

## Ajaväljendite märgendamise juhised

### Sissejuhatus

Käesolevas dokumendis toodud ajaväljendite märgendusviis lähtub ISO-TimeML märgenduskeelest [1], mis omakorda põhineb varasematel märgenduskeeltele TimeML [2] ja TIMEX2 [3]. Seda märgendusviisi kasutab ka eesti keelele loodud automaatne ajaväljendite tuvastaja [4].

Käesolev märgendusjuhise on üles ehitatud eelkõige ajaväljendite semantikat silmas pidades. Seega, kui tekstis esineb mõni potentsiaalne ajaväljend, on esimeseks küsimuseks olnud – *kas ja kuidas saab selle ajalist tähendust TimeML-is edasi anda*, muud kriteeriumid, sh morfoloogia ja süntaks, on märgendamisel olnud mõnevõrra tagaplaanil.

Märgendamisel ümbritsetakse ajaväljendifraas TIMEX-märgenditega ning märgendi atribuutides tuuakse välja ajaväljendi semantika.

Näide annoteeritud tekstilõigust. Teksti loomise ajaks on 2009-12-07.

Linna maksutulu võib <TIMEX type="DATE" value="2010">tuleval aastal</TIMEX> langeda kuni 300 miljonit krooni.

Märgendamise käigus määratakse ajaväljendi liik (atribuudis `type`), esitatakse ajaväljendi normaliseeritud semantika (atribuudis `value`) ning vajadusel semantika täpsustus (atribuudis `mod`). Atribuute `beginPoint`, `endPoint` ja `anchorTimeID` kasutatakse ajaväljenditevaheliste seoste esiletoomiseks (vahemikuseosed ja sõltuvusseosed); ajaliste korduvuste semantika esitamiseks kasutatakse eraldiseisvaid atribuute `quant` ja `freq`. Atribuut `temporalFunction` näitab, kas tegu on relatiivse ajaväljendiga ning kas selle semantika tuleb leida arvutuslikul teel.

Atribuuti `functionInDocument` kasutatakse dokumendis erilist staatust omavate väljendite äramärgimiseks. Praegu on kasutusel vaid üks eristaatus – `CREATION_TIME`, mis märgib dokumendi loomise kuupäeva.

Problemaatiliste ajaväljendite puhul antakse atribuudis `comment` lühikommentaari probleemi kohta.

### Ajaväljendite tüübid

Ajaväljendi tüüp pannaks kirja atribuuti `type`. Atribuudi kasutamine on kohustuslik (igal ajaväljendil peab olema tüüp) ning sellel võivad olla järgmised väärtused:

- `DATE` – ajateljele paigutatav kalendriaeg („toimumisaeg“), mis võib olla antud aasta-, kuu-, nädala- või päeva täpsusega. Kas ajateljele paigutamine toimub punkti või intervallina, ei täpsustata. Nt *järgmisel reedel, 2004. aastal*.
- `TIME` – ajateljele paigutatav kalendriaeg („toimumisaeg“), mis on antud kellaajalise täpsusega (siia alla kuuluvad ka umbmäärased kellaajad – päevaosad). Näiteks: *esmaspäeva hommikul, järgmisel reedel kell 14.00*.
- `DURATION` – ajaline kestvus, ajavahemik. Alguspunkt ja lõpp-punkt võivad olla määramata, nt *3 tundi*, varjatult määratud, nt *viimase viie päeva jooksul*, või ilmutatult määratud, nt *aastatel 2001-2004*.
- `SET` – ajaline korduvus, „aegade hulk“. Näiteks: *neljapäeviti, hommikuti*

Ilmutatud ajavahemike puhul märgendatakse kõik otspunktid eraldiseisvate ajaväljenditena (edaspidi käsitleme täpsemalt).

## Ajaväljendite semantika esituskujud

Ajaväljendi kalendripõhine semantika esitatakse atribuudis `value`. Atribuudi kasutamine on kohustuslik – kui tekstis leidub ajaväljend, mille semantikat ei saa avaldada ühelgi alljärgneval viisil, tuleks see ajaväljend jätta pigem märgendamata, kui jätta `value` määramata.

### Standardised ajaformaadid

Atribuudil `value` on vastavalt ISO aja märkimise standardile 4 üldist 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. Ainult kellaaega sisaldav formaat: `Thh:mm`
4. Ajalise kestvuse formaat: `PniYn2Mn3Wn4DTn5Hn6M`  
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 ajateljel paigutuvate aegade (sh ajavahemike otspunktide) ja ajaliste korduvuste semantika esitamisel (`DATE`, `TIME`, `SET`); formaat 3 leiab kasutust juhtudel, kui kellaajaga seotud kuupäeva pole võimalik täpsustada (`TIME`). Formaati 4 on mõeldud ajaliste kestvuste (`DURATION`) jaoks. NB! Kui ajaväljendi semantika on esitatav ühtviisi nii kuupõhises kui ka nädalapõhises formaadis, eelistatakse alati kuupõhist formaati.

### Kokkuleppelised eritähised

Lisaks arvudele võib formaatides 1, 2 ja 3 kasutada ka 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

Formaatides 1 ja 2 võib kasutada järgmisi kokkuleppelisi eritähiseid:

Nädalavahetus/tööpäev - kasutame nädalapõhises formaadis, nädalapäeva ("`wd`") asemel;

WD – *workday* – tööpäev (NB! ISO-TimeML ei kasuta seda tähistust)

WE – *weekend* – nädalalõpp

Aastaajad – kasutame 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 – kasutame 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

## Aja detailsuse määramine

Kuupõhist ja nädalapõhist semantika esitust võib paremast otsast lühendada, vastavalt sellele, millise granulaarsusega ajalist informatsiooni ajaväljend sisaldab. Näiteks:

(ankrupunkt<sup>1</sup> on 2009-12-17):

```
<TIMEX type="DATE" value="2009-12" mod="START">  
selle kuu alguses  
</TIMEX>
```

Eeltoodud näites ei sisalda ajaväljend kellaaeg- ning kuupäev-granulaarsusega informatsiooni, seetõttu jäetakse need granulaarsused täpsustamata. Aasta-granulaarsus tuuakse välja ankrupunkti põhjal.

Aastaarvu paremast otsast võib numbreid ära jätta, andmaks edasi aastakümneid ja sajandeid, nt:

```
<TIMEX type="DATE" value="199" mod="END"> 1990ndate lõpus </TIMEX>  
<TIMEX type="DATE" value="17"> 18. sajandil </TIMEX>
```

## Enne ja pärast *meie ajaarvamist*

NB! Kui ajaväljendiks on näiteks „*aastal 17. pKr*“, kasutatakse esituskuju `value="0017"`.

Märkimaks aastaid ja sajandeid *enne meie ajaarvamist*, liidetakse aastaarvule prefiks BC:

```
<TIMEX type="DATE" value="BC05">VI sajandist e. m. a</TIMEX>
```

## Ajalise kestvuse detailsuse määramine

Ajaliste kestvuste korral (formaad 3) tuuakse välja vaid väljendis mainitud granulaarsused, nt:

```
<TIMEX type="DURATION" value="PT3H"> kolm tundi </TIMEX>
```

NB! Tuleks jälgida, et minutid ja kuud segamini ei läheks (mõlemad on tähisega M) – minutite puhul peab `value` sisaldama kellaaeg-detailsuse tähist T:

```
<TIMEX type="DURATION" value="P5M"> viis kuud </TIMEX>  
<TIMEX type="DURATION" value="PT2M"> kaks minutit </TIMEX>
```

## Ajalise korduvuse semantika

Kui ajalise korduvuse edasiandmisel saab kasutada formaate 1, 2 või 3, siis seda ka tehakse, kattes sealjuures X-sümbolitega kinni mittekorduvad ajalised granulaarsused, nt

```
<TIMEX type="SET" value="XXXX-WXX-2"> teisipäeviti </TIMEX>
```

Teatud liiki ajalisi korduvusi pole aga võimalik eeltoodud viisil väljendada (nt ei saa selliselt anda edasi ajaväljendite *igal tunnil*, *kaks korda nädalas* semantikat). Sellisel juhul kasutatakse kestvusega sarnast esitusviisi: korduvust hõlmava perioodi pikkus tuuakse välja atribuudis `value`, atribuudis `quant` tuuakse välja korduvuse kvantori inglisekeelne nimetus ning atribuudis `freq` täpsustatakse täisarvuline kordumissagedus koos selle granulaarsusega. Näide:

```
<TIMEX type="SET" value="P1M" quant="EVERY" freq="3D">  
Kolm päeva igas kuus  
</TIMEX>
```

Kui kordumissagedust pole võimalik ajalise granulaarsusega siduda, kasutatakse sümbolit X märkimaks teadmata granulaarsust.

```
<TIMEX type="SET" value="P1M" quant="EVERY" freq="3X">  
Kolm korda igas kuus  
</TIMEX>
```

## Mittekonkreetne kalendriline semantika

Kui ajaväljend sisaldab mingi granulaarsusega kalendrilit informatsiooni, ent granulaarsuse täpne väärtus on umbmäärane või raskestimääratletav, kasutatakse semantika esitamisel granulaarsuse kinnikatmist X sümbolitega. Näited:

```
<TIMEX type="DATE" value="XXXX-XX-XX">
```

---

1 Ankrupunkt – ajapunkt, mille suhtes relatiivse ajaväljendi semantika leitakse (enamasti *dokumendi loomise aeg*, v.a mõned erijuhud).

```

Ühelkenal päeval
</TIMEX>
<TIMEX type="DATE" value="XXXX-03-XX">
Ühel märtsikuu päeval
</TIMEX>
<TIMEX type="DURATION" value="PTXH"> mitu tundi </TIMEX>

```

### Ligikaudsed viited minevikule, olevikule ja tulevikule

Kui ajaväljendis ei leidu kalendrilist informatsiooni, küll aga on tegu viitega minevikule, olevikule või tulevikule, kasutatakse value osas vastavalt kokkuleppelisi väärtuseid PAST\_REF (minevikuviide), PRESENT\_REF (olevikuviide) ja FUTURE\_REF (tulevikuviide). Näiteks:

```

<TIMEX type="DATE" value="PAST_REF" anchorTimeID="t0">
hiljuti
</TIMEX>
<TIMEX type="DATE" value="FUTURE_REF" anchorTimeID="t0">
tulevikus
</TIMEX>

```

Atribuudiga anchorTimeID täpsustatakse, millisest ankrupunktist ligikaudne viide lähtub (anchorTimeID="t0" tähendab, et ankrupunktiks on *dokumendi loomise aeg*). Tüübiks peaks olema alati DATE.

### Ajaväljendite semantika täpsustamine

Ajaväljendi märgenduse ulatus peaks hõlmama kõiki väljendi ajalist tähendust täpsustavaid ees- ja järeltäiendeid (nt eestäiendid *umbes*, *rohkem kui* ning järeltäiendid *lõpus*, *keskosas*). Täpsustuse korral kasutatakse atribuuti mod täpsustuse edasiandmiseks. Atribuudil võivad olla järgmised väärtused:

- START – ajaväljend viitab ajastatava kalendriaaja algusosale. Näiteks:

```

<TIMEX type="DATE" value="2009" mod="START">
2009. aasta alguses
</TIMEX>
<TIMEX type="DATE" value="2007-06" mod="START">
juuni alguseks 2007. aastal
</TIMEX>

```
- MID – ajaväljend viitab ajastatava kalendriaaja keskosale.
- END – ajaväljend viitab ajastatava kalendriaaja lõpuosale.
- FIRST\_HALF – märgib, et mõeldakse ajastatava kalendriaaja väikseima ilmutatud granulaarsuse<sup>2</sup> „esimest poolt“. Näiteks:

```

<TIMEX type="DATE" value="2009" mod="FIRST_HALF">
2009. aasta esimesel poolel
</TIMEX>

```

NB! TIMEX2 ja ISO-TimeML näe ette selle väärtuse kasutamist. Poolaastate märkimiseks kasutatakse küll märgendeid H1 ja H2, ent need ei laiene teistele granulaarsustele.

- SECOND\_HALF – märgib, et mõeldakse väikseima ilmutatud ajalise granulaarsuse „teist poolt.“ *TIMEX2 ja ISO-TimeML ei toeta selle väärtuse kasutamist.*
- APPROX – märgib, et toodud ajaväljendi semantika ei ole täpne. Enamasti kasutatakse

<sup>2</sup> Granulaarsuste võrdlemine: (siin: väiksem granulaarsus = väiksema ajaühikuga granulaarsus) ehk siis nt *kuu*-granulaarsust loetakse väiksemaks kui *aasta*-granulaarsust.

umbmäärasusele viitavate sõnade (*umbes, ligikaudu* jms) edasiandmiseks. Näiteks:

```
<TIMEX type="DURATION" value="P4Y" mod="APPROX">
umbes 4 aastat
</TIMEX>
```

Standardi järgi on lubatud veel järgmised täpsustused: BEFORE, AFTER, ON\_OR\_BEFORE, ON\_OR\_AFTER, LESS\_THAN, MORE\_THAN, EQUAL\_OR\_LESS, EQUAL\_OR\_MORE.

Näiteks:

```
<TIMEX type="DURATION" value="P4Y" mod="LESS_THAN">
peaaegu 4 aastat
</TIMEX>
```

## Ajavahemike semantika esitamine

Ajavahemike korral märgendatakse iga otspunkt eraldiseisvalt ning seejärel tuuakse välja nende kaudu tekkiv ajaline kestvus. Järgnevas näites on toodud ajaväljendi "12-15 märts 2009" otspunktide ja kestvuse märkimine:

```
<TIMEX tid="t14" type="DATE" value="2009-03-12">
12
</TIMEX>
-
<TIMEX tid="t15" type="DATE" value="2009-03-15">
15 märts 2009
</TIMEX>
<TIMEX tid="t16" type="DURATION" beginPoint="t14" endPoint="t15"/>
```

Igale tekstis märgendatud ajavahemiku otspunktile lisatakse unikaalne identifikaator (atribuut `tid`), mille väärtus peab olema teksti piires unikaalne. Märgenduse lõppu (või teksti lõppu) lisatakse ilma tekstilise sisuta ajaline kestvus (näites ajaväljend identifikaatoriga `t16`), mis ühendab väljatoodud otspunktid ühtseks vahemikuks (atribuutide `beginPoint` ja `endPoint` abil otspunktidele viidates).

Kui ajavahemik avaldub eelkõige kestvusena, mille otspunktid pole tekstis ilmutatud kujul välja toodud, aga need on võimalik tuletada (fraasi sisu ja ankrupunkti järgi), tuuakse välja kestvus ning selle otspunktid avaldatakse sisuta märgenduste kujul. Järgnevas näites on ankrupunktiks 2009-12-10:

```
<TIMEX tid="t17" type="DURATION" value="P3M" beginPoint="t18"
endPoint="t19">
tänavu kolme esimese kuu jooksul
</TIMEX>
<TIMEX tid="t18" type="DATE" value="2009-01"/>
<TIMEX tid="t19" type="DATE" value="2009-03"/>
```

NB! Käsitsi märgendades võib piirduda ainult tekstis ilmutatud kujul oleva informatsiooni märgendamisega. Ehk siis, kahe eelneva näite puhul märgendada vaid väljendid `t14`, `t15` ja `t17`.

## Relatiivsed ajaväljendid ja *temporalFunction*

Relatiivseteks ajaväljenditeks nimetame DATE ja TIME ajaväljendeid, mille ajateljele

paigutamiseks ei piisa ajaväljendis sisalduvast informatsioonist ning on tarvis täiendavat konteksti-informatsiooni (ja kalendriarvutusi). Sellised ajaväljendid tuuakse esile, lisades märgendusse atribuudi/väärtuse `temporalFunction="true"` (ehk öeldakse, et tegemist on nõ „arvutamist nõudva väärtusega“). Näiteks

(ankrupunktiks on 2009-12-17)

```
<TIMEX type="DATE" value="2009-11" temporalFunction="true" mod="END">
eelmise kuu lõpus
</TIMEX>
```

Ja ka vastupidi – `temporalFunction="false"` märgib jälle seda, et tegemist on nn absoluutse (ilma arvutusteta ajateljele paigutatava) ajaväljendiga. Nt

```
<TIMEX type="DATE" value="2004-SU" temporalFunction="false">
2004. aasta suvel
</TIMEX>
```

NB! Käsitsi märgendades pole `temporalFunction` väärtuste määramine prioriteetne ning need võib ka ära jätta.

### Relatiivsete väljendite ajalise detailsuse määramine

Kui tekstis kasutatakse nt ajaväljendit „*kolm kuud tagasi*“ ning teksti loomise aeg on antud kuupäeva täpsusega, võib tekkida küsimus, kas arvutuslikult leitud ajaväljendi semantika tuleks esitada *kuupäev*- või *kuu*-täpsusega. Mina olen sellistel juhtudel võtnud konservatiivse lähenemise ning toonud välja pigem vähem detaile kui rohkem – ehk antud näite puhul – esitanud ajaväljendi *kuu*-täpsusega. Ei ole muidugi välistatud, et kontekstist sõltuvalt võib olla tarvis minna ka detailsemaks.

### Abiks kalendriarvutuste tegemisel

Käsitsi märgendamisel võib olla relatiivsetele ajaväljenditele vastavate kalendriaegade arvutamisel abi olla matemaatikute otsingumootorist <http://www.wolframalpha.com>

- Kui trükkida otsingumootoris ISO formaadis kuupäev (nt 2012-01-10), saab selle kohta mitmesugust detailset kalendri-informatsiooni, sh öeldakse, millise nädalapäevaga on tegu;
- Otsingumootoris võib sisestada kalendriarvutusi, nt 2012-01-10 + 30 days;
- Otsingumootorist saab infot ajaliste perioodide pikkuse kohta, nt  
2012-01-13 to 2012-03-21

### Ankurdatavad ajaväljendid

Relatiivsete ajaväljendite erijuht on *ankurdatavad ajaväljendid*, mille semantika leitakse mingi teise tekstis esineva ajaväljendi semantika alusel. Näiteks lauses „*Detsembris oli keskmine temperatuur kaks korda madalam kui kuu aega varem*“ leitakse ajaväljendi „*kuu aega varem*“ semantika ajaväljendi „*Detsembris*“ semantika alusel. Sellisel juhul kasutatakse atribuuti `anchorTimeID` et viidata ankrupunktiks olevale ajaväljendile. Näide:

```
<TIMEX tid="t21" type="DATE" value="2009-12">
Detsembris
</TIMEX>
oli keskmine temperatuur kaks korda madalam kui
<TIMEX tid="t22" type="DATE" value="2009-11" temporalFunction="true"
anchorTimeID="t21">
kuu aega varem
</TIMEX>
```

Eeltoodud näiteks kasutatakse väljendi `t22` semantika leidmisel ankrupunktina ajaväljendit `t21`.

### Spetsiifilised ja problemaatilised ajaväljendid

Järgnevalt toome eraldi välja mõned spetsiifilised/problemaatilised ajaväljendid. Problemaatiliste ajaväljendite puhul tuuakse välja strateegiad, mida on kasutatud nende automaatsel märgendamisel.

Parema loetavuse mõttes kasutatakse enamikus ajaväljendi-näidetes vaid *aasta*-granulaarsust, kuigi potentsiaalselt võidakse kasutada suvalist granulaarsust (*kuu, nädal, minut* jms). Näidete lühendamiseks kasutatakse mõnikord TIMEX-märgendite asemel sümboleid [ ja ].

### **spetsiifilised fraasikonstruktsioonid**

- *neljapäeval, 17. juunil*

Kuupäev koos täpsustava nädalapäevaga märgendatakse tervikliku ajaväljendina: [ *neljapäeval, 17. juunil* ].

- *täna, 100 aastat tagasi*

Märgendatakse ühe tervikliku fraasina [ *täna, 100 aastat tagasi* ];

- *neljapäeva öösel vastu reedet*

Märgendatakse ühe tervikliku fraasina: [ *neljapäeva öösel vastu reedet* ].

### **omadussõnalised ajaväljendid**

Omadussõnaliste ajaväljendite all mõeldakse väljendeid, mille peasõna on *-ne* lõpuline omadussõna, näiteks *kolme aasta tagune, möödunud aastane, eilne, praegune*. Kuna selliste ajaväljendite semantika on sageli esitatav TimeML raamistikus, siis leiavad need ka automaatse ajaväljendite tuvastaja poolt märgendamist.

Käitsi märgendamine: kuna süntaktiliselt on sageli tegemist eestäienditega, võib märgendamisel tunduda küsitav, kas antud ajaväljendeid tuleks üldse märgendada, eriti kui fraasi peasõna pole „sündmus“, nt „[*homses*] *Postimehes*“, „[*22-aastane*] *Aivar*“. Eelistatud lähenemine oleks need märgendada vaid siis, kui ajaväljend laiendab mingit sündmusena märgendatud sõna (nt [ *neljapäevane* ] [ *koosolek* ]).

NB! Teatud juhtudel ei ole omadussõnalised ajaväljendid fraasi koosseisus, vaid esinevad iseseisvalt, nt „*Saksamaa maksab [praeguseni] I maailmasõja võlgu*“. Sellisel juhul on need tüüpiliselt finiiitse verbiga reksioonisuhetes ning tuleks kindlasti märgendada.

### **grammatilistes käänetes toimumisaeg-väljendid**

Siin all mõeldakse DATE/TIME ajaväljendeid, mille peasõna on grammatilises käändes, nt *2003. aasta, eelmise nädala, kevad, möödunud aastat*. Sellistel juhtudel võib samuti tekkida küsimus, kas ajaväljend tuleks üldse märgendada, kui see esineb subjekti/objekti rollis (nt lauses „*Seevastu [pühapäev] sobib hästi suhtlemiseks*“) või on mõne fraasi täiend (nt [ *elmise nädala* ] *koosolekut*). Käitsimärgendamisel tuleks sellised ajaväljendid märgendada, kui need saab siduda mingi tekstis märgendatud sündmusega.

Automaatne ajaväljendite tuvastaja märgendab kõik sellised juhud valimatult. Probleemid tekivad semantika määramisel (nt tuvastaja eeldab alati, et semantika on konkreetne, kuigi lauses „[ *pühapäev* ] *sobib hästi suhtlemiseks*“ ei pruugita mõelda ühtegi konkreetset pühapäeva).

### **viimastel, esimestel koos suurema ilmutatud granulaarsusega**

- *aasta viimastel kuudel, nädalatel, päevadel*
- *aasta esimestel kuudel, nädalatel, päevadel*

Märgendatakse terviklikuna (nt [ *aasta esimestel kuudel* ]). Tüübiks saab DATE või TIME; semantika esitamisel tuuakse välja vaid suurem granulaarsus (*aasta*) ning väiksema granulaarsuse (*kuudel, nädalatel, päevadel*) semantika esitatakse täpsustusena (mod=“START“ või mod=“END“).

### **... jooksul ning {kvantiteet} jooksul konstruktsioonid**

Kaassõnaga „*jooksul*“ ajaväljendifraaside semantika esitamine võib olla keeruline, kuna ilmutatud

kujul semantika tundub olevat midagi toimumisaja ja kestvuse vahepealset. Näiteks kui „viimase aasta jooksul” võib veel esitada toimumisajana (DATE+viimane aastanumber), siis „viimase kolme aasta jooksul” on ilmutatud kujul pigem kestvus (DURATION+3 aastat).

Ilmselt tuleks ühikulise kvantiteediga juhtudel (nt „esmaspäeva jooksul”, „järgneva aasta jooksul” jne) püüda semantika määramisel lähtuda kontekstist; samas suurema kui ühikulise kvantiteetiga väljendite (nt „viimase kolme kuu jooksul”, „järgneva kahe aasta jooksul” jne) korral on eelistatud märgendamine DURATION-väljendina.

DURATION-väljendina märgendamise puhul on võimalik välja tuua ka kestvuse algus- ning lõpp-punkt. Näide:

(ankrupunkt on 2010-11-17)

```
<TIMEX tid="t25" type="DURATION" value="P3Y" beginPoint="t0"
endPoint="t26">
lähima kolme aasta jooksul
</TIMEX>

<TIMEX tid="t26" type="DATE" value="2013" temporalFunction="true"
comment="value problemaatiline" />
```

(Eelnevas näites märgib t0 ankrupunkti ehk dokumendi loomise ajahetke)

NB! Ajaväljendi fraasipiiride määramisel: alternatiivne võimalus on jätta sõna *jooksul* ajaväljendist välja ning märgendada kui ajalise seose märgusõna ehk „signaal” (vt täpsemalt dokumendist „seoste märgendamine”), nt

```
<TIMEX tid="t40" type="DURATION" value="P7Y">
Seitsme kuu
</TIMEX>
<SIGNAL sid="s1">
jooksul
</SIGNAL>
vajus tiib 1,3 sentimeetrit.
```

Eelnevas lauses märgib sõna *jooksul* ilmutatud kujul ajalist seost ajaväljendi „seitse kuud” ning sündmuse „vajus” vahel.

### ***lähi-... liitsõnakonstruktsioon***

- *lähipäevil/-kuudel/-aastatel*, aga ka „*lähiajal*”  
Ajakirjandustekstides on eeldatud, et tegu on tulevikuviitega ning on kasutatud märgendit FUTURE\_REF; Loomulikult tekib teatud semantikakadu, kuna *lähipäevil* ≠ *lähiaastail*;

### ***eelmistel, järgmistel / eelnenud, järgnenud ... konstruktsioonid***

- *eelmistel/möödunud/minevatel/eelnevatel/... aastatel*
- *järgmistel/tulevatel/eelolevatel/... aastatel*  
Automaatsel märgendamisel esitatakse umbmäärase kestvusena (DURATION: PXY), kus kestvuse alguspunkt on dokumendi loomise aeg ning lõpp-punktiks on ilma sisuta märgend, kus konkreetne väärtus on välja arvatamata. Näiteks:  
<TIMEX tid="29" type="DURATION" value="PXM" beginPoint="t0" endPoint="t30"> eelolevatel kuudel </TIMEX>  
<TIMEX tid="t30" type="DATE" temporalFunction="true" comment="value problemaatiline" />
- *eelnenud aastate jooksul*
- *järgnenud aastate jooksul*  
Esitatakse umbmäärase kestvusena (DURATION: PXY), mille puhul tuuakse välja ka lõpp-punkt ja alguspunkt. Ankurdada tuleb tõenäoliselt mingi dokumendis esineva ajaväljendi



külge. Ligikaudne näide:

```
<TIMEX tid="t31" type="DATE" value="2007"> aastal 2007. </TIMEX>
. . .
<TIMEX tid="t32" type="DURATION" value="PXM" beginPoint="t31"
endPoint="t33">
eelnenud aastate jooksul
</TIMEX>
<TIMEX tid="t33" type="DATE" temporalFunction="true"
comment="problemaatiline value" />
```

### **varasem, hilisem ... konstruktsioonid**

- *varasemad aastad, hilisemad aastad*  
Kasutatakse umbmäärast viidet (PAST\_REF, FUTURE\_REF). Alternatiiv oleks kasutada perioodi, aga see oleks üsnagi määramatu periood (teadmata nii kvantiteet kui ka paiknemine ajateljel);
- *N varasemat aastal, N hilisemat aastat*  
Avaldatakse kestvusena (DURATION: PNY), algus ja lõpp-punkt jäävad määramata;

### **vahemikuväärtustega varem, hiljem konstruktsioonid**

- *viie-kuue aasta pärast, kahe-kolme aasta tagune*  
Märgendada tuleks mõlemad otspunktid eraldi, nt *[kahe]-[kolme aasta tagune]*. Seeläbi on võimalik ka mõlema ajapunkti semantika eraldi välja tuua. (Automaatne märgendaja jätab tüüpiliselt alguspunkti märgendamata).

### **umbmäärased tagasi, pärast, varem, hiljem konstruktsioonid**

- *aastaid tagasi; aastate eest/tagune; aastaid varem;*
- *aastate pärast; aastaid hiljem;*  
Automaatsel märgendamisel kasutatakse umbmäärast viidet (PAST\_REF, FUTURE\_REF). Seega jääb semantika mõnevõrra ebatäielikuks (st ei tehta vahet, kas „aastaid tagasi“ või „päevi tagasi“). Atribuudis anchorTimeID tuuakse välja ajahetk, mille suhtes viide rakendub. Järgnevas pealiskaudses näites ankurdub „aastate pärast“ väljendi „1999“ külge:  
<TIMEX tid="t34" type="POINT" value="1999">1999</TIMEX>  
. . .  
<TIMEX tid="t35" type="POINT" value="FUTURE\_REF" anchorTimeID="t34">  
aastate pärast </TIMEX>
- *mõned/mitmed aastad tagasi/varem; mõni aasta tagasi/varem;*
- *mõned/mitmed aastad hiljem; mõne/mitme aasta pärast;*  
Analoogselt eelmise punktiga: kasutatakse umbmäärasuse tähiseid (PAST\_REF, FUTURE\_REF).

## **Kirjandus**

[1] ISO-TimeML standard (*mustand*)

[http://lirics.loria.fr/doc\\_pub/SemAFCD24617-1Rev12.pdf](http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf) (07.10.2011)

[2] TimeML raamistik, versioon 1.2.1

<http://timeml.org/site/publications/specs.html> (07.10.2011)

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

[http://projects ldc.upenn.edu/ace/docs/English-TIMEX2-Guidelines\\_v0.1.pdf](http://projects ldc.upenn.edu/ace/docs/English-TIMEX2-Guidelines_v0.1.pdf) (07.10.2011)

[4] S.Orasmaa. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. ERÜ aastaraamat, 2012.

<http://rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/13/0>