doi:10.5128/ERYa8.10

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 8, 153–169

AUTOMAATNE AJAVÄLJENDITE TUVASTAMINE EESTIKEELSETES TEKSTIDES

Siim Orasmaa

Ülevaade. Artikkel käsitleb eestikeelsete tekstide arvutianalüüsi alamprobleemi: ajaväljendite automaatset tuvastamist tekstist. Ülesanne on püstitatud kaheosalisena: tekstist tuleb üles leida ajaväljendid (piiritleda ajaväljendifraasid) ning normaliseerida leitud ajaväljendite semantika (st esitada semantika eeldefineeritud märgenduskeele raamides). Artiklis kirjeldatakse ajaväljendite tuvastamisel kasutatavat märgenduskeelt ning piiritletakse vaadeldavate ajaväljendite hulk lähtuvalt märgenduskeele (aga ka praktilise analüüsi) võimalustest. Antakse ülevaade loodud reeglipõhise ajaväljendite tuvastaja tööpõhimõtetest ajaväljendite leidmisel ning semantika normaliseerimisel kasutatavatest strateegiatest. Programmi testimiseks moodustatakse Tartu Ülikooli koondkorpuse tekstidest u 70 000-sõnaline korpus, millel parandatakse käsitsi automaatse tuvastamise vead ning hinnatakse tuvastaja töö kvaliteeti.*

Võtmesõnad: arvutilingvistika, reeglipõhine keeletöötlus, semantiline märgendus, annoteerimine, eesti keel

1. Sissejuhatus

Ajaväljendite tuvastamine on tekstide automaatse analüüsi alamprobleem, mis seisneb ajaväljendite leidmises tekstist ning nende semantika kirjeldamises fikseeritud märgendusviisi alusel. Probleemi lahendamine aitab kaasa mitmete keeletehnoloogia rakenduste (nt automaatne küsimustele vastamine, sisukokkuvõtete tegemine, dialoogisüsteemid) arengule. Samuti võib ülesannet käsitleda laiemas kontekstis, tekstide ajasemantilise analüüsi alamülesandena, kus laiemaks probleemiks on sündmuste tuvastamine tekstides ning sündmuste ajalise järgnevuse määramine.

Viimase aastakümne jooksul on tekstide ajasemantiline analüüs pälvinud automaatse infoeraldamise (ingl *information extraction*) kontekstis laialdast tähelepanu. Välja on töötatud ajasemantika märgendamise keeli, nt spetsiaalselt

* Artikli valmimist on toetanud Euroopa Regionaalarengufond Eesti Arvutiteaduse Tippkeskuse kaudu. Autor tänab Tarmo Vainot nõu eest koondkorpuse märgendamise küsimustes ja anonüümseid retsensente kasulike kommentaaride eest.

ajaväljendite märgendamiseks loodud keel TIMEX2 (Ferro jt 2005) ning ajaväljendite, sündmuste ja nende vaheliste seoste märgendamiseks loodud keel TimeML (Pustejovsky jt 2003). Inglise ja prantsuse keele jaoks on loodud ka TimeML märgendusega tekstikorpused (nn TimeBank korpused[1]).

Eesti keeles on ajaväljendite tuvastamise ülesannet käsitletud kõige põhjalikumalt dialoogisüsteemide kontekstis (Treumuth 2008), kus eesmärgiks on olnud kasutajasisendist ajaväljendite leidmine ning nendele vastavate infopäringute tegemine (nt kinokavast etteantud kuupäeva(de)l linastuvate filmide leidmine).

Artikkel toetub autori magistritööle (Orasmaa 2010) ning käsitleb ajaväljendite tuvastamist tekstide (ajasemantilise) märgendamise ülesandena. Esmalt antakse ülevaade kasutatavast märgenduskeelest ning tutvustatakse lühidalt ajaväljendite tuvastamist teistes keeltes. Seejärel esitatakse ülevaade loodud süsteemist, pööratest eraldi tähelepanu ajaväljendite eraldamisele ning normaliseerimisele. Artikli lõpuosas kirjeldatakse süsteemi hindamiseks valitud korpust ning analüüsitakse tuvastaja töö kvaliteeti selle korpuse põhjal.

2. Märgendatavad ajaväljendid ja märgenduskeel

Ajaväljendite märgendamisel on siin võetud aluseks märgenduskeel TimeML (Pustejovsky jt 2003)[2]. Ajaväljendite kirjeldamise vahendid keeles TimeML on robustsed: need on orienteeritud eelkõige kalendrilise ja absoluutse semantika märgendamisele ning võimaldavad relatiivse semantikaga[3] väljendite lahendamist teatud määral edasi lükata (st esitatakse vaid osaline semantika lahendus ning lõplik interpreteerimine jäetakse lõppkasutaja hooleks). Järgnevalt antakse märgenduskeelest lühiülevaade, põhjalikuma käsitluse võib leida TimeML märgendusjuhistest (Saurí jt 2006).

Antud märgenduskeele järgi jagatakse ajaväljendid järgmisteks liikideks:

- **kalendrilised toimumisajad** (ingl *date*), mille alla kuuluvad ajaväljendid võib paigutada ajateljele ning need sisaldavad *aasta*, *kuu*, *nädala* või *päeva* detailsusega ajalist informatsiooni, nt ajaväljendid *1999. aastal*, *eelmine kuu*, *kolmandal nädalal* ning *22. veebruariks*. Siia alla ei kuulu kellaaja detailsusega ajalist informatsiooni sisaldavad ajaväljendid;
- **kellaajalised toimumisajad** (ingl *time*), mille alla kuuluvad ajaväljendid võib paigutada ajateljele ning need võivad sisaldada kuupäevalise detailsusega informatsiooni ja peavad sisaldama kellaaja detailsusega ajalist informatsiooni (nt *reedel kell 13.45*, *poole kolmeks*) või päevaosa detailsusega ajalist informatsiooni (nt *järgmise hommikuni*);
- **ajalised kestused** (ingl *duration*), mille alla kuuluvad ajaväljendid avalduvad eelkõige intervallidena, nt *kolm päeva*, *8 kuu jooksul*. Siia alla loetakse ka nn suunaga kestused (nt *eelneva 6 kuu jooksul*, *järgmised kolm aastat*), kuna esmalt on selliste väljendite semantika normaliseerimisel võimalik välja tuua kestus (kuigi ka ajateljele paigutamine võib olla teatava analüüsi järel võimalik);

[1] Vt http://timeml.org/site/timebank/timebank.html, http://www.linguist.univ-paris-diderot.fr/~abittar/french-timebank/ (26.09.2011).

[2] Teatud erijuhtudel minnakse käesolevas töös TimeML märgendusviisist lahku. Need erijuhud on dokumenteeritud ja huvi korral autori kaudu kättesaadavad.

[3] Absoluutse semantikaga ajaväljendid on üheselt ajateljel piiritletavad (nt *2009. aastal*, *22. oktoober 2011*); relatiivse semantika korral nõuab väljendi ajateljele paigutamine täiendavat kontekstiinformatsiooni (nt *reedel*, *järgmisel kuul*).

- **ajalised korduvused** (ingl *set of times*), mille alla kuuluvad ajaväljendid viitavad (sündmus-)aegade kordumisele, nt *igal aastal*, *kolm korda nädala jooksul*.

Selline liigitusviis ühtib osaliselt ka "Eesti keele grammatikas" (EKG II: 76–86) kasutatud ajamääruste liigitusviisiga. Seal jagatakse ajamäärused nelja rühma: 1) toimumisaega väljendavad ajamäärused (nt *homme*), 2) ajapiiri väljendavad ajamäärused (nt *hommikust*, *lõunani*), 3) kestust näitavad ajamäärused (nt *seitse tundi*) ja 4) korduvust väljendavad ajamäärused (nt *öösiti*). Kuigi TimeML ei näe ette ajapiiride väljatoomist eraldiseisva ajaväljendiliigina, on seal vahendid ajavahemike märgendamiseks hierarhiliselt, kombineerides toimumisaegade ning kestuste märgendusi (sellest edaspidi täpsemalt).

Erinevalt EKG II käsitlusest piirdub käesolevas töös vaadeldavate ajaväljendite hulk väljenditega, mis rahuldavad vähemalt üht kahest kriteeriumist:

a) ajaväljend sisaldab *aasta*, *kuu*, *nädala*, *päeva*, *tunni* või *minuti* detailsusega ajalist informatsiooni. Vaatluse alt jäävad välja näiteks mittekonkreetsed ajaväljendid (*omal ajal*, *lühikese aja jooksul*, *aeg-ajalt*) ning valdkonna-spetsiifilised ajaväljendid (nt *sel hooajal*, *teiseks poolajaks*, *varase kiviaja lõpuosas*, *nõukogude ajal*);
b) ajaväljend viitab minevikule, olevikule või tulevikule, lähtuvalt teksti kirjutamise hetkest (nt *hiljuti*, *praegu*, *varsti* prototüüpses kasutuses) või mõnest märgendamisele kuuluvast ajaväljendist (nt *varem*, *tookord*, *hiljem* prototüüpses kasutuses). Seega jäävad vaatluse alt välja tekstis kirjeldatud sündmustele toetuvad ajalised viited.

Toodud kitsendused tingib eelkõige praktiline nõue: märgendatav väljend peaks olema ka normaliseeritav antud märgenduskeele raamides ning normaliseerimine ei tohiks nõuda olulist valdkonna- või taustateadmiste rakendamist.

Märgendus näeb ette väljendi ümbritsemist TIMEX-märgenditega[4], mille atribuutides tuuakse välja ajaväljendi unikaalne identifikaator (atribuudis *tid*), liik (atribuudis *type*), normaliseeritud kujul semantika (atribuudis *value*) ning olemasolul semantika mõjutajad (atribuudis *mod*). Näited märgendatud lausetest:

(1) **1929. aastal** toodeti USAs 5,4 miljonit autot.
<TIMEX tid="t1" type="DATE" value="1929">1929. aastal</TIMEX> toodeti USAs 5,4 miljonit autot.

(2) **1999. aasta lõpul** esilinastunud Macbeth on teeninud meedias erakordset tähelepanu.
<TIMEX tid="t2" type="DATE" value="1999" mod="END">1999. aasta lõpul</TIMEX> esilinastunud Machbeth" on teeninud meedias erakordset tähelepanu.

Semantika normaliseerimiskujud (atribuudi *value* väärtused) baseeruvad rahvusvahelisel kalendriaegade esitamise standardil (ISO-8601) ning jagunevad kolmeks:

- **kuupõhised toimumisaja esitused** (nt *15. veebruaril 2009* esitatakse kujul 2009-02-15);

[4] Rangelt TimeML formaati järgides peaks märgendi nimi olema TIMEX3. Käesolevalt kasutatakse nimetust TIMEX, kuna käesolev märgenduskeel ei ühildu täielikult TimeML formaadiga.

- **nädalapõhised toimumisaja esitused** (nt eeldades, et kõnehetk viitab 2010. aasta 35. nädalale, võib ajaväljendi *järgmise nädala teisipäeval* normaliseerida kujul 2010-W36-2);
- **kestuste esitused** (nt *kolm aastat* esitatakse kujul P3Y).

Kuupõhist ja nädalapõhist normaliseerimiskuju võib paremast otsast pikendada või lühendada vastavalt sellele, millise ajalise detailsusega informatsiooni ajaväljend sisaldab. Näiteks väljendi *2009. aasta maikuus* puhul tuuakse välja vaid aasta ja kuu: 2009-05, samas väljendi *2009. aasta 20. mai hommikul kell 6* puhul saab normaliseerimisega minna kuni tunni täpsuseni kellaajas: 2009-05-20T06.

Lisaks näeb märgendusformaat ette erimärgiste kasutamist aastaaegade, kvartalite, poolaastate, nädalalõppude ning päevaosade semantika normaliseerimisel, nt *2009. aasta suvel* esitatakse kujul 2009-SU (SU = ingl *summer*) ning *2011, 6. mai õhtu* esitatakse kujul: 2011-05-06TEV (EV = ingl *evening*).

Ajavahemiku normaliseerimisel märgendatakse eraldi vahemiku otspunktid ning nendega määratud kestus. Selle järgi võib eristada ilmutatud otspunktidega vahemikke (3) ning varjatud otspunktidega vahemikke (4).

(3) Maailma Kirikunõukogus töötas Tutu **aastail 1972–1975**.
Maailma Kirikunõukogus töötas Tutu
<TIMEX tid="t3" type="DATE"value="1972">aastail 1972</TIMEX>–
<TIMEX tid="t4" type="DATE" value="1975" >1975.</TIMEX>
<TIMEX tid="t5" type="DURATION" value="P3Y" beginPoint="t3" endPoint="t4"/>

(4) Flaami Bloki populaarsus on märkimisväärselt kasvanud just **viimase kümne aasta jooksul**.
Flaami Bloki populaarsus on märkimisväärselt kasvanud just <TIMEX tid=t6 type="DURATION" value="P10Y" beginPoint="t7" endPoint="t0">viimase kümne aasta jooksul.</TIMEX>
<TIMEX tid="t7" type="DATE"value="1990" />

Näites (3) on ajavahemiku otspunktid tekstis kujul *1972* ja *1975* ning seda vahemikku kattev kestus tuuakse välja ilma tekstilise sisuta märgendis, mille identifikaator on *t5*. Kestus viitab otspunktidele atribuutide *beginPoint* ja *endPoint* kaudu. Näites (4) on tekstis ilmutatud kujul kestus (*viimase kümne aasta jooksul*) ning *beginPoint* ja *endPoint* kaudu viidatakse selle varjatud otspunktidele. Otspunkt indentifikaatoriga *t7* on esitatud kui ilma tekstilise sisuta märgend ning otspunkt identifikaatoriga *to* viitab vaikimisi alati teksti kirjutamise ajale (näite 4 tekst on kirjutatud kuupäeval 2000-10-10).

Hägusa või mittetäieliku semantika esitamiseks on kolm viisi. Esiteks, kui ajaväljend sisaldab semantikat mõjutavat sõna või fraasi (nt *1990ndate* ***alguses***, *aprilli* ***lõpus***, ***rohkem kui*** *5 päeva*), tuuakse semantika mõjutaja ingliskeelne tähis välja atribuudis *mod*, vt (2). Teiseks, kui ajaväljend näitab ainult ajalise viite suunda, täpsustamata ajalist detailsust (kriteeriumi *b* rahuldavad väljendid), kasutatakse atribuudi *value* väärtustena eritähiseid PAST_REF (minevikuviide), PRESENT_REF (olevikuviide) ja FUTURE_REF (tulevikuviide). Kolmandaks, kui ajaväljendis avaldub mingi ajaline detailsus, aga selle väärtus on mittekonkreetne,

kasutatakse atribuudi *value* esituses konkreetsete väärtuste asendamist tähtedega X (nt *ühel päeval* esitatakse atribuudi *value* väärtusena XXXX-XX-XX ning *aastateks* atribuudi *value* väärtusena PXY).

Ajaliste korduvuste märgenduslaad tugineb kestuse märgenduslaadile. Korduvust hõlmava perioodi pikkus tuuakse välja atribuudis *value* ning atribuudis *freq* täpsustatakse täisarvuline kordumissagedus, millega võib olla seotud ajaline detailsus (5).

(5) Seda vahemaad läbib Heimonen **vähemalt neli korda päevas**.
Seda vahemaad läbib Heimonen
<TIMEX type="SET" value="P1D" mod="EQUAL_OR_MORE" freq="4X">
vähemalt neli korda päevas</TIMEX>.

3. Ajaväljendite tuvastamisest teiste keelte arvutianalüüsis

Automaatset ajaväljendite tuvastamist on sageli vaadeldud kahest alamülesandest koosnevana: ajaväljendite eraldamine tekstis (ingl *recognition of temporal expressions*) ning ajaväljendite semantika normaliseerimine (ingl *normalization of temporal expressions*).

Lähenemised, mis keskenduvad ainult ajaväljendite eraldamisele, kasutavad sageli statistilist masinõpet. Juhendatud masinõppe puhul õpitakse käsitsi märgendatud treeningkorpuselt statistiline ajaväljendite keelemudel, mida on seejärel võimalik ennustavalt rakendada uute tekstide märgendamisel (Kolomiyets, Moens 2009). Osaliselt juhendatud masinõppe puhul kasutatakse väikest arvu positiivseid näiteid, et õppida märgendamata korpusest uute ajaväljendite tekstilist kuju ja fraasistruktuuri kirjeldavad mustrid (Craveiro jt 2009).

Lähenemised, mis lisaks ajaväljendite eraldamisele seavad eesmärgiks ka ajaväljendite kalendrilise semantika normaliseerimise, on seni valdavalt kasutanud reeglipõhiseid meetodeid. Ühe näitena sellistest lähenemistest võib vaadata süsteemi Chronos (Negri, Marseglia 2004), kus kasutatakse ligi 1000 põhireeglit ajaväljendite eraldamiseks ja kontekstiinfo kogumiseks, väikest arvu lisareegleid märgenduse ülekattuvuse probleemide lahendamiseks ning heuristilisi eeskirju ajaväljendite ankurdamisel ja lõpliku normaliseerimisstrateegia valimisel.

4. Ajaväljendite tuvastamine eestikeelses tekstis

4.1. Ülevaade süsteemist

Ajaväljendite tuvastaja on programm, mis saab sisendiks loomuliku keele teksti ning teksti loomise aja (nn kõnehetke). Programmi väljundiks on tekst, milles on märgendatud ajaväljendid, juhindudes eelnevalt kirjeldatud märgenduskeelest.

Programmi detailse kirjelduse leiab tööst (Orasmaa 2010). Selles ja järgnevates alapeatükkides kirjeldatakse olulisemaid aspekte programmi toimeloogikas, laskumata tehnilistesse detailidesse.

Siinne ajaväljendite tuvastaja on üles ehitatud kahte liiki reeglitele. Primaarsed reeglid on **tuvastamisreeglid**, mis kirjeldavad ühelt poolt ajaväljenditele vastavaid fraase[5] tekstis **fraasimustrite** abil ning teiselt poolt annavad edasi operatsioonide jada, mis tuleb ajaväljendi semantika leidmiseks läbi viia. Sekundaarsed reeglid on **liitumisreeglid**, mis täpsustavad, kuidas kõrvuti paiknevad ajaväljendid ühendatakse pikemateks fraasideks.

Tuvastamisreeglite abil eraldatakse tekstist võimalikult lühikesed terviklikku semantikat kandvad fraasid (näiteks *järgmisel neljapäeval*, *aprillis*, *kell pool kuus*). Enamasti annavad sellised fraasid edasi ajalist informatsiooni ühe detailsuse (ajaühiku) piires. Samuti kasutatakse tuvastamisreegleid semantikat mõjutavate fraaside eraldamiseks (nt *rohkem kui*, *umbes*). Liitumisreeglid näitavad, kuidas tuvastamisreeglite poolt eraldatud fraase võib liita pikemate ajaväljendifraaside moodustamiseks (nt *järgmisel neljapäeval* + *kell pool kuus*). Liitumisreeglite kasutamine võimaldab vähendada tuvastamisreeglite hulka (ei tule defineerida eraldi tuvastamisreegleid kõikvõimalike ajaliste täpsustuste jaoks, nt *mullu*, *mullu aprillis*, *mullu 26. aprillil* jne) ning samuti pakub osalist[6] lahendust sõnajärje probleemile (ei tule defineerida eraldi reegleid kõikvõimalike fraasijärjestuste nagu *täna kell 8 hommikul*, *kell 8 täna hommikul*, *täna hommikul kell 8* tabamiseks).

Iga eraldatud ajaväljendiga seotakse semantika kirjeldus, mis on defineeritud **operatsioonide** (arvutuskäskude) jadana. Semantika normaliseerimisel rakendatakse neid operatsioone ettemääratud järjekorras. Võimalikud operatsioonid võib jagada kolmeks liigiks: 1) kalendriaritmeetika operatsioonid, 2) ankurdamise operatsioonid, 3) märgenduse atribuutide täitmise operatsioonid. Kalendriaritmeetika operatsioonid võimaldavad teha kalendrimudelil arvutusi (nt liita või lahutada ajaühikuid). Vaikimisi rakendatakse kalendriarvutusi programmile sisendiks antud kuupäeval (teksti loomise aeg), ankurdamise operatsioonid võimaldavad aga muuta kalendriarvutuste alust: näiteks võib arvutamisel aluseks võtta tekstis eelneva ajaväljendi normaliseeritud semantika. Märgenduse täitmise operatsioonid võimaldavad seada ajaväljendi atribuutide väärtuseid (nt atribuudi *mod* väärtust).

Ajaväljendiga seotud semantikaoperatsioonide defineerimisel tuleb arvestada ka ajaväljendite liitumisega, st semantika lahenduskäik on oluliselt erinev tekstis üksikult esineva ajaväljendi puhul (nt *esmaspäeval*) ning teise ajaväljendiga liitunud ajaväljendi puhul (nt *järgmise nädala* + *esmaspäeval*). Probleemi lahenduseks saab iga semantikaoperatsiooniga siduda rea **kontekstikitsendusi**, mis võimaldavad enne operatsiooni rakendamist kontrollida ajaväljendite liitumise ja ankurdamise tulemusi.

4.2. Ajaväljendite eraldamine

Ajaväljendite eraldamisele eelneb teksti eeltöötlus. Selles etapis viiakse läbi sisendteksti automaatne osalausestamine, morfoloogiline analüüs ja ühestamine (Filosofti programmi *t3mesta* abil) ning leitakse järgmistel analüüsietappidel olulised tunnused (nt arvsõnafraaside paiknemine ja normaliseeritud semantika, verbide grammatilised ajad).

[5] Fraasi all on siinses kirjutises mõeldud ühest või mitmest sõnast koosnevat sõnade järjendit, esitamata sellele mingeid täpsemaid süntaktilisi kriteeriume.

[6] Lahendus on osaline, kuna ei aita juhtudel, kui sõnajärg varieerub normaliseerimise seisukohast terviklikes fraasides, nt *viimase kolme kuu* ja *kolme viimase kuu*.

Ajaväljendite eraldamiseks tekstis kasutatakse tuvastamisreeglite alla kuuluvaid **fraasimustreid**. Fraasimuster sisaldab üldistavat fraasikirjeldust ning sarnaneb toimeloogika poolest lõplikule mittedetermineeritud automaadile. Fraasimustrile antakse ükshaaval sisendteksti sõnu ette ning iga kord, kui järjestikku etteantud sõnad vastavad mingile mustri poolt kirjeldatud fraasile (terves fraasi ulatuses), teostatakse uue **ajaväljendikandidaadi** eraldamine.

Fraasimustris on sõnad kirjeldatud **sõnamallide** abil. Käesolevas töös eristatakse nelja sõnamalliliiki: **regulaaravaldisega** määratud sõnamall, **algvormiga** määratud sõnamall, **arvsõnafraasi** mall ning **sõnaklass**. Regulaaravaldisega määratud sõnamalle kasutatakse eelkõige numbrimustrite või vähese arvu erinevate sõnavariantide kirjeldamiseks. Algvormiga määratud sõnamallid on mõeldud kasutamiseks juhtudel, kui tahetakse tabada sõna kõiki võimalikke vorme. Arvsõnafraasi malli abil saab kirjeldada sõnadega edasi antud täis- või murdarve, seega võib malli poolt tabatav olla nii üksiksõna kui ka fraas. Kirjeldamisel antakse ette arvu liik (põhiarvsõna, järgarvsõna või murdarvsõna) ning vajadusel ka võimalike väärtuste vahemik. Sõnaklass võimaldab kasutada teisi sõnamalle alamosadena (elementidena) ning seeläbi koondada eri sõnade kirjeldused ühe malli alla. Näiteks võib koostada sõnaklassi NÄDALAPÄEV, mille elementideks on eri nädalapäevanimetusi algvormide kaudu kirjeldavad sõnamallid: *{esmaspäev, teisipäev, kolmapäev, neljapäev, reede, laupäev, pühapäev}*.

Ajaväljendite kirjeldamisel on lisaks positiivsete juhtude kirjeldamisele tarvis määratleda ka negatiivsed juhud: juhud, mil üleliigselt eraldatud ajaväljendikandidaat tuleb kustutada. Kustutamine järgneb ajaväljendikandidaatide eraldamisele ning koosneb kahest alametapist: ülekaetud kandidaatide kustutamine ning kandidaatide kustutamine negatiivsete mustrite alusel. Ülekaetud kandidaatide kustutamine viiakse läbi alati, kui üks kandidaat on täielikult üle kaetud pikema kandidaadi poolt. Näiteks kuupäevana eraldatud ajaväljendikandidaadiga *30. jaanuar* paralleelselt eraldatakse ka ainult kuunimetusest koosnev kandidaat *jaanuar* ning just viimane kustutatakse kui täielikult ülekaetud kandidaat. Negatiivsete mustrite alusel kustutamine viiakse läbi juhul, kui fraasimustriga on seotud mitte-ajaväljendilise konteksti kirjeldus (sisuliselt sõnu kirjeldavate regulaaravaldiste jada) ning eraldatud fraasi lähiümbrus rahuldab seda kirjeldust. Näiteks kasutatakse negatiivseid mustreid, et piirata ajaväljendite (kuunimetuste) eraldamist isikunimedest (*Karl August Hermann*, *August Mälk*).

Kui ajaväljendikandidaatide seast on välja jäetud üleliigsed kandidaadid, viiakse läbi kõrvuti paiknevate kandidaatide liitmine. See toimub kahel tasemel: **fraasi** tasemel ning ajavahemike tasemel. Fraasi tasemel liitmist kontrollivad liitumisreeglid, mis kirjeldavad, kuidas liita kõrvuti paiknevad erineva ajalise detailsusega kandidaadid (nt *järgmise nädala* + *esmaspäeval*) või kuidas liita ajaväljendikandidaadiga semantika täpsustus (nt *2009. aasta* + *lõpus*). **Ajavahemike** tasemel liitmine toetub kahe eelmise taseme (kandidaatide eraldamine ja fraasi tasemel liitmine) tulemustele ning toimub heuristiliste eeskirjade alusel. Esimene eeskiri eraldab vahemiku seestütlevas ja rajavas käändes sõnade olemasolu põhjal (nt *reede**st***+*pühapäeva**ni***), teine eeskiri tuvastab vahemiku, kui kandidaadist paremal või vasakul esineb lühikujul vahemiku teine otspunkt (nt *aastatel 2007* +***kuni 2009***, ***1.–*** + *3. juunil*).

4.3. Ajaväljendite normaliseerimine

Ajaväljendite normaliseerimise seisukohast on oluline toimumisaegade jagunemine absoluutseteks ja relatiivseteks ajaväljenditeks:

- **absoluutsete ajaväljendite** normaliseerimine on triviaalne ning seisneb kalendriväärtuste ümberkirjutamises märgenduskeele esituskujusse. Sellised väljendid sisaldavad *aasta* detailsusega absoluutset ajalist informatsiooni, nt *20. mai 2009*, *12.03.2011*;
- **relatiivsete ajaväljendite** normaliseerimine nõuab kalendriaritmeetika rakendamist: ajaväljendi normaliseeritud väärtus tuleb leida mingi teise ajapunkti suhtes. Käesolevas töös eristatakse kaht liiki relatiivseid ajaväljendeid:
  - **teksti loomise aja järgi lahenduvad väljendid** – sellisteks on vahetult deiktilised väljendid, nt *täna, homme, järgmisel nädalal, mullu* ning ilma täpsustava *aasta*-detailsuseta väljendid nt *24. aprillil*, *mai lõpus*;
  - **teise ajaväljendi külge ankurduvad ajaväljendid** – nt ajaväljend *aasta varem* ankurdub eelneva aastaarvu külge lauses ***1996. aastal*** *oli puudujääk 5% väiksem kui* ***aasta varem***.

Toimumisaegade semantika normaliseerimisel võetakse vaikimisi aluseks teksti loomise aeg (enamasti antud kuupäeva täpsusega) ning eesmärgiks on semantikaoperatsioonide rakendamise teel teisendada alusaega seni, kuni on jõutud ajaväljendi semantikale vastava kalendriajani.

Lihtsaim kalendriaritmeetika operatsioon on **omistamisoperatsioon**, mis sisuliselt kirjutab alusaja mingi kalendrivälja[7] väärtuse üle uue väärtusega. Omistamisoperatsioone kasutatakse absoluutsete ajaväljendite semantika kirjeldamisel.

Liitmis- ja **lahutamisoperatsioonid** lubavad alusaja mingit kalendrivälja suurendada või vähendada etteantud arvu ajaühikute võrra, võimaldades nn suunaga ajaväljendite (nt *eelmisel nädalal* ja *järgmisel nädalal*, *viis aastat tagasi* ja *viie aasta pärast*) lahendamist. Teatud suunaga ajaväljendite puhul muutub aga liitmis- ja lahutamisoperatsioonide rakendamine problemaatiliseks. Näiteks väljend *eelseisval reedel* võib viidata nii *selle nädala* kui ka *järgmise nädala reedele*, seega ei saa anda ühest eeskirja, kas alusaja väljale *nädal* tuleks üks nädal liita või mitte.

Saksa keele ajaväljendite tuvastamist uurinud Frank Schilder ja Christopher Habel (2001) on pakkunud välja liitmis- ja lahutamisoperatsioonidele alternatiivse strateegia, mida võib tinglikult nimetada *otsimisoperatsiooniks*. Sisuliselt võimaldab otsimisoperatsioon leida etteantud suunalt (minevik või tulevik) alusajale lähima, nõutud kalendriväljade konfiguratsioonile vastava kalendriaja. Eelneva näite puhul võib otsimisoperatsiooni defineerida selliselt, et leitakse alusajale lähim reede tulevikust, seega lahenduskäik järgib märksa täpsemalt keeles väljendatavat tähendust. Käesolevas töös kasutatakse otsimisoperatsioone peamiselt selliste ajaväljendite normaliseerimisel, kus suund on edasi antud oleviku kesksõna kujulises täiendis (nt *eelneval reedel*, *eelseisval nädalavahetusel*, *tuleval kevadel*).

Suunaga relatiivsete väljendite juures on problemaatilised veel juhud, kus väljendi kasutamisel ollakse mainitud ajatsükli alguses või lõpus. Näiteks kui pühapäeval kasutatakse väljendit *viimasel nädalal*, ei ole selge, kas mõeldakse seda

[7] Kalendriväljade all mõeldakse kalendriaja erinevaid tahke: *aasta*, *kuu*, *nädal*, *nädalapäev*, *kuupäev*, *tund*, *minut*.

või eelmist nädalat, samas nt esmaspäeval antud väljendit kasutades mõeldakse tõenäoliselt eelmist nädalat.

Suunaga väljendite kasutus eeltoodud probleemide lõikes vajab ilmselt põhjalikumat uurimist, mis aga ei mahu käesoleva töö raamidesse.

Omaette probleemina võib käsitleda ilma täpsustava suunata relatiivsete väljendite, s.o tekstis üksikult esinevate nädalapäevade, kuude ja kuupäevade[8] lahendamist. Selliste väljendite lahendamiseks on välja pakutud kaks strateegiat. Esimene strateegia näeb ette ajaväljendile (lause piires) lähima verbi leidmist ning suuna otsustamist verbi grammatilise aja põhjal. Sellist strateegiat on inglise keele suunata ajaväljendite lahendamisel kasutanud näiteks Inderjeet Mani ja George Wilson (2000) ning Jannik Strötgen ja Michael Gertz (2010). Alternatiivne strateegia (Baldwin 2002) näeb ette alusaja ümber kalendrivälja unikaalsete väärtuste akna loomist ning otsitava väärtuse valimist akna seest. Strateegiat illustreerib kõige paremini nädalapäevade lahendamise näide (6): kui kõnehetk on 2010-03-28 (*pühapäev*) ning otsitakse lahendust väljendile *teisipäeval*, võib moodustada seitsme unikaalse nädalapäeva akna järgmiselt (tärn märgib kõnehetke):

(6)	N	R	L	P	E	T	K
	25	26	27	28 *	29	30	31

Toodud näite puhul on lahenduseks akna sisse jääv teisipäev: 2010-03-30. Strateegia rakendamisel üksikute kuude või kuupäevade lahendamiseks jääb üks kuu alati aknast välja, seega tuleb kasutada mingit varustrateegiat (nt võib eeldada, et väljajääv kuu on käesoleva aasta kuu).

Kahte strateegiat on inglise keele üksikute nädalapäevade lahendamisel võrrelnud Pawel Mazur ja Robert Dale (2008), kes leidsid, et transkribeeritud tekstide korpusel (telefonivestluste, jutusaadete, uudiste jm transkribeeringud) andis Jennifer Baldwini 7 päeva aken korrektse tulemuse 94,28% juhtudel ning verbi grammatilisele aja heuristik viis korrektse lahenduseni 92,64% juhtudel. Käesoleva töö autor on neid strateegiaid võrrelnud eestikeelsete ajakirjandustekstide (täpsemalt: päevalehtede) korpusel (Orasmaa 2010) ning on leidnud, et üksikute nädalapäevade lahendamisel andis kõrgemaid tulemusi verbi grammatilise aja heuristik (90,2% korrektseid tulemusi), samas üksikute kuude ja kuupäevade lahendamisel osutus valitud korpusel parimaks Baldwini akna heuristik (vastavalt 94,7% ja 92,7% korrektseid tulemusi). Süsteemi praegune reeglikonfiguratsioon juhindub nendest tulemustest, kuigi probleem iseenesest vajab edasist uurimist.

Kui valdava osa relatiivseid ajaväljendeid võib ankurdada teksti loomise aja külge (eeldus, mis näib kehtivat vähemalt suure osa ajakirjandustekstide puhul), siis teatud võtmesõnade olemasolu ajaväljendifraasis nõuab ankurdamisstrateegia muutmist. Näiteks võtmesõnade *following* 'järgmine', *previous* 'eelmine', *same* 'sama', *that* 'too', *before* 'varem' ja *later* 'hiljem' olemasolu inglise keele ajaväljendifraasides nõuab väljendi ankurdamist tekstis eelneva toimumisaja külge. Matteo Negri ja Luca Marseglia (2004) kasutavad seda tüüpi väljenditele ankurväljendi valimisel ajalise detailsuse kitsendust: võrreldes ankurdatavaga peab ankruks valitav väljend olema suurema või sama ajalise detailsusega. Näiteks võib väljend *three days later* 'kolm päeva hiljem' ankurduda küll sama detailsusega väljendi *on*

[8] Laiemas plaanis mõeldakse siin relatiivseid väljendeid, mille puhul ei saa ainuüksi ajaväljendifraasi põhjal otsustada, millisest suunast (minevik, olevik, tulevik) tuleks mainitud nädalapäeva, kuud või kuupäeva otsida. Siia kuuluvad nii ühest sõnast koosnevad ajaväljendid (nt *teisipäeval*, *märtsis*) kui ka suuremat detailsust sisaldavad liitväljendid (nt *teisipäeva õhtul*).

Monday 'esmaspäeval' külge, aga mitte väiksema ajalise detailsusega väljendi *on this month* 'sellel kuul' külge.

Käesolevas töös kasutatakse toimumisaja tüüpi ajaväljendite ankurdamisel samuti heuristilisi strateegiaid, mis tuginevad autori senisele (ajakirjandus)korpuste uurimise kogemusele ning mille efektiivsust eraldi mõõdetud pole. Tekstis eelneva ajaväljendi külge ankurdatakse võtmesõnu *varem*, *hiljem*, *sama* ja *too* sisaldavad ajaväljendifraasid, ajalise detailsuse kitsendusi ankurdamisel ei rakendata. Tekstis eelneva, *päeva* detailsusega ajaväljendi külge ankurdatakse kellaaja ja päevaosa detailsusega ajaväljendid. Kõikidel ankurdamise juhtudel ei vaadata sobiva ankru otsimisel kaugemale kui kolm eelnevat lauset.

5. Süsteemi testimine

5.1. Testimiseks valitud korpus

Ajaväljendite tuvastaja hindamiseks kasutati Tartu Ülikooli koondkorpuse tekste. Koondkorpuse kirjakeele osa kogumaht on umbes 250 miljonit sõna, millest 84% moodustavad ajakirjandustekstid, 5,9% riigikogu stenogrammid, 4,7% seadustekstid, 2,9% ilukirjandus ning 2,5% teadustekstid. Mitmed koondkorpuse alamosad on heterogeensed ning representatiivse testkorpuse koostamine seetõttu problemaatiline: paratamatult tuleb teha ka subjektiivseid valikuid. Heterogeensuse tõttu on vaatluse alt välja jäetud ilukirjanduse ja teadustekstide alamkorpused; seadustekstide ja ajakirjanduse korpustest on vaadeldud vaid teatud kriteeriumidele vastavaid alamkorpusi.

Ajakirjanduskorpuses püüti eristada alamkorpusi vastavalt sellele, millistesse rubriikidesse olid artiklid jagatud. Paraku kasutavad erinevad meediaväljaanded erinevaid rubriikidesse jagamise viise ning isegi ühe väljaande piires võib rubriigiline jaotus oluliselt muutuda aastate lõikes. Kitsendades valikut kolmele väljaandele – Postimees (aastatest 1995–2000), Eesti Päevaleht (aastatest 1995–2007) ning ajakiri Luup (aastatest 1996–2002) –, oli võimalik eristada sagedasemaid rubriike: "Eesti uudised", "Välisuudised", "Arvamused", "Sport", "Majandus" ja "Kultuur"[9].

Eesmärgiga hinnata süsteemi toimimist ajalugu käsitlevate tekstide märgendamisel moodustati eraldi alamkorpus ajakirja Horisont ajalooteemalistest artiklitest. Seadustekstide korpusest võeti vaatluse alla Eesti seaduste alamkorpus.

Testkorpust suurendati terviktekstide kaupa, kuna ajaväljendite normaliseerimisel tuleb lähtuda laiemast kontekstist kui üks lause. Ajakirjanduskorpuses oli terviktekstiks üks artikkel, riigikogu stenogrammides loeti terviktekstiks ühe päevakorrapunkti arutamine ning seadustekstides moodustas tervikteksti üks seadus.

Juhuslikult valitud tekstidest moodustati 70 209 sõna suurune korpus, mille proportsioonid esitab tabel 1. Ajaväljendite märgendamine korpuses viidi läbi poolautomaatselt: esmalt rakendati ajaväljendite tuvastajat ning seejärel parandati käsitsi automaatsel tuvastamisel tekkinud vead. Vigu parandas üks inimene (käesoleva töö autor).

Pärast vigade parandamist oli testkorpuses 1900 tekstilise sisuga ajaväljendimärgendit; ajaväljendite jagunemist alamkorpuste vahel kirjeldab tabel 2. Kuigi

[9] Autori poolt pandud üldistavad nimed, rubriikide täpsed nimetused erinevad väljaannete lõikes.

tuvastaja väljundis on ka ilma tekstilise sisuta märgendid (nagu eespool näidetes (3) ja (4)), on sellise märgenduse arvestamine hindamisel mõnevõrra keerukas ning siinses töös seda ei tehta.

Tabel 1. Testkorpuse alamkorpuste proportsioonid

Alamkorpus		Sõnu[10]	Tekste	Osakaal kogu korpusest
Ajakirjandus	Eesti uudised	17271	31	24,6%
	Välisuudised	7724	15	11,0%
	Arvamused	7024	15	10,0%
	Sport	6981	16	9,9%
	Majandus	5205	12	7,4%
	Kultuur	5102	12	7,3%
Ajalugu (ajakirjast Horisont)		7088	6	10,1%
Riigikogu stenogrammid		6950	5	9,9%
Eesti seadused		6864	3	9,8%

Tabel 2. Ajaväljendite arv testkorpuses

Alamkorpus		Ajaväljendeid
Ajakirjandus	Eesti uudised	553
	Välisuudised	155
	Arvamused	130
	Sport	160
	Majandus	177
	Kultuur	142
Ajalugu (ajakirjast Horisont)		229
Riigikogu stenogrammid		109
Eesti seadused		245

5.2. Süsteemi hindamine

Käesoleva artikli autor oli ajaväljendite tuvastajat varem arendanud ja testinud ajakirjanduskorpuste peal; ajakirjanduskorpustel väljatöötatud reegleid ja heuristikuid kasutati ka käesolevas eksperimendis. Kuna aga varem ei olnud tehtud süstemaatilist alamkorpuste eristamist, oligi üheks huvipakkuvaks küsimuseks, mil määral erinevad tulemused alamkorpuste lõikes.

Programmi töö hindamisel mõõdeti ajaväljendite eraldamise saagist ja täpsust (rangelt ja mitterangelt fraasipiire arvestades) ning atribuutide *type* ja *value* määramise täpsuseid. Mitterangel fraasipiiride arvestamisel loeti automaatne eraldamine korrektseks ka juhul, kui eraldatud fraas kattis käsitsi eraldatud ajaväljendifraasi vaid osaliselt (vähemalt ühe sümboli ulatuses). Rangel fraasipiiride arvestamisel

[10] Sõnade hulka loetakse ka punktuatsioon.

nõuti automaatse ning käsitsi märgendatud ajaväljendi fraasipiiride täielikku kattumist. Ajaväljendite eraldamise saagis ja täpsus defineeriti järgmiselt:

$$saagis = \frac{korrektselt\ eraldatud\ ajaväljendite\ arv}{käsitsi\ eraldatud\ ajaväljendite\ arv}$$

$$täpsus = \frac{korrektselt\ eraldatud\ ajaväljendite\ arv}{automaatselt\ eraldatud\ ajaväljendite\ arv}$$

Atribuutide normaliseerimise täpsust hinnati vaid korrektselt eraldatud ajaväljendite puhul, mitterangelt fraasipiire arvestades. Normaliseerimise täpsus arvutati järgmise valemi järgi:

$$täpsus = \frac{korrektselt\ määratud\ atribuudiväärtuste\ arv}{automaatselt\ määratud\ atribuudiväärtuste\ arv}$$

Tabelis 3 on esitatud hinnangud programmi töö tulemustele ajaväljendite eraldamisel.

Tabel 3. Ajaväljendite eraldamise tulemused

Alamkorpus		**Saagis**	**Täpsus**	**Saagis (*range*)**	**Täpsus (*range*)**
Ajakirjandus	Eesti uudised	447/553 (80,8%)	447/447 (100%)	421/553 (76,1%)	421/447 (94,2%)
	Välisuudised	135/155 (87,1%)	135/139 (97,1%)	130/155 (83,9%)	130/139 (93,5%)
	Arvamused	104/130 (80%)	104/107 (97,2%)	95/130 (73,1%)	95/107 (88,8%)
	Sport	146/160 (91,2%)	146/149 (98%)	138/160 (86,2%)	138/149 (92,6%)
	Majandus	137/177 (77,4%)	137/141 (97,2%)	127/177 (71,8%)	127/141 (90,1%)
	Kultuur	112/142 (78,9%)	112/116 (96,6%)	105/142 (73,9%)	105/116 (90,5%)
Ajalugu (ajakirjast Horisont)		166/229 (72,5%)	166/168 (98,8%)	130/229 (56,8%)	130/168 (77,4%)
Riigikogu stenogrammid		99/109 (90,8%)	99/103 (96,1%)	96/109 (88,1%)	96/103 (93,2%)
Eesti seadused		212/245 (86,5%)	212/214 (99,1%)	204/245 (83,3%)	204/214 (95,3%)
Kogu korpus		1558/1900 (82%)	1558/1584 (98,4%)	1446/1900 (76,1%)	1446/1584 (91,3%)

Võib öelda, et ajaväljendite eraldamisel on koostatud reeglid on kaldu täpsuse kasuks – kogu korpusel mõõdeti täpsuseks 98,4% (rangelt fraasipiire arvestades 91,3%), samas kui saagis oli vaid 82% (rangelt 76,1%).

Üldist madalat saagist aitavad selgitada mõned reeglite koostamisel tehtud valikud. Reeglite loomisel keskenduti konkreetsetele ajaväljenditele, madala prioriteediga oli nn lühiväljendite ja umbmääraste väljendite kirjeldamine. Lühiväljenditeks loeti ühest sõnast koosnevaid väljendeid nagu aasta ja kuus (tähendusega 'ühes kuus'), üksikuid aastaarve ja minimalistlikke kuupäevi (nt 6.3.). Selliste väljendite puhul on problemaatiline mitmetähenduslikkus, sh ka võimalus, et tegemist võib olla mitteajaväljenditega. Umbmäärasteks loeti väljendid, mis jätavad kvantiteedi või kalendrilise informatsiooni lahtiseks, nt mõni aeg tagasi, mõneaastane, aastakümnete jooksul, aasta-paari pärast ja mitu päeva.

Kõige madalam saagis (mitterangelt 72,5%) mõõdeti alamkorpuses "Ajalugu", kus ühe põhjusena võib välja tuua täpsustusega *enne Kristust* aastaarvude eraldamata jätmise. Üksikute ajakirjanduses esinenud näidete alusel koostatud reeglid ei suutnud katta selliste väljendite variatiivsust.

Alamkorpusespetsiifiliseks ilminguks võiks lugeda ka kümnenditele viitavate lühiväljendite (nt *seitsmekümnendate*) suhteliselt sagedase esinemise alamkorpuses "Kultuur"; selliste väljendite eraldamata jätmine põhjustas antud korpuses keskmisest madalamat saagist (mitterangelt 78,9%).

Peatükis 2 seatud kriteeriumid ei võimaldanud alati tõmmata selget piiri märgendatava ning mittemärgendatava väljendi vahele. Näiteks alamkorpustes "Majandus" ja "Eesti seadused" tuli kriteeriumi (a) järgi lugeda märgendatavateks ka valdkonnaspetsiifilised väljendid *viimase nelja börsipäevaga* ja *eelarveaasta alguseks*. Samas ei kuulunud alamkorpuses "Sport" sagedasti esinevad valdkonnaspetsiifilised väljendid nagu *teise poolaja alguses* või *viimane hooaeg* selle kriteeriumi järgi märgendamisele, mis selgitab ka keskmisest kõrgemat eraldamise saagist (mitterangelt 91,2%) antud korpuses.

Üleminekul mitterangelt mõõtmiselt rangele langesid eraldamise saagis ja täpsus keskmiselt 6–7%. Ajaväljendifraasi piiride määramisel olid sagedased vead *arv ja pool* -tüüpi fraaside poolik eraldamine (nt *kolm ja [pool aastat]*[11], *ligemale kahe- ja [pooletunnises]*) ning umbmäärasusele viitavate sõnade väljajätmine fraasist (nt *mitu [tuhat aastat]*).

Tabelis 4 on toodud hinnangud programmi töö tulemustele ajaväljendi liigi määramisel (atribuudi *type* täitmine) ning semantika normaliseerimisel (atribuudi *value* täitmine). Ajaväljendi liigi määramine üldiselt probleeme ei tekitanud (täpsus kogu korpuses 97,2%); sagedased olid eksimused vaid alamkorpuses "Ajalugu" (täpsus 87,3%), kus liigi valesti määramine oli põhjustatud fraasipiiride mittetundmisest (nt *u. [7500 aastat] e.Kr.*).

Semantika normaliseerimisel mõõdeti keskmiseks täpsuseks 87,4%. Sagedasti põhjustas vigu suutmatus eristada väljendi kasutust üldises ja konkreetses tähenduses. Nt *täna* konkreetses tähenduses viitab käesolevale kuupäevale, üldises tähenduses aga *tänapäevale*. Sagedasti eksiti veel pooliku eraldamise ning vale ankurdamisstrateegia tõttu. Suhteliselt suur vigade arv alamkorpuses "Ajalugu" (täpsus 61,8%) oli peamiselt tingitud poolikust eraldamisest.

Kõige kõrgemad täpsused semantika normaliseerimisel mõõdeti alamkorpustes "Riigikogu stenogrammid" (93,9%), "Välisuudised" (93,3%) ning "Eesti seadused" (92,9%). Riigikogu stenogramme uurides võis täheldada suhteliselt lihtsalt lahenda-

[11] Kandilised sulud märgivad automaatselt eraldatud ajaväljendi piire.

tavate kuupäevaliste toimumisaegade (nt *1. juuli*) domineerimist (ligi 42% kõigist märgendatud ajaväljenditest olid normaliseeritavad kuupäevaks).

Alamkorpuses "Eesti seadused" olid eranditult kõik relatiivsed väljendid (nt *järgneva aasta 1. veebruariks*) absoluutsel ajaskaalal mitteajastatavad, seega ajakirjanduses kasutatavad normaliseerimisstrateegiad seal ei toiminud. Kuna aga korpuses olid ülekaalus absoluutsed väljendid (seaduste jõustumise kuupäevad ning aastaarvud Riigi Teataja allikaviidetes), siis eksimused lõplikku tulemust oluliselt ei mõjutanud.

Välisuudiste korpuses võis täheldada suhteliselt suurt olevikuviidete (*praegu*, *nüüd*) osakaalu ajaväljendite seas (ligi 20%, võrdluseks Eesti uudistes oli olevikuviidete osakaal ligi 12%); selliste väljendite lihtsa lahendatavuse võib lugeda üheks kõrgema täpsuse põhjuseks.

Tabel 4. Ajaväljendite liigitamise ning semantika normaliseerimise tulemused

Alamkorpus		**Ajaväljendite liigitamise täpsus**[12]	**Semantika normaliseerimise täpsus**
Ajakirjandus	Eesti uudised	441/447 (98,7%)	405/447 (90,6%)
	Välisuudised	132/135 (97,8%)	126/135 (93,3%)
	Arvamused	103/104 (99%)	89/104 (85,6%)
	Sport	140/145 (96,6%)	125/145 (86,2%)
	Majandus	134/137 (97,8%)	124/137 (90,5%)
	Kultuur	108/110 (98,2%)	97/110 (88,2%)
Ajalugu (ajakirjast Horisont)		144/165 (87,3%)	102/165 (61,8%)
Riigikogu stenogrammid		99/99 (100%)	93/99 (93,9%)
Eesti seadused		208/211 (98,6%)	196/211 (92,9%)
Kogu korpus		1509/1553 (97,2%)	1357/1553 (87,4%)

6. Kokkuvõte

Artikkel tutvustab ajaväljendite automaatse tuvastamise ülesannet ning kirjeldab eesti keele jaoks loodud ajaväljendite tuvastajat. Ajaväljendite tuvastamise ülesanne jagati kaheks alamosaks: ajaväljendite tekstist eraldamine ning ajaväljendite semantika normaliseerimine valitud märgenduskeele raamides. Märgenduskeele võimalustest lähtuvalt kitsendati ka vaadeldavat ajaväljendite hulka ning keskenduti eelkõige kalendriliselt normaliseeritavatele väljenditele. Kasutatav märgenduskeel tugineb sündmuste ja ajaväljendite märgenduskeelele TimeML.

Kasutatud on reeglipõhiselt üles ehitatud ajaväljendite tuvastajat. Tuvastamisreeglite abil kirjeldatakse võimalikult lühikesi terviklikku semantikat kandvaid ajaväljendifraase ning liitumisreeglite abil kontrollitakse pikemate fraaside (sõnajärjest sõltumatut) moodustamist. Fraaside kirjeldamisel kasutatakse lõpliku automaadi tehnikat, üksiksõnade kirjeldamine toetub morfoloogilise analüüsi ja ühestamise tulemustele, arvsõnade tuvastamisele ning regulaaravaldistele. Aja-

[12] Teatud juhtudel esines süsteemi väljundis märgendeid, millel puudus liiki määrav atribuut *type*. Neid juhte arvestades tuleks välja tuua ka liigi määramise saagis. Täpsustava hinnangu saab lisada järgmistes alamkorpustes: "Kultuur" (saagis 108/112; 96,4%), "Ajalugu" (saagis 144/166; 86,7%), "Eesti seadused" (saagis 208/212; 98,1%) ning kogu korpuses (saagis 1509/1557; 96,9%).

väljendi semantika esitatakse normaliseerimiseks vajalike operatsioonide jadana. Artiklis on käsitletud erinevaid heuristilisi meetodeid mitmese semantikaga väljendite normaliseerimiseks.

Tuvastajat arendati ajakirjandustekstidel. Programmi hindamiseks moodustati u 70 000-sõnaline testkorpus, mis koosnes Tartu Ülikooli koondkorpuse kirjakeele osast juhuslikult valitud ajakirjandustekstidest, ajalugu käsitlevatest artiklitest, riigikogu stenogrammidest ning Eesti seadustest. Programmi koondtulemuseks ajaväljendite eraldamisel oli saagis 82% ning täpsus 98,4% (fraasipiire mitterangelt arvestades), ajaväljendite liigitamisel saavutati täpsus 97,2% ning semantika normaliseerimisel täpsus 87,4%.

Programmi arendamisel on keskendutud konkreetsete ajaväljendite kirjeldamisele, ent edaspidi tuleks rohkem tähelepanu pöörata ka mitmese semantikaga lühiväljendite ning umbmääraste väljendite kirjeldamisele, kuna nende eraldamata jätmine on üheks oluliseks madala saagise põhjuseks. Semantika normaliseerimise seisukohast vajab mitmete ajaväljendite kasutus eraldiseisvat ja süstemaatilist uurimist; laiemateks probleemideks on relatiivsete ajaväljendite korrektne ankurdamine ning konkreetse ja üldise tähenduse eristamine.

Tulevikus rakendatakse ajaväljendite tuvastajat Tartu Ülikooli kirjakeele korpuse terviklikul märgendamisel ning märgendatud ajaväljenditega korpus muutub kättesaadavaks ka Keeleveebi[13] kaudu.

Viidatud kirjandus

Baldwin, Jennifer A. 2002. Learning temporal annotation of French news. Magistritöö. Graduate School of Arts and Sciences. Georgetown University, Washington, DC. www.jenniferbaldwin.com/jb/documents/Thesis.pdf (24.09.2011).

Craveiro, Olga; Macedo, Joaquim; Madeira, Henrique 2009. Use of Co-occurrences for Temporal Expressions Annotation. – Jussi Karlgren, Jorma Tarhio, Heikki Hyyrö (Eds.). Proceedings of the 16th International Symposium on String Processing and Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 156–164.

EKG II = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi 1993. Eesti keele grammatika II. Süntaks. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Ferro, Lisa; Gerber, Laurie; Mani, Inderjeet; Sundheim, Beth; Wilson, George 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions. http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf (24.09.2011).

Kolomiyets, Oleksandr; Moens, Marie-Francine 2009. Meeting TempEval-2: Shallow Approach for Temporal Tagger. – Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (DEW 2009). Stroudsburg, USA: Association for Computational Linguistics, 52–57.

Mani, Inderjeet; Wilson, George 2000. Robust temporal processing of news. – Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL 2000). Stroudsburg, USA: Association for Computational Linguistics, 69–76.

Mazur, Paweł; Dale, Robert 2008. What's the date?: High accuracy interpretation of weekday names. – COLING 2008 Proceedings of the 22nd International Conference on Computational Linguistics, Vol. 1. Stroudsburg, USA: Association for Computational Linguistics, 553–560.

[13] http://www.keeleveeb.ee/ (26.09.2011).

Negri, Matteo; Marseglia, Luca 2004. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Tehniline raport. ITC-irst, Trento. http://www.lsi.upc.edu/~nlp/meaning/documentation/3rdYear/WP3.6.pdf (20.09.2011).

Orasmaa, Siim 2010. Ajaväljendite tuvastamine eestikeelses tekstis. Magistritöö. Tartu Ülikool, matemaatika-informaatikateaduskond.

Pustejovsky, James; Castaño, José M.; Ingria, Robert; Saurí, Roser; Gaizauskas, Robert J.; Setzer, Andrea; Katz, Graham; Radev, Dragomir R. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. – Mark T. Maybury (Ed.). New Directions in Question Answering. Papers from 2003 AAAI Spring Symposium, Stanford University. Stanford, CA: AAAI Press, 28–34.

Saurí, Roser; Littman, Jessica; Knippen, Bob; Gaizauskas, Robert; Setzer, Andrea; Pustejovsky, James 2006. TimeML Annotation Guidelines. http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf (07.09.2011).

Schilder, Frank; Habel, Christopher 2001. From temporal expressions to temporal information: Semantic tagging of news messages. – Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, ACL-2001. Toulose, 65–72.

Strötgen, Jannik; Gertz, Michael 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. – Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010). Stroudsburg, USA: Association for Computational Linguistics, 321–324.

Treumuth, Margus 2008. Normalization of temporal information in Estonian. – Petr Sojka, Aleš Horak, Ivan Kopecek (Eds.). Proceedings of the 11th international conference on Text, Speech and Dialogue. Brno, Czech Republic, September 8-12. Lecture Notes in Computer Science, 5246. Springer, 211–218. http://dx.doi.org/10.1007/978-3-540-87391-4_28

Siim Orasmaa (Tartu Ülikool) on erialalt informaatik. Uurimisvaldkondadeks on infootsingu meetodid ning tekstide automaatne ajasemantiline analüüs.
siim.orasmaa@gmail.com

AUTOMATIC RECOGNITION AND NORMALIZATION OF TEMPORAL EXPRESSIONS IN ESTONIAN LANGUAGE TEXTS

Siim Orasmaa
University of Tartu

In this article, I give an overview of the task of *temporal expression recognition* and normalization in natural language texts and describe a rule-based system, which was designed to solve this task in Estonian language texts.

A TimeML-based markup language was used to annotate temporal expressions. Following the markup possibilities in TimeML, I define two rough criteria for deciding whether a temporal expression is markable or not: 1) a markable expression should contain temporal information of *year*, *month*, *week*, *day*, *hour* or *minute* granularity, or 2) a markable expression should contain a reference to the past, present or future, anchored to the time of creation of a document or to some other markable expression.

I use a finite state technique for *temporal expression recognition* and define recognition rules as phrase patterns consisting of templates for words (describing a word by lemma or by regular expression) and descriptions of numeral phrases. Recognition rules were mostly used to extract the smallest phrases that carried separate temporal meanings and had unchangeable word order. Another set of rules was combined with built-in heuristics to join extracted phrases into longer temporal expressions.

The semantics of an expression were defined as a sequence of instructions, which needed to be executed in order to normalize the expression. Three types of instructions were distinguished: 1) calendar arithmetic instructions, 2) anchoring instructions, and 3) markup changing instructions. I also discuss special heuristics, which can be used to interpret semantically ambiguous expressions.

The system was tested on a 70,000-word subcorpora of the Estonian Reference Corpus. The test corpus consists of newspaper texts from six subgenres, legal texts, parliamentary transcripts and articles describing historical events or periods. I measured a recall of 82% and a precision of 98.4% as overall performance on recognition of temporal expressions. Normalization accuracy of the attribute "type" was 97.2% and for the attribute "value" was 87.4%.

Keywords: computational linguistics, semantic markup, annotation, temporal expressions, Estonian