirjeldamiseks piisab konkreetse sõnavormi kirjakuju edasiandmisest. Regulaaravaldisega määratud sõnamallid ei kasuta samuti morfoloogiat ning on mõeldud numbrimustrite või vähese arvu erinevate sõnavariantide edasiandmiseks. Algvormiga määratud sõnamalle kasutatakse juhtudel, kui sõna kirjeldamisel tahetakse arvestada kõiki võimalikke sõnavorme.

Lisaks tavateksti, regulaaravaldise ning algvormiga määratud sõnamallidele loetakse mõneti erandlikult sõnamallide alla kuuluvaiks ka arvsõnafraaside mallid ja sõnaklassid.

- Arvsõnafraasi mall – kirjeldab arvsõna või arvsõnadest koosnevat fraasi, mis annab edasi ühe täisarvu või murdarvu semantikat. Arvsõnafraasi defineerimisel on võimalik täiendava kitsendusena määrata arvu liik (põhiarv `_N_`, järgarv `_O_` või murdarv `_F_`) ning täisarvulised piirid arvu semantikale. Arvsõnafraaside tähistamiseks kasutatakse edaspidi ka kuju `arvsona{liik}{lõiguAlgus-lõiguLõpp}`, kus `liik`, `lõiguAlgus` ja `lõiguLõpp` on mittekohustuslikud kitsendused.

  Näide 4. Sõnamallile `arvsona{_O_}{1-31}` vastavad arvsõnafraasid *kolmandal*, *viieteistkümnes*, *kahekümne viiendal, kolmekümne esimene* jms.

- Sõnaklass – kirjeldab mallile vastavate sõnade hulka, kasutades teisi sõnamalle alamosadena (elementidena).[19] Sõna loetakse sõnaklassiga sobitunuks, kui see rahuldab ühte sõnaklassi alla kuuluvat sõnamalli. Sõnaklasside defineerimisel ei ole lubatud rekursiivsus, st üks sõnaklass ei või sisaldada teist sõnaklassi elemendina.

18 Kasutatakse Java standardteegi regulaaravaldiste formaati. Vt täpsemalt http://java.sun.com/j2se/1.5.0/docs/api/java/util/regex/Pattern.html (16.03.2010)

19 Sõnaklassi-kontseptsiooni loomisel on võetud eeskuju A.Berglundi ajaväljendeid ja sündmuseid tuvastavast süsteemist [18], kus kasutatakse sarnast kontseptsiooni fraasi ja liitsõna alamosade kirjeldamisel.

Edaspidi kasutatakse sõnaklassi nime tähistamiseks läbivate suurtähtedega kirjakuju.

Näide 5. Kasutades algvormiga määratud sõnamalle `|esmaspäev|`, `|teisipäev|`, `|kolmapäev|`, `|neljapäev|`, `|reede|`, `|laupäev|` ja `|pühapäev|`, võib defineerida sõnaklassi NADALAPAEV, mida rahuldavad kõik nädalapäevade nimetused erinevates käänetes.

Tekstiüksuste sobitamisel sõnamallidega pööratakse täiendavat tähelepanu sõna ümbritsevatele kirjavahemärkidele. Kui sobitatav tekstiüksus on sõnaline ning selle alguses või lõpus on kirjavahemärk, eemaldatakse märk enne sõnamalliga sobitamist üksuse küljest. Numbriliste üksuste puhul eemaldatakse kirjavahemärgid ainult algusest ning jäetakse lõppu alles.

3.4.3 Ajaväljendite eraldamise algoritm

Ajaväljendite eraldamise üldist algoritmi kirjeldab Pseudokood 1. Fraasimustrite sobitamise põhikeerukus on koondatud meetodisse `kontrolliMustrileVastavust` (Pseudokood 1, rida 16), kus kontrollitakse, kas etteantud sõna lubab järjega mustris edasi liikuda või tuleb mustri sobitamist algusest peale alustada. Kontrollimise käigus jäetakse jooksvalt meelde positiivselt sobitatud sõnad (nn *rahuldatud sõnamallid*) ning kui sobitamine on õnnestunud ühe mustri poolt kirjeldatava fraasi terves ulatuses, eraldatakse meetodis uus ajaväljendikandidaat ning seotakse fraasi kuuluvate sõnadega.[20]

Pseudokood 1. Ajaväljendifraaside eraldamise üldine algoritm. Sümbolipaar // märgib kommentaari algust.

```
1  //
2  // Sisend:
3  //  reeglid – tuvastamisreeglite nimestik
4  //  sonad – sisendtekst kui sõnade nimestik
5  //  klassid – paisktabel, kuhu pannakse paarid:
6  //          (kontrollitud sõnaklass, kontrollimise tulemus)
7  //
8  for (i = 0; i < sonad.size(); i++){
9     AjavtSona sona = sonad.get(i); // i. sõna tekstis
10    klassid.tyhjenda(); // eemaldame kõik võtmed ja väärtused
```

20 Tehniliselt: iga sõna juures hoitakse viitasid ajaväljendikandidaatidele, mille alla sõna kuulub, ning samuti sisaldab iga ajaväljendikandidaat viitasid fraasi koosseisu kuuluvatele sõnadele.

```
11    for (j = 0; j < reeglid.size(); j++) {
12       Reegel reegel = reeglid.get(j); // j. reegel
13       FraasiMuster fraasimuster = reegel.getFraasiMuster();
14       // Kontrollime, kas tekstis i-ndal positsioonile olev
15       // sõna rahuldab j-inda reegli fraasimustrit.
16       (fraasimuster).kontrolliMustrileVastavust(sona, klassid);
17       // Kui oleme jõudnud teksti lõppu, sulgeme poolelioleva
18       // eraldamise
19       if (i == sonad.size() - 1){
20          (fraasimuster).sulgePooleliOlevFraas();
21       } // if-lõpp
22    } // for-lõpp
23 } // for-lõpp
```

Sõnaklasside kasutamine võimaldab fraasimustri alamosade taaskasutust, st mitu erinevat fraasimustrit võivad kasutada oma definitsioonis ühte ja sama sõnaklassi. Samuti võimaldab sõnaklasside kasutamine muuta sobitamist efektiivsemaks: ühe sisendsõna (Pseudokood 1: `sona`) vastavust mingile sõnaklassile kontrollitakse ainult üks kord ning jäetakse tulemus meelde (salvestatakse paisktabelisse `klassid`) – teised fraasimustrid, mis on järjega sama sõnaklassi peal, ei pea kontrollimist enam uuesti läbi viima, vaid saavad tugineda esimese kontrolli tulemusele.

Meetodi `sulgePooleliOlevFraas()` eesmärgiks on lõpetada pooleliolev eraldamine, kui sobitamisega on jõutud teksti lõppu.

3.4.4 Valikulised sõnamallid ja väljajäetavad sõnad

Fraasimustri defineerimisel võib märkida osad sõnamallid *valikulisteks*, st lubada nende vahelejätmist mustriga sobitamisel, ning osad sõnamallid *väljajäetavateks*, st nõuda nendega sobitunud sõnade eemaldamist eraldatud ajaväljendifraasi koosseisust. Järgnevalt toome näited valikuliste sõnamallide ja väljajäetavate sõnade kasutamisest fraasimustri definitsioonis.

Näide 6 toob ühe valikulisi sõnamalle kasutava fraasimustri definitsiooni. Suurtähtedega mustriosad märgivad sõnaklasse ning küsimärk mustriosa lõpus tähistab valikulisust. Toodud fraasimustrile peaksid vastama näiteks fraasid *aasta aega tagasi, kümne kuu eest, 30 ööpäeva tagune*.

Näide 6.Valikulisi sõnamalle kasutava fraasimustri XML-definitsioon.

```
<Muster>
    ARV_LOENDA_VAIKE_A?  YHIK_2  AEGA?  TAGASI_EEST_TAGUNE
</Muster>
```

Näide 7. Sõnaklasside XML-definitsioonid: `ARV_LOENDA_VAIKE_A`, `YHIK_2`, `AEGA` ja `TAGASI_EEST_TAGUNE`. Toodud on ainult ajaväljendite eraldamist puudutav osa, välja on jäetud sõnaklasside semantilise osa kirjeldus.

```
<SonaKlass nimi="ARV_LOENDA_VAIKE_A">
    <Element tyyp="reg"
             vaartus="([0-5]?[0-9]?[0-9])...[21]" />
    <Element tyyp="arvSona"
             arvuPiirang="0-599" arvuLiik="_N_|_F_" />
</SonaKlass>
<SonaKlass nimi="YHIK_2">
    <Element tyyp="algv" vaartus="tund|h"      />
    <Element tyyp="algv" vaartus="minut|min"   />
    <Element tyyp="algv" vaartus="ööpäev|päev" />
    <Element tyyp="algv" vaartus="nädal"       />
    <Element tyyp="algv" vaartus="kuu"         />
    <Element tyyp="algv" vaartus="aasta|a"     />
</SonaKlass>
<SonaKlass nimi="AEGA">
    <Element tyyp="tekst" vaartus="aega" />
</SonaKlass>
<SonaKlass nimi="TAGASI_EEST_TAGUNE">
    <Element tyyp="algv" vaartus="tagasi" />
    <Element tyyp="algv" vaartus="eest"   />
    <Element tyyp="algv" vaartus="tagune" />
</SonaKlass>
```

Näide 7 toob fraasimustris (Näide 6) kasutatud sõnamallide definitsioonid. Märgend `Element` defineerib sõnaklassi alla kuuluva sõnamalli: atribuut `tyyp` määrab sõnamalli liigi (`reg` - regulaaravaldis, `algv` – algvorm, `arvSona` – arvsõnafraas ja `tekst` - tava-tekst) ning `vaartus` toob sõnamalli sisu (tekst, regulaaravaldis või sõna algvorm(id)). Algvorm-sõnamallides saab tuua mitu sobivat algvormikandidaati, kasutades püstkriipsu

21 Lühendatud, täispikkuses väärtus on:
`([0-5]?[0-9]?[0-9])-?(n)?(da)?(t|sse|le|ks|ta|s|l|ni|ga|st|lt|na)?`

eraldajana (nt `ööpäev|päev`). Arvsõnafraasi mallis kasutatakse täiendavaid atribuute: `arvuPiirang` piiritleb lubatud arvuliste väärtuste lõigu ning `arvuLiik` täpsustab lubatud arvsõnaliigid.

Näide 8 toob ühe väljajäetavat sõna kasutava fraasimustri definitsiooni. Toodud fraasimustris kasutatakse ainult regulaaravaldistega määratud sõnamalle ning hüüumärk esimese sõnamalli ees tähistab malliga sobituva sõna väljajätmist ajaväljendifraasist. Toodud fraasimustrile vastavad näiteks fraasid *sündinud 1983, sünd 1993* ning *s. 2004* (eeltoodud näidetes märgib allajoonitud sõna tegelikult eraldatud ajaväljendikandidaati).

Näide 8.Väljajäetavat sõna kasutava fraasimustri XML-definitsioon.

```
<Muster>
   !/(sündinud|s|snd|sünd)/   /([1-2][0-9][0-9][0-9])\)?/
</Muster>
```

3.4.5 Negatiivsed mustrid

Ajaväljendi defineerimisel ei piisa alati positiivsete juhtude võimalikult täpsest kirjeldamisest sõnamallide ja fraasimustri abil. Näiteks ei ole sellisel viisil võimalik täielikult välistada tervituste (*tere hommikust / õhtust*) ja kõnekäändude (nt *nagu öö ja päev*) seest ajaväljendite eraldamist. Eesmärgiga vähendada valesid eraldusi, lubatakse tuvastamisreegli all lisaks fraasimustrile (mis on nn positiivne muster) defineerida ka *negatiivsed mustrid*, mis kirjeldavad positiivset mustrit rahuldavaid mitte-ajaväljendeid. Kui tekstis ettetulev fraas on fraasimustriga edukalt sobitatud, kontrollitakse enne ajaväljendikandidaadi eraldamist ka negatiivsete mustrite sobitumist antud fraasiga. Ajaväljendikandidaat jäetakse eraldamata, kui see sobitub vähemalt ühe negatiivse mustriga.

Negatiivsed mustrid defineeritakse sarnaselt positiivsetele mustritele, kirjeldades sõna-sõnahaaval fraasi alamosi. Erinevalt fraasimustritest kasutatakse negatiivsete mustrite defineerimisel ainult regulaaravaldisi.

Näide 9 kirjeldab negatiivseid mustreid, mille eesmärgiks on vältida isikunimede (*August, Juuli, Mai*) eraldamist kuunimedena (*august*, *juuli, mai*). Toodud mustrites tehakse heuristiline eeldus, et kui suure algustähega „kuunimele“ (*August, Juuli, Mai*) eelneb või järgneb suurtähega algav sõna, on tõenäoliselt tegemist nime, mitte ajaväljendiga. Negatiivse mustri atribuut `startPos` määrab, milline on negatiivse mustri alguspositsioon fraasimustri poolt eraldatud fraasi alguspositsiooni suhtes, nt 0 märgib seda, et algused on

kohakuti ja -1 märgib seda, et negatiivne muster algab üks sõna enne positiivse mustri poolt eraldatud fraasi.

Näide 9. Negatiivsed mustrid, mis peaksid vältima isikunimede eraldamist kuunimedena. Toodud on mustrite XML-definitsioonid.

```
<NegMuster startPos="0">
      /(Mai|Juuli|August)\w*/   /\p{Lu}\w+/
</NegMuster>
<NegMuster startPos="-1">
      /\p{Lu}\w+/    /(Mai|Juuli|August)\w*/
</NegMuster>
```

3.4.6 Ajaväljendikandidaatide liitmine

Ajaväljendimustrite (fraasimustrite) koostamisel on püütud jälgida põhimõtet, et mustrid oleksid võimalikult lühikesed ning pikemad ajaväljendifraasid moodustatakse tekstis kõrvutipaiknevate fraaside liitmisel. Ajaväljendite liitmisel eristatakse kahte alametappi:

1. Ajaväljendikandidaatide ühendamine fraasideks liitumisreeglite alusel ning eraldiseisvana ajalist tähendust mitteomavate kandidaatide eemaldamine;
2. Ajaväljendikandidaatide ühendamine ajavahemikeks eeltöötluse käigus leitud tunnuste alusel;

Liitumisreeglid võimaldavad määrata, millised fraasimustrite poolt eraldatud ajaväljendikandidaadid võivad omavahel liituda ning kas liitumisel on oluline, et kandidaadid oleksid kindlas järjekorras. Liitumisreeglite ja fraasimustrite sidumiseks kasutatakse *mustritähiseid.* Tuvastamisreegli all seotakse fraasimustri mingi alamosaga mustritähis (läbivate suurtähtedega sõne), mis läheb kaasa ka eraldatud ajaväljendikandidaadile. Liitumisreeglites kasutatakse samu mustritähiseid, et kirjeldada, millised kõrvutiseisvad ajaväljendikandidaadid võib ühendada fraasiks.

Tuvastamisreeglid võivad ka kirjeldada fraase, mis ei oma eraldiseisvana ajalist tähendust, kuid vahetult eelneva või järgneva ajaväljendikandidaadi külge liidetuna muudavad kandidaadi tähendust (nt sõnad *orienteeruvalt, umbes, peaaegu, alguses, lõpus* ja fraasid *esimesel poolel, teiseks pooleks*). Mustritähisega sidumisel saab fraasi märkida *mitteeraldiseisvaks*, mis tähendab, et kui liitumisreeglite rakendamisel ei õnnestu fraasi teiste (nn *eraldiseisvate*) ajaväljendikandidaatidega liita, eemaldatakse antud fraas ajaväljendikandidaatide hulgast.

Näide 10 toob ühe liitumisreegli XML-definitsiooni. Toodud reegel lubab kahe ajapunkti kirjeldava fraasi liitmist: *aasta*-granulaarsusega väljendile lubatakse liita aastaaja nime sisaldav väljend, seadmata mingeid piiranguid fraaside järjekorrale. Mustritähise `AASTA_TAIS_AP` alla kuuluvad nii konkreetset aastaarvu edasiandvad ajaväljendifraasid (nt *aastal 2004, 2007*), kui ka kontekstist sõltuva semantikaga *aasta*-fraasid (nt *mullu, tänavu, eeloleval aastal, sama aasta*). Mustritähise `AASTAAEG_POOL_AP` alla kuuluvad eestäienditeta aastaajanime sisaldavad fraasid (nt *varakevadel, sügisel, talve jooksul*).[22] Kuna liidetavate alamfraaside järjekorrale piirangut ei ole seatud, lubab antud reegel fraasiks liita nii sõnajärjendi *kevadel, 1999* kui ka fraasijärjendi *aasta 2004, hilissügisel*.

Näide 10. Ühe liitumisreegli XML-definitsioon.

```
<LiitumisReegel tase="FRAAS">
       AASTA_TAIS_AP   AASTAAEG_POOL_AP
</LiitumisReegel>
```

Vältimaks üleliigsete liitmiste läbiviimist, lubatakse fraasi (vasakult) esimese liikme mustritähist liitumisreegli rakendamisel ühe fraasi piires kasutada ainult üks kord. Näiteks, kui eeltoodud liitumisreegli (Näide 10) abil liidetakse fraasiks *aasta 2004 + kevadest*, ei ole võimalik sama liitumisreeglit rakendada, et liita fraasi paremasse otsa väljend *hilissügiseni* (st viia läbi liitmine (*aasta 2004 + kevadest*) + *hilissügiseni*). Sellised liitmised hoitakse ära, kuna ajavahemike moodustamine on süsteemi töös eraldiseisev alametapp.

Ajavahemike moodustamine tugineb liitumisreeglite rakendamise tulemustele ning toimib järgnevate heuristikute alusel:

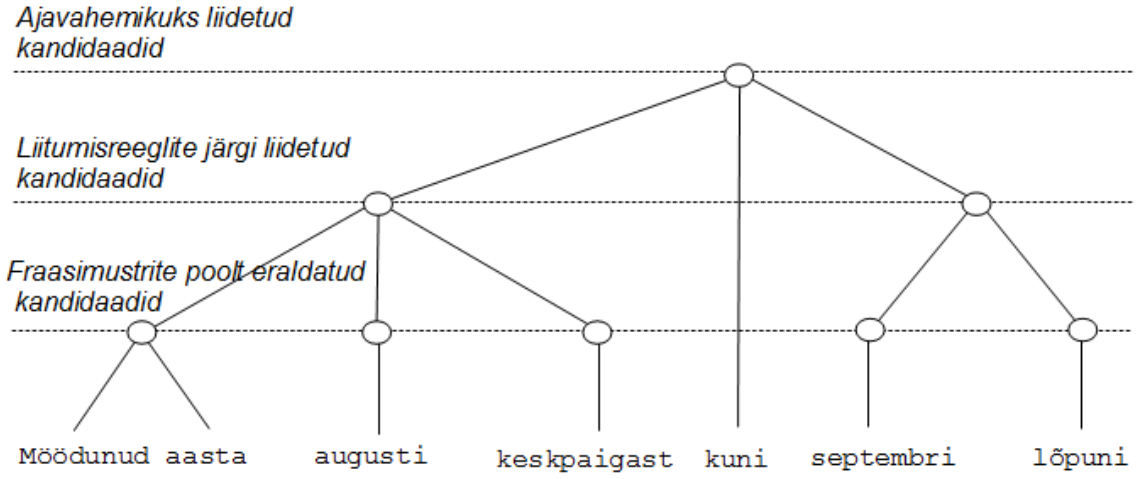
a) Kaks kõrvutipaiknevat ajaväljendikandidaati liidetakse ajavahemikuks, kui esimeses kandidaadis on seestütlevas käändes sõna (nt *eelmise aasta detsembri**st***) ning teises kandidaadis on sõna rajavas käändes (nt *selle aasta jaanuari**ni***). On ka lubatud, et sellisel kujul kandidaatide vahel paikneb sõna *kuni* või sidekriips.
b) Kui ajaväljendikandidaadi viimased sõnad moodustavad arvu (kas siis arvsõnafraasi või mingi numbrikombinatsiooni) ning kandidaadile järgneb samuti üks arv, sidekriipsu või sõnaga *kuni* eraldatult, moodustatakse ajavahemik. Enne liitmist kontrollitakse arvude ühildumist: mõlemad peavad olema kas arvsõnafraasid või

22 Mustritähiste nimetamisel kasutati infiksit `POOL` juhul, kui tähise alla kuuluv fraas võis liituda nii ajalise granulaarsuse poolest suurema kui ka väiksema fraasiga; samas kasutati tähises infiksit `TAIS` märkimaks fraase, mis enam suurema ajalise granulaarsusega fraasiga liituda ei saanud.

numbritega edasi antud arvud. Toodud heuristiku järgi moodustatakse vahemik näiteks tekstiüksustest *aastatel 2007* ja *kuni 2009*.

c) Analoogselt heuristikule b) – kui ajaväljendikandidaadi esimesed sõnad moodustavad arvu (kas siis arvsõnafraasi või mingi numbrikombinatsiooni) ning sellele eelneb ühilduv arv, sidekriipsu või sõnaga *kuni* eraldatult, moodustatakse ajavahemik.

Tehniliselt tähendab kahe või enama ajaväljendikandidaadi liitmine uue ajaväljendikandidaadi moodustamist ning sidumist vastavate sõnade ja liidetavate kandidaatidega. Liitmisel loodavate sidemete järgi moodustatakse nn kandidaatide puu, milles eristatakse kolme kandidaatide taset: 1) fraasimustrite poolt eraldatud kandidaadid, 2) liitumisreeglite järgi fraasiks liidetud kandidaadid ja 3) ajavahemikuks liidetud kandidaadid. Joonis 2 toob näite kandidaatide puust, mille juurtipp on ajavahemik-kandidaat „*Möödunud aasta augusti keskpaigast kuni septembri lõpuni*“.



Joonis 2. Näide ajaväljendite eraldamisel ja liitmisel moodustatavast ajaväljendikandidaatide puust. Ringid (puu sisetipud) tähistavad ajaväljendikandidaate.

3.5 Ajaväljendite semantika normaliseerimine

3.5.1 Märgendusformaat

Käesolevas töös kasutatakse ajaväljendite märgendusformaadina TIMEX2 standardi alamosa, millesse on teatud kohtades toodud sisse ka modifikatsioonid. Kasutatava mär-

genduskeele põhjalik kirjeldus ning erinevused TIMEX2 standardist on toodud töö lisas (vt Lisa 2). Järgnevalt antakse märgendusformaadist lühiülevaade.

Märgendusformaadis tuuakse välja ajaväljendi liik, normaliseeritud kujul semantika ning semantika modifikatsioonid. Ajaväljendi semantika võimalikke esituskujusid on neli:

1. Kuupõhine esitus. Näide:

```
<TIMEX TYPE="POINT" VALUE="2009-03-04TXX:XX">
4. märtsil 2009
</TIMEX>
```

2. Nädalapõhine esitus. Näide:

   Referentsajaks on 12. nädal aastal 2010.

```
<TIMEX TYPE="POINT" VALUE="2010-W13-4TXX:XX">
Järgmise nädala neljapäeval
</TIMEX>
```

3. Ajalise kestvuse esitus. Näide:

```
<TIMEX TYPE="DURATION" VALUE="P2Y6M">
kaks ja pool aastat
</TIMEX>
```

4. Mittekonkreetne mineviku-, oleviku- või tulevikuviide. Näide:

```
<TIMEX TYPE="POINT" VALUE="FUTURE_REF">
tulevikus
</TIMEX>
```

Esituskujusid 1 ja 2 kasutatakse ajapunktide, ajavahemike ja ajaliste korduvuste semantika edasi andmisel. Esituskuju 3 annab edasi ajalise kestvuse ning esituskujul 4 väljendatud semantika loetakse kokkuleppeliselt ajapunkti alla kuuluvaks (kuigi potentsiaalselt võib olla tegu ka viitega nt ajavahemikule minevikus, olevikus või tulevikus).

Ajaväljendi semantika esitamisel tuuakse välja (***avatakse***) vaid need ajalised granulaarsused, mille kohta leidub informatsiooni väljendis endas või väljendiga seotud referentsajas. Ülejäänud granulaarsused kaetakse X sümbolitega (esituskujudes 1-2) või jäetakse välja toomata (esituskuju 3).

Semantika esituse täpsustamiseks kasutatakse kahte viisi:

- Ajapunkti alamosa täpsustamine (algus-, kesk- või lõpuosa, esimene või teine pool)

  Näiteks

```
<TIMEX TYPE="POINT" VALUE="2009-03-XXTXX:XX" MOD="START">
2009. a märtsi alguses
</TIMEX>
```

- Umbmäärasuse väljatoomine

  Näiteks

```
<TIMEX TYPE="DURATION" VALUE="P3Y" MOD="APPROX">
umbes kolm aastat
</TIMEX>
```

Kokkuleppeliselt kasutatakse umbmäärasuse tähist `MOD="APPROX"` ka väikseima avatud granulaarsuse „varieerumise“ tähistamiseks. Järgnevas näites märgib umbmäärasuse tähis seda, et toodud kalendrikuu väärtus võib varieeruda.

Referentsajaks on 2009-12-20

```
<TIMEX TYPE="POINT" VALUE="2010-02-XXTXX:XX" MOD="APPROX">
Paari kuu pärast
</TIMEX>
```

3.5.2 Ajaväljendikandidaadi semantiline osa

Ajaväljendikandidaadi eraldamisel fraasimustri poolt lisatakse kandidaadi külge ka selle semantika kirjeldus. Ajaväljendi semantika esitatakse kui käskude/operatsioonide jada, mis tuleb väljendi normaliseerimiseks läbi viia. Neid operatsioone nimetatakse *semantika-reegliteks*.

Ajaväljendi semantika leidmisel sorteeritakse väljendiga seotud semantikareeglid ette-määratud järjekorda ning viiakse läbi operatsioonide järjestikune rakendamine. Esimene operatsioon saab sisendiks dokumendi loomise aja (kuupäev ja kellaaeg) ning iga järgmine operatsioon saab sisendiks eelmise operatsiooni rakendamise tulemuse. Vaikimisi eelda-takse, et iga konstrueeritav ajaväljend on ajapunkt-tüüpi; tüübi muutmist võimaldavad spetsiifilised operatsioonid.

3.5.3 Semantikareeglid

Semantikareegel pannakse kokku kahte liiki atribuutidest: a) reegli rakendamiseks sobivaid kontekstitingimusi täpsustavad atribuudid ning b) reegli sisu (s.o käsku) kirjeldavad atribuudid. Semantikareegli konteksti täpsustamist kirjeldatakse jaotistes 3.5.4 ning 3.5.6, käesolevas jaotises vaadeldakse semantikareegli sisu kirjeldamist.

Semantikareegli sisu kirjeldavad järgmised atribuudid:

- `priority` – määrab operatsiooni rakendamise järjekorra;
- `op` – annab edasi operatsiooni nime;
- `semField` – täpsustab, millisel ajalisel granulaarsusel operatsioon rakendub;

- `semValue` – näitab, kuidas operatsioon rakendub (nt milline peab olema granulaarsuse uus väärtus);
- `semLabel` – kasutatakse `semValue` asemel, kui määratav väärtus on kalendriväline (nt väärtus `SP` (*spring*) tähistamas aastaaega „kevad“);
- `direction` – operatsioonispetsiifiline atribuut.

Näide 12 toob ühe semantikareegli XML-kirjelduse. Toodud operatsiooni `SET` nimetatakse omistamisoperatsiooniks ning selle sisuks on kalendrivälja `semField` ülekirjutamine väärtusega `semValue`.

Näide 12. Semantikareegel, mis muudab aastaarvu väärtuseks 2007.

```
<SemReegel priority="1" op="SET" semField="YEAR" semValue="2007" />
```

Lisaks omistamisoperatsioonile kasutatakse kalendriaritmeetika teostamisel veel liitmis- ja lahutamisoperatsioone (`ADD` ja `SUBTRACT`) ning erinevaid otsimisoperatsioone (`SEEK`, `SEEK_IN` ja `BALDWIN_WINDOW`).

Liitmisoperatsioon `ADD` suurendab kalendrivälja `semField` väärtust `semValue` võrra, lahutamisoperatsioon `SUBTRACT` vähendab kalendrivälja `semField` väärtust `semValue` võrra. Kui konstrueerimisel on ajapunkt-tüüpi ajaväljend, võivad toodud operatsioonide rakendumisel muutuda ka teised kalendriväljad, vastavalt sellele, kuidas on kalendris „liigutud“ (näiteks, kui sisendkuupäevaks on 2010-03-28, annab kalendrivälja *päev* 4 võrra suurendav operatsioon tulemuseks kuupäeva 2010-04-01).

Otsimisoperatsioone kasutatakse juhtudel, kui on tarvis leida referentsajale lähim kindlate omadustega ajapunkt. Operatsioonid `SEEK` ja `SEEK_IN` nõuavad täiendava argumendina otsimissuunda `direction` (-1 minevik, +1 tulevik) ning leiavad antud suunast referentsajale lähima ajapunkti, milles kalendrivälja `semField` väärtuseks on `semValue`. Erinevus kahe operatsiooni vahel on järgmine: operatsioon `SEEK` välistab referentsaja lähima ajapunkti kandidaatide seast, samas operatsioon `SEEK_IN` lubab ka referentsaega kandidaadiks.

Operatsiooni `SEEK` võib kasutada näiteks ajaväljendi *eelseisval kolmapäeval* semantika leidmiseks. Kuna *eelseisev kolmapäev* võib olla nii *selle nädala kolmapäev* kui ka *järgmise nädala kolmapäev* (sõltuvalt sellest, millisele nädalapäevale viitab referentsaeg),

ei saa selle väljendi semantikat ainuüksi `ADD` ja `SET` operatsioonidega edasi anda, vaid tuleb leida referentsajale lähim *kolmapäev* tulevikust.

Operatsiooni `SEEK_IN` võib kasutada üksikult esineva nädalapäeva (nt *neljapäeval*) semantika leidmiseks. Kuna ajaväljend ei sisalda suunda täpsustavat eestäiendit (nt *järgmisel, eelmisel*), tuleb kõigepealt kindlaks määrata otsimissuund. Üheks võimaluseks suuna määramisel on lähtuda ajaväljendile (lause piires) lähima verbi[23] grammatilisest ajast: seda tehakse, määrates operatsiooni argumendiks `direction="VERBI_AEG"`. Kui lähim verbivorm kuulub ajakategooriate lihtminevik, täisminevik, enneminevik või üldminevik alla, või on tegu kesksõna minevikuvormiga, valitakse otsimissuunaks -1; verbi olevikuvormi puhul on otsimissuunaks +1. Kui lause piiridest sobivat verbikandidaati ei leita, rakendub operatsioon tavalise `SET` operatsioonina.

Alternatiivne võimalus üksikult esineva nädalapäeva (aga ka kuu, aastaaja vms ilma suunda täpsustava eestäiendita ajaväljendi) lahendamiseks on operatsiooni `BALDWIN_WINDOW`[24] kasutamine. Operatsiooni rakendamisel moodustatakse ainult unikaalseid kalendrivälja `semField` väärtuseid sisaldav ajaaken, mille keskpunktiks on referentsaeg. Kui akna sees leidub ajapunkt, milles kalendrivälja `semField` väärtuseks on `semValue`, valitakse see lahenduseks. Näiteks, kui referentsaeg viitab kuupäevale 2010-03-28 (pühapäev) ning otsitakse lahendit ajaväljendile *teisipäeval*, moodustatakse 7 unikaalse nädalapäeva aken järgmiselt

```
N    R    L    P    E    T    K
25   26   27   28   29   30   31
               *
```

ning valitakse lahendiks akna sisse jääv teisipäev (2010-03-30). Problemaatilisemad on juhud, mil terviklik kalendriring sisaldab paaris arvu liikmeid: sellisel juhul jääb üks võimalik väärtus aknast välja. Näiteks, kui referentsajaks on 2009-10-16 ning otsitakse lahendit ajaväljendile *aprillis*, saab moodustada 11 unikaalse kuu akna järgmiselt:

```
05  06  07  08  09  10  11  12  01  02  03
                    *
```

23 Kaugust lähima verbini mõõdetakse sõnades. Kui leitakse kaks võrdsel kaugusel olevat verbikandidaati (ajaväljendile eelnev ja järgnev verb), valitakse ajaväljendile eelnev verb.

24 Järgnevalt kirjeldatavat strateegiat kasutas selle töö autorile teadaolevalt esimest korda J. A. Baldwin töös [11]. „Baldwini aknaks“ on strateegiat nimetatud hiljem töös [19].

Kuna kuud 2009-04 ja 2010-04 on võrdsel kaugusel referentsaja kuust 2010-10, jäävad need aknast välja. Aknast väljajäävate väärtuste korral muutub `BALDWIN_WINDOW` tavaliseks `SET` operatsiooniks. Eeltoodud näite puhul saadakse `SET` operatsiooni rakendamisena tulemuseks 2009-04.

Ülejäänud operatsioonid kas suunavad arvutuste käiku (nt käsk `ANCHOR_TIMEX` võtab uueks referentsajaks tekstis eelneva või järgneva ajaväljendi lahenduse) või muudavad ülejäänud TIMEX-atribuutide väärtuseid (nt `ASSIGN_TYPE` muudab ajaväljendi tüüpi ning `SET_MOD` omistab TIMEX-atribuudile `MOD` mingi väärtuse). Lisa 3 toob operatsioonide täieliku loendi koos täpsema kirjeldusega.

3.5.4 Semantilise osa lisamine ajaväljendikandidaadi külge

Ajaväljendikandidaadi külge kinnitatav semantiline osa (semantikareeglite jada) võib tulla kahest kohast: sõnaklasside elementide küljest ja tuvastamisreeglite alt. Semantikareeglid võivad olla defineeritud selliselt, et need sisaldavad lünkasid ning täieliku semantikareegli saamiseks tuleb ühendada sõnaklassi all paiknev osa tuvastamisreegli all paikneva osaga.
Näide 13 toob sõnaklasside `AASTA_LOENDA_VAIKE_A` ja `YHIK_2` definitsioonid, milles elemente on täiendatud semantilise osaga (vrd Näide 7, kus on toodud ainult ajaväljendite eraldamist kontrolliv osa). Võib täheldada, et elementidega seotud semantikareeglid ei ole täielikud: kõikidest elementidest on puudu `op` ja `priority` atribuudid, klassi `ARV_LOENDA_VAIKE_A` elementidel puuduvad `semField` atribuudid ning klassi `YHIK_2` elementidel `semValue` atribuudid. Atribuudi `semValue` väärtus `REF:1` tähistab seda, et tegelik väärtus parsitakse sõnamalliga sobitunud sõnast (täpsemalt: regulaaravaldise esimesest sulgudega ümbritsetud alamosast) või arvsõnafraasist (täisarvuline või murdarvuline väärtus).

Näide 13. Sõnaklasside `ARV_LOENDA_VAIKE_A` ja `YHIK_2` XML-definitsioonid koos semantilise osaga.

```
<SonaKlass nimi="ARV_LOENDA_VAIKE_A">
     <Element tyyp="reg"
```

```
                vaartus="([0-5]?[0-9]?[0-9])...[25]"
                semValue="REF:1" />
        <Element tyyp="arvSona"
                arvuPiirang="0-599"
                arvuLiik="_N_|_F_"
                semValue="REF:1" />
</SonaKlass>

<SonaKlass nimi="YHIK_2">
        <Element tyyp="algv" vaartus="tund|h"
                semField="HOUR_OF_HALF_DAY" />
        <Element tyyp="algv" vaartus="minut|min"
                semField="MINUTE" />
        <Element tyyp="algv" vaartus="ööpäev|päev"
                semField="DAY_OF_MONTH" />
        <Element tyyp="algv" vaartus="nädal"
                semField="WEEK_OF_YEAR" />
        <Element tyyp="algv" vaartus="kuu"
                semField="MONTH" />
        <Element tyyp="algv" vaartus="aasta|a"
                semField="YEAR" />
</SonaKlass>
```

Näide 14 toob tuvastamisreegli, kus kasutatakse sõnaklasse `ARV_LOENDA_VAIKE_A`, `YHIK_2` (Näide 13), `AEGA` ja `TAGASI_EEST_TAGUNE` (Näide 7). Tuvastamisreegli all olevad semantikareeglid (`SemReegel`-elemendid) täidavad puuduolevad lüngad sõnaklasside all olevates semantikareeglites, täpsustades operatsioonikirjeldustest puuduvad väärtused. Atribuut `seotudMustriosa` kirjeldab nõudeid fraasimustri poolt eraldatud fraasile (millised sõnamallid peavad olema või ei tohi olla rahuldatud) ning ühtlasi määrab ka sõnaklassi, mille semantilist osa täiendatakse (sulgudega ümbritsetud sõnaklass).

Näide 14. Ühe tuvastamisreegli XML-definitsioon.

```
<Reegel>
        <Muster>
           ARV_LOENDA_VAIKE_A?  YHIK_2  AEGA?  TAGASI_EEST_TAGUNE
        </Muster>
```

25 Lühendatud, täispikkuses väärtus on:
`([0-5]?[0-9]?[0-9])-?(n)?(da)?(t|sse|le|ks|ta|s|l|ni|ga|st|lt|na)?`

```
        <MustriTahis seotudMustriosa="1"
                     tahised="TAGASI_EEST_AP,VOTAB_KVANT_EESLIITE" />
        <SemReegel priority="1"
                seotudMustriosa="ARV_LOENDA_VAIKE_A (YHIK_2)"
                op="SUBTRACT" semValue="REF_VAL:ARV_LOENDA_VAIKE_A" />
        <SemReegel priority="1"
                seotudMustriosa="^0 (YHIK_2)"
                op="SUBTRACT" semValue="1" />
    </Reegel>
```

Mõlemad toodud semantikareeglid (Näide 14) täiendavad sõnaklassi `YHIK_2` all olevat semantilist osa, lisades puuduolevad atribuudid `priority`, `semValue` ja `op`. Esimene semantikareegel nõuab, et mustriga sobitunud fraasil leiduks sõnaklassi `ARV_LOENDA_VAIKE_A` rahuldav alamosa, ning võtab `semValue` väärtuseks sõnaklassi `ARV_LOENDA_VAIKE_A` rahuldatud elemendi vastava atribuudi (`semValue`) väärtuse (nt, kui sõnaklassis `ARV_LOENDA_VAIKE_A` on rahuldatud regulaaravaldisega element, võetakse väärtus regulaaravaldise esimeste sulgudega sobitunud osast).

Teine semantikareegel nõuab, et leitud fraasist oleks puudu esimest sõnamalli rahuldav sõna (st `^0` tähistab sõnaklassi `ARV_LOENDA_VAIKE_A` puudumist). Seega on kaks tuvastamisreegli all toodud semantikareeglit teineteist välistavad: esimene kirjeldab ajaväljendeid, kus on täpsustatud lahutatav ajakvantiteet (*kümne aasta eest, 11 kuud tagasi* jms), ning teine kirjeldab ajaväljendeid, kus kvantiteeti pole määratud, seega eeldatakse, et lahutada tuleb üks ajaühik (nt väljendid *aasta tagasi, ööpäeva eest*, aga ka *nädalaid tagasi*).

3.5.5 Semantilise osa morfoloogiline filtreerimine

Kuna fraasimustrid sisaldavad üldistavaid fraasikirjeldusi, on mõningatel juhtudel pärast fraasi leidmist tarvilik läbi viia täiendav filtreerimine, et semantiline osa saaks seotud ainult kindlaid tunnuseid kandva fraasiga. Selleks on loodud süsteemi kaks võimalust: 1) tuvastamisreegli all võib defineerida *filtrid*, mis lubavad ainult teatud morfoloogiliste tunnustega fraaside külge semantikareeglite sidumist, ning 2) semantikareegli kirjelduses saab nõuda, et operatsiooni rakendamiseks peavad olema rahuldatud *täiendavad kontekstitingimused* (nt ajaväljendite liitmisel või ankurdamisel on saadud mingid kindlad tulemused). Selles jaotises antakse ülevaade morfoloogiliste filtrite kasutamisest, järgnevas jaotises käsitletakse täiendavate kontekstitingimuste määramist.

Näide 15 täiendab eelmises jaotises toodud tuvastamisreeglit (Näide 14), lisades morfoloogilise filtriga tõkestatud „umbmäärasuse“ semantika. Reegli alla kuuluv element `Filter` nõuab, et fraasimustri teist liiget (`YHIK_2`) rahuldav sõna oleks mitmuses (`pl`). Kui filtri tingimus on rahuldatud, seotakse ajaväljendikandidaadiga kaks täiendavat semantika-reeglit, mis annavad kokkuleppeliselt edasi umbmääraste väljendite nagu *aastaid tagasi*, *nädalate eest* jms semantika.

Näide 15. Tuvastamisreegli (Näide 14) definitsioon täiendatud kujul.

```
<Reegel>
     <Muster>
        ARV_LOENDA_VAIKE_A?  YHIK_2  AEGA?  TAGASI_EEST_TAGUNE
     </Muster>
     <MustriTahis seotudMustriosa="1"
                  tahised="TAGASI_EEST_AP,VOTAB_KVANT_EESLIITE" />
     <SemReegel priority="1"
             seotudMustriosa="ARV_LOENDA_VAIKE_A (YHIK_2)"
             op="SUBTRACT" semValue="REF_VAL:ARV_LOENDA_VAIKE_A" />
     <SemReegel priority="1" seotudMustriosa="^0 (YHIK_2)"
             op="SUBTRACT" semValue="1" />
     <Filter morfTunnused="_ {pl} _ _">
          <SemReegel priority="2"
                  seotudMustriosa="^0 (YHIK_2)"
                  op="SUBTRACT" semValue="1" />
          <SemReegel priority="3"
                  seotudMustriosa="3"
                  op="SET_MOD" semValue="APPROX" />
     </Filter>
</Reegel>
```

Üks morfoloogiliste tunnuste grupp (loogeliste sulgude vahel) võib sisaldada ka mitut nõutud tunnust: sellisel juhul eraldatakse tunnused üksteisest komaga.

Lisaks ainsuse ja mitmuse eristamisele on filtrit rakendatud ka käänete eristamisel (näiteks kellaaeg-väljenditest eraldatakse ühe fraasimustriga väljendid *kolmeteist minuti pärast kolm* ja *kolmteist minutit pärast kolme* ning nende semantika eristamisel kasutatakse filtreerimist) ja sõnaliikide eristamisel (põhiarvsõna ja järgarvsõna eristamine).

3.5.6 Semantikareeglite täiendavad kontekstitingimused

Semantikareegli täiendavad kontekstitingimused võimaldavad piirata reegli rakendamist vastavalt sellele, millised on ajaväljendite liitmise või ankurdamise tulemused.

Kontekstitingimuste arvestamine on vajalik näiteks nädalapäevade semantika lahendamisel. Kui nädalapäev-ajaväljend esineb lauses üksikuna (nt *esmaspäeval*), saab semantika lahendamisel rakendada üht operatsioonidest `SEEK_IN` või `BALDWIN_WINDOW`. Kui aga ajaväljendikandidaat on fraasiks liidetud nädalat määrava kandidaadiga (nt *järgmise nädala* + *esmaspäeval*), tuleb arvestada, et kandidaadi *järgmise nädala* lahendamisel pannakse õige nädal juba paika ning seega annab väljendile *esmaspäeval* korrektse lahenduse ainult `SET` operatsiooni rakendamine.

Täiendavad kontekstitingimused antakse edasi elemendi `SemReegel` atribuudis `seotudKontekst`, mille sisuks on kontekstikitsenduste jada, eraldatud "`&`" märkide abil. Semantikareegel leiab rakendamist ainult juhul, kui sellele pole seatud ühtegi kontekstitingimust või kui kõik kontekstitingimused on rahuldatud. Kasutada võib järgmisi tingimusi:

- `KORVALFRAASI_GRAN: gran`$_1$`, gran`$_2$`, ..., gran`$_n$

  Nõuab, et ajaväljendikandidaadiga (fraasiks või vahemikuks) liidetud kandidaadil (nn *kõrvalkandidaadil*) oleks vähemalt üks loetletud granulaarsustest `gran`$_1$`, ..., gran`$_n$. Kui ajaväljendikandidaadil on mitu kõrvalkandidaati, kontrollitakse neid kõiki. Kõrvalkandidaat loetakse granulaarsust `gran`$_i$ (`1<=`i`<=`n) omavaks, kui sellel leidub granulaarsust `gran`$_i$ muutev semantikareegel. Granulaarsuse ees võib kasutata märki `^` eituse tähistamiseks: sellisel juhul ühtegi eitusega granulaarsust kõrvalpaikneval kandidaadil olla ei tohi.

- `KORVALFRAAS_PUUDUB`

  Nõuab, et ajaväljendikandidaat ei oleks ühegi teise kandidaadiga fraasiks liidetud. Vahemikuks liitmised on lubatud.

- `ANKURDAMINE_LABIVIIDUD`

  Nõuab, et eelnevalt oleks rakendatud ankurdamist teostavat semantikareeglit `ANCHOR_TIMEX` ning saadud positiivne tulemus (st, ankurdatud on sellise ajaväljendi külge, millele semantika on lahendamine on õnnestunud).

- `NUM_VAHEM_OTSPUNKT_JARGNEB`

  Nõuab, et ajaväljendikandidaat oleks ühendatud vahemikuks järgneva numbrikombinatsiooni või arvsõnafraasiga (st, rakendatud alampeatükis 3.4.6 toodud ajavahemiku loomise heuristikut b)). Kui tingimus on täidetud, kasutatakse semantikareeglit ühtlasi vahemiku lõpp-punkti semantika lahendamisel.

- NUM_VAHEM_OTSPUNKT_EELNEB

  Nõuab, et ajaväljendikandidaat oleks ühendatud vahemikuks eelneva numbrikombinatsiooni või arvsõnafraasiga (st, rakendatud alampeatükis 3.4.6 toodud ajavahemiku loomise heuristikut c)). Kui tingimus on täidetud, loetakse semantikareegel ühtlasi vahemiku alguspunkti semantika lahenduseks.

Toodud kitsendustingimusi KORVALFRAASI_GRAN, KORVALFRAAS_PUUDUB ja ANKURDAMINE_LABIVIIDUD saab ka eitada, pannes nende ette märgi ^.

3.5.7 Semantika lahendamine

Ajaväljendikandidaatide semantika lahendamine kogu sisendtekstil sooritatakse nelja etapi käigus:

1) Ankurdamise etapp.
2) Esimene semantikareeglite rakendamise etapp: lahendatakse ankurdamist mittevajavad kandidaadid.
3) Teine semantikareeglite rakendamise etapp: lahendatakse ankurdamist vajavad kandidaadid.
4) Granulaarsuste jagamise ja avamise etapp.

Järgnevalt kirjeldatakse neid etappe detailsemalt.

1) Ankurdamise etapis seotakse iga ajaväljendikandidaadiga lähima verbi grammatiline aeg ning kui kandidaat nõuab ankurdamist mõne teise ajaväljendi külge, leitakse sobiv ankurkandidaat ning luuakse kahe kandidaadi vahele semantika ülevõtmist lubav link.

Kui kandidaadi küljes olev semantikareegel vajab rakendumiseks verbi grammatilist aega (st omab atribuuti direction="VERBI_AEG"), leitakse lause piiridest kandidaadile lähim verb, millele on eeltöötluse käigus edukalt määratud grammatiline aeg, ning võetakse sealt operatsiooni rakendamise suund (vastavalt peatükis 3.5.3 kirjeldatud loogikale). Verbi otsimist piirab täiendav heuristik: lähima verbi otsingul ei ole lubatud mööduda jutumärkidest. See peaks tagama, et tsiteeringu sees kasutatavad ajaväljendid seotakse ainult tsiteeringu sees kasutatud verbidega.

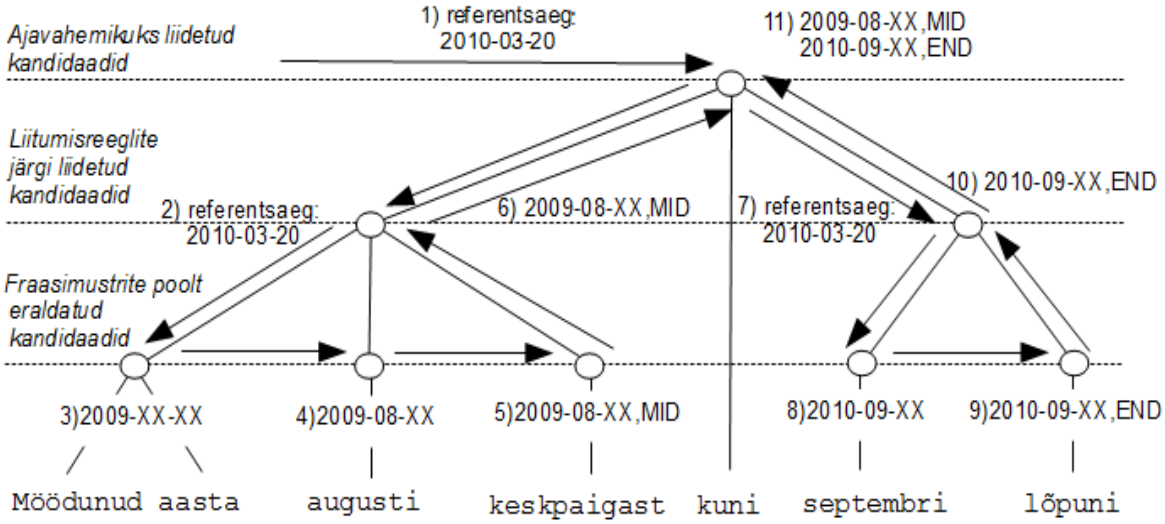
Kui kandidaadi küljes on ankurdamisoperatsioon ANCHOR_TIMEX, otsitakse atribuudis direction määratud suunalt (tekstis kandidaadile eelnemas või järgnemas) kandidaat, mis rahuldab ankurdamise tingimusi. Otsimisele on kehtestatud heuristiline piirang: kandidaadist ei minda kaugemale kui 100 sõna.

2) Esimeses semantikareeglite rakendamise etapis lahendatakse teise ajaväljendi külge ankurdamist mittevajavad kandidaadid. Kandidaatide lahendamisel lähtutakse ajaväljendite liitmisel moodustatud kandidaatide puust (vt nt Joonis 2): kui kandidaat kuulub kandidaatide puusse, leitakse kõigepealt puu juurtipp (fraas või ajavahemik) ning seejärel lahendatakse puusse kuuluvad kandidaadid lõppjärjestuses puud läbides. Kui kandidaatide puus leidub vähemalt üks ankurdamist vajav kandidaat, jäetakse kogu puu esimesel semantikareeglite rakendamise etapil lahendamata.

Fraas-tasemel kandidaadi alamkandidaatide lahendamisel saab esimene alamkandidaat sisendiks dokumendi loomise kuupäeva ning iga järgmine alamkandidaat saab sisendiks eelmise kandidaadi lahenduse. Eesmärgiga vältida kalendrivälistele kuupäevadele (nt 2010-02-31) sattumist, järjestatakse alamkandidaadid ümber selliselt, et kalendriväljade muutmisel liigutakse suuremate granulaarsuste poolt väiksemate poole. Juhtudel, kui selline järjestamine ei ole võimalik (nt kandidaadid sisaldavad kattuvaid granulaarsusi), jääb lahendamise järjekorraks siiski kandidaatide esinemisjärjekord tekstis ning seda tuleb ka reeglite koostamisel arvestada.

Kalendrivälju muutvate semantikareeglite rakendamisel leiab eraldi käsitlust kellaaegade ja päevaosade vaheline seos. Kui kellaaja tundi muutvas semantikareeglis ei täpsustata, kas on tegu ennelõunase või pärastlõunase ajaga (nt, kas *„kell 5“* märgib hommikust või õhtust aega), ent kellaaeg-fraasiga on *fraas*-tasemel liidetud päevaosa-fraas (nt *„varahommikul“*), arvestatakse päevaosa-fraasis olevat lisainformatsiooni kellaaja täpsel väljatoomisel.

Joonisel 3 tuuakse näide ajaväljendikandidaadi *„möödunud aasta augusti keskpaigast kuni septembri lõpuni“* lahendamise tulemusest esimeses semantikareeglite rakendamise etapis. Joonisel tuuakse välja ajaväljendikandidaatide puu ning sammude järjekord lahendamisel. Igal sammult tuuakse välja kas kandidaadile etteantud referentsaeg või lahendamise tulemus (kandidaadi külge kinnitatav lahendus). Kuna vahemiku mõlema otspunkti lahendamisel võetakse aluseks juurtipule etteantud referentsaeg (joonisel punkti 1 referentsaeg), on kandidaatide puu läbimisel saavutatud tulemus (punkti 11 tulemus) puudulik: on tõenäolisem, et vahemiku teise otspunkti all mõeldakse kuud 2009-09, mitte kuud 2010-09. Antud viga parandatakse granulaarsuste jagamise etapis.



Joonis 3. Näide esimese semantikareeglite rakendamise etapi tulemusest ajaväljendi *„möödunud aasta augusti keskpaigast kuni septembri lõpuni“* kandidaatidepuus. Ringid (puu tipud) tähistavad ajaväljendikandidaate.

3) Teises semantikareeglite rakendamise etapis lahendatakse eelmises etapis lahendamata jäänud (st ankurdatud) kandidaadid ning nendega samasse puusse kuuluvad kandidaadid. Lahendamise loogika on analoogne etapis 2) kasutatud loogikale: puu läbitakse lõppjärjestuses liikudes. *Fraas*-tasemel kandidaadi alamate ümberjärjestamisel pööratakse täiendavat tähelepanu ankurdatud kandidaatidele: luuakse selline järjekord, et ankurdatud alamkandidaat leiab lahendamist kõige esimesena ning ülejäänud alamkandidaatide lahendused toetuvad ankurdatud kandidaadi lahendusele.

4) Semantikareeglite rakendamise etapile järgneb granulaarsuste jagamise ja avamise etapp.

Granulaarsuste jagamine toimub ainult ajavahemikuks liidetud kandidaatide vahel ning näeb ette puuduvate granulaarsuste kandmist ühelt vahemiku otspunktilt teisele. Käesolevas süsteemis toimub see heuristilise eeskirja alusel: ülekanne tehakse ainult ajavahemiku esimeselt otspunktilt teisele ning üle kantakse granulaarsused, mis on suuremad teise otspunkti suurimast muudetud granulaarsusest. Näiteks, joonisel 3 kujutatud lahenduses on teise otspunkti suurimaks muudetud granulaarsuseks *kuu*-granulaarsus, seega kantakse esimeselt otspunktilt üle *aasta*-granulaarsuse väärtus (2009). Toodud heuristik ei paku korrektset lahendust sugugi kõigile võimalikele vahemikele ning

süsteemi edasisel arendamisel tuleb leida töökindlam viis granulaarsuste jagamiseks otspunktide vahel.

Granulaarsuste avamine on tarvilik seetõttu, et süsteemis hoitakse lahus semantika-reeglitega muudetav kalendripunkt ning semantika lõplik esituskuju. Ajapunktide ja aja-vahemike otspunktide semantika esituskujuks on vaikimisi `XXXX-XX-XXTXX:XX`. Kui normaliseerimisel rakendatav semantikareegel muudab mingit kalendrivälja, avatakse esituskujus ainult muudetud kalendriväli. Näiteks, kui referentsajaks on 2010-04-01 ning rakendatakse operatsiooni

```
<SemReegel op="SET" semField="MONTH" semValue="5" />
```

saadakse uueks esituskujuks `XXXX-05-XXTXX:XX`. Granulaarsuste avamisel tuuakse nähta-vale ka kõik muudetud granulaarsusest suuremad granulaarsused, võttes nende väärtused arvutamisel aluseks olnud kalendripunktist (referentsajast). Eelneva näite puhul saadakse normaliseerimiskujuks pärast granulaarsuste avamist `2010-05-XXTXX:XX`.

Granulaarsuste avamise etapiga jõuab ajaväljendite semantika lahendamine lõpuni. Järeltöötluse etapis tuuakse leitud ajaväljendikandidaatidest välja ainult need, mille semantika lahendamisel on edukalt lahenduseni jõutud. Kui ajaväljendikandidaat kuulub kandidaatide puusse, tuuakse välja ainult puu juurtipu lahendus.

4 Testimine ja tulemuste analüüs

Käesolevas peatükis anname ülevaate loodud süsteemi testimisest ning analüüsime probleeme, mis antud lähenemisega kaasnesid. Süsteemi arendamisel ja testimisel keskenduti eelkõige ajaväljendite tuvastamisele ajakirjandustekstides, kuna ka suur osa senisest teiste keelte jaoks tehtud uurimustööst kasutab testimisel just ajakirjanduskorpuseid. Kuna ajaväljendite kasutus teistes tekstiliikides (nt ilukirjanduses) võib oluliselt erineda kasutusest ajakirjandusvaldkonnas, ei ole oodata, et siin mõõdetud tulemused kanduksid edasi teistesse valdkondadesse.

Järgmises alampeatükis tutvustame eksperimente, mille käigus võrreldi erinevaid heuristikuid mõningate spetsiifiliste ajaväljendite semantika lahendamisel. Seejärel kirjeldame testitava süsteemi konfiguratsiooni. Alampeatükis 4.3 anname ülevaate süsteemi arendamisel kasutatud korpusest, hindame süsteemi tööd sellel korpusel ning võrdleme saadud tulemusi vana süsteemi tulemustega. Süsteemi tööst adekvaatsema pildi saamiseks märgendame uue korpuse ning alampeatükis 4.4 hindame süsteemi tööd tundmatul tekstil. Alampeatükis 4.5 vaatleme olulisemaid probleeme, mis süsteemi töös ette tulid ning pakume võimalikke lahendusi. Alampeatükis 4.6 vaatleme võimalusi süsteemi edasisel arendamisel ning peatükis 4.7 käsitleme süsteemi kasutusvõimalusi.

4.1 Eksperimendid spetsiifiliste ajaväljendite semantika lahendamisel

Üksikult esineva nädalapäevanime, kuupäeva või kuu[26] semantika lahendamisel on võimalik kasutada erinevaid heuristikuid (nt käesolevas töös implementeeritud heuristikuid `SET`, `SEEK`, `SEEK_IN` ja `BALDWIN_WINDOW`).

Nädalapäev-ajaväljendite lahendamise heuristikute kohta on teistes keeltes teada ka mõningad eksperimentaalsed tulemused. J. A. Baldwin katsetas antud ajaväljendite lahendamist inglise- ja prantsusekeelsetel ajakirjanduskorpustel [11] ning leidis, et 96,20% prantsusekeelsetest ning 96,97% inglisekeelsetest nädalapäev-ajaväljenditest jäävad 7-päeva akna sisse (käesoleva töö heuristik `BALDWIN_WINDOW`). P.Mazur ja R.Dale võrd-

26 Laiemas plaanis mõeldakse siin ajaväljendeid, mille puhul ei saa ainuüksi ajaväljendifraasi põhjal üheselt otsustada, millisele nädalapäevale, kuule või kuupäevale väljend viitab. Siia alla loetakse nii ühest sõnast koosnevad väljendid (nt *teisipäeval, märtsis*), kui ka väiksemaid granulaarsusi sisaldavad liitväljendid (nt *teisipäeva / 20. jaanuari hommikul kell 10*). Täpsustava eestäiendiga ajaväljendid (nt *järgmisel reedel*) või suuremaid granulaarsusi sisaldavad liitväljendid (nt *möödunud aasta 31. märtsil*) arvatakse siit välja.

lesid erinevaid heuristikuid töös [19] ning mõõtsid Baldwini akna korrektsuseks 94,28% ja verbi grammatilisele ajale toetuva heuristiku[27] korrektsuseks 92,64%. Kombineerides kahte eeltoodud heuristikut, lõid autorid ka algoritmi, mis lahendas nädalapäev-väljendite semantika 95,91% korrektsusega.

Hindamaks käesolevas süsteemis implementeeritud heuristikute efektiivsust eestikeelsete nädalapäev-, kuupäev- ja kuu-ajaväljendite lahendamisel, loodi ajakirjandustekstidest koosnev testkorpus. Korpus moodustati 6-8 päeva pikkuste perioodide tervikväljaannetest: tekstides tuvastati automaatselt vastavad ajaväljendid ning parandati käsitsi automaatsel tuvastamisel tehtud vead. Tabel 1 toob loodud korpuse täpsema kirjelduse.

Tabel 1. Nädalapäev-, kuupäev- ja kuu-ajaväljendite semantika lahendamise heuristikute katsetamiseks loodud korpus.

	Eesti Päevaleht	Postimees	SL Õhtuleht	Kokku ajaväljen-deid
nädalapäev-ajaväljendid	Periood: 2007-02-07(K)– 2007-02-14(K) 122 ajaväljendit	Periood: 2000-05-22(E)– 2000-05-27(L) 92 ajaväljendit	Periood: 2004-09-06(E)– 2004-09-11(L) 112 ajaväljendit	326
kuupäev-ajaväljendid	Periood: 2007-01-08(E)– 2007-01-14(P) 136 ajaväljendit	Periood: 2000-07-03(E)– 2000-07-08(L) 111 ajaväljendit	Periood: 2004-09-20(E)– 2004-09-25(L) 96 ajaväljendit	343
kuu-ajaväljendid	Periood: 2006-12-11(E)– 2006-12-17(P) 123 ajaväljendit	Periood: 2000-03-13(E)– 2000-03-18(L) 86 ajaväljendit	Periood: 2004-10-18(E)– 2004-10-23(L) 57 ajaväljendit	266

Tabelis 1 kirjeldatud korpused on võetud Eesti keele koondkorpusest.[28] Automaatselt tuvastatud ajaväljendite seast jäeti välja vale-eraldused, ajavahemike koosseisu kuuluvad ajaväljendid ning üldises tähenduses kasutatud ajaväljendid. Lisa 5 (materjalide CD) sisaldab loodud testkorpust.

Loodud korpuste peal katsetati järgmisi semantika lahendamise mudeleid:

- **mudel 0** ehk nn baasmudel. Ajaväljendite lahendamiseks kasutati operatsiooni `SEEK_IN` vaikimisi määratud suunaga `-1` (minevik) – st, kõikide ajaväljendite

27 Täpsemalt, süsteemis TempEx kasutatud heuristik: kirjeldatud alampeatükis 2.1.1 kui heuristik c).
28 vt http://www.cl.ut.ee/korpused/segakorpus/ (viimati vaadatud 20.04.2010)

semantika lahendust otsiti lähiminevikust (lähim referentsajale eelnev nädalapäev või kuu) või olevikust (lahenduseks võis olla ka referentsaja nädalapäev või kuu).

- **mudel 1**. Ajaväljendite lahendamiseks kasutati operatsiooni `SET`. Sisuliselt tähendas see, et lahenduseks valiti nädalapäev, mis jäi referentsajaga samasse nädalasse, või kuu, mis jäi referentsajaga samasse aastasse.
- **mudel 1.1**. Ajaväljendite lahendamiseks kasutati operatsiooni `SEEK_IN`, milles ajaväljendi otsimise suund otsustati ajaväljendile lähima verbi grammatilise aja põhjal. Kui lause piiridest verbi ei leitud, lahendati semantika tavalise `SET` operatsiooniga (vt mudel 1).
- **mudel 1.2**. Ajaväljendite lahendamiseks kasutati operatsiooni `BALDWIN_WINDOW`, st nädalapäeva otsiti referentsaja kuupäeva ümbritseva 7-päeva akna seest ning kuud otsiti referentsaja kuud ümbritseva 11-kuu akna seest. Kui otsitav kuunimi ei kuulunud akna 11-kuu hulka, lahendati ajaväljendi semantika tavalise `SET` operatsiooniga (vt mudel 1).

Referentsajana kasutati kõikides mudelites ajaleheartikli ilmumise kuupäeva. Eeltoodud mudelite rakendamise tulemused toob Tabel 2.

Tabel 2. Erinevate heuristikute tulemused nädalapäev-, kuupäev- ja kuu-ajaväljendite semantika lahendamisel. Toodud on korrektselt lahendatud ajaväljendite suhtarv ja protsent.

	mudel 0	**mudel 1**	**mudel 1.1**	**mudel 1.2**
nädalapäev-ajaväljendid	253/326 (77,6%)	167/326 (51,2%)	294/326 (**90,2%**)	263/326 (80,7%)
kuupäev-ajaväljendid	229/343 (66,8%)	294/343 (85,7%)	303/343 (88,3%)	318/343 (**92,7%**)
kuu-ajaväljendid	182/266 (68,4%)	214/266 (80,5%)	248/266 (93,2%)	252/266 (**94,7%**)

Käesoleva eksperimendi tingimustes leiti, et erinevalt inglise- ja prantsusekeelsetest nädalapäeva-ajaväljenditest jääb oluliselt vähem eestikeelseid nädalapäev-väljendeid referentsaega ümbritseva 7-päeva akna sisse (80,7%) ning nende lahendamisel annab paremaid tulemusi verbi grammatilisele ajale toetuv heuristik (90,2%). Kuupäev- ja kuu-ajaväljendite puhul oli protsentuaalne erinevus Baldwini akna ja verbi ajale toetuva heuristiku vahel väiksem, kuigi mõlema ajaväljendiliigi puhul andis paremaid tulemusi

Baldwini akna kasutamine. Ilmselt on saadud tulemused siiski korpusespetsiifilised ning nende põhjal ei ole võimalik suuremaid üldistusi teha.

4.2 Süsteemi konfiguratsioon

Käesoleva alampeatükis anname lühiülevaate testitava süsteemi konfiguratsioonist.

Süsteemi poolt kasutatavas reeglifailis on defineeritud 97 sõnaklassi, 213 tuvastamisreeglit ning 37 liitumisreeglit.

Juhindudes eelmises alampeatükis kirjeldatud eksperimendi tulemustest, kasutatakse üksikult esinevate nädalapäev-väljendite semantika lahendamisel verbi grammatilise aja heuristikut (mudel 1.1) ning üksikute kuupäev-väljendite ja kuu-väljendite lahendamisel toetutakse Baldwini akna heuristikule (mudel 1.2).

Valdava osa ajaväljendite lahendamisel kasutatakse referentsajana dokumendi loomise kuupäeva. Teise ajaväljendi külge ankurdamist kasutatakse ainult järgmistel juhtudel:

- Üksikult või koos kellaajaga esinevad päevaosa-väljendid (nt *hommikul, õhtul kell 17*) ankurdatakse tekstis eelneva, lähima *päev*-granulaarsusega väljendi külge.
- Üksikult esinevad kellaaeg-ajaväljendid ankurdatakse tekstis eelneva, lähima *päev*- või *tund*-granulaarsusega ajaväljendi külge.
- Määrsõnu *varem* ja *hiljem* sisaldavad ajaväljendid (nt *kaks päeva varem*) ankurdatakse tekstis eelneva, lähima ajaväljendi külge.
- Asesõnu *sama* ja *too* sisaldavad ajaväljendid (nt *samal päeval, tollel aastal*) ankurdatakse tekstis eelneva, lähima ajaväljendi külge.

Ankruks valitav ajaväljend peab olema kas ajapunkt, ajavahemik või ajaline korduvus. Ajavahemiku puhul valitakse ankruks vahemiku lõpp-punkt.

4.3 Testimine arenduskorpusel

Süsteemi testimiseks loodi spetsiaalne programm, mis võrdles käsitsi märgendatud korpust automaatselt märgendatud korpusega, tõi välja märgenduste erinevused ning väljastas hinnangu süsteemi tööle (saagis, täpsus).

Süsteemi arendamise käigus koguti testimiseks ajakirjandustekstidest koosnev korpus[29] (artiklid Eesti Päevalehest ja Postimees võrguväljaandest), mis sisaldas 239 artiklit, 47652 sõna ja 1700 ajaväljendit (edaspidi arenduskorpus). Tabelis 3 on toodud tekstide arvud

29 Kogutud arenduskorpus on eraldiseisev alampeatükis 4.1 tutvustatud katsekorpusest, mida leidis kasutust nädalapäev-, kuupäev- ja kuu-ajaväljendite lahendamise heuristikute hindamisel.

korpuses tekstide loomisaastate lõikes.

Lisa 4 annab ülevaate arenduskorpuses esinenud ajaväljendite sagedusprofiilist. Toodud lisas on ajaväljendid grupeeritud semantiliste esituskujude järgi. Lisa 5 sisaldab kirjeldatud arenduskorpust.

Tabel 3. Kogutud arenduskorpus. Tabelis on toodud artiklite arv loomisaastate lõikes.

	2000	2007	2008	2009	2010	Kokku tekste
Eesti Päevaleht	–	146	–	–	–	146
Postimees	64	1	3	19	6	93

Arenduskorpuse ajaväljenditest 173 (10,2% korpusest) olid varustatud lühikommentaaridega (st, ajaväljendil oli määratud atribuut `COMMENT`). Ajaväljendi kommenteerimist kasutati juhtudel, kui märgendusformaat ei võimaldanud väljendi semantikat täpselt edasi anda või ei suutnud märgendaja olemasoleva konteksti põhjal semantikat täpselt määrata.

Programmi töö hindamine viidi läbi analoogselt TERN 2004 hindamisele [20]: mõõdeti ajaväljendite eraldamise saagis ja täpsus, fraasipiiride määramise saagis ja täpsus ning atribuutide `TYPE`, `VALUE`, `MOD`, `VALUE2` ja `MOD2` määramise saagised ja täpsused.

Ajaväljendite eraldamise saagis defineeriti kui korrektselt eraldatud[30] ajaväljendite arv jagatuna käsitsi märgendatud ajaväljendite arvuga. Eraldamise täpsus defineeriti kui korrektselt eraldatud ajaväljendite arv jagatuna kõigi eraldatud ajaväljendite arvuga. Fraasipiiride määramise saagis näitab korrektsete fraasipiiridega ajaväljendite[31] arvu suhet käsitsi märgendatud ajaväljendite arvuga. Fraasipiiride määramise täpsuseks loetakse korrektsete fraasipiiridega ajaväljendite arvu suhet kõigi automaatselt eraldatud ajaväljendite arvuga.

Atribuutide määramise saagised ja täpsused leiti ainult korrektselt eraldatud ajaväljendite puhul. Atribuudi määramise saagise saamiseks jagati automaattuvastamisel korrektselt määratud atribuudiväärtuste arv käsitsi määratud atribuudiväärtuste arvuga. Atribuudi määramise täpsus saadi, kui jagati automaattuvastamisel korrektselt määratud atribuudiväärtuste arv automaattuvastamisel läbiviidud atribuudimääramiste koguarvuga.

30 Korrektseks eraldamiseks loeti ka seda, kui automaatselt eraldatud ajaväljend kattus käsitsi eraldatud väljendiga ainult osaliselt.

31 Korrektsete fraasipiiridega ajaväljend – automaatselt eraldatud ajaväljend, mille fraasipiirid langesid täpselt kokku käsitsi eraldatud ajaväljendi fraasipiiridega.

Saamaks ülevaadet sellest, milline on käesolevas töös loodud süsteemi areng võrreldes bakalaureusetöös loodud süsteemiga, hinnati esmalt vana süsteemi tööd arenduskorpusel. Kuna vana süsteem on nii ülesehituselt kui ka kasutatava märgendusformaadi poolest uuest erinev, ei olnud võimalik otsene võrdlus: võrdlemiseks tuli vana süsteemi reeglid ümber kirjutada uue süsteemi reeglite formaati ning kasutada neid uue süsteemi koosseisus.
Tabelis 4 tuuakse vana süsteemi reeglitega märgendamise tulemused arenduskorpusel. Tabelis märgib `TIMEX` eraldamise mõõdikuid, `EXTENT` fraasipiiride määramise mõõdikuid ning `TYPE`, `VALUE`, `MOD`, `VALUE2` ja `MOD2` vastavate atribuutide määramise mõõdikuid.

Tabel 4. Arenduskorpuse märgendamine vana süsteemi reeglitega.

	Saagis	Täpsus
`TIMEX`	857/1700 (50,4%)	857/936 (91,6%)
`EXTENT`	633/1700 (37,2%)	633/936 (67,6%)
`TYPE`	834/857 (97,3%)	834/857 (97,3%)
`VALUE`	565/845 (66,9%)	565/856 (66%)
`MOD`	0/47 (0%)	–
`VALUE2`	10/35 (28,6%)	10/18 (55,6%)
`MOD2`	–	–

Tabelis 5 on toodud uue süsteemi töö tulemused arenduskorpusel. Märgendamisel on kasutatud alampeatükis 4.2 tutvustatud konfiguratsiooni.

Tabel 5. Uue süsteemi töö tulemus arenduskorpuse märgendamisel.

	Saagis	Täpsus
`TIMEX`	1428/1700 (84%)	1428/1450 (98,5%)
`EXTENT`	1322/1700 (77,8%)	1322/1450 (91,2%)
`TYPE`	1399/1428 (98%)	1399/1428 (98%)
`VALUE`	1202/1388 (86,6%)	1202/1393 (86,3%)
`MOD`	72/89 (80,9%)	72/78 (92,3%)
`VALUE2`	30/54 (55,6%)	30/45 (66,7%)
`MOD2`	–	0/1 (0%)

4.4 Testimine uuel korpusel

Arenduskorpust kasutati jooksvalt süsteemi testimisel: uute reeglite sissetoomisel ja

vanade reeglite muutmisel ning seetõttu ei pruugi tulemused arenduskorpusel anda piisavalt adekvaatset pilti süsteemi tööst. Adekvaatsema hinnangu saamiseks vaadeldi süsteemi tööd ka uuel korpusel (edaspidi testkorpus).

Uueks korpuseks valiti Eesti keele koondkorpusest üks Eesti Päevalehe tervikväljaanne (ilmumiskuupäev 2006-08-12), mis sisaldas 86 artiklit ning 10617 sõna. Korpus märgendati automaatselt ning seejärel parandati käsitsi automaatsel märgendamisel tehtud vead. Kokku leiti korpusest 385 ajaväljendit. Lisa 4 toob uues testkorpuses esinenud ajaväljendite sagedusprofiilid ning Lisa 5 sisaldab uut korpust terviklikul kujul.

Süsteemi töö tulemustest uuel korpusel annab ülevaate Tabel 6.

Tabel 6. Uue süsteemi töö tulemus testkorpuse märgendamisel.

	Saagis	Täpsus
TIMEX	290/385 (75,3%)	290/294 (98,6%)
EXTENT	274/385 (71,2%)	274/294 (93,2%)
TYPE	284/290 (97,9%)	284/290 (97,9%)
VALUE	259/283 (91,5%)	259/283 (91,5%)
MOD	11/19 (57,9%)	11/11 (100%)
VALUE2	10/14 (71,4%)	10/11 (90,9%)
MOD2	–	–

4.5 Probleemid

Järgnevalt käsitleme põhjalikumalt probleeme, mis on tulnud esile ajaväljendite tuvastaja testimisel, ning pakume võimalikke lahendusi süsteemi edasisel arendamisel.

4.5.1 Eraldamata jäävad ajaväljendid

Testimistulemused näitasid, et ajaväljendite eraldamisel on problemaatiline just suhteliselt madal eraldamise saagis (84% arenduskorpuses, 75,3% testkorpuses). Antud alampeatükis analüüsitakse, millised ajaväljendid jäävad loodud süsteemi poolt eraldamata.

Tabelis 7 tuuakse välja arenduskorpuses kõige sagedamini eraldamata jäänud ajaväljendite kirjeldused. Tulemusi analüüsides püüti ajaväljendeid grupeerida mingite ühiste tunnuste või eraldamata jätmise põhjuste alusel, kuigi alati polnud selline grupeerimine võimalik (vigadel oli rohkem kui üks põhjus/tunnus). Tabelis tuuakse välja just sellised ajaväljendid, mis oli võimalik selgelt ühe tunnuse alla liigitada.

Tabel 7. Kümme kõige sagedasemat eraldamata jäänud ajaväljendigruppi arendus-korpuses. Toodud on ajaväljendi kirjeldus ning vea esinemissagedus korpuses.

Ajaväljendi kirjeldus	Esinemissagedus
Valdkonnaspetsiifilised/valdkonnateadmisi nõudvad ajaväljendid, nt: *viimase sõja ajal, rootsi ajal, teise poolaja 23. minutil, ajaga 13,89.*	64 (23,5%)
Nn lühikujul sõnalised väljendid, nt *kvartaliga* (vrd *kolmanda kvartaliga*, *ühe kvartaliga*), *aasta* (vrd *üks aasta, käesolev aasta*)	38 (14,0%)
Liitsõnakoosseisu kuuluvad ajaväljendid, nt *kinoleviaastat 2009, esmaspäevahommikuti* (vrd *esmaspäeva hommikutel*), *läinud-nädalane* (vrd *läinud nädala*).	22 (8,1%)
Nn lühikujul numbrilised väljendid, nt *2000* (vrd *aastal 2000, 2000 töötajat*), *10.00* (vrd *kell 10.00* või *Küüslauguleib(10.00)*).	19 (7,0%)
Ligikaudse semantikaga ajaväljendid, nt *mõni aeg tagasi, mitmeid aastaid, mõneks päevaks.*	18 (6,6%)
Puuduoleva reegli tõttu eraldamata jäävad väljendid.	17 (6,2%)
Sidekriipsuga numbrilise arvuvahemiku esimene pool, nt *6-[8 kuud], 10-[12 tunniks]*.[32]	10 (3,7%)
Vaatluse alt välja jäänud granulaarsuseid (peamiselt *sajand* ja *sekund*) sisaldavad ajaväljendid, nt *18. sajandil, 3 sekundit.*	8 (2,9%)
Ajaväljendid, mille sees on mitteajalist tähendust kandev sõna või fraas, nt *viimase päikeselise päevaga, peaaegu kümnest lasteaja-kirja Hea Laps peatoimetajana töötatud aastast.*	7 (2,6%)
Sidekriipsuga mittenumbrilised ajaväljendid, nt *aasta-poolteist*, *laupäeviti-pühapäeviti*.	6 (2,2%)

Kõige sagedamini eraldamata jäänud ajaväljenditeks olid valdkonnaspetsiifilised aja-väljendid (23,5% kõigist eraldamata jäänud väljenditest). Kuna selliste ajaväljendite semantika esitamine on sageli problemaatiline, ei koostatud nende eraldamiseks ka tuvastamisreegleid, küll aga märgendati need korpustes, et oleks võimalik hinnata nende osakaalu jooksvas tekstis. TIMEX2 standardi alusel kuuluvad sellised ajaväljendid siiski märgendamisele, kuigi nende semantika väljatoomine ei ole kohustuslik.

Teiseks suuremaks vigade allikaks oli nn. lühikujul väljendite eraldamata jätmine. Lühikujul väljenditeks loeti ainult ajaühiku nimest koosnevad väljendid (nt *aasta*, *päevaga*) ning ainult aastarvust või kellaajast/kuupäevast koosnevad väljendid (nt *2000, 04.04*). Kuna selliste ajaväljendite semantikale on sageli mitu võimalikku tõlgendust ning

32 Kandilised sulud märgivad eraldatud ajaväljendi piire.

õige valimisel tuleb arvestada laiemat konteksti, ei ole tuvastamisreeglid antud ajaväljendite jaoks välja ehitatud.[33] Süsteemi edasisel arendamisel on üheks võimalikuks lahenduseks lühikujul väljendite jaoks siiski reeglid lisada, mis tooks aga tõenäoliselt kaasa (nii eraldamise kui semantika lahendamise) täpsuse languse, kuna selliste väljendite konteksti on raske reeglites kirjeldada. Alternatiivset lahendust võiks pakkuda masinõppe meetodid, mis võimaldaksid paljude tunnuste abil ennustada, milline on (aja-)väljendi kasutuskontekst.

Kolmandaks suuremaks eraldamata jäänud ajaväljendite grupiks arenduskorpuses olid liitsõna kujul esitatud ajaväljendid (nt *esmaspäevahommikuti, läinudnädalane*) ning ajalist tähendust mittekandvate sõnadega liitsõnaks liitunud ajaväljendid (nt fraasides *kuue teenistuskuu jooksul, kolmekümne eluaasta*). Kuna morfoloogilise analüüsi käigus määratakse kindlaks ka liitsõnade piirid, on selliste ajaväljendite puhul potentsiaalseks lahenduseks fraasimustri loogika laiendamine, lubades fraasimustri sõnamallide sobitamist ka liitsõna sees: üks sõnamall kirjeldaks sellisel juhul liitsõna alamsõna. See võimaldaks kirjeldada ka sidekriipsuga mittenumbrilisi ajaväljendeid (nt *jaanuaris-veebruaris*), mis annavad edasi konkreetset ajavahemikku.

Kui sidekriipsuga ajaväljend ei anna edasi konkreetsete otspunktidega vahemikku (nt *6-8 kuud, laupäeviti-pühapäeviti*), muutub selle semantika terviklik edasiandmine käesoleva töö märgendusformaadi alusel problemaatiliseks. Samuti leidub väljendeid, mille võimalikuks lahendiks on pigem kaks eraldiseisvat ajapunkti kui ajavahemik (nt *kahe-kolme nädala eest = kahe nädala eest või kolme nädala eest*). Juhindudes TIMEX2 formaadist, toodi käsitsi märgendamisel selliste väljendite „mõlemad otspunktid“ eraldi välja (nt *[kahe-][kolme nädala eest], [6-][8 kuud]*), misläbi oli võimalik ka semantikat eraldi kirjeldada. Tuvastamisreeglid eraldatud fraasi sellist poolitamist ei võimalda ning kirjeldavad ainult fraasi lõppu, seetõttu sai esimese otspunkti eraldamata jätmine sagedaseks vigade allikaks automaatsel eraldamisel.

Ajaväljendite kirjeldamise seisukohast on kõige problemaatilisemad ajaväljendifraasid, mis sisaldavad mitteajalist tähendust kandvaid sõnu või alamfraase (nt *viie haiglas veedetud kuu jooksul*). Selliste väljendite korrektne kirjeldamine eeldaks süntaksianalüüsi kasutuselevõtmist.

Arenduskorpuses eraldamata jäänud väljenditest 6,2% olid sellised, mille semantika

33 Ajaühiku nime sisaldavad lühikujul väljendid leiavad siiski eraldamist, kui nendega on liitunud semantikat täpsustav sõna, nt *aasta alguses*, *kuu lõpus*. Semantika jääb sellistel juhtudel aga poolikuks.

kirjeldamine oleks käesoleva süsteemi vahenditega võimalik: nende eraldamine nõuab ainult vastavate reeglite lisamist. Väljajäänud granulaarsusi (*sajand* ja *sekund*) sisaldavate ajaväljendite täielikuks kirjeldamiseks aga ei piisa ainult reeglite koostamisest, vaid tuleb teha muutuseid ka süsteemi tasemel: luua liides uute kalendriväljade ning kalendrimudeli vahele.

Tabelis 8 tuuakse testkorpuses kõige sagedamini eraldamata jäänud ajaväljendite kirjeldused. Tabel katab kõik testkorpuses eraldamata jäänud ajaväljendid. Võib täheldada, et gruppideks jaotus on suures osas sama (vrd Tabel 7), kaheks uueks grupiks on ajalised korduvused, mille semantikat pole kasutatud märgendusformaadiga võimalik edasi anda, ning katkise reegli tõttu eraldamata jäänud ajaväljendid. Lühikujul numbriliste ajaväljendite ning vaatluse alt välja jäänud (*sekund*-)granulaarsusega väljendite suurem esinemissagedus korpuses on tingitud spordiuudiste suuremast osakaalust.

Tabel 8. Kümme eraldamata jäänud ajaväljendigruppi uues testkorpuses.
Toodud on ajaväljendi kirjeldus ning vea esinemissagedus korpuses.

Ajaväljendi kirjeldus	Esinemissagedus
Valdkonnaspetsiifilised/valdkonnateadmisi nõudvad ajaväljendid.	35 (36,8%)
Puuduoleva reegli tõttu eraldamata jäävad väljendid.	13 (13,7%)
Nn lühikujul numbrilised väljendid.	13 (13,7%)
Vaatluse alt välja jäänud granulaarsuseid sisaldavad ajaväljendid.	12 (12,6%)
Nn lühikujul sõnalised väljendid.	5 (5,3%)
Sidekriipsuga numbrilise arvuvahemiku esimene pool.	4 (4,2%)
Puuduolevad ajalised korduvused, nt *iga tunni, iga mõne tunni järel*.	4 (4,2%)
Katkise reegli tõttu eraldamata jäävad ajaväljendid.	4 (4,2%)
Liitsõnakoosseisu kuuluvad ajaväljendid.	3 (3,2%)
Ligikaudse semantikaga ajaväljendid.	2 (2,1%)

4.5.2 Valeeraldused

Testimistulemused näitasid, et ajaväljendite eraldamise täpsus oli suhteliselt kõrge (~98% nii arendus- kui ka testkorpusel), seega ei saa lugeda valede eralduste tegemist kriitiliseks probleemiks. Siiski toome välja testimisel ilmnenud juhud, kus tuvastaja eraldab mitte-ajaväljendi ajaväljendina. Toodud näidetes tähistavad kandilised sulud automaatselt eralda-

tud ajaväljendi piire.

- Valesid eraldamisi võib põhjustada lause tegelike fraasipiirde mittetundmine, nt:
  - *Tallinna halduskohus ei rahuldanud kaebust tühistada riigikogu kantselei [poolt nelja] auto kasutusrendilepingu riigihanke võitjaks tunnistanud AS Reval Auto pakkumine.*[34]

  Selliste vale-eralduste vältimine nõuaks korrektset fraasipiiride kindlaksmääramist kogu lause ulatuses.
- Mittekonkreetseid olevikuviited (nt *nüüd, hetkel*) ei viita alati käesolevale hetkele, nt lauses *„Kuni selleni välja, et "remiksid" oma varasemaid luuletusi uuteks tekstideks kokku ja esitad need [nüüd] murdmata ridadena.“* Eeltoodud lauses võiks väljendi *nüüd* ka eraldamata jätta, kuna ajaviidet kasutatakse pigem üldises tähenduses kui viitena olevikule.
- Nimede eraldamine ajaväljenditena on endiselt problemaatiline, nt lauses „*Praegu on [Mai] psühholoogi abiga leidnud tookord märkamata jäänud ohumärke juba algusaastatest, mees soovis kõike teada, kõike kontrollida.“* Kui isikunimi on esitatud täisnimena (eesnimi ja perekonnanimi), on võimalik negatiivsete mustrite abil isikunimede eraldamist kuunimedena vältida (vt nt alampeatükis 3.4.5). Eeltoodud näites aga negatiivseid mustreid kasutada ei saa: lisaks suurtähelisusele tuleks kontrollida ka seda, et kustutatav kandidaat ei paikneks lause alguses.

4.5.3 Ajaväljendifraaside ulatuse määramine

Käesolevas alampeatükis toome välja peamised juhud, kus automaatsel tuvastamisel eksitakse ajaväljendipiiride määramisega. Poolikult eraldatakse järgmised ajaväljendi-fraasid:

- Arvuvahemikku sisaldavad ajaväljendifraasid. Näiteks: *kahe-kolme [nädala eest]*, *6-[8 kuud]*.
- Mittetoetatud arvukuju sisaldavad fraasid. Nt: *miljoneid [aastaid tagasi], kaks ja [pool aastat], 10 [000 aasta eest]*.
- Umbmäärase semantikaga ajaväljendid. Nt: *kümmekond [aastat tagasi], mõne [kuu eest]*.
- Sündmuse külge ankurdatud ajaväljendid. Nt: *[Poolteist nädalat] enne kohalike omavalitsuste valimisi, [kaks päeva] enne kaevurite päeva.*

34 Käesolev lõik ning järgnevad tekstilõigud pärinevad Eesti Päevalehe 2007 a korpusest.

- Eestäienditega mittekonkreetsed ajalised viited. Nt: *Kaugemas [tulevikus], Järgmisel [hetkel]*.
- Keerukamad fraasikonstruktsioonid. Nt: *[12. detsembri öösel kella 01.00] ja 02.00 vahel, 11., 13. ja [14. märtsil]*.
- Trükivigu või ebaharilikku joondust (puuduv tühik) sisaldavad ajaväljendid. Näiteks: *eisipäeva [õhtul], 2006.a. [jaanuaris], 3.juunist [1996. a]*.

Osa ilmnenud vigadest on lahendatavad eeltöötluse protsessi täpsemaks muutes: näiteks võiks arvsõnafraaside tuvastaja üles leida ning semantiliselt määratleda veel ligikaudseid kvantiteete (nt *kümmekond, mitu*), *ja*-konstruktsiooniga arve (nt *kaks ja pool*) ning liitsõna kujul toodud arvuvahemikke (*kahe-kolme*). Süsteemi endisel kujul laiendades pole aga võimalik lahendada näiteks ajaväljendite ankurdamist sündmuste külge ning mittekonkreetsete ajaliste viidete (kõikvõimalike) eestäiendite leidmist: need tegevused nõuaksid paratamatult süntaksianalüüsi kaasamist.

Rinnastusseoses ajaväljendite (nt *11., 13. ja 14. märtsil*) ning arvuvahemikku sisaldavate fraaside (nt *kahe-kolme nädala eest*) terviklikul eraldamisel muutub problemaatiliseks ka nende semantika esitamine, mistõttu on võimalikuks alternatiiviks eraldada need fraasid eraldiseisvate väljenditena (nt *[11.], [13.] ja [14. märtsil]*).

4.5.4 Vead semantika normaliseerimisel

Käesolevas peatükis käsitletakse semantika normaliseerimise vigu detailsemalt. Kui vaadata semantika normaliseerimist eelkõige kui atribuudi `VALUE` väärtuste määramist, saavutas süsteem suhteliselt kõrgeid tulemusi nii arenduskorpusel (saagis ja täpsus ~86%) kui ka uuel testkorpusel (saagis ja täpsus ~91%). Siinkohal tuleb aga rõhutada, osa ajaväljendeid, mille semantikat märgendusformaat edasi ei võimaldanud anda, jäid `VALUE` väärtuseta ega peegeldu selles tulemuses. Problemaatiline on ka ajavahemiku otspunktide semantika määramine: ajavahemiku teise otspunkti semantika määramist (`VALUE2`) iseloomustavad suhteliselt madal saagis (arenduskorpusel 55,6%, testkorpusel 71,4%) ja täpsus (arenduskorpusel 66,7%, testkorpusel 90,9%).

Tabelis 9 tuuakse kümme kõige sagedasemat eksimust semantika (st `VALUE` väärtuste) määramisel arenduskorpuses. Tabelis on toodud juhud, mil oli võimalik üheselt määrata eksimuse põhjus, välja jäävad mitmete erinevate põhjuste kokkulangemisel tekkinud vead.

Tabel 9. Kümme kõige sagedasemat eksimust semantika määramisel arendus-korpuses. Toodud on eksimuse kirjeldus ning esinemissagedus korpuses.

Eksimuse kirjeldus	Esinemissagedus
Ajaväljendi kasutus üldises tähenduses. Näiteks lauses: *Suurema osa oma lapsepõlvest veetsin vana-vanemate kodus ning loomulikult oli [suve] lahutamatuks osaks talutöö.*	51 (24,8%)
Eksiti ajaväljendi ankurdamisel teise ajaväljendi külge (üleliigne, puuduv või valesti teostatud ankurdamine).	31 (15,0%)
Semantika määrati valesti pooliku eraldamise tõttu.	28 (13,6%)
Ajaühiku nime sisaldava lühiväljendi (nt *aasta alguses, kuu lõpus*) semantika oli määramata.	10 (4,9%)
Semantika korrektne väljendamine nõuab ajaväljenditevaheliste seoste väljatoomist, näiteks lauses: *Alates [14. veebruarist] kordab ETV juba eetris olnud Lasteekraani osasid, [neli korda nädalas] ja [kaks osa päevas], [teisipäevast reedeni] [vahemikus 15.00 - 16.00] ning seda kuni [maikuu keskpaigani]*.	10 (4,9%)
Ainuüksi artiklis sisalduva informatsiooni põhjal ei olnud võimalik täpselt kindlaks määrata, millisele konkreetsele kuupäevale väljendid (nt *„Järgmisel päeval kella 15ks“, „Hommikul“*) viitavad.	7 (3,4%)
Kellaaeg anti täpsustava informatsiooni puudumise tõttu edasi poolikult (nt kas *kell seitse* on T07.00 või T19.00?).	6 (2,9%)
Väljendi lahendamine eeldas mingi sündmuse toimumisaja teadmist (nt spordivõistluse või telesaate toimumisaeg).	5 (2,4%)
Semantika määrati valesti üleliigse eraldamise tõttu. Nt lauses: *Tramme ja trolle sõidab [sügisel mullusest] märgatavalt vähem* eraldati kahe ajaväljendi asemel üks ning semantikat interpreteeriti kui väljendi *mullu sügisel* semantikat.	4 (1,9%)
Semantika määramata, kuna puudub vajalik arvutusoperatsioon. Nt *maikuu teisel pühapäeval, jaanuari viimasel nädalavahetusel.*	4 (1,9%)

Kõige sagedasemaks eksimuseks semantika lahendamisel (24,8% vigadest) oli suutmatus tuvastada juhtumeid, kus ajaväljendit kasutatakse üldises tähenduses (nt väljendit *täna* tähenduses *tänapäeval*). Kuna üldine tähendus tuleneb sageli ümbritsevast kontekstist, mitte väljendist endast, pakuvad käesolevas süsteemis implementeeritud vahendid vähe võimalusi üldise tähenduse eristamiseks. Võimalik, et probleemi aitaks lahendada masinõppe tehnikad, mis lubaksid mitme erineva kontekstitunnuse järgi eristada üldise tähenduse kasutust konkreetse tähenduse kasutusest (nagu katsetati töös [9]).

Suuruselt teiseks vigade allikaks arenduskorpuses oli eksimine ajaväljendi ankurdamisel teise ajaväljendi külge (31 viga, 15% vigadest). Üle poole vigadest (19 viga) moodustasid juhud, kui ankurdamine oli teostamata jäetud (nt üksikult esinevad kuud ja aastaajad nõudsid ankurdamist tekstis eelneva aastaarvu külge), 7 viga moodustasid juhud, ajaväljend oli valesti ankurdatud (st vale ajaväljendi külge) ning 5 juhul oli teise väljendi külge ankurdamine üleliigne (st väljendi semantika oleks tulnud leida dokumendi loomise kuupäeva suhtes). Ajaväljendite ankurdamise tehnika nõuab edasist uurimist, et selgitada välja juhud, millal on teise ajaväljendi külge ankurdamist võimalik (minimaalselt eksides) teostada.

Kolmas sage vigade põhjus arenduskorpusel oli ajaväljendite poolik eraldamine (28 viga, 13,6% vigadest). Oluliselt vähem vigu põhjustab ajaväljendikandidaatide üleliigne eraldamine (st kokkuliitmine, kui tegelikult peaksid ajaväljendid olema eraldiseisvad), andes 1,9% vigadest. Mõningatel juhtudel on reeglite lisamise ning eeltöötluse komponentide täiendamisega on võimalik pooliku ja üleliigse eraldamise vigu vähendada, kuid selliste vigade vältimine laiemas perspektiivis eeldab siiski süntaksianalüüsi olemasolu.

Vigu põhjustab ajavahemiku nimest koosnevate lühiväljendite (nt *aasta, nädala*) semantika määramata jätmine liitfraasides (nt *nädala keskel, aasta algul*): reeglid nende väljendite lahendamiseks ei ole süsteemis välja ehitatud.

Oluliseks probleemide allikaks on ka kasutatava märgendusformaadi mittetäielikkusest tulenevad vead: kasutatava märgendusformaadi järgi ei ole võimalik ajaväljendite-vaheliste seoste välja-toomine ning mõningate ajaliste korduvuste semantika edasiandmine. Näiteks pole võimalik väljendada ühe ajavahemiku kuulumist teise ajavahemiku sisse (*[6.-8. veebruarini] [kell 11.00-17.00]*) ning kordumiste arvu mingil ajaperioodil (nt *kaks korda nädalas*). Kui problemaatiliste korduvuste puhul jäeti käsitsi märgendamisel semantika määramata, siis üksteise sees paiknevate vahemike korral toodi välja sisemise vahemiku ilmutatud kujul semantika, kattes puuduva osa X sümbolitega, näiteks (referentsajaga 2010-02-01):

```
<TIMEX TYPE="INTERVAL" VALUE="2010-02-06TXX:XX"
VALUE2="2010-02-08TXX:XX"> 6.-8. veebruarini </TIMEX>
<TIMEX TYPE="INTERVAL" VALUE="XXXX-XX-XXT11:XX"
VALUE2="XXXX-XX-XXT17:XX"> kell 11.00-17.00 </TIMEX>
```

Sellise märgendusviisi tõttu peegeldub ka ajavahemike-vaheliste seoste määramatajätmine

hindamistulemustes.

Märgendusformaadi vigade alla saab ka liigitada katsed edasi anda väljendite *viimase viie aasta jooksul, lähima kahe aasta jooksul* jms semantikat ajavahemikena: selliste ajavahemike otspunkte pole alati võimalik üheselt määratleda. Näiteks, eeldades, et ajaväljend *lähima kahe aasta jooksul* on tulevikusuunaline ning referentsajaks on 2010-01-05, võib kasutada kahte märgendusviisi:

```
<TIMEX TYPE="INTERVAL" VALUE="2010-XX-XXTXX:XX"
VALUE2="2012-XX-XXTXX:XX"> lähima kahe aasta jooksul </TIMEX>
<TIMEX TYPE="INTERVAL" VALUE="2011-XX-XXTXX:XX"
VALUE2="2013-XX-XXTXX:XX"> lähima kahe aasta jooksul </TIMEX>
```

Problemaatiline on mõningate semantikat täpsustava eestäiendiga väljendite (nt *viimasel nädalal, eelmisel ööl, möödunud esmaspäeval*) lahenduskäik: selle kirjeldamine kõnehetkest sõltumatult ei pruugi alati olla võimalik. Näiteks, kui ajaväljendit *viimasel nädalal* kasutatakse esmaspäeval, siis mõeldakse tõenäoliselt *„lõppenud nädalat“*, ent sama väljendit nädala lõpus kasutades võidakse mõelda ka *„käesolevat“* ehk *„lõppevat nädalat.“* Käesolevas süsteemis pole võimalik luua semantikareegleid, mis sõltuksid mingi tingimuse täidetusest referentsajal (nt sellest, kas referentsaeg viitab nädala algusele või lõpule).

Süsteemis puuduvad ka operatsioonid ajaväljendite nagu *maikuu teisel pühapäeval*, *eelviimasel märtsikuu päeval*, *jaanuari viimasel nädalavahetusel* semantika leidmiseks. Laiemas plaanis eeldab selliste väljendite lahendamine kalendriseoseid täpsemalt modelleerivat ajamudelit (mudelit, mis võimaldaks nt leida *esimese, teise*, ..., *eelviimase* ja *viimase* nädalapäeva / kuupäeva suvalises kuus).

Tabelis 10 on toodud vead semantika määramisel uues testkorpuses. Tabel katab kõik testkorpuses tehtud vead. Kolm kõige sagedasemat vigade põhjust on jäänud samaks (vrd Tabel 9), küll on erinev nende suhteline sagedus: kõige sagedasem semantikavigade põhjustaja uues korpuses on poolik eraldamine ning ajaväljendite kasutust üldises tähenduses esineb oluliselt vähem. Uute vealiikidena täheldati uues testkorpuses katkistest reeglitest tulenevaid vigu ning morfoloogilise ühestamise viga.

Tabel 10. Kaheksa veatüüpi semantika määramisel uues testkorpuses.

Toodud on eksimuse kirjeldus ning esinemissagedus korpuses.

Eksimuse kirjeldus	Esinemissagedus
Semantika määrati valesti pooliku eraldamise tõttu.	8 (32%)
Eksiti ajaväljendi ankurdamisel teise ajaväljendi külge (üleliigne, puuduv või valesti teostatud ankurdamine).	7 (28%)
Ajaväljendit kasutati üldises tähenduses.	4 (16%)
Väljendi korrektne lahendamine eeldas mingi sündmuse toimumisaja teadmist.	2 (8%)
Semantika määrati valesti katkise reegli tõttu.	2 (8%)
Semantika valesti määramine oli põhjustatud morfoloogilise ühestamise veast.	1 (4%)
Ajaühiku nime sisaldava lühiväljendi semantika jäi määramata.	1 (4%)

4.5.5 Vead semantika täpsustamisel

Semantikat täpsustava atribuudi MOD määramise suhteliselt madal saagis (80,9% arenduskorpusel ning 57,9% uuel testkorpusel) on peamiselt põhjustatud semantika täpsustamata jätmisest sõna *paar* sisaldavates väljendites (nt *paari minutiga*) ning liitsõnalistes ajaväljendites (nt *varakevadel, hilisõhtul*). Probleem on tingitud sõnaklassi semantilise osa defineerimise piiratud võimalustest: sõnaklassi elemendi all saab defineerida ainult ühe semantikareegli. Seetõttu lisatakse näiteks sõna *varahommikul* semantikat defineerivasse elementi ainult kõige olulisem: *hommiku*-definitsioon (MO märgendina) ning jääb välja semantikat täpsustava liite *vara-* definitsioon.

4.5.6 Verbi grammatilise aja heuristiku eksimused

Erinevaid semantika lahendamise mudeleid võrdlevad eksperimendid (alampeatükis 4.1) näitasid, et kõige stabiilsemaid tulemusi andis otsimisoperatsiooni SEEK_IN kombineerimine verbi grammatilise ajaga – korrektsete lahenduste protsent jäi kõigi ajaväljendiliikide puhul vahemikku 88,3%-93,2%. Seega on tegu heuristikuga, mis väärib kindlasti edasist arendamist. Käesolevas alampeatükis vaatleme juhte, mil verbi grammatilise aja heuristik ei andnud korrektset lahendust. Selleks toome välja heuristiku vead, mis tehti nädalapäev-väljendite semantika leidmisel alampeatükis 4.1 kirjeldatud eksperimendis.

Kokku tehti verbi grammatilise aja heuristikut kasutades 32 viga, mis jagunesid järgmiselt:

- Kõige rohkem vigu (9 viga) tehti juhtudel, kui artiklis kasutati mineviku-sündmustest rääkimisel oleviku-aega. Nt. tekstikatkes *...nagu arvab ekslikult {persooni nimi} [teisipäevases] Postimehes...* on ajaväljendile lähimaks verbiks olevikku edasiandev *arvab*. Toodud näites viitab selgelt minevikule aga see, et ajaväljend annab edasi juba ilmunud ajalehe ilmumiskuupäeva.
- 8 viga tehti juhtudel, kui tulevikule viitavale ajaväljendile lähim verb oli minevikuajas ning seega otsiti ajaväljendile lahendust minevikust. Nt lausekatkes *[Pühapäevani] jäänud aega pidasid nad liiga lühikeseks...* . Tehtud vigadest 5 olid seotud täismineviku-ajaga (kasutati liitvorme *on oodatud, on avatud*), mis viitab sellele, et täismineviku kasutus tulevikusündmuste edasiandmisel vajab edasist uurimist.
- 7 viga tehti nädalapäev-ajaväljendi ankurdamata jätmisel. Näiteks lauses „*[Eelmise nädala esmaspäeval] teadsid sõudjad Pärnus, et [neljapäeva hommikuks] tuleb 3100 km kaugusel Brive'is mandaadis olla“* tuleb ajaväljend „*neljapäeva hommikuks“* ankurdada väljendi „*Eelmise nädala esmaspäeval“* külge. Tehtud vigadest 4 olid juhud, kus nädalapäev-väljend tuli ankurdada sellega kõrvutiasuva kuupäev-väljendi külge (nt „*reedel, 16. veebruaril“*).
- 6 viga tehti nädalapäev-väljendi sidumisel vale verbiga. Näiteks lause „*Kultuuriminister Raivo Palmaru teeb valitsusele ettepaneku eraldada [püha-päeval] Los Angeleses Grammy muusikaauhinnaga pärjatud Eesti muusikutele kokku 1 202 000 krooni*“ analüüsil tuleks ajaväljend ankurdada mineviku kesksõna *„pärjatud“* külge (ankurdati aga oleviku verbi *„teeb“* külge).
- 1 viga tehti morfoloogilise ühestamise eksimuse tõttu (ajaväljendile lähimat verbi tõlgendati nimisõnana).
- 1 viga tehti tsitaadi tõttu: ajaväljendi ja õiget grammatilist aega kandva verbi vahele jäi tsitaat, tsitaadini jõudes aga katkestab heuristik lähima verbi otsimise.

4.6 Edasiarendamisvõimalused

Käesolevas peatükis anname lühiülevaate võimalustest süsteemi edasisel arendamisel.

- Käesolevas töös kasutatud märgendusformaat ei ole täielik: problemaatiline on ajaliste korduvuste, ligikaudse semantikaga ajaväljendite ning ajaväljendite-

vaheliste seoste (nt ajavahemik ajavahemiku sees) semantika edasiandmine. Üheks potentsiaalseks arenguvõimaluseks on standardse märgendusviisi (TIMEX2 või TimeML) täiemahuline kasutuselevõtmine. See aitaks küll lahendada eeltoodud semantika esitamisega seotud probleeme, ent tuvastaja väljundi praktiliseks kasutamiseks oleks endiselt tarvis täiendavat interpreteerimist (nagu seda on tarvis ka käesoleva süsteemi väljundi kasutamisel). Praktilisemad rakendused võivad aga nõuda hoopis tuvastaja väljundi konkretiseerimist (nt päevaosade esitamist kellaajavahemikena, et neid saaks võrrelda muude kellaaegadega).

- Ajaväljendite eraldamise süsteemi on võimalik täiendada järgmiselt:
  1) Ajaväljend-liitsõnade (nt *esmaspäevahommikune*) tuvastamiseks on võimalik laiendada fraasimustri loogikat selliselt, et oleks võimalik üldistavalt kirjeldada liitsõna selle alamsõnade kaudu (nt `NADALAPAEVA` + `PAEVAOSANE`).
  2) Arvsõnafraaside tuvastamise etappi saaks laiendada selliselt, et leitakse üles ka umbmäärased kvantiteedid (*mõni, kümmekond* jms) ning harvemini ajaväljendites kasutatavad kvantiteedid (nt *10-miljonit*).
  3) Liitumisreegleid on võimalik täiendada selliselt, et ka ajavahemike moodustamine oleks kontrollitav liitumisreeglite abil.
  4) Automaatse süntaksianalüüsi kasutamine võiks lahendada paljud fraasipiiride ebatäpsest määramisest tulenevad vead.
- Ajaväljendite semantika avaldumisel on võimalik uurida:
  1) Ajaväljendi semantika ankurdamist teise ajaväljendi semantika külge.
  2) Ajaväljendi kasutust üldises ja konkreetses tähenduses.

     Mõlema eeltoodud probleemi puhul on oluline on leida, kas mingid ajaväljendi konteksti tunnused võimaldavad ennustada vastavalt siis väljendi ankurdamise vajadust või üldises tähenduses kasutust. Potentsiaalselt saaks siin rakendada masinõppe meetodeid, mis aga eeldaksid suuremahuliste treeningkorpuste loomist.
  3) Võimalusi verbi grammatilise aja heuristiku parendamiseks.
- Ajaväljendite semantika lahendamise süsteemi on võimalik täiendada järgmiselt:
  1) Lisada *sajand*- ja *sekund*-granulaarsuste edasiandmise võimalused.
  2) Kasutatavasse semantikareeglite hulka tuleks lisada operatsioonid, mis võimaldaksid konkreetse ajapunkti/ajalõigu osadeks jagamist ning *n*-inda alamosa

leidmist (nt väljendite *„mai teisel laupäeval“* või *„2009 aasta eelviimasel nädalal“* semantika leidmisel).

4.7 Kasutusvõimalused

Käesolevas peatükis anname lühiülevaate ajaväljendite tuvastaja kasutusvõimalustest. Ajaväljendite tuvastamist võib kasutada:

- Keeleressursside loomisel. Ajaväljendite tuvastajat saab kasutada ühe abivahendina semantiliselt märgendatud tekstikorpuse loomisel.
- Infootsingutel. Dokumente võib grupeerida vastavalt selle, milliste perioodidele viitavad dokumentides kasutatud ajaväljendid ning kasutaja saab seejärel infopäringus täpsustada, milliste perioodide kohta ta informatsiooni soovib.
- Automaatsel küsimustele vastamisel. Kasutades korpust, milles on märgendatud ajaväljendid, võib automaatne küsimustele vastav süsteem vastata näiteks *millal*- ja *kuna*- küsimustele.
- Dialoogisüsteemides. Tuvastajat võib kasutada modulaarse komponendina dialoogisüsteemi koosseisus, eesmärgiga leida kasutajasisendist ajaväljendeid.
- Automaatsel sisukokkuvõtete tegemisel. Automaatsel sisukokkuvõtete tegemisel on üheks võimaluseks pöörata rohkem tähelepanu just (konkreetseid) ajaväljendeid sisaldavatele lausetele, kuna need võivad pakkuda kirjeldatavate sündmuste käigu kohta olulist informatsiooni. Ajaväljendite tuvastamine koos sündmuste tuvastamisega võimaldaks automaatset kronoloogiate koostamist.

Kokkuvõte

Käesoleva töö eesmärgiks oli luua automaatne ajaväljendite tuvastaja eesti keelele. Süsteem on ülesehituselt reeglipõhine ning toetub automaatse morfoloogilise analüüsi ja ühestamise tulemustele. Reeglite koostamisel lähtuti eeskätt ajaväljendite kasutusest aja-kirjandustekstides.

Antud töö teoreetilises osas kirjeldati erinevaid ajaväljendite liigitusaluseid ning tutvustati ajaväljendite märgendamiseks kasutatavaid keeli. Samuti käsitleti erinevaid lähenemisi, mida on kasutatud teistes keeltes ajaväljendite tuvastajate loomisel.

Töö tuuma moodustab praktiline osa, milles arendati edasi autori bakalaureusetöös alustatud ajaväljendite tuvastajat. Ajaväljendite eraldamist teostavat osa süsteemis muudeti paindlikumaks ja efektiivsemaks (nt toodi sisse taaskasutatavad ja valikulised reegli alam-osad). Eesmärgiga leida eraldamisreeglite koostamiseks näiteid, katsetati ka aja-väljendifraaside automaatset kaevandamist korpustest.

Ajaväljendite semantika esitamiseks koostati TIMEX2 standardile tuginedes uus märgen-duskeel, milles oli võimalik esitada paindlikumalt mitmete konventsionaalsete ajaväl-jendite (päevaosade nimed, aastaajad) semantika ning eristada uusi ajaväljendiliike (ajali-sed kestvused ja korduvused). Siiski olid käesolevas töös kasutatud märgenduskeele väljendusvõimalused mõneti piiratud: problemaatiline oli umbmääraste ajaväljendite semantika edasiandmine, ajaliste korduvuste semantika edasiandmine ning ajaväljendite vaheliste seoste väljatoomine.

Ajaväljendite semantika täpsemaks lahendamiseks võeti süsteemis kasutusele uusi heuristikuid (nt verbi grammatilise aja arvestamine ning ajaväljendi semantika ankur-damine tekstis eelneva ajaväljendi semantika külge). Mitmese semantikaga nädalapäev-, kuu- ja kuupäev-ajaväljendite semantika lahendamiseks leiti ka heuristikute kombi-natsioon, mis eksperimendiks kogutud katsekorpusel lahendas antud väljendite semantika üle 90% korrektsusega.

Ajaväljendite tuvastaja jooksvaks testimiseks märgendati 47652-sõnaline ajakirjanduskorpus (1700 ajaväljendit) ning lõplikuks testimiseks valiti ajalehe tervikväljaanne (10617 sõna, 385 ajaväljendit). Testimisel leiti, et loodud süsteemi juures oli kõige problemaatilisem ajaväljendite eraldamise suhteliselt madal saagis (arenduskorpusel 84%, tundmatul tekstil 75,3%), ajaväljendite eraldamise täpsus oli aga suhteliselt kõrge (mõlemal korpusel ~98%). Ajaväljendite semantika normaliseerimise tulemused olid võrreldavad teiste keelte jaoks loodud süsteemide tulemustega: arenduskorpusel mõõdeti saagiseks ja täpsuseks ~86%, testkorpusel olid saagis ja täpsus ~91%.

Resolution of Estonian Temporal Expressions

Master thesis
Siim Orasmaa
Abstract

The purpose of this Master thesis was to develop an automatic temporal expression recognizer and resolver for Estonian language texts. This work is the continuation of author's Bachelor Thesis, where first steps towards building this system were made.

A rule-based approach was used to extract temporal expressions from text and normalize expressions's semantics. Estonian morphological analyzer and disambiguator was used to preprocess the input text. Extraction rules were defined as phrase patterns consisting of templates for words (describing a word by lemma or regular expression) and descriptions of numeral phrases. The word templates could be defined as reusable across the rules, which helped to reduce the number of patterns. Extraction rules were mostly used to recognize phrases with unchangeable word order and another set of rules – called compounding rules – were combined with built-in heuristics to join extracted phrases into longer temporal expressions. In order to gather examples of Estonian temporal expressions, a phrase mining technique was used.

A subset of TIMEX2 standard was used as a basis of annotation scheme. Some deviations from the standard were made, most importantly, the ways how to represent semantics of temporal intervals and fuzzy temporal expressions were changed.
The semantics of an expression were defined as a sequence of instructions, which needed to be executed in order to normalize the expression. Instructions allowed to perform calendar arithmetics (based on the Java library Joda Time), anchor temporal expressions and set attribute values of the annotation. Also, special heuristics were developed, which allowed interpretation of semantically ambigious weekday names, month names and dates with relatively high accuracy (90%) on a corpus of newspaper texts.

Measurements on a test corpus (Estonian newspaper articles) showed that on detection of

temporal expressions, the system achieved recall 75.3% and precision 98.6%, while normalization of temporal expressions (filling an attribute similar to TIMEX2 attribute VAL) was achieved with both recall and precision 91.5%.

Kirjandus

[1] E.Saue. *Eestikeelsete ajaväljendite automaatne eraldamine.* Bakalureusetöö. Juhendaja: M.Treumuth. Tartu Ülikool, Matemaatika-informaatikateaduskond, 2007.

[2] E.Saquete, R. Muñoz ja P. Martínez-Barco. *TERSEO: Temporal Expression Resolution System Applied to Event Ordering*. Proceedings of the 6th International Conference, TSD 2003, Text, Speech and Dialogue. Ceske Budejovice, Czech Republic, lk. 220-228, 2003.

[3] B. Han, D. Gates and L. Levin. *From Language to Time: A Temporal Expression Anchorer.* Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning, 2006.

[4] M. T. Vicente-Díez, D. Samy and P.Martínez. *An empirical approach to a preliminary successful identification and resolution of temporal expressions in Spanish news corpora*. Proceedings of the Sixth International Language Resources and Evaluation. Marrakech, Morocco, 2008.

[5] L.Ferro, L.Gerber, I.Mani, B.Sundheim, G.Wilson. *TIDES 2005 Standard for the Annotation of Temporal Expressions*, 2005.
http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf
(Viimati vaadatud 22.02.2010)

[6] I.Mani, *Recent Developments in Temporal Information Extraction.* Proceedings of Recent Advances in Natural Language Processing ’03, 2004.

[7] R.Saurí, J.Littman, B.Knippen, R.Gaizauskas, A.Setzer, J.Pustejovsky. *TimeML Annotation Guidelines*, 2006.
http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf
(Viimati vaadatud 22.02.2010)

[8] M.Erelt, R.Kasik, H.Metslang, H.Rajandi, K.Ross, H.Saari, K.Tael, S.Vare. *Eesti keele grammatika 2., Süntaks*. Eesti Teaduste Akadeemia, Keele ja Kirjanduse Instituut, Tallinn, 1993, lk 76-86.

[9] I.Mani, G.Wilson. *Robust temporal processing of News*. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong Kong, 2000.

[10] M.Negri, L.Marseglia. *Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004*. Tehniline raport. ITC-irst, Trento, 2004.
http://tcc.itc.it/people/negri/papers/TERN-2004/Final-TERN-irst.pdf
(Viimati vaadatud: 12.02.2010)

[11] J. A. Baldwin. *Learning temporal annotation of French news*. Magistritöö. Graduate School of Arts and Sciences. Georgetown University, Washington, DC, 2002.
www.jenniferbaldwin.com/jb/documents/Thesis.pdf
(Viimati vaadatud 25.02.2010)

[12] D.Ahn, J. van Rantwijk, M. de Rijke. *A Cascaded Machine Learning Approach to Interpreting Temporal Expressions.* Proceedings of the Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, 2007.

[13] O.Craveiro, J.Macedo, H.Madeira. *Use of Co-occurrences for Temporal Expressions Annotation*. Proceedings of the 16th International Symposium on String Processing and Information Retrieval. Saariselkä, Finland, 2009.

[14] M.Treumuth. *Normalization of Temporal Information in Estonian*. Proceedings of the 11th international conference on Text, Speech and Dialogue. Brno, Czech Republic, 2008.

[15] S.Orasmaa. *Ajaväljendite tuvastamine eestikeelses tekstis*. Bakalaureusetöö. Juhendaja: M.Treumuth. Tartu Ülikool, Matemaatika-informaatikateaduskond, 2008. http://lepo.it.da.ut.ee/~soras/bakatoo.pdf (Viimati vaadatud 20.04.2010)

[16] Heiki-Jaan Kaalep, Tarmo Vaino. *Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine*. Keel ja Kirjandus 1/1998, lk 30-38.

[17] M.Erelt, T.Erelt, K.Ross. *Eesti keele käsiraamat*. Eesti Keele Sihtasutus, Tallinn, 1997.
http://julia.eki.ee/books/ekkr/m86.html
http://julia.eki.ee/books/ekkr/m87.html
http://julia.eki.ee/books/ekkr/m88.html
http://julia.eki.ee/books/ekkr/m89.html
http://julia.eki.ee/books/ekkr/m90.html
(Viimati vaadatud: 08.03.2010)

[18] A.Berglund. *Extracting Temporal Information and Ordering Events for Swedish*. Magistritöö. Lund University, Lund, 2004.

[19] P.Mazur, R.Dale. *What's the date?: high accuracy interpretation of weekday names*. COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1, 2008.

[20] *The TERN 2004 Evaluation Plan: Time Expression Recognition and Normalization*, 2004.
http://fofoca.mitre.org/tern_2004/tern_evalplan-2004.29apr04.pdf
(Viimati vaadatud: 03.05.2010)

Lisad

Lisa 1 Näide kaevandatud fraasidest

Lisa 2 Kasutatud märgendusformaat

Lisa 3 Semantikareeglite liigid

Lisa 4 Ajaväljendite sagedusprofiilid korpustes

Lisa 5 Materjalide CD

Lisa 1 Näide kaevandatud fraasidest

Järgnevalt on toodud näide regulaaravaldise `.*(kevad).*` järgi kaevandatud fraasidest tasakaalustatud korpuse ajalehti sisaldavast alamkorpusest.[1] Kaevandamisel on fraasi esinemissageduse alampiiriks valitud 5 ning fraasi maksimaalne pikkus sõnades on samuti 5. Looksulud märgivad võtmesõna (st sõna, millega regulaaravaldis sobitus). Fraasi ees sulgudes olev number on fraasi esinemissagedus korpuses.

```
 {kätkevad}
(7) {kätkevad}

 {kevad}
(5) ja {kevad}
(82) {kevad}
(6) {kevad} hiinas
(5) {kevad} [d][d][d][d]

 {kevade}
(6) praha {kevade}
(81) {kevade}
(5) {kevade} poole

 {kevadega}
(11) {kevadega}

 {kevadeks}
(5) [d][d][d][d]. aasta {kevadeks}
(7) aasta {kevadeks}
(28) {kevadeks}

 {kevadel}
(7) eelmise aasta {kevadel}
(6) selle aasta {kevadel}
(88) [d][d][d][d]. aasta {kevadel}
(6) möödunud aasta {kevadel}
(121) aasta {kevadel}
(6) [d][d][d][d] {kevadel}
(5) [d][d].a. {kevadel}
(9) läinud {kevadel}
(14) eelmisel {kevadel}
(8) tuleval {kevadel}
(5) ütles {kevadel}
(10) igal {kevadel}
(18) mullu {kevadel}
(37) tänavu {kevadel}
(14) järgmisel {kevadel}
(6) ja {kevadel}
(59) sel {kevadel}
(13) möödunud {kevadel}
(680) {kevadel}
(5) {kevadel} tegi
(6) {kevadel} eesti
(13) {kevadel} [d][d][d][d]
```

1 Tasakaalus korpus: http://www.cl.ut.ee/korpused/grammatikakorpus/ (Viimati vaadatud: 09.02.2010)

(15) {kevadel} ja
(7) {kevadel} ja suvel

{kevadeni}
(24) {kevadeni}

{kevadesse}
(6) {kevadesse}

{kevadest}
(10) aasta {kevadest}
(50) {kevadest}
(9) {kevadest} saadik

{kevadet}
(8) enne {kevadet}
(38) {kevadet}

{kevadine}
(27) {kevadine}

{kevadise}
(32) {kevadise}

{kevadised}
(5) {kevadised}

{kevadisel}
(14) {kevadisel}

{kevadisest}
(12) {kevadisest}

{kevadisi}
(10) {kevadisi}

{kevadist}
(11) {kevadist}

{kevadiste}
(13) {kevadiste}

{kevadistel}
(7) {kevadistel}

{kevaditi}
(14) {kevaditi}

{kevadpäevad}
(7) {kevadpäevad}

{kevadpäevade}
(17) {kevadpäevade}

{kevadpäevi}
(5) {kevadpäevi}

{kevadringi}
(6) {kevadringi}

```
 {kevadsuvel}
(6) {kevadsuvel}

 {kevadtalvel}
(5) [d][d][d][d]. aasta {kevadtalvel}
(5) aasta {kevadtalvel}
(13) {kevadtalvel}


 {puhkevad}
(7) {puhkevad}


 {varakevadel}
(21) {varakevadel}
```

Lisa 2 Kasutatud märgendusformaat

Ajaväljendite märgendamisel kasutatud formaat tugineb TIMEX2-standardile[1], ent ei laienda seda täielikult, vaid kasutab ainult alamosa, teatavate modifikatsioonidega. Järgnevalt kirjeldatakse kasutatud märgendusformaati täpsemalt ning tuuakse välja erinevused TIMEX2 standardist.

1. Märgendus

Ajaväljendite annoteerimiseks kasutatakse `TIMEX`-märgendeid. Tekstist leitud ajaväljendifraas ümbritsetakse `TIMEX`-märgenditega ning märgendi atribuutides tuuakse välja ajaväljendi semantika (vt Näide 1).

Näide 1. Annoteeritud tekstilõik. Teksti loomise ajaks on 2009-12-07.

```
Linna maksutulu võib <TIMEX TYPE="POINT" VALUE="2010-XX-XXTXX:XX">
tuleval aastal</TIMEX> langeda kuni 300 miljonit krooni.
```

Märgenduse ulatus peaks hõlmama kõiki väljendi ajalist tähendust täpsustavaid ees- ja järeltäiendeid (nt eestäiendid *umbes, rohkem kui* ning järeltäiendid *lõpus, keskosas*). Erinevalt TIMEX2 formaadist, ei kasutata pesastatud (*nested*) märgendamist (s.o märgendusviis, kus ühe märgenduse sees tuuakse veel välja alammärgendusi).
Märgendamisel määratakse ajaväljendi liik (atribuudis `TYPE`), esitatakse ajaväljendi semantika (atribuudis `VALUE`) ning vajadusel semantika täpsustus (atribuudis `MOD`). Ajavahemike puhul kirjeldavad atribuudid `VALUE` ja `MOD` ajavahemiku alguspunkti ning lõpp-punkti kirjeldamiseks kasutatakse täiendavaid atribuute `VALUE2` ja `MOD2`. Problemaatiliste ajaväljendite puhul antakse atribuudis `COMMENT` lühikommentaar probleemi kohta (käsitsi märgendamisel).

2. Ajaväljendite tüübid

Ajaväljendi liigi määrab atribuut `TYPE`. Atribuudi kasutamine on kohustuslik ning sellel võivad olla järgmised väärtused:

- `POINT` – ajapunkt. Ajapunktide granulaarsusastmed võivad olla erinevad, seega loetakse ajapunktiks ühtviisi nii ajaväljend *järgmisel reedel kell 14.00* kui ka

1 L.Ferro, L.Gerber, I.Mani, B.Sundheim, G.Wilson. *TIDES 2005 Standard for the Annotation of Temporal Expressions*, 2005.

ajaväljend *2004. aastal*.

- `INTERVAL` – ajavahemik, mis on defineeritud kahe ajapunkti kaudu. Näiteks: *esmaspäeva hommikust kolmapäeva pärastlõunani.*
- `RECURRENCE` – ajaline korduvus. Näiteks: *neljapäeviti*, *hommikuti*
- `DURATION` – ajaline kestvus, mille alguspunkt ja lõpp-punkt on määramata. Näiteks *3 tundi ja 14 minutit.*
- `UNK` – tundmatu väljend. Reserveeritud juhtudeks, kui semantika lahendamine süsteemis ebaõnnestub.

TIMEX2 standard ajaväljendi liiki eraldi esile ei too, seal määrab „liigi" atribuudi `VALUE` formaat.

3. Semantika esituskuju

Põhiosa ajaväljendi semantikast esitatakse atribuudis `VALUE`. Kui ajaväljendi tüüp on `INTERVAL`, tuuakse atribuudis `VALUE` ajavahemiku alguspunkti semantika ning atribuudis `VALUE2` ajavahemiku lõpp-punkti semantika. Konkreetse semantika esitamisel kasutatakse kolme formaati:

1. Kuupõhine formaat:

   `yyyy-mm-ddThh:mm`

   `yyyy` – 4-kohaline aastaarv `hh` – tund päevas (00-23)

   `mm` – kuu aastas (01-12) `mm` – minut tunnis (00-59)

   `dd` – kuupäev (01-31)

2. Nädalapõhine formaat:

   `yyyy-Wnn-wdThh:mm`

   `nn` – nädal aastas (01-53)

   `wd` – päev nädalas (1-7, kus 1 on esmaspäev ja 7 on pühapäev)

3. Ajalise kestvuse formaat:

   $Pn_1Yn_2Mn_3Wn_4DTn_5Hn_6M$

   kus n_i märgib arvu ning `Y`, `M`, `W`, `D`, `H`, `M` vastavat ajaühikut/granulaarsust (aasta, kuu, nädal, päev, tund, minut);

Formaate 1 ja 2 kasutatakse ajapunktide, ajavahemike ja ajaliste korduvuste semantika esitamisel, formaat 3 on mõeldud ajaliste kestvuste jaoks. Kui ajaväljendi semantika on esitatav ühtviisi nii kuupõhises kui ka nädalapõhises formaadis, eelistatakse kuupõhist formaati.

Lisaks arvudele kasutatakse formaatides 1 ja 2 järgmisi kokkuleppelisi eritähiseid:

- Päevaosad – kasutatakse kellaaja (`hh:mm`) asemel
  - `MO` – *morning* – hommik
  - `AF` – *afternoon* – pärastlõuna
  - `EV` – *evening* – õhtu
  - `NI` – *night* – öö (kui on antud ka kuupäev/nädalapäev, mõeldakse ööd päeva alguses)
  - `DT` – *daytime* – päevane aeg
- Nädalavahetus/tööpäev - kasutatakse nädalapõhises formaadis, nädalapäeva (`wd`) asemel;
  - `WD` – *workday* – tööpäev (TIMEX2 ei kasuta seda tähistust)
  - `WE` – *weekend* – nädalalõpp
- Aastaajad – kasutatakse kuupõhises formaadis kuu (`mm`) asemel:
  - `SP` – *spring* – kevad
  - `SU` – *summer* – suvi
  - `FA` – *fall* – sügis
  - `WI` – *winter* – talv (kui on antud aastaarv, mõeldakse talve aasta alguses)
- Kvartalid – kasutatakse kuupõhises formaadis kuu (`mm`) asemel:
  - `Q1` – *1st quarter* – 1. kvartal
  - `Q2` – *2nd quarter* – 2. kvartal
  - `Q3` – *3rd quarter* – 3. kvartal
  - `Q4` – *4th quarter* – 4. kvartal
  - `QX` – teadmata kvartal

Semantika esituses tuuakse välja (***avatakse***) vaid granulaarsused, mille kohta leidub piisavalt informatsiooni kas väljendis endas või sellega seotud referentsajas. Ülejäänud granulaarsused jäävad `X` sümbolitega kaetuks (***suletuks***). Näiteks, ajaväljendis *„selle kuu alguses“* ei peitu informatsiooni kuupäeva ning kellaaja kohta, seetõttu on need granulaarsused semantika esitamisel suletud:

(Referentsajaks on 2009-12-17)

```
<TIMEX TYPE="POINT" VALUE="2009-12-XXTXX:XX" MOD="START"> selle kuu
```

```
alguses </TIMEX>
```

Referentsaja põhjal avatakse enamasti vaid granulaarsused, mis on suuremad ajaväljendis ilmutatud kujul olevatest granulaarsustest,[2] nagu *aasta*-granulaarsus eeltoodud näites.

Aastaarvu alamosade kinnikatmisega antakse edasi aastakümneid ja sajandeid, nt:

```
<TIMEX TYPE="POINT" VALUE="199X-XX-XXTXX:XX" MOD="END"> 1990ndate
lõpus </TIMEX>
<TIMEX TYPE="POINT" VALUE="17XX-XX-XXTXX:XX"> 18. sajandil </TIMEX>
```

Ajaliste korduvuste puhul määratakse `VALUE`-väärtus vaid siis, kui seda saab väljendada formaatides 1 või 2. Ning suletud on siis mittekorduvad ajalised granulaarsused, nt

```
<TIMEX TYPE="RECURRENCE" VALUE="XXXX-WXX-2TXX:XX"> teisipäeviti
</TIMEX>
```

Kui ajalist korduvust ei ole võimalik eeltoodud viisil väljendada (nt ei saa selliselt anda edasi ajaväljendite *igal tunnil, kaks korda nädalas* semantikat), jäetakse semantika määramata.

Puuduolevate ajaliste granulaarsuste käsitlemisviis erineb TIMEX2 märgendusformaadis kasutatust: seal jäetakse puuduolevad ajalised granulaarsused `VALUE` osast üldse välja. Puuduolevate granulaarsuste väljajätmist kasutab käesolev märgendusformaat vaid ajaliste kestvuste korral, nt:

```
<TIMEX TYPE="DURATION" VALUE="PT3H"> kolm tundi </TIMEX>
```

4. Semantika täpsustamine

Ajaväljendi semantikat on võimalik täpsustada mittekohustuslikus atribuudis `MOD`. Atribuudil võivad olla järgmised väärtused:

- `START` – ajaväljend viitab ajapunkti algusosale (st – mõeldakse ajapunkti väikseima avatud granulaarsuse algusosa).

  Näiteks:

  ```
  <TIMEX TYPE="POINT" VALUE="2009-XX-XXTXX:XX" MOD="START">
  2009. aasta alguses </TIMEX>
  <TIMEX TYPE="POINT" VALUE="2007-06-XXTXX:XX" MOD="START">
  2007 juuni algus </TIMEX>
  ```

- `MID` – ajaväljend viitab ajapunkti keskosale.
- `END` – ajaväljend viitab ajapunkti lõpuosale.

2 Erandiks on juhud, mil semantikat on täpsustatud, nt: *täpselt kuu aega tagasi, täpselt 100 aastat hiljem*.

- FIRST_HALF – märgib, et mõeldakse ajapunkti väikseima avatud ajalise granulaarsuse „esimest poolt“. Näiteks:

```
<TIMEX TYPE="POINT" VALUE="2009-XX-XXTXX:XX"
 MOD="FIRST_HALF"> 2009. aasta esimesel poolel </TIMEX>
```

  TIMEX2 ei toeta selle väärtuse kasutamist. Poolaastate märkimiseks kasutatakse küll märke H1 ja H2, ent need ei laiene teistele granulaarsustele.

- SECOND_HALF – märgib, et mõeldakse väikseima avatud ajalise granulaarsuse „teist poolt.“ TIMEX2 ei toeta selle väärtuse kasutamist.

- APPROX – märgib, et toodud ajaväljendi semantika ei ole täpne, vaid varieerub väikseima VALUE-osas avatud või olemasoleva ajalise granulaarsuse suhtes. Näiteks:

```
<TIMEX TYPE="DURATION" VALUE="P4Y" MOD="APPROX">
umbes 4 aastat </TIMEX>
```

  Tähist APPROX kasutatakse ka kokkuleppeliselt mõningate hägusate väljendite semantika edasiandmisel, seda käsitletakse jaotises 6.

5. Ajavahemike käsitlemine

TIMEX2 standard ei toeta ajavahemiku märgendamist tervikuna, vaid nõuab otspunktide väljatoomist eraldiseisvatena, nagu on toodud järgnevas näitelauses:

```
That emergency clinic is open from <TIMEX2 VAL="T19:00">7
p.m.</TIMEX2> to <TIMEX2 VAL="T07:00">7 a.m.</TIMEX2>
```

Käesolevas töös kasutatud märgendusviis esitab tervikliku ajavahemikuna kõik ajaväljendifraasid, mis sisaldavad ilmutatud kujul vahemiku alguspunkti ja lõpp-punkti. Näiteks:

```
<TIMEX TYPE="INTERVAL" VALUE="2009-03-12TXX:XX"
VALUE2="2009-03-15TXX:XX"> 12-15 märts 2009 </TIMEX>
```

Lisaks ilmutatud kujul otspunkte sisaldavatele fraasidele kasutatakse ajavahemik-esituskuju ka fraaside puhul, kus on võimalik otspunktid fraasi sisu ja referentsaja järgi tuletada. Järgnevas näites on referentsajaks 2009-12-10:

```
<TIMEX TYPE="INTERVAL" VALUE="2009-01-XXTXX:XX"
VALUE2="2009-03-XXTXX:XX"> tänavu kolme esimese kuu jooksul
</TIMEX>
```

Ainult ajaliste kestvuste vahemike märgendamisel tuuakse otspunktid eraldi välja, näiteks ajaväljend „*6-8 kuud*“ märgendatakse kujul:

```
<TIMEX TYPE="DURATION" VALUE="P6M">6-</TIMEX>
<TIMEX TYPE="DURATION" VALUE="P8M">8 kuud</TIMEX>
```

6. Hägusa semantika esitus

Kui ajaväljendis ei leidu kalendrilist informatsiooni, küll aga saab eristada, kas on tegu viitega minevikule, olevikule või tulevikule, kasutatakse `VALUE` osas vastavalt kokkuleppelisi väärtuseid `PAST_REF`, `PRESENT_REF` ja `FUTURE_REF`. Näiteks:

```
<TIMEX TYPE="POINT" VALUE="PAST_REF"> hiljuti </TIMEX>
<TIMEX TYPE="POINT" VALUE="FUTURE_REF"> tulevikus </TIMEX>
```

Kokkuleppeliselt loetakse sellised viited ajapunktide alla kuuluvateks. TIMEX2 lubab ka täpsustada ankurhetke, millelt need viited lähtuvad. Käesolevas töös kasutatud märgendusviis ankurhetke välja ei too, vaikimisi eeldatakse, et ankurhetkeks on dokumendi loomise kuupäev.

Kui hägus ajaväljend sisaldab informatsiooni mingisuguse ajalise granulaarsuse kohta, püütakse märgenduse `VALUE` osas vastav granulaarsus ka avada, ning kasutatakse tähist `MOD="APPROX"` märkimaks granulaarsuse väärtuse „varieeruvust.“ TIMEX2 formaadist on see märgendusviis erinev: seal tähistab `APPROX` peamiselt umbmäärasusele viitavate märksõnade *approximately, around* jms esinemist fraasis.

Järgnevalt tuuakse ajaväljendid, mille puhul kasutatakse tähist `MOD="APPROX"` varieeruvuse märkimiseks, ning kirjeldatakse nende semantika kokkuleppelist esituskuju:

- *aastaid/kuid/nädalaid/... tagasi; aastate/kuude/nädalate/... eest/tagune;*
- *aastaid/kuid/nädalaid/... varem;*
- *aastate/kuude/nädalate/... pärast; aastaid/kuid/nädalaid hiljem;*
    - Semantika „lahendamisel“ lahutatakse referentsajast (liidetakse referentsajale) vastava granulaarsuse kaks ühikut ning semantika esitamisel tähistatakse väärtustusega `MOD="APPROX"` granulaarsuse varieeruvust. Järgnevates näidetes on referentsajaks 2010-03-25:

      ```
      <TIMEX TYPE="POINT" VALUE="2010-01-XXTXX:XX" MOD="APPROX">
      kuude tagune </TIMEX>
      <TIMEX TYPE="POINT" VALUE="2012-XX-XXTXX:XX" MOD="APPROX">
      aastate pärast </TIMEX>
      ```
- *mõned/mitmed aastad/kuud/nädalad/... tagasi/varem;*
- *mõni aasta/kuu/nädal/... tagasi/varem;*

- *mõned/mitmed aastad/kuud/nädalad/... hiljem;*
- *mõne/mitme aasta/kuu/nädala/... pärast;*
    - Analoogselt eelmise punktiga: semantika lahendamisel lahutatakse referents-ajast (liidetakse referentsajale) vastava granulaarsuse kaks ühikut ning semantika esitamisel tähistatakse väärtustusega `MOD="APPROX"` granulaarsuse varieeruvust.
- *eelmistel/möödunud/minevatel/eelnevatel/... aastatel/kuudel/nädalatel/...*
- *järgmistel/tulevatel/eelolevatel/... aastatel/kuudel/nädalatel/...*
    - Semantika lahendatakse selliselt, nagu oleks tegemist ainsuses variandiga ajaväljendist (nt *eelmisel päeval, tuleval aastal*), seejärel kasutatakse atribuudi väärtustust `MOD="APPROX"` varieeruvuse märkimiseks. Järgnevas näites on referentsajaks 2010-03-25:

```
<TIMEX TYPE="POINT" VALUE="2010-04-XXTXX:XX" MOD="APPROX">
eelolevatel kuudel </TIMEX>
```

Eeltoodud hägusa semantika esitusviisid tähise `MOD="APPROX"` abil on ajutised ning süsteemi edasise arendamise käigus on oluline leida paremad esitusviise, mis jätaksid hägusate ajaväljendite interpretatsiooni lõppkasutaja otsustada.

Lisa 3 Semantikareeglite liigid

Alljärgnev tabel toob kasutatud semantikareeglite liigid, operatsioonide poolt nõutud atribuudid ning operatsioonide kirjeldused.

Kalendriaritmeetikat teostavad operatsioonid toetuvad Java teegi Joda Time[1] klassidele `org.joda.time.LocalDate` ning `org.joda.time.LocalTime`.

Kõik tabelis olevad operatsioonid nõuavad järjekorda määrava atribuudi `priority` olemasolu, seetõttu ei ole seda tabelis eraldi välja toodud.

Operatsioon	Atribuutide funktsioonid	Kirjeldus
`SET` (omistamis-operatsioon)	`semField` – muudetav kalendriväli. `semValue` – kalendrivälja uus väärtus.	Kirjutab kalendrivälja `semField` väärtuse üle väärtusega `semValue`.

`ADD`
(liitmis-operatsioon)
`SUBTRACT`
(lahutamis-operatsioon)
	`semField` – kalendriväli, mille väärtust suurendatakse.
	`semValue` – kalendriväljale liidetav väärtus.
		Suurendab kalendrivälja `semField` väärtust `semValue` võrra. Operatsiooni `SUBTRACT` korral korrutatakse `semValue` enne liitmist läbi väärtusega -1.

`SEEK` (referentsaega välistav otsimis-operatsioon; otsimissuuna mitteleidmisel rakendub `SET` operatsioonina)	`semField` – muudetav kalendriväli; `direction` – otsimissuund: -1 – minevik, +1 – tulevik, VERBI_AEG – suund võetakse lähima verbi grammatilisest ajast; `semValue` – otsitav kalendrivälja väärtus.	Leiab etteantud suunast (`direction`) referentsajale lähima ajahetke, kus kalendrivälja `semField` väärtuseks on `semValue`. Referentsaeg ise arvatakse sobivate ajahetkede seast välja. Kui suunaks on grammatiline aeg, aga lausest ei leita grammatiliste ajatunnustega verbi, rakendub operatsioon kui vastavate atribuutidega `SET` operatsioon.
`SEEK_IN` (referentsaega mittevälistav otsimis-operatsioon; otsimissuuna	`semField` – muudetav kalendriväli; `direction` – otsimissuund: -1 – minevik, +1 – tulevik, VERBI_AEG – suund võetakse lähima verbi	Leiab etteantud suunast (`direction`) referentsajale lähima ajahetke, kus kalendrivälja `semField` väärtuseks on `semValue`. Ka referentsaeg võib olla

1 Vt http://joda-time.sourceforge.net/ (27.04.2010)

mitteleidmisel rakendub `SET` operatsioonina)	grammatilisest ajast; `semValue` – otsitav kalendrivälja väärtus.	sobivaks kandidaadiks. Kui suunaks on grammatiline aeg, aga lausest ei leita grammatiliste ajatunnustega verbi, rakendub operatsioon kui vastavate atribuutidega `SET` operatsioon.
`BALDWIN_WINDOW` (spetsiifiline otsimisalgoritm)	`semField` – muudetav kalendriväli; `semValue` – otsitav kalendrivälja väärtus.	Moodustab ainult unikaalseid `semField` väärtuseid sisaldava ajaakna, mille keskpunktiks on referentsaeg. Lahenduseks otsib akna seest ajapunkti, kus kalendrivälja `semField` väärtuseks on `semValue`. Kui akna seest nõutud omadustega ajapunkti ei leia, rakendub operatsioon kui vastavate atribuutidega tavaline `SET` operatsioon.
`BEGIN_INTERVAL` (ajavahemiku alguspunkti loomine)	(*atribuute ei nõua*)	Viib semantika ajapunkti kujult ajavahemiku kujule. Kui ajapunkti on juba eelnevalt muudetud, moodustatakse vahemiku alguspunkt ja lõpp-punkt muudetud ajapunkti kopeerides. Käsule järgnevad semantikareeglid (kuni reeglini `END_INTERVAL`) rakenduvad ajavahemiku alguspunkti peal.
`END_INTERVAL` (ajavahemiku lõpp-punkti loomine)	(*atribuute ei nõua*)	Suunab järgnevad semantikareeglid aja-vahemiku lõpp-punkti muutma.
`ASSIGN_TYPE` (ajaväljendi tüübi muutmine)	`semValue` – ajaväljendi uus tüüp (`DURATION` või `RECURRENCE`);	Muudab ajaväljendi tüübi ajaliseks kestvuseks või korduvuseks.
`SET_MOD` (TIMEX atribuudi `MOD` ülekirjutamine)	`semValue` – omistatav `MOD` väärtus (`FIRST_HALF`, `SECOND_HALF`, `START`, `MID`, `END` või `APPROX`);	Kirjutab TIMEX atribuudi `MOD` üle väärtusega `semValue`.
`SET_VAL` (TIMEX atribuudi	`semValue` – omistatav `VALUE` väärtus (`PAST_REF`, `PRESENT_REF` või	Kirjutab TIMEX atribuudi `VALUE` üle väärtusega

VALUE ülekirjutamine)	FUTURE_REF);	semValue. Erinevalt SET operatsioonist ei pea uus väärtus olema kalendrivälja väärtuseks sobiv, vaid võib olla suvaline sõne.
COVER_VAL (TIMEX atribuudi VALUE alamosa kinnikatmine)	semField – muudetav kalendriväli. semValue – kinnikaetavate sümbolite arv.	Katab teatava osa kalendriväljast X sümbolitega. Kinnikaetavate sümbolite arvuks on semValue sümbolit ning kinnikatmist alustatakse vasakult.
ANCHOR_TIMEX (ajaväljendi ankurdamine teise ajaväljendi külge)	direction – suund tekstis, millelt ankruks olevat ajaväljendit otsitakse: -1 käesolevast väljendist eespool, 1 käesolevast väljendist tagapool semField – ankruks sobivalt väljendilt nõutav granulaarsus.	Otsib tekstist suunalt direction esimese ajaväljendi, mis sisaldab kalendrivälja semField, muutvaid operatsioone, ning ankurdab käesoleva ajaväljendikandidaadi külge. Lisaks kalendrivälja olemasolule kitsendavad ankurdamist järgmised tingimused: 1) ankruks olev kandidaat ei või kuuluda käesoleva kandidaadiga samasse kandidaatidepuusse; 2) ankruks olev kandidaat ei või olla tüüpi DURATION (st, ei tohi sisaldada vastavat tüüpi seadistavat semantikareeglit); Pärast edukat ankurdamist võetaks käesoleva ajaväljendikandidaadi semantika lahendamisel aluseks ankruks oleva kandidaadi semantika lahendus.

Täiendavad märkused semantikareeglite kasutamise kohta:

- Atribuudi semValue asemel kasutatakse atribuuti semLabel, kui määratav väärtus on kalendriväline (nt väärtus SP (*spring*) tähistamas aastaaega kevad).
- Atribuudi semValue väärtustena võib kasutada järgmisi viiteid:
  - REF:*n* – väärtuseks võetakse sõnamalliga sobitunud sõna või selle alamosa. Kasutatakse regulaaravaldis-sõnamalli ning arvsõnafraas-sõnamalli semantilise osa defineerimisel. Regulaaravaldise puhul tähistab *n* seda, mitmendast

alamavaldisest[2] tuleb väärtus võtta. Arvsõnafraas-sõnamalli puhul saab väärtuseks arvsõnafraasi semantika (täisarv või murdarv), *n* väärtusena oodatakse vaikimisi väärtust 1.

- PARSE_FROM_SELF:*n* – väärtuseks võetakse sõnamalliga sobitunud sõna või selle alamosa. Kasutatakse ainult regulaaravaldis-sõnamalli puhul ning väljaspool sõnaklassi (st tuvastamisreegli all semantikareegleid defineerides). Analoogselt viitele REF tähistab *n* seda, mitmendast alamavaldisest tuleb väärtus võtta.
- PARSE_FROM_NUMERAL:*n* – väärtuseks parsitakse sõnamalliga sobitunud liitsõna arvsõnalisest alamsõnast murdarvu või täisarvu väärtus. Kasutatakse ainult regulaaravaldis-sõnamalli puhul ning väljaspool sõnaklassi. Tähis *n* viitab arvsõna sisaldavale alamavaldisele.
- REF_VAL:*klass* – väärtus võetakse sõnaklassi *klass* semantilisest osast, atribuudist semValue.
- REF_LAB:*klass* – väärtus võetakse sõnaklassi *klass* semantilisest osast, atribuudist semLabel.

2 Alamavaldisteks loetakse regulaaravaldises sulgude vahel olevaid osi. Kui loendada avanevaid sulge regulaaravaldise vasakust servast alates, saab teada alamavaldise järjekorranumbri.

Lisa 4 Ajaväljendite sagedusprofiilid korpustes

Järgnevalt tuuakse arenduskorpuses ja uues testkorpuses käsitsi (või poolautomaatselt) märgendatud ajaväljendite sagedusprofiilid. Märgendatud ajaväljendid on grupeeritud sarnaste semantika esituste alusel. Grupitähises tuuakse välja ajaväljendi tüüp (atribuudi `TYPE` väärtus), semantika esituskuju (atribuutide `VALUE` ja `VALUE2` väärtused) ning semantikat täpsustava atribuudi `MOD` (`MOD2`) väärtus (aga viimane vaid juhul, kui see on `APPROX`).

Semantika esituskujudes olevad väärtused agregeeritakse: number asendatakse tähisega `d`, päevaosa tähised koondatakse tähise `pod` alla ning aastaaja tähised koondatakse tähise `se` alla. Ajaliste kestvuste korral kasutatakse täiendavat agregeerimist: `d+` märgib ühte või mitut järjestikku paiknevat numbrit, `X+` märgib ühte või mitut järjestikku paiknevat `X`-i ning tähiste `(YMWD)` ja `(HMS)` korral sisaldab esialgne esituskuju ühte sulgudes olevatest ajaühiku tähistest. Semantika esituskuju puudumist märgib `NULL`.

Grupitähise ees sulgudes tuuakse välja vastava grupi alla kuuluvate ajaväljendite sagedus korpuses (loendus ja protsent kõigist ajaväljenditest). Grupitähisele järgneval real tuuakse näide ühest esituskuju alla kuulunud ajaväljendist.

Arenduskorpus (1700 ajaväljendit)

```
(269, 15.8%) POINT|dddd-XX-XXTXX:XX
             üle-eelmisel aastal.
(264, 15.5%) POINT|dddd-dd-ddTXX:XX
             ülehomme
(150, 8.8%) DURATION|Pd+(YMWD)
            üle 100 aasta
(147, 8.6%) POINT|PRESENT_REF
            «Praegused
(128, 7.5%) POINT|dddd-dd-XXTXX:XX
            veebruaris
(96, 5.6%) POINT|NULL
           Ühel ilusal päeval
(68, 4.0%) POINT|dddd-dd-ddTpod
           üleeile öösel
(64, 3.8%) POINT|dddd-dd-ddTdd:dd
           Üleeile pärastlõunal kell veerand neli
(55, 3.2%) POINT|dddd-se-XXTXX:XX
           ületuleva aasta kevadel
(37, 2.2%) DURATION|PTd+(HMS)
           – 8 tunni jooksul
(35, 2.1%) RECURRENCE|NULL
           üle päeva
```

```
(28, 1.6%) POINT|dddd-Wdd-XXTXX:XX
           viimase nädala jooksul
(27, 1.6%) POINT|dddd-dd-ddTdd:XX
           üks öösel,»
(27, 1.6%) POINT|XXXX-XX-XXTpod
           üks öö
(25, 1.5%) POINT|PAST_REF
           äsja
(24, 1.4%) POINT|FUTURE_REF
           uute
(19, 1.1%) POINT|dddd-Qd-XXTXX:XX
           teises kvartalis
(18, 1.1%) INTERVAL|dddd-dd-XXTXX:XX|dddd-dd-XXTXX:XX
           viimase tosina aasta maikuude
(18, 1.1%) POINT|dddd-XX-XXTXX:XX|APPROX
           üle kahe aasta tagasi
(18, 1.1%) INTERVAL|dddd-XX-XXTXX:XX|dddd-XX-XXTXX:XX
           viimasel kahel aastal

  (Vahesumma: 89.2% korpusest)

(16, 0.9%) DURATION|PX+(YMWD)
           nädalaid
(14, 0.8%) POINT|dddX-XX-XXTXX:XX
           üheksakümnendatel aastatel
(11, 0.6%) INTERVAL|dddd-dd-ddTXX:XX|dddd-dd-ddTXX:XX
           Nelja päeva jooksul
(9, 0.5%) POINT|ddXX-XX-XXTXX:XX
          kaks sajandit tagasi
(9, 0.5%) RECURRENCE|XXXX-WXX-dTXX:XX
          orienteerumisneljapäevakute
(8, 0.5%) POINT|XXXX-se-XXTXX:XX
          uuel talihooajal
(8, 0.5%) RECURRENCE|XXXX-XX-XXTpod
          öötundidel
(7, 0.4%) DURATION|Pd+(YMWD)|APPROX
          umbes kuuajase
(7, 0.4%) RECURRENCE|XXXX-se-XXTXX:XX
          talviti
(6, 0.4%) POINT|XXXX-XX-XXTdd:XX
          pärast kella kaheksat
(5, 0.3%) INTERVAL|NULL
          viimastel hooaegadel
(5, 0.3%) POINT|XXXX-XX-XXTdd:dd
          õhtul pärast kella 20t,
(5, 0.3%) DURATION|NULL
          periood
(4, 0.2%) DURATION|PdYdM
          poolteiseks
(4, 0.2%) POINT|dddd-dd-XXTXX:XX|APPROX
          viimastel kuudel
(4, 0.2%) POINT|XXXX-dd-ddTXX:XX
          jaanipäeva ajal
(4, 0.2%) RECURRENCE|XXXX-WXX-WETXX:XX
          nädalavahetustel
(4, 0.2%) POINT|dddd-Wd-XXTXX:XX
          sel nädalal
(3, 0.2%) POINT|dddd-Wdd-WETXX:XX
          nädalavahetusel
```

(3, 0.2%) POINT|XXXX-WXX-XXTpod
sama päeva hilisõhtul.
(2, 0.1%) POINT|XXXX-WXX-dTpod
ööl vastu esmaspäeva
(2, 0.1%) POINT|dddd-Wd-WETXX:XX
Möödunud nädalavahetusel
(2, 0.1%) RECURRENCE|XXXX-WXX-WDTXX:XX
tööpäeviti
(2, 0.1%) POINT|dddd-Wdd-XTXX:XX
terve nädala jooksul
(2, 0.1%) POINT|dddd-Wdd-XXTXX:XX|APPROX
viimased nädalad
(2, 0.1%) DURATION|PTXM
minutite
(2, 0.1%) POINT|XXXX-XX-XXTXX:XX
aasta alguse
(2, 0.1%) DURATION|PTXH
tundide jooksul.
(2, 0.1%) POINT|dddd-dd-ddTXX:XX|APPROX
viimastel päevadel
(1, 0.1%) POINT|dddd-XX-XXTdd:XX
«Õhtul enne kella kuut,
(1, 0.1%) INTERVAL|dddd-dd-ddTdd:XX|dddd-dd-ddTdd:XX
hommikul kella üheteistkümnest neljani pärastlõunal
(1, 0.1%) RECURRENCE|XXXX-dd-XXTXX:XX
jaanuar
(1, 0.1%) INTERVAL|XXXX-dd-XXTXX:XX|XXXX-dd-XXTXX:XX
esimese viie kuuga
(1, 0.1%) INTERVAL|dddd-se-XXTdd:XX|dddd-se-XXTdd:XX
kell 17-19
(1, 0.1%) DURATION|PTd.dddS
2,039 sekundiga
(1, 0.1%) RECURRENCE|XXXX-XX-XXTdd:XX
iga päev kell 14
(1, 0.1%) POINT|dddd-dd-ddTdd:dd|APPROX
paar tundi varem
(1, 0.1%) DURATION|PTd+(HMS)|APPROX
paari minutiga
(1, 0.1%) POINT|dddd-XX-XXTpod
pärastlõunaks
(1, 0.1%) INTERVAL|XXXX-XX-XXTdd:dd|XXXX-XX-XXTdd:XX
kell 16.30-18.
(1, 0.1%) POINT|dddd-WX-XXTXX:XX
mõned nädalad hiljem
(1, 0.1%) DURATION|Pd.dY
1,8 aastat
(1, 0.1%) POINT|XXXX-WXX-dTXX:XX
pühapäeva
(1, 0.1%) POINT|dddd-dd-XXTdd:XX
aprillis kell 20
(1, 0.1%) INTERVAL|dddd-dd-ddTXX:XX|APPROX|dddd-dd-ddTXX:XX
paaril eelmisel päeval
(1, 0.1%) RECURRENCE|XXXX-WXX-WDTpod
argiõhtuti
(1, 0.1%) POINT|XXXX-WXX-WETXX:XX
nädalavahetuseks
(1, 0.1%) INTERVAL|XXXX-XX-XXTdd:dd|XXXX-XX-XXTdd:dd
hommikul kella kuuest kümneni
(1, 0.1%) INTERVAL|XXXX-XX-XXTpod|XXXX-XX-XXTpod

```
          mingist päevast mingi ööni
(1, 0.1%) INTERVAL|XXXX-WXX-XTdd:dd|XXXX-WXX-XTdd:dd
          15.00 - 16.00
(1, 0.1%) POINT|dddd-dd-ddTdd:XX|APPROX
          umbes kell 12
(1, 0.1%) DURATION|PTXS
          mõnesekundilised
(1, 0.1%) INTERVAL|XXXX-WXX-dTXX:XX|XXXX-WXX-dTXX:XX
          teisipäevast reedeni
(1, 0.1%) POINT|XXXX-WXX-dTdd:dd
          esmaspäeval kell 5.50
(1, 0.1%) INTERVAL|XXXX-WXX-XXTdd:dd|XXXX-WXX-XXTdd:dd
          päeval kell 15.00-16.00
(1, 0.1%) INTERVAL|dddd-Wd-XTXX:XX|APPROX|dddd-Wd-XTXX:XX
          lähema kahe nädala jooksul
(1, 0.1%) DURATION|PdWdD|APPROX
          - poolteist
(1, 0.1%) POINT|dddd-dd-XTXX:XX
          kuu lõpus
```

Uus testkorpus (385 ajaväljendit)

```
(64, 16.6%) POINT|dddd-XX-XXTXX:XX
            viis aastat varem
(57, 14.8%) POINT|dddd-dd-ddTXX:XX
            üleeile
(41, 10.6%) POINT|PRESENT_REF
            tänapäeval
(36, 9.4%) DURATION|Pd+(YMWD)
           vähemalt 10 päeva
(17, 4.4%) POINT|dddd-dd-XXTXX:XX
           veebruarist
(15, 3.9%) POINT|NULL
           tolleaegset
(12, 3.1%) DURATION|PTd.dd,ddS
           8.52 , 69.
(11, 2.9%) DURATION|PTd+(HMS)
           nelja tunniga
(10, 2.6%) POINT|dddd-dd-ddTpod
           õhtul
(9, 2.3%) POINT|dddd-se-XXTXX:XX
          sügisel
(9, 2.3%) RECURRENCE|NULL
          päevas
(9, 2.3%) POINT|dddd-dd-ddTdd:dd
          kell 14.30
(8, 2.1%) POINT|dddX-XX-XXTXX:XX
          Dekaadi alguses
(7, 1.8%) DURATION|PTdd.dd,dS
          23,63 , 8.
(7, 1.8%) POINT|dddd-XX-XXTXX:XX|APPROX
          viis aastat hiljem
(6, 1.6%) POINT|PAST_REF
          äsja
(5, 1.3%) DURATION|PTdd.dS
          42,7 sekundit
(4, 1.0%) INTERVAL|dddd-dd-ddTXX:XX|dddd-dd-ddTXX:XX
```

```
          Augusti viiel esimesel päeval
(4, 1.0%) POINT|dddd-Wdd-XXTXX:XX
          Sel nädalal
(4, 1.0%) INTERVAL|dddd-XX-XXTXX:XX|dddd-XX-XXTXX:XX
          viimase 50 aasta
(4, 1.0%) DURATION|PX+(YMWD)
          aastatepikkune

  (Vahesumma: 87.8% korpusest)

(3, 0.8%) DURATION|PTd.ddS
          2.38 , 2
(3, 0.8%) DURATION|PTdd.ddS
          ajaga 14,48
(3, 0.8%) POINT|XXXX-se-XXTXX:XX
          talvised
(3, 0.8%) INTERVAL|dddd-dd-XXTXX:XX|dddd-dd-XXTXX:XX
          samal ajal möödunud aastal
(3, 0.8%) DURATION|PTd.dS
          1,3 sekundiga
(3, 0.8%) POINT|dddd-dd-ddTdd:XX
          eile kella 23 paiku kohaliku aja järgi
(3, 0.8%) POINT|FUTURE_REF
          lähipäevil
(2, 0.5%) INTERVAL|NULL
          viimasest kaheksast-üheksast treeninguvabast päevast
(2, 0.5%) POINT|dddd-Wdd-WETXX:XX
          sellel nädalavahetusel
(2, 0.5%) POINT|dddd-Wdd-XXTXX:XX|APPROX
          paari nädala taguses
(2, 0.5%) POINT|XXXX-XX-XXTXX:XX
          12
(2, 0.5%) POINT|ddXX-XX-XXTXX:XX
          eelmise sajandi keskpaigani
(2, 0.5%) POINT|XXXX-XX-XXTdd:dd
          12.00
(2, 0.5%) DURATION|PTd+(HMS)|APPROX
          paarisekundilise
(2, 0.5%) DURATION|PTdHddM
          pooleteise tunniga
(2, 0.5%) POINT|XXXX-XX-XXTpod
          öösel
(2, 0.5%) DURATION|Pd+(YMWD)|APPROX
          paariks päevaks
(1, 0.3%) INTERVAL|dddd-dd-ddTdd:XX|dddd-dd-ddTdd:XX
          kella 10-16
(1, 0.3%) RECURRENCE|XXXX-XX-XXTpod
          öösiti
(1, 0.3%) INTERVAL|dddd-dd-ddTdd:dd|dddd-dd-ddTdd:dd
          Üleeile õhtul kella poole üheksast kuni üheteistkümneni
(1, 0.3%) INTERVAL|dddX-XX-XXTXX:XX|dddX-XX-XXTXX:XX
          20-50-ndatel
(1, 0.3%) POINT|dddd-dd-XXTXX:XX|APPROX
          Mõned kuud tagasi
```

Lisa 5 Materjalide CD

Tööga on kaasa pandud materjalide CD, mis sisaldab järgnevaid tulemeid (toodud kaustade loetelu):

- `fraasikaevandamine` – fraasikaevandamise programm ning selle töö tulemused Tasakaalus korpusel. Täiendavat informatsiooni leiab kataloogis asuvast failist `loemind.txt`.
- `sem_eksperimendid` – nädalapäev-, kuupäev- ja kuu-ajaväljendite semantika leidmise eksperimendid: korpused ning eksperimentide tulemused. Täiendavat informatsiooni leiab kataloogis asuvast failist `loemind.txt`.
- `testkorpused` – ajaväljendite tuvastaja hindamisel kasutatud korpused: arenduskorpus ja uus testkorpus ning hindamise tulemused nendel korpustel. Täiendavat informatsiooni leiab kataloogis asuvast failist `loemind.txt`.

CD ei sisalda ajaväljendite tuvastaja lähtekoodi. Tuvastaja lähtekoodi või demo saamiseks palun võtta ühendust autoriga (`siim . orasmaa (ät) gmail . com`).