Lukijalle saatteeksi: Twollatun tekstiaineiston disambiguointi

Vesa Heikkinen, Mikko Lounela & Satu Leisko-Järvinen

31.3.2012

Teksti Twollatun tekstiaineiston disambiguointi on alkujaan tarkoitettu Kotuksen eli Kotimaisten kielten tutkimuskeskuksen (nyk. Kotimaisten kielten keskus) sisäiseen käyttöön. Kyse on ohjeistavasta käyttötekstistä: Kotuksen korpus- ja tekstintutkijoiden aineistoja tehtiin ja analysoitiin vuosituhannen vaihteessa paljolti tilapäisin harjoittelijavoimin, jolloin korostui ohjeistuksen ja dokumentoinnin tarve. Tekstissä käytetäänkin muun muassa puhe- ja arkityylisiä ilmauksia, kuten otsikon *twollata*, jolla viitataan *Fintwol*-nimiseen morfologiseen jäsentimeen.

Tämä erityisesti ”Tekstistä korpukseksi” -aineistojen eli Teko-aineistojen tekijöille tarkoitettu ohje ilmestyi ensimmäisen kerran Kotuksen sisäisessä verkossa Apajassa 20.8.2007 (Leisko-Järvisen nimi oli tuolloin Uosukainen). Sisältöä on päivitetty viimeksi 31.12.2007. Ohje perustuu vuosien mittaan kerättyihin kokemuksiin tämäntapaisen aineistotyön, korpuslingvistiikan ja tekstintutkimuksen parissa. Kun ohje valmistui, ajateltiin, että samantapainen aineisto- ja analyysityö jatkuisi Kotuksessa. Todellisuudessa mainitunlainen toiminta oli juuri hiipumassa, kun ohje vihdoin saatiin tähän vaiheeseen.

Teko-projektin aikana valmistui useita aineistoja ja tutkimuksia. Näitä esitellään tarkemmin muun muassa kirjassa *Genreanalyysi – tekstilajitutkimuksen käsikirja* (Gaudeamus 2012) olevissa Vesa Heikkisen, Mikko Lounelan ja Eero Voutilaisen kirjoittamissa artikkeleissa *Aineistot ja niiden käyttö tekstilajitutkimuksessa* ja *Automaattinen analysaattori tekstilajitutkimuksessa*.

Twollausteksti julkaistaan tässä lähes alkuperäisessä muodossaan. Joitakin pieniä lyöntivirheitä on korjattu, muutamia muotoilumuutoksia on tehty, ja parissa kohdassa on selkeytetty sanontaa. Tekstin loppuun on liitetty Kotuksessa työelämään tutustujana työskennelleen Viljami Haakanan (2011) kokoama luettelo niistä merkitsimistä, joita ei mainittu jäsentimen dokumentaatiossa eikä tullut vastaan eri aineistojen analyyseissa vastaan.

Teksti kertoo konkreettisesti niistä ratkaisuista, joita (puoli)automaattista analysaattoria käyttävä tutkija joutuu työssään tekemään. Tässä mielessä sillä on nähdäksemme yleispätevämpääkin merkitystä.

Asiallinen tapa viitata tekstiin *Twollatun tekstiaineiston disambiguointi* on tämä:

Vesa Heikkinen, Mikko Lounela & Satu Leisko-Järvinen (ent. Uosukainen) (2012 [2007]). *Twollatun tekstiaineiston disambiguointi*. Helsinki: Kotimaisten kielten keskus. www.kotus.fi Haettu [päivämäärä].

Twollatun tekstiaineiston disambiguointi

Vesa Heikkinen, Mikko Lounela & Satu Leisko-Järvinen

1. Johdanto

Tämä on ohje tekstimateriaalien disambiguoimiseksi eli morfologisten tulkintojen yksiselitteistämiseksi. Yksiselitteistäminen tarkoittaa käytännössä tarpeettomien vaihtoehtojen poistamista niistä tulkinnoista, joita sanat saavat Fintwol-analysaattorilla tehdyssä automaattisessa analyysissa.

Ohje tulee tarpeeseen. Yhteisiä pelisääntöjä tarvitaan, jotta eri ihmisten erilaisista aineistoista tekemät kvantitatiiviset tutkimukset olisivat mahdollisimman vertailukelpoisia. Tässä *Johdanto*-luvussa esitellään aineistojen käsittelyn vaiheita ja työn dokumentointia. Lisäksi käydään läpi niitä Kotuksessa tehtyjä tutkimuksia, joissa on sovellettu Fintwol-analyysia. Luvussa 2 käsitellään sellaisia ilmiöitä, jotka on koettu ongelmallisiksi, kun on poistettu sanojen ylimääräisiä tulkintoja. Luvussa 3 käydään läpi niitä tietoja, joilla Fintwolin antamaa analyysia täydennetään. Luku 4 sisältää ohjeet siitä, mitä tehdään sellaiselle disambiguoitavalle aineistolle, jossa on virheitä. Luvussa 5 on tietoa lisämerkinnöistä, joita aineistoihin voidaan disambiguoinnin yhteydessä tehdä. Huomaa, että ohjeen eri osissa on jonkin verran päällekkäisyyksiä, ja joitakin asioita käsitellään useammassa kuin yhdessä kohtaa ohjetta. Lukuun 6 on koottu morfologisissa tulkinnoissa käytettävät merkinnät. Merkinnät tehdään xml-merkitsimillä, joita myös kutsutaan tägeiksi.

1.1. Tekstiaineiston koostamisesta ja käsittelemisestä

Aineistojen puoliautomaattinen morfologinen analyysi on monivaiheista. Ensin aineistot **kerätään** jostakin halutusta paikasta. Aineistoja on koottu esimerkiksi vanhoista aikakauslehdistä, nykyaikaisista sanomalehdistä ja verkkosivuilta. **Tallennusvaiheessa** tekstimateriaali muunnetaan sähköiseen muotoon, jos sellaista ei valmiiksi ole. Esimerkiksi sanomalehtiaineistoja on **skannattu** kuva- ja tekstitiedostoiksi. Skannattu tekstiaineisto **oikoluetaan**, sillä tekstiin saattaa skannausvaiheessa tulla virheitä. Aineistosta korjataan vain skannauksessa tulleet virheet. Alkuperäistekstin "virheitä" ei korjata, koska tutkimuksen kannalta on tärkeää säilyttää aineiston alkuperäiset ominaisuudet muuttumattomina.

Seuraavaksi tiedostot **nimetään** informatiivisesti ja **lajitellaan** kansioihin ja alakansioihin, jos on tarpeen. Jos aineistona on esimerkiksi vanha sanomalehti, jonka artikkelit halutaan analysoida, on järkevää tallentaa kaikki artikkelit samaan kansioon, joka on nimetty lehden numeron mukaan. Näin käsiteltäväksi voidaan myöhemmin ottaa saman lehden muitakin numeroita. Näistä ja muista tallennuskäytännöistä sovitaan aineiston käyttäjän kanssa.
Tallentamisen jälkeen on vuorossa aineiston **rakenteistaminen**: tavallisena tekstinä tallennettuun materiaaliin lisätään xml-koodaus. Koodauksessa muun muassa tekstikappaleet ympäröidään p-merkitsimellä (`<p>`) ja virkkeet s-merkitsimellä (`<s>`). Koodiin voidaan lisätä myös muita merkitsimiä tarpeen mukaan. Jos esimerkiksi nimiä halutaan tutkia tarkemmin, ne voidaan erottaa tekstistä ja luokitella name-merkitsimen avulla (`<name tyope="prop">Erisnimi</name>`).
TEI tarkoittaa tekstiaineistojen rakenteistuksessa käytettävän xml-koodauksen standardia. TEI:ssä siis määritellään, millaisia merkitsimiä rakenteistuksessa voidaan käyttää. Kotuksessa TEI:stä käytössä oleva versio on TEI P4. Lisätietoja TEI:stä on saatavilla mm. sivustolta http://www.tei-

c.org/P4X/SG.html (*A Gentle Introduction to XML*). Mikäli et ole aiemmin tutustunut xml:ään, lue se heti.

Rakenteistamisen jälkeen tekstit **twollataan** eli ajetaan Fintwol-analysaattorin läpi. Fintwol on Kimmo Koskenniemen ja Lingsoftin kehittämä morfologinen analysaattori suomen kielelle. Twol antaa automaattisesti sille syötetyn tekstisanan eli saneen kaikki eri luennat. Yhteen luentaan kuuluu syötteenä annettu sanamuoto palautettuna perusmuotoonsa ja morfologisia merkitsimiä eli tägejä, jotka kertovat sanan morfologisista ominaisuuksista, esimerkiksi sellaisista kuin sanaluokka, persoona, luku ja sijamuoto. Morfologiset merkitsimet eroavat xml-merkitsimistä niin muotonsa kuin merkityksensäkin puolesta. Esimerkkinä xml-merkitsimestä voisi olla virkkeen aloittava `<s>`, morfologinen merkitsin `INE` puolestaan esiintyy vain `<w>`-merkitsimen msd-attribuutin arvona ja tarkoittaa inessiivimuotoa.

```
<s>
<w lemma="talo" norm="talossa" type="N" msd=" INE SG ">talossa
</s>
```

Disambiguoimattomassa materiaalissa tekstin jokaisen saneen jokainen mahdollinen morfologinen Fintwol-tulkinta on esillä. Virke *Olin tiennyt tehneeni tyhmästi alusta asti.* näyttää twollattuna mutta disambiguoimattomana tältä:

```
<w lemma="olla" norm="olin" type="V" msd=" COP PAST ACT SG1 ">Olin</w>
<w lemma="olka" norm="olin" type="N" msd=" INS PL ">Olin</w>
<w lemma="ola" norm="olin" type="N" msd=" PROP INS PL ">Olin</w>

<w lemma="tietää" norm="tiennyt" type="V" msd=" PAST ACT NEG SG ">tiennyt</w>
<w lemma="tietää" norm="tiennyt" type="PCP2" msd=" ACT POS NOM SG ">tiennyt</w>

<w lemma="tehdä" norm="tehneeni" type="V" msd=" REF PAST ACT 1SG ">tehneeni</w>
<w lemma="tehdä" norm="tehneeni" type="PCP2" msd=" ACT POS NOM SG 1SG
">tehneeni</w>
<w lemma="tehdä" norm="tehneeni" type="PCP2" msd=" ACT POS GEN SG 1SG
">tehneeni</w>
<w lemma="tehdä" norm="tehneeni" type="PCP2" msd=" ACT POS NOM PL 1SG
">tehneeni</w>
<w lemma="tehnyt" norm="tehneeni" type="A" msd=" ACT PCP2 POS NOM SG 1SG
">tehneeni</w>
<w lemma="tehnyt" norm="tehneeni" type="A" msd=" ACT PCP2 POS GEN SG 1SG
">tehneeni</w>
<w lemma="tehnyt" norm="tehneeni" type="A" msd=" ACT PCP2 POS NOM PL 1SG
">tehneeni</w>

<w lemma="tyhmä" norm="tyhmästi" type="ADV" msd=" POS MAN ">tyhmästi</w>

<w lemma="alku" norm="alusta" type="N" msd=" ELA SG ">alusta</w>
<w lemma="alusta" norm="alusta" type="N" msd=" NOM SG ">alusta</w>
<w lemma="alustaa" norm="alusta" type="V" msd=" PRES ACT NEG ">alusta</w>
<w lemma="alustaa" norm="alusta" type="V" msd=" IMPV ACT SG2 ">alusta</w>
<w lemma="alustaa" norm="alusta" type="V" msd=" IMPV ACT NEG SG ">alusta</w>
<w lemma="alunen" norm="alusta" type="N" msd=" PTV SG ">alusta</w>
<w lemma="alus" norm="alusta" type="N" msd=" PTV SG ">alusta</w>

<w lemma="asti" norm="asti" type="ADV" msd="">asti</w>
<w lemma="asti" norm="asti" type="PSP" msd="">asti</w>
<w lemma="." norm="." type="PUNCT" msd=
" FULLSTOP ">.</w>
```

Twollauksen jälkeen materiaali on valmis **disambiguoitavaksi**. Disambiguoinnissa Fintwolin antamista tulkinnoista valitaan yksi, kontekstissaan ”oikea” tulkinta. Muut vaihtoehdot eli ambiguiteetti poistetaan.

Disambiguointi voidaan jakaa kahteen osaan: ei-toivottujen tulkintojen poistamiseen ja Fintwol-analyysista puuttuvien tietojen lisäämiseen. Esimerkin kuuden sanan lause saa yli kaksikymmentä mahdollista sanantulkintaa. Suurin osa näistä on kontekstissaan selvästi ei-toivottuja, ja ne voidaan poistaa ilman ongelmia.

Virkkeen aloittava *Olin* ei selvästikään tässä yhteydessä viittaa *olkaan* eikä erisnimeen *Ola*. Virkkeessä ei ole kyse myöskään *alustamisesta* tai *alustasta*. Vaikeampaa on päättää, onko sane *asti* adverbi vai postpositio. Disambiguoijan on päätettävä myös, mikä on *tehneeni*-saneen oikea tulkinta tässä tekstissä.

Yleensä tulkinnoista valitaan se, jonka type-attribuutin arvo (sanaluokka) ja lemma (perusmuoto) ovat oikein. Sopivista tulkinnoista valitaan täsmällisin sopiva. Jos sana saa vain yhden hyväksyttävän tulkinnan, se säilytetään. Fintwolin sanoille antamia analyyseja (lemma-, type- ja msd-kenttien arvoja) ei muokata muuten kuin tämän ohjeen mukaisesti. (Lemma tarkoittaa sanan perusmuotoa, norm sanan normalisoitua muotoa ja type sanaluokkaa. Msd-attribuutti sisältää muut morfologiseen tulkintaan liittyvät tiedot.)

Disambiguoituna esimerkki näyttää tältä:

```
<w lemma="olla" norm="olin" type="V" msd=" COP PAST ACT SG1 " function=" PL
">Olin</w>
<w lemma="tietää" norm="tiennyt" type="PCP2" msd=" ACT POS NOM SG " function="
PL ">tiennyt</w>
<w lemma="tehdä" norm="tehneeni" type="V" msd=" REF PAST ACT 1SG ">tehneeni</w>
<w lemma="tyhmä" norm="tyhmästi" type="ADV" msd=" POS MAN ">tyhmästi</w>
<w lemma="alku" norm="alusta" type="N" msd=" ELA SG ">alusta</w>
<w lemma="asti" norm="asti" type="PSP" msd="">asti</w>
<w lemma="." norm="." type="PUNCT" msd=" FULLSTOP ">.</w>
```

Aineistojen disambiguointi pyritään tekemään mahdollisimman yhtenäisesti. Uudet aineistot ja niiden erikoisominaisuudet saattavat kuitenkin vaatia uusia päätöksiä ja tämän ohjeen päivittämistä. Mahdolliset tästä ohjeesta poikkeavat tulkinnat tai disambiguointipäätökset vaativat keskustelua, ja ne on dokumentoitava huolellisesti.

1.2. Dokumentoinnista

Disambiguoijien on dokumentoitava päätökset, joita disambiguointityön aikana tehdään. Tämä tapahtuu siten, että jokainen disambiguoija täyttää **dokumentointiraportin** ja lähettää sen työnsä ohjaajalle.

Kesäharjoittelijoiden ja muiden määräaikaisten tutkimusapulaisten on hyvä tehdä työstään myös vapaamuotoinen **työselostus**. Se saa olla pidempi kuin dokumentointiraportti, joka on tarkoitettu vain keskeisimpien tietojen keräämiseen. Työselostuksestakin on hyvä ilmetä seuraavat perusasiat: harjoittelijan nimi, työn ajankohta (myös vuosi), disambiguoidut tai muulla tavoin käsitellyt tiedostot (joita dokumentaatiossa esitellyt päätökset koskevat) sekä käsiteltyjen tiedostojen nimeämistapa ja tallennuspaikka. Selostuksessa voi keskittyä kuvaamaan esimerkiksi työn ongelmakohtia ja etenemistä sekä mahdollisia aineiston erityispiirteitä. Tällöin selostus voi toimia

myös harjoittelijan muistin tukena ja kuvata hänen edistymistään työn parissa. Hyvä käytäntö oman työselostustiedoston nimeämiseen on esimerkiksi *tyoselostus_uosukainen.rtf*. Tällöin selostusten lukijat näkevät jo tiedoston nimestä, kenen tekemä selostus on kyseessä.

Erityisesti dokumentointiasiakirjoihin kannattaa lisätä esimerkkejä ongelmallisista tapauksista ja mahdollisista erikoistapauksista. Jos työstettyyn aineistoon on sovellettu jotain poikkeuspäätöstä, pitää tällaisista tapauksista esittää jokin esimerkki ja sen lähde (missä tiedostossa ko. kohta on). Jos raportti kattaa useamman kuin yhden disambiguoijan työn, siinä on hyvä mainita, kuka mitkäkin aineiston osat on käsitellyt (esimerkiksi disambiguoinut tietyn presidentin puheet tai tietyn lehden numeron kaikki artikkelit).

1.3. Kotuksessa tehtyjä tutkimuksia

Fintwol-analyysia alettiin tehdä Kotuksessa 1990-loppupuolella. Ensimmäinen tutkimus, jossa sitä sovellettiin, oli Vesa Heikkisen väitöskirja *Ideologinen merkitys kriittisen tekstintutkimuksen teoriassa ja käytännössä* (1999). Heikkistä avustivat käytännön analyyseissa Mikko Lounela ja Outi Lehtinen. Tutkimuksessa twollattiin 100 sanomalehtien pääkirjoitusta. Aineistot ja twollaukset ovat tallennettuina VH:n hallussa.

Kokeiluja jatkettiin Vesa Heikkisen, Pirjo Hiidenmaan ja Ulla Tiililän akvaariohankkeessa, jossa tutkittiin muun muassa opetuslautakunnan esityslistoja ja pöytäkirjoja. Näitä analysoitiin puoliautomaattisesti noin 76 000 tekstisanan verran (disambiguoijina Vesa Heikkinen ja tutkimusapulainen Laura Vesa). Laskelmia on julkaistu teoksessa *Teksti työnä, virka kielenä*. Aineisto on tallennettuna sähköisessä muodossa (VH:n hallussa), mutta twollatut versiot ovat kadonneet.

Myöhemmin aloitetuista aineistoista pisimmälle on viety analyysit tasavallan presidenttien uudenvuodenpuheista (Vesa Heikkinen ja Mikko Virtanen), maakuntalehtien pikku-uutisista (Vesa Heikkinen, Mikko Lounela, Outi Lehtinen ks. Lisätietoja uutistutkimukseen (http://www.kotus.fi/index.phtml?s=2271), rakennusviraston tiedotteista (Salli Kankaanpää), kuljetuspalvelupäätöksistä (Ulla Tiililä), säädösaineistosta (Aino Piehl) ja kirkkohallituksen kirjeistä (Riitta Kolehmainen). Kaikkien aineistojen tietokonelingvistinen vastuu on ollut Mikko Lounelalla. Useimpien aineistojen käsittelyssä on ollut mukana myös tutkimusapulainen Mikko Virtanen. Disambiguointia ovat tehneet mainittujen lisäksi tekstintutkimushankkeiden tutkimusapulaiset Marianne Laaksonen, Anna Ontero, Helena Ässämäki, Heidi Vapaa, Satu Uosukainen, Leena Maria Heikkola, Jussi Keinänen, Tomi Visakko ja Marika Karhunpesä sekä siviilipalvelusmies Tuure Hurme.

Tekstiaineistojen puoliautomaattisen analyysin rakenne on vähitellen vakiintunut. Mikko Lounelan ja Vesa Heikkisen yhteistyönä on aloitettu ”Teko-projekti”, jonka puolivirallisia syntysanoja lausuttiin syksyllä 2003. *Teko* tulee sanoista ”tekstistä korpukseksi”. Teko-aineistot koostetaan, käsitellään ja analysoidaan siten, että niistä saatavat kvantitatiiviset tulokset ovat vertailukelpoisia. Teko-aineistojen tekijät soveltavat tämän ohjeen suosituksia.

2. Ei-toivottujen tulkintojen poistaminen

2.1. Partisiipit

Fintwol tarjoaa partisiippimuodoille kolme tulkintaa, joista yhdessä sanaluokkana (type) on verbi (ty-pe="V"), toisessa partisiippi (type="PCP") ja kolmannessa adjektiivi (type="A"):

```
<w lemma="tehdä" norm="tehty" type="V" msd=" PAST PSS NEG ">tehty</w>
<w lemma="tehdä" norm="tehty" type="PCP2" msd=" PSS POS NOM SG ">tehty</w>
<w lemma="tehty" norm="tehty" type="A" msd=" PSS PCP2 POS NOM SG ">tehty</w>
```

Yleissääntönä partisiippien disambiguoinnissa voidaan pitää sitä, että adjektiivitulkinta (type="A") valitaan silloin, kun sana on komparatiivissa (*pysyvämpi ratkaisu*) tai superlatiivissa (*elämän kestävimmät siteet*). Tällöin voidaan varmasti puhua adjektiivisesta leksikaalistumasta. Adjektiivitulkinta annetaan myös, jos partisiippi määrittää substantiivia ja esiintyy hakusanana Perussanakirjassa tai Kielitoimiston sanakirjassa. Näissä tapauksissa adjektiivitulkinnan on oltava käyttöyhteydessään mielekäs: *loistava tulevaisuus* vs. *pitkälle loistava valo* tai *kauan kestävä kehitys* vs. *kestävä kehitys*. Jos nämä ehdot eivät täyty, valitaan partisiipin sanaluokaksi partisiippi (PCP1/PCP2) tai verbi (V).

Fintwolin ykköspartisiipille (PCP1) antamista tulkinnoista adjektiivi- ja partisiippitulkinnan välillä valitaan seuraavan säännön mukaan: Adjektiivitulkinta valitaan vain, jos kyseessä oleva partisiippimuoto esiintyy adjektiivina sanakirjassa. Muissa tapauksissa valitaan aina partisiippitulkinta. Seuraavien esimerkkien tapauksista *kuoleva* ei ole sanakirjassa, *haastava* sen sijaan on:

```
<w lemma="kuolla" norm="kuoleva" type="PCP1" msd=" ACT POS NOM SG ">kuoleva</w>
<w lemma="joutsen" norm="joutsen" type="N" msd=" NOM SG ">joutsen</w>

<w lemma="haastava" norm="haastava" type="A" msd=" POS NOM SG ">haastava</w>
<w lemma="tehtävä" norm="tehtävä" type="N" msd=" NOM SG ">tehtävä</w>

<w lemma="riita" norm="riitaa" type="N" msd=" PTV SG ">riitaa</w>
<w lemma="haastaa" norm="haastava" type="PCP1" msd=" ACT POS NOM SG
">haastava</w>
<w lemma="tyyppi" norm="tyyppi" type="N" msd=" NOM SG ">tyyppi</w>
```

Ensimmäisessä esimerkissä sanan tulkinnaksi valitaan partisiippi, sillä *kuoleva* ei ole leksikaalistunut adjektiiviksi eikä sitä löydy sanakirjasta. Toisen esimerkin *haastava* sen sijaan saa adjektiivitulkinnan. Kolmannessa esimerkissä käytetään adjektiiviksi leksikaalistunutta sanaa, mutta tässä tekstiyhteydessä se ei kuitenkaan saa adjektiivitulkintaa.

Kakkospartisiippeja (PCP2) voi tekstissä esiintyä (ainakin) kolmessa tehtävässä: Osana temporaalista tekstiketjua (perfektiä tai pluskvamperfektiä, *on **tehty***), osana kielteistä imperfektiivistä verbiketjua (*ei **onnistunut***), tai etumääreenä (*vuosi sitten **valmistuneet** ohjeet*). Partisiippisanaluokkainen vaihtoehto valitaan temporaalisissa verbiketjuissa. Adjektiivisanaluokkainen vaihtoehto valitaan etumääreissä, mutta vain jos sana on *Perussanakirjassa* tai *Kielitoimiston sanakirjassa* adjektiivina. Esimerkkejä:

```
<w lemma="olla" norm="oli" type="V" msd=" COP PAST ACT SG3 ">oli</w>
<w lemma="tehty" norm="tehty" type="A" msd=" PSS PCP2 POS NOM SG ">tehty</w>

<w lemma="ei" norm="ei" type="V" msd=" NEGV SG3 ">ei</w>
<w lemma="tehdä" norm="tehty" type="V" msd=" PAST PSS NEG ">tehty</w>

<w lemma="kuollut" norm="kuollut" type="A" msd=" ACT PCP2 POS NOM SG
">kuollut</w>
<w lemma="mies" norm="mies" type="N" msd=" NOM SG ">mies</w>
```

Valittaessa adjektiivi- ja partisiippitulkintojen väliltä on erityistä huomiota kiinnitettävä siihen, millaisesta rakenteesta on kyse ja kuinka tarkasteltavaa ilmaisua on siinä käytetty. Esimerkiksi luonteeltaan verbaalisissa ja produktiivisissa rakenteissa on valittu partisiippitulkinta siitä huolimatta, että sanat ovat sanakirjassa itsenäisinä lekseemeinä: *toivottua paremmin, kuten tunnettua, olla näkyvillä, tulla näkyville*. Kyse on siis siitä, kumpi tulkintavaihtoehto näissä tapauksissa on mielekkäämpi, adjektiivi vai partisiippi.

Partisiipeista johdetut *sti*-adverbit: perusmuodoksi valitaan partisiippi, jos sana on leksikaalistunut adjektiiviksi (eli on sanakirjassa). Muussa tapauksessa perusmuodoksi valitaan verbin infinitiivi. Esimerkiksi *huolestuttavasti –> huolestuttava* ei löydy sanakirjasta, joten valitaan lemma="huolestuttaa", samoin *odotetusti* –> lemma="odottaa". Toisaalta kuitenkin: *ratkaisevasti* –> lemma="ratkaiseva".

Substantiiveiksi leksikaalistuneet partisiipit (*syytetty*, *valtuutettu*, *syötävä*): Fintwol tarjoaa adjektiivi- ja partisiippitulkinnat, mutta ei substantiivitulkintaa. Näissä kohdissa Fintwolin antamista tulkinnoista hyväksytään partisiippitulkinta, johon lisätään function-attribuutti, joka saa arvon "N" (function="N"). Näin tapaukset ovat myöhemmin helposti löydettävissä ja – jos niin päätetään – muunnettavissa.

```
<w lemma="syyttää" norm="syytetty" type="PCP2" msd=" PSS POS NOM SG " function="
N ">syytetty</w>
<w lemma="valtuuttaa" norm="valtuutettu" type="PCP2" msd=" PSS POS NOM SG "
function=" N ">
valtuutettu</w>

<w lemma="syödä" norm="syötävä" type="PCP2" msd=" PSS POS NOM SG " function=" N
">syötävä</w>
```

Säännöstä poiketaan vain erityistapauksissa, joten kaikki päätökset kannattaa dokumentoida. Asiaa arvioidaan jatkossa uudelleen.

Agenttipartisiipit. Fintwol antaa *-ma*/*-mä* -päätteisille verbijohdoksille yleensä sanaluokan "DV-MA" (johdostieto). Joskus se kuitenkin tarjoaa myös tulkintoja "V" tai "N". Substantiivitulkinnat on suhteellisen helppo erottaa toisistaan.

```
<w lemma="johto" norm="johdon" type="N" msd=" GEN SG ">johdon</w>
<w lemma="arvostaa" norm="arvostamista" type="DV-MA" msd=" ELA PL
">arvostamista</w>
<w lemma="asia" norm="asioista" type="N" msd=" ELA PL ">asioista</w>

<w lemma="tämä" norm="tämä" type="PRON" msd=" DEM NOM SG ">tämä</w>
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 ">on</w>
<w lemma="arvostaminen" norm="arvostamista" type="N" msd=" DV-MINEN PTV SG
">arvostamista</w>

<w lemma="puu#seppä" norm="puusepän" type="N" msd=" GEN SG ">puusepän</w>
<w lemma="tehdä" norm="tekemään" type="DV-MA" msd=" ILL SG ">tekemään</w>
<w lemma="tuoli" norm="tuoliin" type="N" msd=" ILL SG ">tuoliin</w>

<w lemma="tulla" norm="tulin" type="V" msd=" PAST ACT SG1 ">tulin</w>
<w lemma="tehdä" norm="tekemään" type="V" msd=" INF3 ILL " >tekemään</w>
<w lemma="työ" norm="työtä" type="N" msd=" PTV SG ">työtä</w>
```

Agenttipartisiippitulkinnan yhteydessä msd-attribuuttiin ei siis tule PCP-merkintää.

2.2. Adverbit

Adverbitulkintaiset sanat saavat usein myös joko substantiivi-, prepositio- ja postpositiotulkintoja. Vaikkapa *esimerkiksi*-sanalla on kaksi käyttöä (”Esimerkiksi minä kelpaan esimerkiksi.”).

```
<w lemma="esi#merkiksi" norm="esimerkiksi" type="ADV" msd="">esimerkiksi</w>
<w lemma="minä" norm="minä" type="PRON" msd=" PERS NOM SG ">minä</w>
<w lemma="kelvata" norm="kelpaan" type="V" msd=" PRES ACT SG1 ">kelpaan</w>
<w lemma="esimerkki" norm="esimerkiksi" type="N" msd=" TRA SG ">esimerkiksi</w>
```

Substantiivin erottaminen lienee yleensä aika selvää. Esimerkkivirkkeessä sana *esimerkiksi* esiintyy sekä substantiivina että adverbina.

Prepositiot, postpositiot, adverbit. Prepositio- ja postpositiotulkintojen erottaminen toisistaan on useimmiten helppoa:

```
<w lemma="sauva" norm="sauvojen" type="N" msd=" GEN PL ">sauvojen</w>
<w lemma="kanssa" norm="kanssa" type="PSP" msd="">kanssa</w>
<w lemma="tai" norm="tai" type="C" msd=" COORD ">tai</w>
<w lemma="ilman" norm="ilman" type="PP" msd="">ilman</w>
<w lemma="sauva" norm="sauvoja" type="N" msd=" PTV PL ">sauvoja</w>
```

Sanat luokitellaan adverbeiksi, kun ne esiintyvät itsenäisesti ja ilmaisevat aikaa, paikkaa, tapaa, määrää tai muuta sellaista. Adverbitulkinta voidaan valita muun muassa silloin, kun sana esiintyy ilman NP-täydennystä (mm. elliptisissä tapauksissa: *ei yö niin pitkä, ettei päivä **perässä***). Tällä perusteella esimerkiksi *ottaa **vastaan*** saa "ADV"-tulkinnan. Samaa logiikkaa mukailevat alla olevat esimerkit:

```
<w lemma="perimä#tieto" norm="perimätiedon" type="N" msd=" GEN SG
">Perimätiedon</w>
<w lemma="mukaan" norm="mukaan" type="PSP" msd=" ILL ">mukaan</w>
VRT.
<w lemma="lähteä" norm="lähde" type="V" msd=" IMPV ACT SG2 ">Lähde</w>
<w lemma="mukaan" norm="mukaan" type="ADV" msd="">mukaan</w>
```

Esimerkeistä ensimmäisessä on kyse postpositiosta, sillä siinä *mukaan* esiintyy genetiivimuotoisen täydennysosan yhteydessä. Alemmassa esimerkissä *mukaan* on adverbi.

Välillä prepositioita, postpositioita tai adverbeja voi olla vaikea luokitella nopeasti ja oikein. Tällaiset tapaukset voi aluksi jättää hautumaan ja niihin voi palata niihin myöhemmin. Vaikeiksi kokemiaan tapauksia voi myös koota omaan dokumentaatioon, jossa ne säilyvät myös muiden luettavaksi.

Esimerkiksi *asti* ja *saakka* ovat *Ison suomen kieliopin* mukaan terminatiivisia partikkeleita, joilla on yhteisiä piirteitä adpositioiden kanssa. Disambiguoitaessa ne on tulkittu postpositioiksi.

```
<w lemma="voida" norm="voi" type="V" msd=" PRES ACT SG3 ">voi</w>
<w lemma="kävellä" norm="kävellä" type="V" msd=" INF1 NOM ">kävellä</w>
<w lemma="perille" norm="perille" type="ADV" msd=" ALL ">perille</w>
<w lemma="asti" norm="asti" type="PSP" msd="">asti</w>
```

Tulkinnat on syytä tarkistaa. Esimerkiksi *yli* ja *alle* esiintyvät adverbeina määrän ilmauksissa, mutta pre- tai postpositioina genetiivitäydennyksen kera.

```
<w lemma="luoto" norm="luotoja" type="N" msd=" PTV PL ">luotoja</w>
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 ">on</w>
<w lemma="yli" norm="yli" type="ADV" msd="">yli</w>
<w lemma="1000" norm="1000" type="NUM" msd="">1000</w>
```

Myös sellaisilla sanoilla kuin *paitsi*, *päällä* ja *lisäksi* on adverbitulkinnan lisäksi pre- ja postpositiotulkintoja. Pre- tai postpositiotulkinta valitaan, jos sana muodostaa pääsanansa kanssa nominilausekkeen.

```
<w lemma="kone" norm="koneen" type="N" msd=" GEN SG ">koneen</w>
<w lemma="päällä" norm="päällä" type="PP" msd=" ADE ">päällä</w>

<w lemma="kone" norm="kone" type="N" msd=" NOM SG ">kone</w>
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 ">on</w>
<w lemma="päällä" norm="päällä" type="ADV" msd=" ADE ">päällä</w>
```

Adpositiolausekkeista kiteytyneet, sanaliiton muodostavat adverbit (*ennen pitkää*, *tavan takaa*, *sitä vastoin*(?)) ovat hankalia, koska Fintwol käsittelee vain yhden sanan pituisia yksiköitä. Näissä tapauksissa valitaan aina adpositiotulkinta, vaikka tieto adverbiluonteesta tällöin katoaakin. Perusteluna on yhtenäisyyden kriteeri, sillä joissakin tapauksissa Fintwol tarjoaa ainoastaan P(S)P-vaihtoehdon. Lisäksi joskus (harvoin) molemmat tulkinnat tuntuvat periaatteessa mahdolliselta. Leksikaalistumisen arviointiperusteena käytetään myös tässä kysymyksessä sanakirjaa.

AD-A:t. Fintwol tarjoaa joillekin sanoille sanaluokan AD-A. Tällöin sana on määreen määre, eli adverbi, joka määrittää adjektiivia tai toista adverbia. Minkään muun sanaluokan sanojen määreet eivät saa tätä tulkintaa. AD-A-sanat eli ad-adjektiivit ja ad-adverbit (*yhtä*, *liian*, *melko*, *varsin*, *erittäin*) ovat osa adverbien sanaluokkaa. TWOL-analyysien AD-A-tulkinta tarkoittaa, että kyseessä on adverbi, jolla on erityisesti AD-A-ominaisuus.

```
<w lemma="todella" norm="todella" type="AD-A" msd="">todella</w>
<w lemma="iso" norm="iso" type="A" msd=" POS NOM SG ">iso</w>

<w lemma="tehdä" norm="teitkö" type="V" msd=" PAST ACT SG2 kO ">Teitkö</w>
<w lemma="se" norm="sen" type="PRON" msd=" DEM GEN SG ">sen</w>
<w lemma="todella" norm="todella" type="ADV" msd="">todella</w>
<w lemma="?" norm="?" type="PUNCT" msd=" QUESTION ">?</w>
```

AD-A-tulkinta valitaan adjektiivin tai adverbin määritteelle, kun TWOL on sitä tarjoaa, ja kun kyseessä on määritteen määrite, eli ad-a-sana määrittää adjektiivia. Muutoin valitaan väljempi ADV-tulkinta (adverbi). AD-A-luokitusta ei kuitenkaan tehdä käsin. Sanaluokkajakaumia laskettaessa AD-A-luokituksen saaneet sanat voidaan yhdistää adverbiluokituksen saaneisiin. AD-A:na pidetään vain adjektiivien (A) ja adverbien (ADV) määritteitä. TWOL saattaa tarjota AD-A-tulkintaa myös esimerkiksi pronomini- (PRON), lyhenne- (ABBR) tai numeraalitulkinnan (NUM) saaneen sanan määritteelle.

2.3. Numerot ja lyhenteet

Yleensä Fintwolin antamat numero- ja lyhennetulkinnat hyväksytään. Tällöin roomalaiset numerot (III, XVII) saavat lyhennetulkinnan samoin kuin taivutetut numerot (3:s). Tämä on käyttämämme analysaattorin ominaisuus (puute). Jos numerot ja lyhenteet halutaan käsitellä tarkasti, merkitään ne aineistoon xml-merkitsimillä, esimerkiksi <num>numero</num> tai <abbr>lyhenne</abbr>, jolloin ne voidaan myös segmentoida ja tyypitellä halutulla tavalla:

```
<num type="rom">cxxvii</num>
```

```
<abbr>puh.</abbr>
```

Toisinaan Fintwol tulkitsee peräkkäiset numerot joko yhteen tai erilleen virheellisesti:

```
<w lemma="31.12.2002 123" norm="31.12.2002 123" type="NUM" msd="">31.12.2002
123</w>
```

Vastaavaa virhettä ei kuitenkaan esiinny silloin, jos päivämäärää seuraisi nelinumeroinen luku, esim. *31.12.2002 1945*. Näissä tapauksissa tulkinta jakautuu oikein kahteen osaan.
Vanhoissa aineistoissa saattaa edelleen esiintyä segmentointivirheitä, joissa pitkät luvut ovat saaneet useamman kuin yhden tulkinnan. Ks. esimerkki alla:

```
<w lemma="400" norm="400" type="ABBR" msd=" NOM SG ">400</w>
<w lemma="000" norm="000" type="ABBR" msd=" NOM SG ">000</w>
<w lemma="markka" norm="markan" type="N" msd=" GEN SG ">markan</w>
<w lemma="palkkio" norm="palkkion" type="N" msd=" GEN SG ">palkkion</w>
```

Yhteen tulkitut peräkkäiset numerot voi erottaa toisistaan disambiguointivaiheessa ja twollata osat uudelleen. Tällöin uutta tulkintaa vaativa kohta erotetaan tekstistä ja twollataan uudelleen (uudelleentwollauksesta tarkemmin luvussa 4.3. Uudelleentwollaaminen). Twollauksen tuloksena saatu tulkintarivi liitetään entisen (tai entisten) paikalle tiedostoon ja tiedosto tallennetaan.
Twollausohjelmaa on Kotuksessa korjattu siten, että se ei enää jaa numeroiden eri osia eri riveille. Jos näin kuitenkin jostakin syystä käy, voi numeron twollata uudelleen, jolloin virhe korjaantuu. Myös vanhoissa aineistoissa esiintyvät tapaukset voi korjata twollaamalla numerot uudelleen. Jos uudelleentwollauksia tehdään disambiguoinnin yhteydessä, ne on aina dokumentoitava yksityiskohtaisesti.

2.4. Ilman tulkintaa jäävät sanat (#UNKNOWN)

Sanat, joille Fintwol ei anna lainkaan tulkintaa, jätetään sikseen. Omien tulkintojen lisääminen niihin tekisi aineistot vertailukelvottomiksi. Tunnistamattomat sanat saadaan listatuksi erikseen analyysivaiheessa, ja niiden edellyttämät varaukset voidaan tehdä laskemiin erikseen.

```
<w lemma="#UNKNOWN" norm="niinkuin" type="#UNKNOWN" msd="">niinkuin</w>
<w lemma="#UNKNOWN" norm="ja/tai" type="#UNKNOWN" msd="">ja/tai</w>
<w lemma="#UNKNOWN" norm="tummeli" type="#UNKNOWN" msd="">Tummeli</w>
```

Tekstissä olevat erikoismerkit jäävät tunnistamatta melko usein, mikä saattaa aiheuttaa myös tunnistettavan sanan tai lyhenteen jäämisen ilman tulkintaa.

```
<w lemma="#UNKNOWN" norm="km²" type="#UNKNOWN" msd="">km²</w>
<w lemma="km" norm="km" type="ABBR" msd=" NOM SG ">km</w>
```
Näissä tapauksissa #UNKNOWN-tulkinta jätetään sikseen ja mahdolliset korjaukset laskelmiin tehdään myöhemmin käsin.

Jos #UNKNOWN-tulkinta johtuu kirjoitusvirheestä alkuperäisessä tekstissä, se jätetään sikseen. Jos virhettä taas ei ole alkuperäistekstissä, vaan se on tapahtunut aineistonkäsittelyprosessin aikana, voidaan korjattu sana twollata uudelleen erikseen ja liittää oikea tulkinta aineistoon.
Katso myös tämän ohjeen lukua 4.1 Fintwolin tekemät virhetulkinnat.

2.5. Omistusliitteelliset sanat

Monitulkintaisten omistusliitteellisten sanojen tapauksessa käytäntönä on tarkastella tekstiyhteyttä:

```
<w lemma="päätös" norm="päätöksemme" type="N" msd=" NOM SG 1PL ">päätöksemme</w>
<w lemma="päätös" norm="päätöksemme" type="N" msd=" GEN SG 1PL ">päätöksemme</w>
<w lemma="päätös" norm="päätöksemme" type="N" msd=" NOM PL 1PL ">päätöksemme</w>
```

Tällaisissa tapauksissa valitaan kontekstin perusteella joko yksikkö tai monikko, mutta ei systemaattisesti vain toista, kuten on tehty joissakin vanhoissa aineistoissa.

3. Puuttuvien tietojen lisääminen

Fintwolin antama tekstisaneiden analyysi on osin puutteellista. Puuttuva mutta tutkimuksessa välttämättömänä pidettävä informaatio lisätään aineistoon käsin disambiguoinnin yhteydessä. Useimmiten lisäykset ovat koskeneet liittomuotoja sekä A/N-sanaluokkatulkinnan saaneiden sanojen käsittelyä.

Tulkintoihin tehtävät lisäykset tarkoittavat joko function- tai status-attribuutin lisäämistä. Tarkat ohjeet erilaisten attribuuttien lisäämisestä löytyvät alta. Lisäyksien kohdalla on kuitenkin erityisen tärkeää kiinnittää huomiota siihen, kuinka attribuutin saama arvo merkintään attribuutin yhteyteen. Function-attribuutin kohdalla attribuutin saaman arvon (esim. P tai PL) ympärille lainausmerkkien sisään lisätään aina välilyönnit (esimerkissä merkitty alleviivauksella):

```
<w lemma="olla" norm="olit" type="V" msd=" COP PAST ACT SG2 "
function=" PL ">olit</w>
```

Status-attribuutin kohdalla välilyöntejä ei lisätä:

```
<w lemma="tanskalainen" norm="tanskalainen" type="N" msd=" PROP DN-LAINEN POS
NOM SG " status="corrected_A/N">tanskalainen</w>
```

Välilyöntien lisääminen arvon yhteyteen määräytyy attribuutin ominaisuuksien mukaan. Kun arvoja voi olla useampia, on niiden ympärille lisättävä välilyönnit, jos taas vain yksi, voidaan arvo lisätä attribuutin yhteyteen ilman välejä. Status-attribuutti voi kerrallaan saada vain yhden arvon, function-attribuutti taas useamman, joten status-attribuutin yhteyteen välilyöntejä ei lisätä, kun taas function-attribuutin yhteyteen ne on aina laitettava. Huomaa: Välit lisätään siinäkin tapauksessa, että function-attribuutti saa vain yhden arvon. Tämä tekee automaattisesta laskennasta huomattvasti helpompaa.

Sama sääntö pätee myös attribuutteihin type ja msd, jotka löytyvät kaikista Fintwolin tulkinnoista. Type voi kerrallaan saada vain yhden arvon (esim. N tai ADV), msd taas saa useampia. Tästä syystä msd:n arvojen ympärillä on välit, kun taas typen arvon ympärillä ei ole.

3.1. Temporaaliset verbiketjut

Fintwol-analyysi ei riitä aikamuotojen tutkimiseen: analysaattori ei merkitse perfektiä eikä pluskvamperfektiä mitenkään. Tämän vuoksi liittomuotojen molempiin osiin (apu- ja pääverbiin) merkitään käsin tieto siitä, että ne muodostavat perfektisen tai pluskvamperfektisen komposition. Lisäys tehdään lisää-mällä tulkintaan attribuutti function, joka saa arvokseen joko " P " (perfekti) tai " PL " (pluskvamperfekti). Function-attribuutti lisätään aina w-merkitsimen loppuun, juuri ennen merkitsimen sulkumerkkiä >. Huomaa, että attribuutin arvon (P tai PL) ympärille lainausmerkkien sisälle lisätään molemmin puolin yksi välilyönti.

Function-attribuuttia käytetään siis esimerkiksi seuraavaan tapaan:

```
<w lemma="olla" norm="olen" type="V" msd=" COP PRES ACT SG1 " function=" P
">olen</w>
<w lemma="tehdä" norm="tehnyt" type="PCP2" msd=" ACT POS NOM SG " function=" P
">tehnyt</w>

<w lemma="olla" norm="olit" type="V" msd=" COP PAST ACT SG2 " function=" PL
">olit</w>
<w lemma="juosta" norm="juossut" type="PCP2" msd=" ACT POS NOM SG " function="
PL ">juossut</w>

<s>
<w lemma="tampere" norm="Tampereella" type="N" msd=" PROP ADE SG
">Tampereella</w>
<w lemma="ohje" norm="ohjeet" type="N" msd=" NOM PL ">ohjeet</w>
<w lemma="päivä#hoito" norm="päivähoidossa" type="N" msd=" INE SG ">päivähoidos-
<lb />sa</w>
<w lemma="oleva" norm="olevien" type="PCP1" msd=" ACT POS GEN PL ">olevien</w>
<w lemma="lapsi" norm="lasten" type="N" msd=" GEN PL ">lasten</w>
<w lemma="infektio" norm="infektioiden" type="N" msd=" GEN PL ">infektioiden</w>
<w lemma="ehkäisy" norm="ehkäisystä" type="N" msd=" DV-U ELA SG ">eh-<lb
/>käisystä</w>
<w lemma="ja" norm="ja" type="C" msd=" COORD ">ja</w>
<w lemma="hoito" norm="hoidosta" type="N" msd=" ELA SG ">hoidosta</w>
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 " function=" P
">on</w>
<w lemma="tehdä" norm="tehty" type="PCP2" msd=" PSS POS NOM SG " function=" P
">tehty</w>
<w lemma="vuosi" norm="vuosi" type="N" msd=" NOM SG ">vuosi</w>
<w lemma="sitten" norm="sitten" type="PP" msd="">sitten</w>
<w lemma="." norm="." type="delimiter">.</w>
</s>
```

Kielteisiin ketjuihin kieltoverbin kohdalle ei kuitenkaan merkitä funktiota:

```
<w lemma="ei" norm="en" type="V" msd=" NEGV SG1 ">en</w>
<w lemma="olla" norm="ollut" type="V" msd=" COP PAST ACT NEG SG " function=" PL
">ollut</w>
<w lemma="juoda" norm="juonut" type="PCP2" msd=" ACT POS NOM SG " function=" PL
">juonut</w>.
```

Konditionaalimuotoisten liittotempusten (*olisi tehty*) aikamuodoksi merkitään perfekti, function=" P
".

3.2. A/N

Yksi Fintwolin antamista sanaluokkatulkinnasta on "A/N". Tällaisista sanoista (esim. *suomalainen*) Fintwol tarjoaa vain yhden tulkintamahdollisuuden, juuri sanaluokan A/N. Se ei siis erota tämän luokan sanojen adjektiivi- ja substantiiviesiintymiä toisistaan. Vain harvoissa tapauksissa voidaan kuitenkin sanoa jonkin sanan kuuluvan aidosti A/N-luokkaan. On siis vähän sellaisia tapauksia, joista ei voida sanoa, ovatko A/N-tulkinnan saavat saneet käyttöyhteydessään adjektiiveja vai substantiiveja.

Pääsääntö on tämä: Mikäli tekstiyhteyden perusteella on selvää, että sana on adjektiivi, tulkintaa korjataan siten, että typestä poistetaan ylimääräiseksi tulkintavaihtoehdoksi jäävä N sekä

kauttaviiva. Tulkintaan lisätään status-attribuutti, josta ilmenee, että kyseistä kohtaa on korjattu (status="corrected_A/N"). Korjattu tulkinta on seuraavanlainen:

```
<w lemma="kuuluisa" norm="kuuluisa" type="A" msd=" POS NOM SG ">kuuluisa</w>
<w lemma="tanskalainen" norm="tanskalainen" type="A" msd=" PROP DN-LAINEN POS
NOM SG " status="corrected_A/N">tanskalainen</w>
<w lemma="kielentutkija" norm="kielentutkija" type="N" msd=" DV-JA NOM SG
">kielentutkija</w>
```

Jos sana on kontekstissaan substantiivi, tehdään korjaus samaan tapaan:

```
<w lemma="kuuluisa" norm="kuuluisa" type="A" msd=" POS NOM SG ">kuuluisa</w>
<w lemma="tanskalainen" norm="tanskalainen" type="N" msd=" PROP DN-LAINEN POS
NOM SG " sta-tus="corrected_A/N">tanskalainen</w>
```

Typen arvoa siis muokataan sen mukaan, onko tarkasteltavana oleva sana adjektiivi vaiko substantiivi. Korjaus kuitenkin jätetään näkyviin lisäämällä tulkintaan siitä merkintä status-attribuutin avulla.

Huomaa, että ennen tämän ohjeen valmistumista tehdyissä aineistoissa A/N-sanaluokkaan luokiteltuja tapauksia on käsitelty eri tavalla kuin tässä ohjeessa. Vanhat aineistot pyritään korjaamaan tämän ohjeen mukaisiksi takautuvasti.

Jos A/N-luokkaan jostakin syystä on luokiteltu sanoja virheellisesti (esim. *taianomainen* on merkitty A/N-tapaukseksi), käsitellään näitä virheinä normaaliin tapaan. (Katso ohjeita virheiden käsittelystä luvussa 4 Virhetapausten käsittely.)

3.3. Substantiiveiksi leksikaalistuneet partisiipit

Substantiiveiksi leksikaalistuneille partisiipeille (*syytetty*, *valtuutettu*, *syötävä*) Fintwol tarjoaa adjektiivi- ja partisiippitulkinnat, mutta ei substantiivitulkintaa. Näissä kohdissa Fintwolin antamista tulkinnoista hyväksytään partisiippitulkinta, johon lisätään function-attribuutti, joka saa arvon "N" (function="N"). Näin tapaukset ovat myöhemmin helposti löydettävissä ja – jos niin päätetään – muunnettavissa.

```
<w lemma="syyttää" norm="syytetty" type="PCP2" msd=" PSS POS NOM SG " function="
N ">syytetty</w>
<w lemma="valtuuttaa" norm="valtuutettu" type="PCP2" msd=" PSS POS NOM SG "
function=" N ">
valtuutettu</w>

<w lemma="syödä" norm="syötävä" type="PCP2" msd=" PSS POS NOM SG " function=" N
">syötävä</w>
```

Säännöstä poiketaan vain erityistapauksissa, joten kaikki päätökset kannattaa dokumentoida. Asiaa arvioidaan jatkossa uudelleen. Myös tämä päätös on uusi, ja sen mukaiset merkinnät puuttuvat vanhoista aineistoista.

4. Virhetapausten käsittely

Aineistonkäsittelyprosessissa tulee vastaan monenlaisia virheitä. Virheet voidaan jakaa kolmeen ryhmään: kopiointi- tai skannausvirheisiin, alkuperäisaineistossa oleviin virheisiin ja twollausvirheisiin.

Kopiointi- tai skannausvirheet korjataan jo ennen rakenteistamisvaihetta. Jos virheet kuitenkin pääsevät disambiguointivaiheeseen, ne korjataan silloin. Virheellinen sane analysoidaan (twollataan) uudestaan erikseen ja sen saamista analyyseista valitaan paras.

Alkuperäisessä tekstissä olevia virheitä ei korjata. Tämä siksi, että on vaikea arvioida, missä kohdissa virhe on todellinen, esimerkiksi näppäily- tai ladontavirhe, missä kohdissa esimerkiksi kirjoittajan jostakin syystä tarkoittama tekstin ominaisuus. Vanhojen tekstien kohdalla tulkintavirheitä tulee erityisesti sellaisten kielenkäänteiden kohdalle, jotka ovat kirjoitushetkellään olleet hyväksyttävää kieltä, mutta joita nykyään ei pidetä kovin yleiskielisinä ilmauksina. Jos virheelliset sanat saavat tulkintoja, niistä valitaan paras. Normaalikäytännön mukaan alkuperäistekstin virheitä ei myöskään merkitä aineistoon mitenkään, vaikka niiden saama tulkinta olisikin virheellinen. Jos tällaiset virheet halutaan aineistoon jostakin syystä merkitä, tarvitaan siihen aineiston haltijan tai käyttäjän lupa. (Merkitsemisohje on luvussa 5 Vapaaehtoisia lisämerkintöjä.)

4.1. Fintwolin tekemät virhetulkinnat

Myös Fintwol tekee virheitä. Tällöin sanalla ei ole lainkaan käyttöyhteyteen sopivaa tulkintaa tai se on jaettu väärin ja saa useampia tulkintoja. Saaduista tulkinnoista valitaan paras tai hyväksytään tarjottu ja lisätään siihen käsin merkintä virheellisestä analyysista (status="false"). Virheellisiksi katsotaan vain ne sanat, joiden sanaluokka (type) ja/tai perusmuoto (lemma) ovat virheellisiä. Näin pyritään pitämään tulkinnat aisoissa ja aineistot yhteismitallisina.

```
<w lemma="hän" norm="hän" type="PRON" msd=" PERS NOM SG ">Hän</w>
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 ">on</w>
<w lemma="hoito#vapaa" norm="hoitovapaalla" type="A" msd=" POS ADE SG "
status="false">hoitovapaalla</w>

<w lemma="herra" norm="herra" type="N" msd=" NOM SG ">Herra</w>
<w lemma="kuroa" norm="kuronen" type="V" msd=" POTN ACT SG1 "
status="false">Kuronen</w>

<w lemma="vuosi" norm="vuodesta" type="N" msd=" ELA SG ">vuodesta</w>
<w lemma="nykyinen" norm="nykyiseen" type="A" msd=" POS ILL SG ">nykyiseen</w>
<w lemma="puoli" norm="puoleen" type="N" msd=" ILL SG "
status="false">puoleen</w>
<w lemma="vuosi" norm="vuoteen" type="N" msd=" ILL SG ">vuoteen</w>
<w lemma="." norm="." type="delimiter">.</w>
```

Näiden esimerkkien virhetapaukset ovat melko selkeitä: *hoitovapaa* ei ole adjektiivi, eikä nimeä *Kuronen* voida tulkita verbiksi. *Puoleen* taas olisi tässä yhteydessä tulkittava lähinnä numeraaliksi. Tämän sanan kohdalla kuitenkin havainnollistuu Fintwollin virhe, sillä *puoli* saa twollauksessa myös numeraalitulkinnan, mutta *puoleen*-sanalle sitä ei tule lainkaan.

Huomaa, että status-attribuutti lisätään aina msd-attribuutin jälkeen eli w-merkitsimen ensimmäisen osan viimeiseksi (ks. yllä olevat esimerkit). Jos samaan saneeseen joskus pitää lisätä sekä function- että status-attribuutit, sijoitetaan näistä function ensiksi ja status sen jälkeen.

Toisinaan Fintwolin antama tulkinta on kuitenkin vain osittain oikein tai väärin. Tällöin huomiota pitää ensisijaisesti kiinnittää tulkinnan perusmuotoon (lemma) ja sanaluokkaan (type), joiden on aina oltava oikein. Mikäli nämä ovat oikein, tulkinta hyväksytään, vaikka msd sisältäisikin virheitä. Esimerkiksi sana *jonakin* saa msd:hen elatiivitulkinnan, mutta sen tulkinta on muuten oikein, eli lemma ja type ovat virheettömät.

```
<w lemma="jokin" norm="jonakin" type="PRON" msd=" Q ELA SG ">Jonakin</w>
```

Tässä tapauksessa tulkinta hyväksytään, sitä ei siis merkitä virheelliseksi status-attribuutilla. Esimerkkejä virheellisistä sanaluokkatulkinnoista (suluissa Fintwolin tarjoama sanaluokka/type) ovat: *luonteenomainen* (N), *Yrjö-Koskinen* (A), *suhdanne-* (N), *asianomainen* (N) ja *mielihyvä* (A). Esimerkkejä vääristä perusmuodoista: *uudenvuosi* ja *pitkänkantama*. Virheellisen tulkinnan saa systemaattisesti myös *usea*, jonka Fintwol tulkitsee adjektiiviksi. Tällaisiin tulkintoihin lisätään status="false"-merkintä.

Sellaisissa tapauksissa kuin *talous- ja sosiaalipolitiikka* tai *kotimaan- ja ulkomaanliikenne* sanaluokkavirheet ovat yleisiä. Jos on rinnastettu kaksi adjektiivia (esim. *talous- ja sosiaalipoliittinen*), ensimmäinen (se jonka perusosa korvattu yhdysmerkillä) on yleensä virheellisesti tulkittu substantiiviksi. Mikäli sanaluokka on kuitenkin oikein (*talous-* (N) *ja sosiaalipolitiikka*), tulkinta hyväksytään, vaikka lemmaksi olisikin merkitty "talous-". A/N-luokkaan kuuluvat sanat voivat eri yhteyksissä olla joko adjektiiveja tai substantiiveja. Fintwolin analyysissa kaikki tällaiset sanat saavat vain yhden tulkinnan (A/N). A/N-luokan mielekkyyttä arvioitaessa kannattaa huomioida, että Fintwol ei käsittele yhtä sanaa pidempiä kokonaisuuksia. Näin ollen se ei tarjoa mahdollisuutta erottaa substantiiveja ja adjektiiveja toisistaan suoraan. Tästä syystä A/N-tapauksia käsitellään hieman poikkeuksellisesti. (A/N-tapausten käsittelystä tarkemmin luvussa *A/N*.) Poikkeuksellisesti menetellään myös substantiiveiksi leksikaalistuneiden partisiippien (esim. *syytetty*, *valtuutettu*) tapauksissa (ks. lukua *Substantiiveiksi leksikaalistuneet partisiipit*).

Ongelmatapauksissa kannattaa tarkistaa, kuinka vastaavat tapaukset on aiemmin tulkittu. Aina kannattaa myös dokumentoida omat päätökset. Näin aineistojen käsittelyssä säilytetään yhtenäinen linja, ja tiedot mahdollisista poikkeustapauksista jäävät tulevienkin disambiguoijien ja aineistojen käyttäjien saataville.

4.2. Segmentointivirheet

Jos disambiguoitaessa tulee vastaan Fintwollin sananjaossa tekemiä virheitä, niistä mahdollisesti johtuva sanojen irto-osat merkitään status="part"-attribuutilla, jolloin ylimääräinen osa jää huomiotta laskentavaiheessa. Twollausohjelmaa on korjattu sen tekemien segmentointivirheiden osalta. Segmentointivirheitä on edelleen vanhoissa tekstiaineistoissa ja niitä korjataan takautuvasti. Status="part"-attribuutti on lisätty myös silloin, kun Fintwol on tulkinnut päivämääriin ja järjestyslukuihin kuuluvat pisteet erilleen sanan kanssa. Tällöin status="part" liitetään pisteen saamaan tulkintaan. Myös joihinkin, yleensä vakiintumattomiin lyhenteisiin (esim. *valok.*) liittyviä pisteitä voi joutua korjaamaan status="part"-merkinnän avulla. Huomaa, että tämä pätee myös vierasperäisiin sanoihin, joita on taivutettu seuraavasti: *ward'in* tai lyhenne *mk.*

Näissä esimerkeissä sana tai jokin kokonaisuus (kuten päivämäärä) on analysoitu virheellisesti numerot ja merkit kerrallaan:

```
<w lemma="7" norm="7" type="ABBR" msd=" NOM SG ">7</w>
```

```
<w lemma="." norm="." type="delimiter">.</w>
<w lemma="7" norm="7" type="ABBR" msd=" NOM SG ">7</w>
<w lemma="." norm="." type="delimiter">.</w>
<w lemma="1999" norm="1999" type="ABBR" msd=" NOM SG ">1999</w>

<w lemma="#UNKNOWN" norm="parast" type="#UNKNOWN" msd="">parast</w>
<w lemma="#UNKNOWN" norm="’" type="#UNKNOWN" msd="">’</w>
<w lemma="aika" norm="aikaa" type="N" msd=" PTV SG ">aikaa</w>
```

Ensimmäisen esimerkin mukaisia tapauksia voi löytyä vanhoista aineistoista, mutta vastaavanlaisia virheitä ei tapahdu enää, vaan päivämäärä tulkitaan yhdeksi kokonaisuudeksi. Aiemmin tulkintoja on tullut monta yhden sijaan (yksi kullekin numerolle ja merkille), ja niistä osa saattoi olla virheellisiä. Vanhoja aineistoja korjatessa tällaiset kohdat twollataan uudelleen, jolloin virhe poistuu.

Segmentointivirheitä on löytynyt myös sellaisista erikoistapauksista kuin *parast’aikaa*. Kyseessä on vanhahtava ilmaisu, jota Fintwol ei osaa käsitellä oikein. Tällaiset kohdat on korjattava status="part"-merkinnällä, minkä lisäksi viimeiseen osaan tulee merkintä status="false", mikäli viimeinen osa on virheellinen (kuten tässä tapauksessa).

```
<w lemma="#UNKNOWN" norm="parast" type="#UNKNOWN" msd=""
status="part">parast</w>
<w lemma="#UNKNOWN" norm="’" type="#UNKNOWN" msd="" status="part">’</w>
<w lemma="aika" norm="aikaa" type="N" msd=" PTV SG ">aikaa</w>
```

Status-attribuutti lisätään siis niiden sanojen kohdalle, jotka sisältävät vähiten oikeaa tulkintaa. Sanan edustavin osa jätetään merkitsemättä. Edustavin osa on tavallisesti kokonaisuuden viimeinen. Viimeinen osa tai sana sisältää usein esimerkiksi taivutuspäätteen, joka on hyvä saada mukaan aineistosta tehtäviin laskelmiin. Jos viimeinen osa on kuitenkin tulkittu virheellisesti segmentoinnin tai muun syyn takia, se merkitään status="false"-attribuutilla. Jos kyseessä on alkuperäistekstin virhe, merkintä jätetään tekemättä. Asiasta on päätettävä tapauskohtaisesti.

Segmentointivirheistä on vielä olemassa erillinen ohje, Segmentointivirheiden korjaaminen, joka on tarkoitettu muistin tueksi vanhojen aineistojen korjaamisessa. Perusdisambiguoinnissa sitä ei tarvita.

4.3. Uudelleentwollaaminen

Joissakin tapauksissa on perusteltua ajaa osa disambiguoitavasta tekstistä uudelleen morfologisen analysaattorin läpi, siis twollata se uudelleen. Näitä tapauksia voivat olla esimerkiksi sellaiset, joissa tekstintunnistus on skannauksen yhteydessä selvästi tulkinnut väärin alkuperäisessä dokumentissa olevaa tekstiä tai ensimmäisessä twollwuksessa ohjelmisto ei ole selvinnyt ihmiselle selkeiden numerosarjojen segmentoinnista. Esimerkkinä jälkimmäisestä tapauksesta voisi olla virke *Tuotanto oli vuonna 1973 400 000 kappaletta*, joka twollauutuu seuraavasti:

```
<w lemma="tuotanto" norm="tuotanto" type="N" msd=" NOM SG ">tuotanto</w>
<w lemma="olla" norm="oli" type="V" msd=" COP PAST ACT SG3 ">oli</w>
<w lemma="vuosi" norm="vuonna" type="N" msd=" ESS SG ">vuonna</w>
<w lemma="1973 400 000" norm="1973 400 000" type="NUM" msd="">1973 400 000</w>
<w lemma="kappale" norm="kappaletta" type="N" msd=" PTV SG ">kappaletta</w>
<w lemma="." norm="." type="PUNCT" msd=" FULLSTOP ">.</w>
```

Tässä tapauksessa twollataan erikseen merkkijonot *1973* ja *400 000* ja virheellisesti segmentoitu lukujono korvataan niillä:

```
<w lemma="1973" norm="1973" type="NUM" msd="">1973</w>
<w lemma="400 000" norm="400 000" type="NUM" msd="">400 000</w>
```

5. Vapaaehtoisia lisämerkintöjä

Vaikka Fintwolin analyysi antaa teksteistä paljon tärkeää perustietoa, analyysi on monin osin riittämätöntä. Aineistoihin voidaankin tehdä lisämerkintöjä, niin paljon kuin kunnianhimo ja aikataulu ynnä muut sellaiset seikat antavat myöten. Jos disambiguointia tekee joku muu kuin aineistoa tutkimuksessaan käyttävä henkilö, on lisämerkinnöistä hyvä sopia aineiston haltijan tai käyttäjän kanssa mahdollisimman varhaisessa vaiheessa. Kaikki poikkeusratkaisut esimerkkeineen on dokumentoitava.

5.1. Erisnimet, sanaliitot ja muut tekstijaksot

Fintwol tulkitsee erisnimiä puutteellisesti. Analyysi antaa yleisimmille erisnimille merkitsimen "PROP". Fintwol ei kuitenkaan ota huomioon esimerkiksi sitä, että nimet koostuvat usein useista sanoista. Tämän vuoksi Teko-aineistoissa ei käytetä hyväksi Fintwolin "PROP"-merkitsintä. Monet tutkimuksen osaongelmat osoittautuvat sellaisiksi, että niiden tarvitsemat merkinnät ylittävät sanarajat. Esimerkiksi nimet, määriin viittaaminen ja idiomit voivat olla useampisanaisia sekvenssejä. Nimet voidaan lisätä aineistoon (<name>nimi</name>-merkitsimellä), samoin kuin mitat (<measure>), osoitteet (<adress>), ajan ilmaukset (<time>) sanaliitot ja muut tekstijaksot, jos tutkimusongelma sitä edellyttää. Tutkimusongelmakohtaiset merkinnät tehdään jo rakenteistamisvaiheessa tekstirakenteeseen (TEI), koska TEI tarjoaa monille maailmaa kuvaaville elementeille valmiit merkintätavat. Niitä ei siis lisätä disambiguoinnin yhteydessä ilman erityistä syytä (jollainen voi olla esimerkiksi tarve tarkentaa merkintöjä jälkeenpäin). Lisätietoa TEI:stä tämän dokumentin lopussa, luvussa Lisätietoja ja linkkejä.

Sanaliiton muodostavat erisnimet (kuten *Turun Sanomat* tai *Pirkanmaan allergia- ja astmaliitto*) voidaan tarvittaessa ympäröidä name-merkitsimellä:

```
<name>
<w lemma="jyväskylä" norm="Jyväskylän" type="N" msd=" PROP GEN SG
">Jyväskylän</w><lb />
<w lemma="hiihto#seura" norm="hiihtoseuran" type="N" msd=" GEN SG
">hiihtoseuran</w>
</name>
<w lemma="," norm="," type="delimiter">,</w>
<name>
<w lemma="lieksa" norm="Lieksan" type="N" msd=" PROP GEN SG ">Lieksan</w>
<w lemma="hiihto#seura" norm="hiihtoseuran" type="N" msd=" GEN SG ">hiihto-<lb
/>seuran</w>
</name>
<w lemma="ja" norm="ja" type="C" msd="COORD ">ja</w>
<name>
<w lemma="tampere" norm="Tampereen" type="N" msd=" PROP GEN SG ">Tampereen</w>
<w lemma="pyrintö" norm="Pyrinnön" type="N" msd=" GEN SG ">Pyrin-<lb />nön</w>
</name>
<w lemma="mäki#hyppy" norm="mäkihypyn" type="N" msd=" GEN SG ">mäkihypyn</w>
<w lemma="ja" norm="ja" type="C" msd=" COORD ">ja</w>
```

```
<w lemma="yhdistää" norm="yhdistetyn" type="PCP2" msd=" PSS POS GEN SG
">yhdistetyn</w>
<w lemma="jaosto" norm="jaostoille" type="N" msd=" ALL PL ">jaostoille</w>
<w lemma="." norm="." type="delimiter">.</w>
</s>
```

5.2. Modaaliset verbiketjut

Modaaliset (esim. nesessiiviset ja mahdollisuutta ilmaisevat) verbiketjut muistuttavat syntaktisesti temporaalisia verbiketjuja. Ne voidaan tarvittaessa lisätä perusanalyysiin function-attribuutilla samaan tapaan kuin perfekti ja pluskvamperfekti. Modaalisia apuverbejä on kuitenkin enemmän, joten apuverbiys merkitään yhdistelmään vielä erikseen, myös kieltoverbiin.

```
<w lemma="olla" norm="on" type="V" msd=" COP PRES ACT SG3 " function=" MOD-NES-
AUX ">on</w>
<w lemma="olla" norm="oltava" type="PCP1" msd=" COP PSS POS NOM SG " function="
MOD-NES ">oltava</w>

<w lemma="saada" norm="saa" type="V" msd=" PRES ACT SG3 " function=" MOD-NES-AUX
">saa</w>
<w lemma="merkitä" norm="merkitä" type="V" msd=" INF1 NOM " function=" MOD-NES
">merkitä</w>

<w lemma="ei" norm="ei" type="V" msd=" NEGV SG3 " function=" MOD-MAH-AUX
">ei</w>
<w lemma="saada" norm="saa" type="V" msd=" PRES ACT NEG " function=" MOD-MAH-AUX
">saa</w>
<w lemma="käsitellä" norm="käsitellä" type="V" msd=" INF1 NOM " function=" MOD-
MAH ">käsitellä</w>
```

5.3. Lauseenvastikkeet

Myös lauseenvastikkeet voidaan merkitä function-attribuutilla. Rakenteita pidetään perinteisesti lauseenvastikkeina silloin, kun ne ovat muutettavissa sivulauseiksi. Lauseenvastikkeista merkitään temporaaliset ja finaaliset (produktiiviset). Fintwol merkitsee referatiiviset (REF: *tiedämme sen olevan täällä*) ja temporaalisista *saatuanne*-tyyppiset (TEMP), joten function-attribuutti on lisättävä temporaalisista vain *lähtiessäsi*- ja *lähdettyäsi*-tyyppeihin (NFC-TEMP) ja finaalisista *voidakseni*-tyyppiin (NFC-FIN). Lauseenvastiketieto lisätään verbitulkintaan function-attribuutilla:

```
<w lemma="olla" norm="olevan" type="V" msd=" COP REF PRES ACT ">olevan</w>

<w lemma="saada" norm="saatuanne" type="V" msd=" TEMP PAST ACT 2PL
">saatuanne</w>

<w lemma="lähteä" norm="lähtiessäsi" type="V" msd=" INF2 ACT INE 2SG "
function=" NFC-TEMP " >lähtiessäsi</w>

<w lemma="voida" norm="voidakseni" type="V" msd=" INF1 TRA 1SG " function=" NFC-
FIN ">voidakseni</w>
```

5.4. Alkuperäistekstien virhetapaukset

Alkuperäistekstissä olleet lyöntivirheet tai muut sellaiset ongelmat voidaan merkitä status-attribuutin arvolla "false/orig", jos halutaan. Normaalikäytännön mukaan tällaisia tapauksia ei merkitä virheeksi lainkaan, vaan ne jätetään sinälleen. Niitä ei myöskään korjata.

5.5. Muut lisäykset

Selvästi semanttisten tulkintojen lisäämistä aineiston sanoihin on suunniteltu. Kokeilumielessä on joihinkin uudenvuodenpuheisiin lisätty mm. tietoja lauseprosesseista. Samoin uudenvuodenpuheisiin on lisätty tekstikappaleittain topiikkitietoja, siis tietoja jokaisen kappaleen keskeisestä puheenaiheesta (intuition ja maailmantiedon perusteella).

6. Merkitsimet

6.1. Lingsoftin lista

(Ks. http://www2.lingsoft.fi/doc/fintwol/intro/tags.html)

Sanaluokka
A adjektiivi (pieni)
ABBR lyhenne (esim.)
AD_A ad-adjektiivi (melkein)
ADV adverbi (hitaasti)
ART vieraskielinen artikkeli (das)
C konjunktio (ja)
INTJ interjektio (hui)
N substantiivi (koira)
NUM numeraali (kaksi)
PP post- tai prepositio (jälkeen, ennen)
PREP vieraskielinen prepositio (de)
PRON pronomini (sinä)
PSP postpositio (vieressä)
Q kvanttori, määrän ilmaisu (moni)
V verbi (tulla)

Komparaatio
POS positiivi (kuuma, hyvä)
CMP komparatiivi (kuumempi, parempi)
SUP superlatiivi (kuumin, paras)

Sija
NOM nominatiivi (koira)
GEN genetiivi (koiran)
PTV partitiivi (koiraa)
ESS essiivi (koirana)
TRA translatiivi (koiraksi)
INE inessiivi (koirassa)
ELA elatiivi (koirasta)
ILL illatiivi (koiraan)

ADE adessiivi (koiralla)
ABL ablatiivi (koiralta)
ALL allatiivi (koiralle)
ABE abessiivi (koiratta)
CMT komitatiivi (koirineen)
INS instruktiivi (koirin)

Yksikkö ja monikko
SG yksikkö (kala)
PL monikko (kalat)

Possessiivisuffiksit
1SG Yksikön 1. persoona (tyttäreni)
2SG Yksikön 2. persoona (tyttäresi)
3 Yksikön tai monikon 3. persoona (tyttärensä)
1PL Monikon 1. persoona (tyttäremme)
2PL Monikon 2. persoona (tyttärenne)

Tapa
IMPV imperatiivi (lue!, mene!)
COND konditionaali (lukisi, menisi)
POTN potentiaali (lukenee, mennee)
Indikatiivimuotoja (lukee, menee) ei merkitä.

Aikamuoto
PRES preesens (haluan)
PAST mennyt aika, imperfekti (halusin)
Perfektit ja pluskvamperfektit tulkitaan partisiipeiksi (PCP1 ja PCP2).

Aktiivi ja passiivi
ACT aktiivi (uin)
PSS passiivi (uidaan)

Luku
SG1 Yksikön 1. persoona (menen)
SG2 Yksikön 2. persoona (menet)
SG3 Yksikön 3. persoona (menee)
PL1 Monikon 1. persoona (menemme)
PL2 Monikon 2. persoona (menette)
PL3 Monikon 3. persoona (menevät)
PE4 passiivin pääte (mennään)

Kielto
NEGV kielteinen verbi (en, et, ei)
NEG kieltomuoto (en tehnyt)

Infinitiivit
INF1 1. infinitiivi (tulla, tullakseni)
INF2 2. infinitiivi (tullessaan, tullessa)
INF3 3. infinitiivi (tulemaan)

INF5 5. infinitiivi (tulemaisillaan)
4. infinitiivi (tuleminen) tulkitaan substantiiviksi (N)

Partisiipit
PCP1 1. partisiippi (lentävä, lennettävä)
PCP2 2. partisiippi (lentänyt, lennetty)

Kliitit
hAn -han/-hän (poikahan)
kA -ka/-kä (eikä)
kAAn -kaan/-kään (poikakaan)
kin -kin (poikakin)
kO -ko/-kö (oletko)
pA -pa/-pä (oletpa)
s -s (onpas)

Muut
FORGN vieraskielinen sana (British)
PROP propri (Mikko) [Ei käytössä teko-projektissa!]
pi -pi (ompi)

6.2. Käytössä löytyneet merkitsimet

Käytössä löytyneitä merkitsimiä on raportoitu erityisesti uudenvuodenpuheaineistosta (mm. Virpi Aaltonen & Leena Maria Heikkola). Näitä on hyvä kirjata muistiin lisää sitä mukaa, kun niitä tulee vastaan.

COORD rinnastuskonjunktio
COP kopula
DA-UUS deadjectival, -UUs-johdin (rikollisuus)
DEM demonstratiivipronomini
DV-LLINEN deverbal -llinen-johdin (ruumiillinen)
DN-INEN denominal -inen-johdin (osainen)
DN-ITTAIN denominal -ittain-johdin (osittain)
DV-MA deverbal -ma-johdin (luoma)
DV-MATON deverbal -maton-johdin (murtumaton)
DV-NTAA deverbal -ntaa-johdin (vähentämällä)
DV-TTA deverbal -tta-johdin (huolestuttava)
DV-U deverbal -u-johdin (rajoitu)
INTG interrogatiivipronomini (mitä)
INTERR interrogatiivinen
MAN tapa-adverbiluokka
PERS persoonapronomini
REF referatiivinen lauseenvastike
REL relatiivipronomini
SUB alistuskonjunktio
TEMP temporaalinen lauseenvastike

6.3. Viljami Haakanan (4.4.2011) merkitsinlöydöksiä

DN-LAINEN (siilinjärveläistäjätär)
DN-LAISTA (siilinjärveläistäjätär)
DV-JA (siilinjärveläistäjätär)
DV-TAR (siilinjärveläistäjätär)
DN-LAISTU (berliiniläistyvissä)
DN-MAINEN (liisimäistä)
DV-NTA (juhlinta)
DV-NA (kohina)
DV-NTI (syönnit, marinointi)
DN-TON (saareton)
PTV/ABE (viidettä)
kOs (lähekkäisetkös)
DV-ELE (johtele, käyskennellen)
DV-ILE (hikoiluta)