## Nimeolemite märgendamise juhend

**Koostaja:** Laura Katrin Leman

**Autorid:** Laura Katrin Leman, Kairit Sirts ([kairit.sirts@ut.ee](mailto:kairit.sirts@ut.ee))

**Ülevaade**

Käesoleva töö eesmärgiks on märgendada eestikeelne korpus nimeolemitega.

Nimeolemid on mingile isikule, objektile, asukohale, organisatsioonile jms viitavad sõnad või sõnade ühendid, s.t pärisnimed. Neid saab lahterdada eri kategooriatesse ning kategooriale vastava lühendiga nimeolemis esinev(ad) sõna(d) tekstis annoteerida.

**Märgendamissüsteem**

Annoteerimiseks kasutame *LabelStudio*-nimelist tarkvara, millega saab tutvuda lehel <https://labelstud.io/>. Märgendamiseks on vajalik, et arvutis oleks Pythoni versioon 3.5 või suurem. Täpsemaid juhiseid LabelStudio kasutamiseks vaata [siit](https://docs.google.com/document/d/1rdOqz-Vx5CNL3UyCueYiejZw20wxYF542QEk-dJOryE/edit?usp=sharing).

Annoteerimiseks tuleb esmalt klikkida ülevalt reast soovitud märgendile ja siis tekstist leitud nimeolem kursoriga valida. Märgendi kustutamiseks tuleb esmalt klikkida märgendil ja seejärel vajutada Backspace-nuppu.

Kindlasti tuleb pärast teksti märgendamist vajutada *Submit*-nuppu, vastasel juhul märgendused ei salvestu.

**Nimeolemite märgendamise üldine juhend**

Nimeolemeid eristab tavalistest nimisõnadest see, et need viitavad mingile konkreetselt identifitseeritavale referendile, mitte üldisele nähtusele. Näiteks on nimeolem sõna *Läänemeri*, aga mitte lihtsalt sõna *meri*.

Meie märgendussüsteemis on nimeolemid jaotatud **11-sse rühma**: isikud (PERsons), asukohad (LOCations), organisatsioonid (ORGanisations), geopoliitilised entiteedid (GPE), tooted (PRODucts), sündmused (EVENT), kuupäevad (DATE), kellaajad (TIME), tiitlid (TITLE), raha (MONEY) ja protsendid (PERCENT).

Nimeolemid võivad esineda kujul, kus olem kuulub tervikuna ühte kategooriasse, kui ka kujul, kus ühe suurema, mitmesõnalise olemi sees on pesastunud väiksem olem. Selle kohta saab lähemalt lugeda [**Hierarhia märgendamise**](#kix.6d4oovaz6xm) alt.

Üldiselt on nimeolemeid võimalik edukalt tuvastada ja kategooriatesse liigitada emakeelekõneleja keelevaistu põhjal. Kui mõne keerukama nimeolemi puhul tekib kahtlus, mis kategooriasse seda lisada, soovitame panna see tüübi alla, mis hetkel kõige intuitiivsem tundub, ning hiljem tagasisides probleemne koht välja tuua.

Järgnevalt on iga kategooria kohta toodud näiteid, mis peaksid märgendamist lihtsustama.

**Nimeolemite tüübid**

Saab jagada 11-ks:

**PERSONS (PER)**

Selle alla kuuluvad kõikvõimalikele reaalsetele ja fiktsionaalsetele isikutele viitavad nimed:

* eesnimed, nt *Jüri, Hillary, Robert*
* perekonnanimed, nt *Ratas, Clinton, Balboa*
* hüüdnimed, nt *Rocky*
* eelnevate kombinatsioonid, nt *Jüri Ratas, Robert "Rocky" Balboa*
* initsiaalid, nt *M.M.*
* mütoloogilised või fiktsionaalsed tegelased, nt *Medusa, Ämblikmees*

Kui isikunimele eelneb tiitel, nt *president Kersti Kaljulaid*, siis tiitel PERSONS märgendi alla ei kuulu.

**GPE**

GPE ehk geopoliitilise entiteedi alla kuuluvad kõik geograafilised asukohad, mis on seotud mingi poliitilise organisatsiooniga. See tähendab nii riike, linnu, impeeriume jne. Näiteks kuuluvad siia

* riigid, nt *Eesti, Hiina Vabariik, Venemaa, Vene Föderatsioon, Vatikan*
* linnad jms asulad, nt *New York, Tallinn, Brüssel, Ahja, Sindi*
* impeeriumid jms ühendused, nt *Euroopa Liit*, *Briti impeerium, Vene keisririik, Nõukogude Liit, Bütsants*
* osariigid, maakonnad jms haldusüksused, nt *Põlvamaa, Arizona*

**LOCATIONS (LOC)**

Selle alla kuuluvad kõik geograafilised asukohad, mis pole seotud konkreetse poliitilise organisatsiooniga nagu GPEd. See tähendab siis näiteks

* mandrid ja maailmajaod, nt *Lõuna-Ameerika, Aafrika*
* geograafilised alad, nt *Lõuna-Euroopa, Kagu-Aasia, mandri-Eesti*
* ehitised, nt *Empire State Building*
* mered, järved, jõed, ojad jms, nt *Läänemeri, Peipsi järv, Emajõgi*
* pargid, nt *Lahemaa rahvuspark, Barclay park*
* mäed, kõrbed jms maastikuvormid, nt *Kilimanjaro, Sahara, Suur Munamägi*
* planeedid, kosmilised objektid, nt *Maa, Jupiter, Kuu*

**ORGANISATIONS (ORG)**

Selle alla kuuluvad kõikvõimalikud selgelt ja üheselt identifitseeritavad organisatsioonid, nii näiteks firmad jms kommertsasutused kui administratiivorganid, täpsemalt näiteks

* firmad, nt *Apple, Microsoft*
* komisjonid, komiteed, bürood, liidud jms, nt *Euroopa Komisjon, Luure Keskagentuur, Kanada Rahvusvahelise Arengu Agentuur, Eesti Leivatootjate Liit*
* kogukonnad, inimeste rühmitused (sh nt bändid), nt *The Beatles, Occupy Wall Street*
* haridus- ja teadusasutused, nt *21. kool, Tartu Ülikool, Kõrgem Sõjakool, Euroopa Kosmoseagentuur, CERN, Riigi Ilmateenistus*
* õigusasutused, nt *Ameerika Ülemkohus, Euroopa Inimõiguste Kohus, Eesti Advokatuur*
* uudistelehed, -agentuurid, -blogid, ajalehed, nt *Delfi, Reuters*
* parteid, nt *Reformierakond, Liberal Democrats*
* avaliku halduse üksused, nt *Vabariigi Valitsus, Maksu- ja Tolliamet, PPA, Pärnu linn, Tartu linnavolikogu*
* spordiliigad ja -ühendused, nt *Eesti Jalgpalli Liit, National Hockey League*
* pangad, börsid, finantsasutused, nt *Swedbank, Londoni börs*
* veebilehed, kui need viitavad organisatsioonile, nt amazon.com, google.ee.

**PRODUCTS (PROD)**

Selle alla kuuluvad kõik nime kaudu selgelt idenfitseeritavad tooted, esemed, teosed jms.

* nimedega esemed, nt *iPhone 11, Hubble’i kosmoseteleskoop, PlayStation*
* kunstiteosed, taiesed, nt *Mona Lisa, Angry Birds*
* seadused, nt *Patriot Act, kooseluseadus*
* platvormid ja teenusepakkujad, nt *Hooandja, Google Play, App Store, Gmail*
* tootesarjad, nt *Harry Potteri raamatusari*
* programmeerimiskeeled, nt *Java, Python, C#*
* projektid, programmid, operatsioonid, nt *Tiigrihüpe, VLC Player, Kõrbetorm*
* avaldatud kujul väljaanded, nt *Postimees* (ajalehena)
* hüüdlaused, nt *“Just Do It”*
* tarkvara, nt *Windows 10, macOS Catalina, Microsoft Office 2013*
* liiklusvahendid, alused, lennukimudelid nt *ISS (International Space Station), Tallink Shuttle Star, Tesla Model 3, Honda Civic* (aga kui *Honda* esineb eraldi, siis võib ta olla ka ORG)***,*** *TU-154M*
* süsteemide, tehnoloogiate nimed, nt *Alipay, Apple Pay, HoloLens, IMAX*. **Märkus**: kui sa ei tea täpselt, mida mingi tehniline termin tähendab, aga see tundub olevat nimeolem, siis proovi seda korraks guugeldada. Kui sa tulemustest kohe selgust ei saa, kas tegu on nimega või ei, siis jäta pigem märgendamata. Eesmärk ei ole väga tehnilise teksti puhul kulutada aega otsimisele, vaid emakeelekõneleja intuitsiooni järgi märgendada.
* veebilehed (kui neid kasutatakse toodetena)

Kui PRODUCT olem sisaldab versiooninumbrit, nt “iPhone 6”, siis märgitakse olemiks kogu väljend, versiooninumber kaasa arvatud. Mõnikord võib mõne PRODUCT olemil olla kasutusel versiooninimi, nt XP (Windows XP asemel), sellisel juhul tuleb ka versiooninimi märkida PRODUCT olemiks.

**EVENT**

Selle alla kuuluvad konkreetse nimega sündmused, nt

* näitused, messid ja ekspositsioonid, nt *Expo maailmanäitus, Intellektika, Motoshow*
* otseselt märgitud sündmused, nt *sTARTUp Day 2020, Teeme Ära talgupäev*

Siia ei käi üldised nimetused sündmuste kohta, nt *kunstinäitused, maailmameistrikad*, vaid ainult nimed, mis viitavad selgelt tuvastatavale sündmusele, nt *Tartu Kunstnike Liidu aastanäitus, 2010. aasta jalgpalli maailmameistrivõistlused*

**DATE**

Selle alla kuuluvad ajaväljendid, nii päev/kuu/aasta tüübis, nt “*3. oktoober*”, “*aastal 2020*”, “*2019*”, “*septembris*”, kui ka üldise väljendina (*eile, eelmisel kuul, tuleval aastal*) juhul, kui väljendil on olemas referent.

See tähendab, et väljendi põhjal peab olema võimalik kindlaks määrata mingi konkreetne punkt ajas ehk siis nt konkreetne aasta või kuu või päev. Seega ei sobi ebamäärased väljendid nagu *mõne aasta pärast*, *paar kuud tagasi*, aga sobivad konkreetsemad *viis aastat hiljem, kolm kuud tagasi, üleeile*.

Samuti ei sobi kestust näitavad väljendid, nt *kuuendat aastat kestev trall*.

DATE-väljendit ei märgita siis, kui kuu- või päevasõna on kasutusel üldises mõttes, nt *juulikuus on tavaliselt palav*, *detsembris tähistatakse jõule*, *reedeti on kõigil tuju hea*.

Kui viidatakse tähtpäevale üldises mõttes, siis seda DATE-väljendiks ei loeta, nt *lastele jagatakse kingitusi jõululaupäeviti*, aga kui tähtpäeva all on viidatud konkreetsele kuupäevale, siis märgitakse see DATE-olemiks, nt *homme on ussimaarjapäev, tänavusel vabariigi aastapäeval* jne.

Märkida tuleb kogu väljend tervikuna, nt *16. augustil 1926* on üks väljend, samuti *aastal 2015*, *5. septembril* jne. Kui väljend viitab mingile ajavahemikule, nt *aastail 1990-2000*, *märtsist augustini*, siis tuleb samuti annoteerida DATE’iga terve väljend.

* *15.05.2008 (jms formaadid)*
* *15. mai*
* *15. mai 2008*
* *mais*
* *maikuus*
* *1. mai*
* *maikuu 10. kuupäeval*
* *16. ja 17. mai*
* *aprillis 2019*
* *2020. aasta jaanuaris*
* *20.-23. veebruar*
* *detsembrist märtsini*
* *aastatel 1989-2011*
* *aastast 1989 kuni (aastani) 2011*
* *2019. aasta juunist septembrini*
* *kümme aastat tagasi*
* *viis kuud hiljem*
* *ülehomme*
* *tunamullu*

**TIME**

Selle alla kuuluvad ajaväljendid, mis viitavad päevast väiksemale üksusele: kellaaegadele ja päeva osadele, millel on olemas referent (DATE-olemitega analoogselt). Seega ei sobi jällegi üldised väljendid nagu *hommikuti on mul raske ärgata*, aga sobivad nt *täna hommikul, kell 17, 19:30* jne. Samuti ei märgendata kestusi (nt *magasin eile lõunast tänase hommikuni, mitu tundi* jne).

Näiteid:

* *19:40*
* *1:15*
* *13:00*
* *kl 13*
* *17 (kellaaeg lihtsalt arvuna esitatud kujul)*
* *täna pärastlõunal*
* *homme õhtupoolikul*

**TITLE**

Siia kuuluvad ametinimetused, seisused, teaduskraadid jms. Tiitlid tuleb märkida ainult siis, kui eelneva teksti põhjal on tuvastav konkreetne isik tiitli taga. Näiteks kui tekstis on esialgu mainitud isikut koos tiitliga *president Kersti Kaljulaid*, ning hiljem viidatakse Kaljulaidile lihtsalt sõnaga *president,* siis tuleb mõlemal juhul *president* tiitliks märkida. Tiitlile vahetult järgnev isikunimi ei ole tiitli osa. Kui tiitlile eelneb ORG-märgend, siis tuleb TITLE’iga tähistada ainult ametinimetus, mitte ORG-sõnad.

Tiitlite näiteid:

* *Põhja politseiprefekt*
* *peaminister*
* *kooli direktori asetäitja*
* *justiitsministeeriumi endine asekantsler , juhtiv riigiprokurör*

**MONEY**

Siia kuuluvad väljendid, mis viitavad konkreetsetele valuutadele ja summadele nendes valuutades. Näiteks

* *USA dollar*
* *euro*
* *jeen*
* *paarkümmend eurtsi*
* *$100*
* *1000 krooni*
* *€500*
* *750 jüääni*
* £300-400

**PERCENT**

Siia kuuluvad protsenti väljendavad olemid. Protsent võib olla esitatud nii märgiga (%) kui ka sõnaliselt.

* *50%*
* *35-45%*
* *sada protsenti*
* *alla viie protsendi*

**Hierarhia märgendamine**

Eeltoodud nimeolemite kategooriad võivad esineda hierarhiliste üksustena, kus on üks suur põhiväljend ning selle kaks või enam alamosa, nt kui defineeritud üksused on

siis väljendis

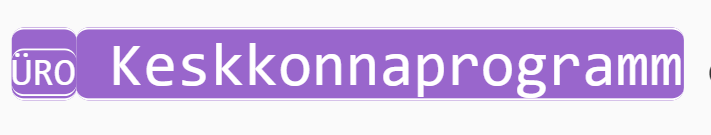
**

on kogu väljend kategooria ORG esindaja, lisaks saab *Tallinna* siin märgendada GPE esindajaks.

Hierarhilise märgenduse puhul võivad olla pesastunud 2 ja enam ORG-märgendit, aga muude märgendite puhul pole selline korduv pesastus lubatud. Näiteks olem *Euroopa Liit* on juba **GPE**, seega ei märgita selle sees veel omakorda sõna *Euroopa* **GPE**ks.

Hierarhilises märgenduses märkige **kuni kolm taset**,mitte rohkem. St isegi kui nimeolemit oleks võimalik jaotada hierarhiliselt neljaks või enamaks osaks, valige välja ainult kolm kõige suuremat taset.

Veel näiteid:

**

1. ÜRO Keskkonnaprogramm (**ORG**) 2. ÜRO (**ORG**)



1. Võrumaa Teataja (**PROD**) 2. Võrumaa (**GPE**)

Rohkem kui kaks taset on nt väljendis



Siin on tervikolem **TITLE**, mille sees *Eesti Kaitseväe* on **ORG** ning selle sees omakorda on *Eesti* **GPE**. Kolm ongi maksimaalne lubatud tasemete arv, seega üle selle ei tohiks ühe nimeolemi sees märgendeid esineda.