**NER-evaluaatioaineiston merkkaamisesta**

Aineiston laadusta: Omorfi tunnistaa aineiston sanoista noin 72 %, joten aineiston laatu vastaa varsin tarkasti koko kokoelman arvioitua laatua (vaikka paikoittain tuntuu siltä, että sotkua on paljon).

1. Käytettävät 8 tagia on sovittu. Niiden käyttöalue/semantiikka on hiukan auki, mutta se täsmennetään n. 2000 sanan malliaineiston luomisessa.
2. Tekstiaineisto on paikoittain raskaasti virheellistä ja sanojen tunnistaminen on vaikeaa. Jos näyttää siltä että sanassa on yli 50 % merkeistä väärin, sanaa ei merkata, vaikka sen voisi tunnistaa merkattavaksi entiteetiksi.
3. Jos sanan kirjoitusasu on vanhastava, 1800-luvun kieltä, mutta sen tunnistaa merkattavaksi entiteetiksi, se merkataan.

**FiNERin taggaus**

Nimiluokkatagit ovat XML-tageja: alkutagi<EnamexPrsHum> kertoo, että kyseisestä sanasta alkaa monisanainen nimi, lopputagi </EnamexPrsHum> kertoo, että kyseiseen sanaan päättyy monisanainen nimi ja tyhjä tagi <EnamexLocPpl/> kertoo, että kyseinen sana on yksisanainen nimi.

|  |  |
| --- | --- |
| **Tagi** | **Merkitys** |
| EnamexPrsHum | henkilö |
| EnamexLocXxx | yleinen paikka |
| EnamexLocGpl | maantieteellinen paikka |
| EnamexLocPpl | "poliittinen alue" (valtio, kaupunki tms.) |
| EnamexLocStr | katu, tie, katuosoite |
| EnamexOrgEdu | koulutusorganisaatio |
| EnamexOrgCrp | yritys, yhdistys, järjestö tms. |
| TimexTmeDat | ajanilmaus |