Ajaväljendite tuvastaja märgendusformaat

Sissejuhatus

Ajaväljendite tuvastaja [4] poolt kasutatud märgendusformaat lähtub ISO-TimeML märgenduskeelest [1], mis omakorda põhineb varasematel märgenduskeeltel TimeML [2] ja TIMEX2 [3]. Käesolevas dokumendis kirjeldatakse märgendusformaati ning ühtlasi tuuakse välja selle mõningad erinevused võrreldes ISO-TimeML märgenduskeelega.

Märgendus

Automaatsel märgendamisel ümbritsetakse ajaväljendifraas `TIMEX`-märgenditega[1] ning märgendi atribuutides tuuakse välja ajaväljendi semantika.

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

```
 Linna maksutulu võib <TIMEX tid="t1" type="DATE" value="2010">tuleval
aastal</TIMEX> langeda kuni 300 miljonit krooni.
```

Märgendamise käigus määratakse ajaväljendi unikaalne identifikaator (`tid`), 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 on leitud arvutuslikul teel.
Atribuuti `functionInDocument` kasutataks dokumendis erilist staatust omavate väljendite äramärkimiseks. Praegu on kasutusel vaid üks eristaatus – `CREATION_TIME`, mis märgib dokumendi loomise kuupäeva.
Kõigi mainitud atribuutide kasutust ja võimalikke väärtusi kirjeldatakse detailsemalt järgnevates sektsioonides.

NB! Atribuudid (sh ka kohustuslikud atribuudid `type` ja `value`) võivad süsteemi poolt väljastatud märgendusest mõningatel juhtudel ka puududa, nt siis, kui ajaväljendi normaliseerimisel on midagi valesti läinud või atribuudi väärtuse määramine on käesoleva märgendusformaadi piirides problemaatiline.

Ajaväljendite tüübid

Ajaväljendi tüübi määrab atribuut `type`. Atribuudi kasutamine on kohustuslik 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

1 Rangelt standardit järgides peaks märgendi nimetus olema TIMEX3. Kuna käesolev märgendus kaldub teatud juhtudel standardist kõrvale, kasutatakse nime TIMEX.

aastatel 2001-2004.

- `SET` – ajaline korduvus, „aegade hulk“. Näiteks: *neljapäeviti, hommikuti*

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

Ajaväljendite semantika esituskujud

Ajaväljendi kalendripõhine semantika esitatakse kohustuslikus atribuudis `value`.

Standardsed ajaformaadid

Atribuudil `value` on vastavalt ISO aja märkimise standardile 4 üldist formaati:

1. Kuupõhine formaat: *yyyy-mm-dd***T***hh: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*-**W***nn-wd***T***hh:mm*
   nn – nädal aastas (01-53)
   wd – päev nädalas (1-7, kus 1 on esmaspäev ja 7 on pühapäev)
3. Ainult kellaaega sisaldav formaat: **T***hh:mm*
4. Ajalise kestvuse formaat: **P***n*$_1$Y*n*$_2$M*n*$_3$W*n*$_4$D**T***n*$_5$H*n*$_6$M
   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`). Formaat 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[2] 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 (formaat 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 kalendrilist informatsiooni, ent granulaarsuse täpne väärtus on umbmäärane või raskestimääratletav, kasutatakse semantika esitamisel granulaarsuse kinnikatmist `X` sümbolitega. Näited:

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

```
<TIMEX type="DATE" value="XXXX-XX-XX">
Ühel kenal 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 kalendriaja 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 kalendriaja keskosale.
- `END` – ajaväljend viitab ajastatava kalendriaja lõpuosale.
- `FIRST_HALF` – märgib, et mõeldakse ajastatava kalendriaja väikseima ilmutatud granulaarsuse1[3] „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.*

3 Granulaarsuste võrdlemine: (siin: väiksem granulaarsus = väiksema ajaühikuga granulaarsus) ehk siis nt *kuu*-granulaarsust loetakse väiksemaks kui *aasta*-granulaarsust.

- APPROX – märgib, et toodud ajaväljendi semantika ei ole täpne. Enamasti kasutatakse 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 on lisatud unikaalne identifikaator (atribuut tid), mille väärtus on 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).

Praegu jätab süsteem ilma tekstilise sisuta kestvused välja arvutamata: tuuakse välja ainult umbmäärasusele viitav normaliseering (nt value väärtused PXXY, PXD).

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! Kui vahemiku otspunktide väljatoomine on probleemaatiline, võidakse kasutada ka tähistust `beginPoint="?"` või `endPoint="?"`. Seega tuleks enne `beginPoint` või `endPoint` kasutamist alati kontrollida, kas need ikka vastavad `tid` formaadile.

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 nö „arvutamist nõudva väärtusega"). Näiteks

(ankrupunktiks on 2009-12-17)

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

Selline täpsustus võimaldab väljundit töötlevatel süsteemidel semantika vajadusel üle arvutada, kasutades mingeid paremaid heuristikuid.

Analoogselt märgib `temporalFunction="false"` seda, et tegemist on nn absoluutse (ilma arvutusteta ajateljele paigutatava) ajaväljendiga. Nt

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

Muud erisused: tüübi `SET` puhul on `temporalFunction` alati `true` ning tüübi `DURATION` puhul valdavalt `false` (v.a juhud, kus kestvus on umbmäärane [nt *aastaid*] või märgib „suunaga" ajavahemikku [nt *lähima viie aasta jooksul*]).

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 näites kasutatakse väljendi `t22` semantika leidmisel ankrupunktina ajaväljendit `t21`.

Märkused

- Dokumendi loomise aja (`t0`) märkimine ankrupunktina on praeguses süsteemi väljundis mittesüstemaatiline, järjekindlalt tehakse seda vaid ligikaudsete viidete korral (`PAST_REF`, `PRESENT_REF`, `FUTURE_REF`).
- Kui ankurdamine on problemaatiline või ebaõnnestunud, on atribuudi `anchorTimeID` väärtuseks `"??";`

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 omadusõna, näiteks *kolme aasta tagune*, *möödunudaastane*, *eilne*, *praegune*. Kuna selliste ajaväljendite semantika on sageli esitatav TimeML raamistikus, siis leiavad need ka automaatse ajaväljendite tuvastaja poolt märgendamist.

Probleemikoht: 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*“. Samas leidub ka omadussõnalisi ajaväljendeid, mis peaaegu alati on iseseisvad (st pole fraasi koosseisus ja alluvad otse finiitverbile), nt „*Saksamaa maksab [praeguseni] I maailmasõja võlgu“*.

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 *[eelmise nädala] koosolekut*).

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**

- *lähima aasta jooksul*
- *Järgnevaks N-ks aastaks ↔ N-ks järgnevaks aastaks*

- *lähima N aasta jooksul ↔ N lähima aasta jooksul*
- *lähima paari aasta jooksul ↔ paari lähima aasta jooksul*

  Eeldatakse, et väljend avaldub dokumendi loomise aja suhtes ja viitab tulevikule. Vt all täpsemalt.

- *eelnenud N aasta jooksul ↔ N eelnenud aasta jooksul*
- *järgnenud N1-N2 aasta jooksul ↔ N1-N2 järgnenud aasta jooksul*
- *viimase N aasta jooksul ↔ N viimase aasta jooksul*
- *viimase paari aasta jooksul ↔ paari/mõne/mitme viimase aasta jooksul*
- *viimase aasta jooksul*

  Eeldatakse, et väljendid lahenduvad dokumendi loomise aja suhtes. Esitatakse kestvusena (`DURATION: PNY`) ning tuuakse välja ka kestvuse algus- ning lõpp-punkt. Kui väljend viitab tulevikule, on `beginPoint="t0"`; minevikuviite puhul on `endPoint="t0"`. Puuduv otspunkt lisatakse kui tekstilise sisuta märgend. 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)

***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*
  Esitatakse umbmäärase kestvusena (`DURATION: PXY`), kus kestvuse alguspunkt on dokumendi loomise aeg ning lõpp-punktiks on ilma sisuta märgend, kus konkreetne väärtus on välja arvutamata. 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 varasemal 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).

konkreetsed *tagasi, pärast, varem, hiljem* konstruktsioonid

- *kolm aastat varem; neli aastat hiljem;*
  Kui ajaväljendi ankurdamine ebaõnnestub (nt ei leita tekstist eespoolt sobivat ajaväljendit, mille külge ankurdada), normaliseeritakse ajaväljend kestvusena ning tuuakse välja kestvuse pikkus; `temporalFunction=“true“` märgib seda, et ajaväljendi lõpliku semantika leidmine vajab täiendavat tööd. Märgenduse ulatus jääb ka sellisel juhul samaks (nt *[kolm aastat varem]*).

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="DATE" value="1999">1999</TIMEX>
  . . .
  <TIMEX tid=“t35“ type="DATE" 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`).

Muud tehnilist

Sõnestamisest

Ajaväljendite tuvastajal on oma sisemine teksti sõnestamiseviis (ingl *tokenization*), mis on mõeldud numbrifraaside viimiseks ühtsemale kujule. Kui juhtub, et see sõnestamisviis ei lange kokku sisendis oleva sõnestusega, siis väljundis sõnestust muutma ei hakata, kuid märgenduses kasutatakse `text` atribuuti, et näidata, millised alamosad sisendsõnest on tegelikult ajaväljendi

koosseisus ning milline oli nende süsteemisisene tükeldus. Nt.

```
<TIMEX text="01. 01. 2001" type="DATE" value="2001-01-01">
jõust.01.01.2001
</TIMEX>
```

Kirjandus

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

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 (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