KAS TEGELIK TEKST ALLUB EESTI KEELE MORFOLOOGILISTELE KIRJELDUSTELE?

Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemus

HEIKI-JAAN KAALEP, KADRI MUISCHNEK, KAILI MÜÜRISEP, ANDRIELA RÄÄBIS, KÜLLI HABICHT

Sissejuhatus

Korpuste morfosüntaktilisele märgendamisele on juba paljude keelte puhul suurel määral energiat kulutatud ning mõneski keeles on tänaseks olemas muljetavaldav kogus märgendatud tekste. Selle taustal võib jääda mulje, et iga uue keele korpuse märgendamine on rutiinne tegevus, kus põhiküsimuseks vaid praktilisus, kiirus ja automatiseeritus – universaalsed märgenduspõhimõtted peaksid kasutuskõlblikud olema. Tingituna keele morfoloogilise süsteemi omapärast, grammatikatraditsioonist ja keele uurituse tasemest, võime siiski leida end olukorrast, kus kõige vajalikumaks osutub korpuse tavaline aeglane ja grammatikanüanssidesse süvenev käsitsi märgendamine ning alles pärast selle etapi läbimist on mõtet pöörata tähelepanu kiirusele ja automaatsusele.

Aeglase käsitsimärgendamise etapi vajalikkus õigustab selle käsitlemist omaette objektina. Võib oletada, et see on etapp, mis tuleb igal juhul läbida nii uue keele kui ka uue märgenduse (süntaktilise, semantilise, pragmaatilise vms) kasutamisel.

Käesolev artikkel kirjeldab George Orwelli "1984" eestikeelse tõlke[1] põhirõhk kvaliteedile. Märgendussüsteemi valikul ei pööratud tähelepanu sellele, kui mugav on seda kasutada märgendamise automatiseerimisel, vaid ennekõike märgendite lingvistilisele informatiivsusele, nn. märgendussüsteemi välisele põhjendatusele.[2] Siinkohal väike näide kasutatud märgenduse kohta (käsitsi märgendamisel tehtud valikut näitab kriipsuke õige vormi kirjelduse ees):

```
Tema
 -   tema+0 //_P_ pers ps3 sg gen //
     tema+0 //_P_ pers ps3 sg nom //
keha
     keha+0 //_S_ com sg gen //
 -   keha+0 //_S_ com sg nom //
     keha+0 //_S_ com sg part //
tundus
     tundu+s //_V_ main indic impf ps3 sg ps af //
 -   tundu+s //_V_ mod indic impf ps3 sg ps af //
olevat
     olev+t //_A_ pos sg part //
     olev+t //_S_ com sg part //
 -   ole+vat //_V_ main quot pres ps af //
     ole+vat //_V_ main quot pres ps neg //
     ole+vat //_V_ aux quot pres ps af //
     ole+vat //_V_ aux quot pres ps neg //
vedel
 -   vedel+0 //_A_ pos sg nom //
     vedel+0 //_S_ com sg nom //
     vedel+0 //_S_ com sg nom //
nagu
     nagu+0 //_D_ //
```

[1] G. Orwell, 1984. – Loomingu Raamatukogu. Tallinn, Perioodika, 1980.

[2] G. Leech, Grammatical tagging. – Corpus Annotation, ed. by R. Garside, G. Leech, A. McEnery. Longman, London and New York, 1997.

```
 -   nagu+0 //_J_ sub //
s&uuml;lt
    s&uuml;lt+0 //_S_ com sg nom //
```

Ka ei püütud märgendamist kuidagi eriliselt automatiseerida – kasutati tavalisi tekstiredaktoreid ja UNIXi skripte. Tööks kasutati äärmiselt detailset märgendussüsteemi, mis peaks eesti keele sõnamuutmist ja sõnaliike adekvaatselt kirjeldama.

Esimene, poolkäsitsi märgendamine tehti COPERNICUSe projekti Multext-East[3] raames 1997. aastal. Hinnates tulemuste usaldusväärsust, selgus, et umbes 3,5% sõnade puhul olid eri filoloogid eri meelt selle suhtes, milline märgend valida. Seejuures oli enamik erimeelsusi tingitud teoreetilise käsitluse erinevustest, näpuvigade osakaal oli tühine.

Nii esimene ühestamisaktsioon "1984ga" kui ka filoloogide arusaamade lahknemise ilmsikstulek viitasid sellele, et suuremahulisele eesti keele ühestamise praktilisele tööle on veel vara asuda. Selle põhjused olid järgmised:

1. Ilmnenud vead morfoloogilises märgenduses (0,1% sõnadest). Kuigi inimeste ülesanne oli valida ainult etteantud variantidest õige, märkasid nad siiski, et vahetevahel õiget alternatiivi ei pakutudki. Sel juhul tuli õige versioon lisada.
2. Filoloogide lahkarvamused õige märgenduse valiku osas, kusjuures ei olnud täpselt ette teada, milles filoloogid eriarvamusel on.

Seetõttu otsustati ennast paremini ette valmistada kahes aspektis:

1) lingvistilises;
2) organisatsioonilis-majanduslikus (et osata efektiivselt tööd planeerida, tuleb teada, kui palju mingi mitteefektiivne tööülesanne aega võtab.)

Selleks tehti katse võimalikult täpselt ja lingvistiliselt põhjendatult märgendada “1984” tekst.

Järgnevalt kirjeldame kaalutlusi, millest lähtusime; töö käiku; eri ülesannete täitmiseks kulunud aega ja esile kerkinud lingvistilisi probleeme.

Eesmärgid

Otsustades märgendada "1984" laitmatu kvaliteediga, lootsime saavutada korraga kolm eesmärki:

1. Luua testmaterjali keeletehnoloogiliste programmide kontrollimiseks.

Kui meil on soov testida morfoloogiaanalüsaatorit, ühestajat või mingit muud programmi, siis peaks testmaterjal olema sajaprotsendiliselt korrektne. Eriti raskesti on leitavad juhtumid, kus morfoloogiaanalüsaator annab küll võimaliku analüüsi, aga see ei sobi kontekstiga. Need on vead, mida muul viisil kui käsitsi ühestades on võimatu leida.

2. Kontrollida, missugune eesti keele morfoloogiasüsteemi kirjeldus vastab tegelikele tekstidele kõige paremini.

Filoloogide raskused mõnel juhul õige märgendi valimisel viitasid sellele, et grammatikakirjeldused pole kas piisavalt detailsed ega/või üheselt mõistetavad. Ainus viis selliste raskete juhtumite klassifitseerimiseks ja ühetaoliseks lahendamiseks on nende ülevaatamine korpuses. “1984” 75 000-sõnaline korpus peaks olema piisav

[3] L. Dimitrova, T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic, D. Tufis, Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European languages. COLING-ACL’98. Proceedings of the Conference, Montreal, 1998, Vol 1, pp. 315-319.

statistika ja järelduste tegemiseks. Korpust saaks sellisel juhul kasutada näidete allikana ka näit. eesti keele morfoloogia õpetamisel.

Valisime küllalt detailse, nn. traditsioonilise morfoloogilise märgenduse. Sel juhul pakub korpus huvi ka väljaspool arvutilingvistikat.

Üks levinud viis korpuste märgendamist lihtsustada on luua selline praktiline märgendussüsteem, mis rasked (nn lootusetud) juhtumid n-ö ära varjab, loobudes näiteks partitsiibi lahutamisest verbiks ja omadussõnaks sõnade nagu *jooksev* või *haukuv* puhul. Meie sellisele lihtsustusele siiski ei läinud, sest ei olnud ette teada, kui palju märgendamisraskusi on tingitud märgendajate ebapädevusest, kui palju eesti keele morfoloogiakirjelduste ebamäärasusest ja kui palju tõelisest mitteühestatavusest. See ei tähenda muidugi, et me põhimõtteliselt sellise lihtsustamise mõttekust teatud juhtudel eitaksime.

3. Luua eeldused korpuste edasiseks märgendamiseks.

Et korpust edukalt märgendada, peab olema kasutada väga hea juhend. Korrektse korpuse ja selgeks vaieldud keeruliste juhtumite põhjal saab koostada detailsema eeskirja raskemate ühestamisjuhtumite jaoks, millest edaspidises töös on kindlasti kasu.

On ette näha, et mõnikord on korrektse märgendi valimise põhjendus nii keeruline, et see ei sobi teistele märgendajatele mõeldud juhendisse: liiga pikki juhendeid lihtsalt ei loetaks. Sel juhul on varem märgendatud korpus selleks näidete kogumiks, kust märgendaja saab juhise oma käesoleva probleemi lahendamiseks. On selge, et selleks peab näidete hulk olema absoluutselt korrektne.

Töö käik

Olukord enne "1984" morfoloogilist märgendamist

Enne "1984" morfoloogilist märgendamist oli olukord järgmine:

1. Olid olemas grammatikakäsitlused, mis sisaldasid erinevaid morfoloogiliste kategooriate süsteeme koos definitsioonide ja näidetega, missugused sõnad mingisse kategooriasse kuuluvad, näiteks Valgma-Remmeli grammatika[4] või eesti keele teaduslik grammatika.[5]
2. Eksisteerisid sõnastikud, mis samuti annavad sõnade kohta teada, millisesse morfoloogilisse kategooriasse need kuuluvad; mõnikord erinevalt, nagu Ü. Viksi vormisõnastik[6] või kirjakeele seletussõnaraamat.[7]
3. Eksisteeris morfoloogia-analüsaator ESTMORF[8], mis põhines ühel sõnastikul – Concise Morphological Dictionary (CMD)[9] ja selle morfoloogiasüsteemil. Seda morfoloogiasüsteemi oli detailiseeritud, et see vastaks Multext-Easti[10] morfoloogiliste

[4] J. Valgma, N. Remmel, Eesti keele grammatika. Valgus, Tallinn, 1970.
[5] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn, 1995.
[6] Ü. Viks, Väike vormisõnastik I. Sissejuhatus ja grammatika; II. Sõnastik ja lisad, Tallinn, 1992.
[7] Eesti kirjakeele seletussõnaraamat. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut, Tallinn, 1988 – 1998.
[8] H.-J. Kaalep, An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. – Computers and the Humanities 31, 1997, pp. 115-133.
[9] Ü. Viks, Väike vormisõnastik I. Sissejuhatus ja grammatika; II. Sõnastik ja lisad, Tallinn, 1992.
[10] L. Dimitrova, T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic, D. Tufis, Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. COLING-ACL'98. Proceedings of the Conference, Montreal, 1998, Vol 1, pp. 315-319.

kategooriate süsteemile. Näiteks CMD adpositsioonid olid jagatud pre- ja postpositsioonideks. Morfoloogiliste kategooriate lubatud kombinatsioone oli 788. ESTMORF andis umbes 46% sõnade puhul mitu analüüsi, ehkki valis liitsõnade ja tuletiste puhul (mis moodustavad 20% tekstist) kõige tõenäolisema sõnaosadeks jaotamise variandi, mis potentsiaalset mitmesust oluliselt vähendas.
4. Eksisteeris ConstraintGrammar-ühestaja [11], mille töö tulemusena jäi mitmeseks 16% sõnadest, kusjuures vigu tehti alla 1%.

Esimene ühestamine

1997.a. analüüsiti "1984" (v.a. Lisa) ESTMORFiga morfoloogiliselt ja ühestati CG-ühestajaga[12]. Järelejäänud mitmesed analüüsid ühestati käsitsi.

Et hinnata inimeste usaldusväärsust, tehti katse, milles kaks filoloogi ühestasid sõltumatult 2000-
Seejuures CG-ühestajat töö lihtsustamiseks ei kasutatud. Selgus, et umbes 3,5% juhtudest olid filoloogid eri meelt, kusjuures enamik erimeelsustest oli tingitud teoreetiliste käsitluste erinevusest; näpuvigade osakaal oli väga väike. See inimeste tehtud märgenduse kattumatus osutus ligikaudu niisama suureks, kui on kirjeldatud TREEBANKi projekti puhul. [13]

Teine ühestamine

1998. aastal märgendati "1984" teist korda, läbides seejuures neli etappi:
1. Otsustati mõnele sõnale lisada täpsem märgendus, et olla morfoloogilises kirjelduses täpsem, näiteks käskiva kõneviisi verbivorm *loe*, mida kasutatakse koos eitussõnaga *ei*, moodustab kindla kõneviisi oleviku eituse, näit. *ei loe*. Algselt oli sõnale *loe* omistatud ainult imperatiivi vorm nii jaatavas kui ka eitavas kontekstis. Nüüd lisati ka kindla kõneviisi oleviku eituse vorm. *Loe* on nüüd jaatavas kontekstis märgendatud kui imperatiivi vorm ning *ei loe* kontekstis kui kindla kõneviisi eitus. Selline möödapääsmatu lisaeristus tegi ühestamise muidugi raskemaks.
2. Ühestati kogu "1984" uuesti, kasutades seekord nn uut märgendust ja kasutamata CG-ühestajat. Seejuures jälgiti, et inimene, kes oli esimesel korral juba mingit osa ühestanud, ei saaks teisel korral sama osa. Kõik rasked juhtumid, trükivead esialgses tekstis ning morfoloogiaanalüsaatori vead märgiti üles ning parandati.
3. Leiti erinevused (v.a uuest märgendusskeemist tingitud erinevused) esimese ja teise märgenduse vahel. Need erinevused vaatas üle teistkordse märgenduse tegija, parandas vead ja märkis üles veel mõned raskusi tekitanud sõnad.
4. Kõik filoloogid, kes teise etapi märgendamisel osalesid, arutasid koos läbi oma raskusi tekitanud sõnad ja valisid metoodika nende märgendamiseks.

Märgendamisel kasutati tavalisi tekstiredaktoreid: õige variandi valimiseks tehti selle juurde märge, õige variandi puudumisel see lisati. Lisaks kasutati ka *ad-hoc* UNIXi skripte märgenduse formaalse korrektsuse kontrollimiseks. Ka eri variantide

[11] T. Puolakainen, Eesti keele kitsenduste grammatika morfoloogiline ühestaja – Keel ja Kirjandus, 1998, nr. 1, lk. 37-46.
[12] Vt ka T. Puolakainen, Eesti keele kitsenduste grammatika morfoloogiline ühestaja. – Keel ja Kirjandus 1998, nr. 1, lk. 37-46.
[13] M. Marcus, B. Santorini, D. Magerman, First steps towards an annotated database of American Ebglish. – Readings for Tagging Linguistic Information in a Text Corpus. Ed by Langendoen and Marcus, Tutorial for the 28th Annual Meeting of the AC, 1990.

võrdlemiseks kasutati UNIXi käske, samuti nagu probleemsete juhtumite puhul kontekstide otsimiseks ja kõrvutamiseks.

Märgenduse lõpliku korrektsuse kontrollimiseks kasutati ka ISSCO ühestajat (_____): pärast ühestatud teksti teisendamist ISSCO ühestaja jaoks vajalikule kujule prooviti ühestajat treenida. Kui sisend ei olnud korrektne, siis ühestaja osutas veale.

Viimasel etapil vajasid muutmist ainult üksikute sõnade või sõnaklasside märgendid. Enamasti olid need sõnad, mille sõnaliik sõltus seniste määrangute alusel tugevasti lause tähendusest. Teise grupi moodustasid sõnad, mille senised märgendid ei vastanud tegelikkusele, näiteks asesõnadel *kes* ja *mis* puudus mitmuse märgend.

Kuigi muudatused haarasid üsna väikest osa korpusest, võttis ühtlustamine oodatust kauem aega, sest lahendust ootasid just kõige keerulisemad probleemid. Näiteks sõnade *kätte*, *käes, käest* sõnaliiki otsustati korduvalt ümber. Esialgses korpuses oli iga lingvist seda määranud oma keeletaju alusel, kuid erinevate inimeste "sisetunne" ei pruugi alati ühte langeda.

Lõplikuks ühtlustamiseks tulid lingvistid kokku ja arutasid, millised võiksid olla probleemsed sõnad ja sõnaklassid, seejärel tehti ühestatud korpusest nende sõnade kohta väljavõte ja vaadati, kuidas need on analüüsitud ning kas nende märgendamise süsteem vajab parandamist või ühtlustamist. Kui parandused olid selgelt vajalikud, sõnastati uued täpsed reeglid ja iga lingvist tegi parandused enda ühestatud korpuse osas.

Lingvistilised probleemid

Eesti keele sõnaliikide ja -vormide määratlemise paljud probleemid kanduvad paratamatult ka automaatanalüüsi valda, olles seda teravamad, et morfoloogia automaatse analüüsi tarbeks peaksid kõik kategooriad olema kuni üksikliikmeteni võimalikult üheselt määratletavad, kuna ühestades tuleb alati valida üks ja kõige õigem võimalus. Tihti teeb aga otsustamise keerukaks n-ö üleminekualade olemasolu aktsepteerimine grammatikatraditsioonis ehk see, et sõnaliikide piirid pole üheselt määratletavad - üks osa sõnu on üleminekustaadiumis kahe sõnaliigi vahel või kasutatavad teise sõnaliigi funktsioonis. See protsess on teoreetiliselt hõlpsasti kirjeldatav tsentri-perifeeria põhimõttel, mis tähendab, et igal sõnaliigil on selle tüüpilisi omadusi esindav tsenter ja ebatüüpilist kasutust esindav hajus perifeeria, millele on oma uurimustes viidanud ka M. Erelt[14], V. Hallap[15]jt. Automaatse morfoloogilise analüüsi tarbeks oleks ideaalne, kui sellist hajuvust ei esineks. See pole aga sõnaliigikategooria semantikaga seotust arvestades ilmselt saavutatav ja nii jääbki ühe osa sõnade puhul probleemseks ja kokkuleppeliseks nende määramine ühte või teise sõnaliiki, kuna määratlus sõltub ka sõna funktsioonist konkreetses lauseümbruses. Seetõttu on oluline ka süntaktiline kriteerium – sõnaliigid peaksid olema kirjeldatavad nii, et nendele toetudes saaks võimalikult otstarbekalt kirjeldada ka süntaksit. Teatud üksikjuhtudel tuleb aga paratamatult arvestada subjektiivse keeletaju ja grupisiseste kokkulepetega.

Morfoloogiline analüsaator lähtub traditsioonilisest 9 sõnaliigist (substantiiv, adjektiiv, pronoomen ja numeraal kui käändsõnad, verb kui pöördsõna ning adverb, pre- ja postpositsioon, konjunktsioon ning interjektsioon kui muutumatud sõnad).

[14] M. Erelt, Ebamäärasusest sõnade liigitamisel. – Keel ja Kirjandus1977, nr. 9, lk. 525-528.
[15] V. Hallap, Sõnaliikide piirimailt. – Keel ja Kirjandus 1984, nr. 1, lk. 30-40.

Arvesse pole võetud eesti keele teadusliku grammatika jaotust 12ks sõnaliigiks, kus adverbi alt on iseseisvate sõnaliikidena välja toodud modaal-, afiksaal- ja proadverbid[16].

Morfoloogilisel ühestamisel tekkinud praktilised probleemid puudutavad praegusel juhul eeskätt piirialasid substantiivide ja adpositsioonide, adjektiivide ja verbide (kesksõnavormid) ning adverbide ja konjunktsioonide vahel, samuti pronoomenite semantilis-funktsionaalsete alaliikide vahel vahetegemist.

Adpositsioonid

Praktilises ühestamistöös kujunes probleemiks vastuolu arusaamaga, et adpositsioonide hulk peaks olema piiratud ja üheselt etteantav. Tegeliku töö käigus selgus, et senistest grammatikakäsitlustest ja sõnastikest koondatud adpositsioonide loendist ei piisanud teksti täpseks märgendamiseks, misjärel tuli adpositsioonide klassi töö käigus laiendada.

Problemaatiline on kaassõnade puhul nende hulga põhimõtteline avatus. On olemas tüüpilisi, väljakujunenud kaassõnu, mis kuuluvad lauses substantiivifraasi juurde, andes sellele käändelõppudega sarnaseid tähendusi. Kuna adpositsioonid nagu adverbidki on enamjaolt tekkinud noomeni teatud käändevormi grammatikaliseerumise tulemusel, mille taustaks on tugev tähendusnihe abstraktsuse suunas, siis on keeles pidevalt olemas nn üleminekustaadiumis sõnu. Teatud kontekstis võivad need esineda veel täistähenduslike nimisõnadena, teisel juhul on nad aga oma algsest leksikaalsest tähendusest kaugenenud ja kinnistunud teatud kasutuses muutumatuteks e suhtesõnadeks. Et see protsess on ka tänapäeva keeles pidev, näitab nii mõnegi tekstisõna määratlemise raskus. Praegusel juhul loeti sõna kaassõnaks siis, kui see esines koos genitiivis või partitiivis noomeniga. Ülejäänud juhtudel kaaluti vabade laiendite paigutamise võimalust noomeni ja võimaliku adpositsiooni vahele ning tugineti teadmisele, et postpositsioonil ei saa olla adjektiivatribuuti. Samuti vaeti sõnade semantilis-süntaktilise seose tugevust, tuginedes seisukohale, et nimisõna ja adverbi seos on hoopis lõdvem kui kaassõna ja nimisõna seos, nagu on rõhutanud ka H. Rätsep[17].

Näiteks tekitas küsitavusi sõnarühm **kätte, käes, käest**, kus tuli arvesse võtta ennekõike sõnadevahelisi süntaktilis-semantilisi seoseid ning kaaluda tähenduse ülekandelisuse astet. Näidetes *tal olid kindad **käes***; *riik käis **käest kätte***; *võttis Winston lusika **kätte*** on *kätte, käes, käest* iseseisva tähendusega nimisõnad; fraasides *ega karju valu **käes***; *on jäänud väikese privilegeeritud kasti **kätte*** on aga nii sisu kui ka vormi põhjal tegemist juba kaassõnadega. Mõnevõrra sarnane oli probleem ka sõnadega **pähe, peas, peast** ning **koju**, **kodus, kodust**, ent kuna nende puhul oli seos otsese tähendusega tihedam, otsustati need lugeda substantiivideks, näit. *tundis end masinate seas **kodus*** või *visake see mõte **peast** välja*. Probleem kerkis ka sõnadega **kombel, moel, viisil**, mis võivad teatud kontekstis esineda nimisõna käändevormidena, teisel juhul aga juba kaassõnadena, kuna sisuline seos nimisõna leksikaalse tähendusega on kadunud või tuhmumas. Praegusel juhul lahendati küsimus nii, et koos genitiivis noomeniga märgiti need sõnad adpositsioonideks, näit. *teenri **kombel***, muude käänetega (põhiliselt adessiiviga) seoses aga substantiivideks, näit. *mingil **kombel***, *keerukamal **moel**, müstilisel **viisil***. Praktilisi probleeme tekitas ka

[16] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn, 1995.

[17] H. Rätsep, Eesti keele lihtlausete tüübid. Tallinn, Valgus, 1978.

sõnarühma **poole, pool, poolt** täpne määratlemine. Tüüpiliselt esinevad *poole*, *pool, poolt* postpositsioonidena koos genitiivis noomeniga, ent noomen võib olla ka kohakäändes, näit. *igale **poole***, *teisel **pool**, kahelt **poolt** tänavat*. Lisaks sellele võib *pool* esineda numeraalina, näit. ***pool** miljonit vangi*, ***pool** sigaretti* ning harva ka substantiivina, tähistades üht partnerit mingis toimingus, näit. *ükskõik, kumb **pool** võidab*, või kaheks jaotatava eseme üht osa, näit. *lehekülje ülemine **pool**, see oli nagu parooli kaks **poolt***. Seega on taolise sisuliselt mitmeti tõlgendatava sõna automaatanalüüs komplitseeritud.

Kuna algselt olemasolev kaassõnade hulk ei rahuldanud täpse morfoloogilise märgendamise vajadusi, tuli töö käigus kaassõnade loendisse sõnu lisada. Nii lisati kaassõnade hulka **hoolimata** ja **vaatamata** kui prepositsioonid, näit. ***hoolimata** tema heidutavast välimusest, **vaatamata** lõpututele arreteerimistele*. Postpositsioonidena olid need sõnad loendis olemas. Lisati ka **seoses** nii pre- kui postpositsioonina (näit. *et niisugust sõna saab temaga **seoses** kasutada*), **vastavalt** prepositsioonina (näit. ***vastavalt** vajadu*sele) ning kvantoriga seoses esinev postpositsioon **tagasi** (näit. *seitse aastat **tagasi***). Kaassõnade hulka tuli lisada ka **suunas** postpositsioonina (näit. *tema prillid välgatasid vaenulikult Winstoni **suunas***), **nimel** postpositsioonina (näit. *manitsesid teda ingsotsi ja Suure Venna **nimel***), **korras** postpositsioonina, näit *erandi **korras***, ja **kaugusel** postpositsioonina (näit. *on meist miljoni valgusaasta **kaugusel***), mis algselt kaassõnade loendist puudusid. Adpositsioonideks otsustati lugeda ka mõned *poole*, *pool, poolt*-põhisõnaga liitsõnad, mis esinevad koos partitiivis noomeniga, näit. *tuleks panna **sissepoole** sulgusid*, ***seespool** koljut*, ***väljaspool** ajalugu, **teispool** seina*.

Seega tuleb tõdeda, et adpositsioonid on eesti keeles tekstisidus avatud sõnaliik, kuhu võib uusi liikmeid lisanduda.

Pronoomenid

Konkreetses tekstis tekitasid täpsel määratlemisel probleeme ka mitmed eesti keele pronoomenid. Asesõnade semantilis-funktsionaalsete rühmade osas on morfoloogilise analüsaatori pakutavad võimalused väga detailsed, kuna arvesse on võetud kõik võimalikud alaliigid. See muutis otsustamise lausekonteksti arvestades sageli raskeks, sest mida täpsem on semantiline jaotus, seda enam kaalu on sisulistel otsustustel, mis võib suurendada otsuste subjektiivsust. Teisalt on täpsusetaotlus muidugi ka analüsaatori tugev külg.

Enim praktilise määratlemise probleeme tekitasid pronoomenid, mis kuuluvad küll tüüpilises kasutuses kindlasse semantilisse alaliiki, ent võivad perifeerses kasutuses saada erinevaid sisulisi varjundeid ja kuuluda sellest lähtuvalt ka pronoomenite eri liikidesse või isegi muudesse sõnaliikidesse. Kõige rohkem küsitavusi tekitas pronoomen **oma**, mis esineb tüüpiliselt possessiivpronoomenina, viidates sama lause elus subjektile ja väljendades sisulist kuulumist tegevuse subjektile, näit. *see jätab igaveseks **oma** jälje*, *Suure Venna **oma***. Samas võib *oma* esineda verbi käänduva laiendina refleksiivpronoomeni funktsioonis, olles sisuliselt asendatav pronoomenitega *enese* ja *enda*, näit *ta võttis juhtimise **oma** kätte* või *keskkiht, kes võidab **oma** poole ka alamkihi*. *Oma* võib esineda ka noomeni käänduva laiendina determinatiivpronoomeni funktsioonis, olles sel juhul sisuliselt asendatav pronoomeniga *ise*, näit. *ka tema **oma** mees*. Lisaks pronoomenile võib aga *oma* olla sekundaarselt kasutusel adjektiivina tähenduses ‘tuttav, omane; iseloomulik; isiklik’, näit. ühendites ***omal** kombel, **omal** ajal, **omal** moel, **omal** jõul, **omal** ajal, **omaks** võtma*. Viimati nimetatud ühendi puhul võiks sisuliselt kõne alla tulla ka substantiivne

tõlgendus, nagu on arvanud Ü. Viks[18], ent praegusel juhul on kõikides sellistes ühendites jäädud siiski adjektiivse määratluse juurde. Teatud kontekstis võib *oma* olla kasutusel ka substantiivina tähenduses 'lähedased (inimesed)', näit. *kui sellega saavad hakkama* ***omad***, *ja mitte vaenlased* või *oli* ***omadega*** *läbi*. Mõnes kontekstuaalses ümbruses võib *oma* esineda isegi adverbina, sõna funktsiooniks on sellisel juhul ligikaudse hulga väljendamine, näit. ***oma*** *tosin korda,* ***oma*** *kolmkümmend teenijat,* ***oma*** *sada aastat vana*.

Teine praktilise määramise seisukohalt keerukas pronoomen on **ise-enese** ja käändevormides eelmisega osaliselt kattuv **enese-enda**. Esimene neist esineb tüüpkasutuses determinatiiv- ja teine refleksiivpronoomenina. Määratlev *ise* esineb koos nimi- ja asesõnadega rõhutavas, võimendavas tähenduses, näit. *õigem oli end* ***ise*** *tappa*, *kõik sõltub teist* ***endast****, meie* ***enda*** *hüvanguks*, *ainult tema* ***enda*** *värisemine*.

Pronoomen *enese/enda* on üldjuhul refleksiivpronoomen, laiendades verbi ja viidates tegevuse subjektile, näit. *ta ei saanud* ***endast*** *võitu*; *mida nad* ***endast*** *kujutavad*. Sama pronoomen võib lauses harvemini täita ka possessiivset funktsiooni, esinedes genitiivatribuudina pronoomeni *oma* asemel, näit. *me toome ta* ***enda*** *leeri*, *partei ei hoia võimu* ***enda*** *käes*. Sellised vahetegemised põhinevad peaasjalikult pronoomenite sisu analüüsil ja on seetõttu raskesti formaliseeritavad.

Probleemseteks osutusid ka sõnad **üks** ja **teine**, mis võivad sõltuvalt kontekstist olla kas numeraalid, väljendades sellisel juhul konkreetset, arvuliselt iseloomustatavat hulka või abstraktset arvu, näit. *tuba* ***üks*** *null* ***üks***. Samas võivad need aga esineda ka indefiniitpronoomenitena tähenduses 'mingi, muu', näit. ***üks teine*** *mürsk kukkus tühermaale*. Teatud kontekstis võib *üks* esineda tähenduses 'sama', olles sellisel juhul kasutusel demonstratiivpronoomenitena, näit. *oli ta* ***ühel*** *meelel, kõik* ***ühe*** *hooga, oli temaga* ***ühes*** *leeris*. Determinatiivne kasutus tuleb kõne alla näitelauses *Vehib* ***teine*** *isegi lõuna ajal tööd teha*.

Samas on numeraalse ja pronominaalse kasutuse vahel paiguti väga raske vahet teha, näit. fraasis *seal oli* ***üks*** *foto* võib sõna *üks* tõlgendada nii arvulisena, st et fotosid oli üks, ent teisalt on mõeldav ka tähendus 'mingi'. Ühestamisel püüti sellised juhtumid lahendada nii, et kui arvuline väljendus oli mõeldav, otsustati numeraali kasuks, ent subjektiivse tõlgendamise moment võis siin teatud juhtudel jääda oma rolli mängima.

Üldjuhul kujunes pronoomenite ühestamisel probleemiks sisuliste alaliikide vahel vahetegemine, ent üksikute asesõnade puhul kerkis probleeme ka vormi kirjeldamisega. Sellised olid interrogatiiv-relatiivpronoomenid *kes* ja *mis* oma käändevormidega. Algselt püüti interrogatiivset ja relatiivset kasutust indoeuroopa keelte eeskujul lahutada, ent kuna see vahetegemine osutus liiga tihedalt konteksti semantikaga seotuks ja seetõttu objektiivselt raskesti määratletavaks, otsustati sellest hiljem loobuda ja jääda traditsioonilise interrogatiiv-relatiivpronoomeni määratluse juurde. Probleemiks oli aga nende lauseosi siduvate pronoomenite ainsuse ja mitmuse vormide üle otsustamine. Küsimus on selles, et kuigi vormiliselt on neil asesõnadel olemas mitmuse paradigma, ei kasutata seda tegelikkuses peaaegu üldse. Siit ka dilemma: kas arvestada määramisel pelgalt vormi ja määrata mitmusliku korrelaadiga sõnad vormi põhjal ainsuslikeks või lähtuda lause sisust ja otsustada mitmusliku korrelaadiga sõnade puhul mitmuse kasuks. Praegu valiti märgendamisel viimane võimalus, mis tundus sisuliselt õigem.

[18] Ü. Viks, Sõnast *oma* eesti keeles. – Keel ja struktuur 1972, nr. 6, lk. 131.

Adverbid ja konjunktsioonid

Vahel tekkis raskusi sõnade *aga, nagu, kui* sõnaliigi määramisel. Enamasti on need sidesõnad, harvem määrsõnad. Lahkarvamusi tekitas ennekõike *aga* lause või osalause keskel. Alati polnud päris selge, kas selle funktsiooniks on vastandamine (sidesõna), näit. *Hetke pärast andis kõrvetus maos **aga** järele* või rõhutamine (määrsõna), näit. *Kollektiivselt omab Partei **aga** kõike Okeaanias*.

Teine probleemne koht oli *aga* iseseisva lause algul. Vahel on vastandus nõrk ning sidesõna *aga* kasutatakse sissejuhatava sõnana, ent samas funktsioonis võib kasutada ka määrsõna *aga*, näit. ***Aga** nüüd on vist aeg minema hakata* (konjunktsioon). ***Aga** ma ehmatasin teda igatahes korralikult* (määrsõna). Sellistel juhtudel püüti vahet teha sõnade semantilise funktsiooni alusel.

Nagu on määrsõna, kui ta väljendab või rõhutab mitmesuguseid modaalsusnüansse, erinevaid suhtumisi jms. Selle äratundmine ei olnud üldiselt raske, vaid üksikud juhud nõudsid ühist arutelu, näit. määrsõnaline kasutus lauses *Ja miski tema hääletoonis näis sellele **nagu** lisavat "see va juhmakas"*.

Sõnaliigi määramisel tekkis vigu, kui määrsõna **kui** alustas kõrvallauseid. Ilmselt ei jälgitud esialgse ühestamise käigus alati hoolikalt, et ka tavapärases sidesõna positsioonis võib *kui* väljendada määra, astet, ulatust ning olla seega määrsõna, näit. *Nüüd alles jõudis tema teadvusse, **kui** suure asja ta on ette võtnud.*

Verbiga seotud ühestamisprobleemid

Eesti verbivormistik on suhteliselt keerukas. Pöördsõna vormid võivad olla finiitsed või infiniitsed, liht- või liitvormid. Finiitsetel sõnavormidel on viis morfoloogilist kategooriat: isik või pööre, tegumood, aeg, kõne ja kõneviis.

Eesti keele teaduslik grammatika (edaspidi EKG) on lisanud traditsioonilisele neljale kõneviisile (indikatiiv, konditsionaal, imperatiiv ja kvotatiiv) viienda - jussiivi. Jussiiv väljendab kaudset (vahendatud) tegevusele õhutamist või käsku, mis a) on suunatud kõnelejalt kõnesituatsioonis mitteosalevale isikule või b) pärineb kõnesituatsioonis mitteosalevalt isikult[19]. "1984" ühestamisel käsitleti imperatiivi ja jussiivi siiski ühe kõneviisi – imperatiivina. Kvotatiivi vormidena on praegu märgendatud ka *vat*-infinitiivi vormid (näit. *mis öeldi **olevat** arve*, *ta paistis **rõõmustavat***). Teadusliku grammatika järgi on *vat*-vormi käsitlemine iseseisva infiniitvormina tingitud tema süntaktilistest omadustest. Vormimoodustuslikult kattub *vat*-infinitiiv täielikult kvotatiivi vormidega[20]. Siiski võiks kogu 75 000-sõnalises tekstis kvotatiivi märgenduse saanud 103 verbivormist 101 märgendada ka *vat*-infinitiivina, so nad ei väljenda sisuliselt mitte vahendatud teadet, vaid on kasutusel kas seoses verbidega *näima, paistma, tunduma* või *et*-objektlause asendajana (*teda kuuldi **ütlevat**, mind arvati **kujutavat***).

[19] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn, 1995, lk. 83.

[20] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn, 1995, lk. 65.

Verbide ja adjektiivide vahekorrast

Siin tekitas probleeme eeskätt mineviku (*-nud* ja *-tud*) kesksõnade ning *mata*-vormi (supiini abessiivi) sõnaliigi üle otsustamine.

Oleviku partitsiip kuulub sõnaliigiliselt kokku adjektiividega ning eesti keele grammatikais käsitletakse seda verbivormina peamiselt vaid traditsiooni tõttu.[21] Oleviku partitsiip seega morfoloogilisel ühestamisel probleeme ei põhjustanud, see loeti alati adjektiiviks, näit. *aususega **kaasnev** eesmärgikindlus, teleekraani kaudu **edasiantav** korraldus*, kuigi morfoloogiline analüsaator annab *v*- ja *tav*-partitsiibile valida ka verbi analüüsi.

Mineviku partitsiibi vormidega on aga seotud mitmed probleemid. Öeldiseks võib mineviku partitsiip olla ainult finiitsete liitvormide komponendina koos verbi *olema* lihtvormiga ja/või eitussõnaga (*oli läinud, ei teinud*). Öeldise osana esinev partitsiip analüüsiti verbi vormiks. Mineviku partitsiip võib lauses olla ka predikatiivi, atribuudi ning seisundimääruse funktsioonis. Sellised partitsiibid märgendati morfoloogilisel analüüsil adjektiivideks. Kuid ka käsitsi ühestades oli raske vahet teha öeldise osaks ja öeldistäiteks olevate mineviku partitsiipide vahel (nt *me oleme **surnud**, see oli äsja ära **kadunud***). Vahet püüti siiski teha ning öeldistäite või täiendina esinevad mineviku partitsiibi vormid loeti adjektiivideks, näit. ***möödunud** aegade utopismist*, *hajameelne **rusutud** pilk*. Analoogiline probleem tekkis *mata*-vormiga, mis loeti täiendina omadussõnaks, näit. ***kirjutamata** seadus*, ent sageli oli raske vahet teha *mata*-vormil öeldistäitena (loeti adjektiiviks) ja öeldise osana (määrati verbiks). Näiteks lauses *osamaks on **maksmata*** on ilmselt tegemist verbiga, lauses *tohututele maa-aladele, mis tegelikult on **asustamata** ja läbi **uurimata*** aga adjektiividega.

Atribuudi funktsioonis olev mineviku partitsiip võib olla nii käändumatu kui ka kääduv omadussõna. Käändumatu omadussõna on mineviku partitsiip eestäiendina, kus ta ei ühildu oma põhisõnaga (nt ***kulunud** portfell*). Kääduv omadussõna on partitsiip järeltäiendina, sest siis ta ühildub fraasi põhjaga (nt. *sigaretid, väga tihedalt ja hästi **topitud**.*

Adjektiivide ja substantiivide vahekorrast

Sarnaselt adjektiiviga võib ka adjektiivselt talitlev partitsiip substantiveeruda ja adjektiivi ning substantiivi vahele jääb siingi üleminekuala. Nt tuleb erinevalt analüüsida lauseid *Peksaks tüdruku kumminuiaga **surnuks***, kus *surnuks* on adjektiiv ja ***Surnutest** olid saanud märtrid*, kus *surnutest* on substantiiv.

Morfoloogiline analüsaator ei anna nimisõna tõlgendust mitte kõigile mineviku partitsiipidele (kõik saavad küll nii kääduva kui ka käändumatu omadussõna tõlgenduse), vaid ainult neile, mis on nimisõnadena sõnastikku kantud.

Siit tekibki küsimus, kas on üldse mõtet morfoloogilise ühestamise käigus eristada verbi ja omadussõna (ja sealtkaudu nimisõna) funktsioonis esinevaid partitsiipe. Põhimõtteliselt võiks küll vähemalt morfoloogilisel analüüsil nimetada neid vorme lihtsalt partitsiipideks, nagu on tehtud näiteks inglise keele kitsenduste

[21] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn, 1995, lk. 67.

grammatikas[22], kuid see tähendaks probleemi "lükkamist" süntaksianalüsaatori lahendada.

Muudest kokkulepetest

Verbi eitavas kõnes liitvormide puhul kasutati algselt sellist märgendussüsteemi, kus kõik verbiahela liikmed said ka eituskategooria märgendi. Töö käigus aga otsustati, et märgendi *neg* saavad ainult need liitajavormi osad, mille vorm on otseselt tingitud eitusest, so abiverbid.

Öeldise infiniitseid komponente aga peab see märgendussüsteem eituse-jaatuse kategooria suhtes neutraalseteks ja neil puudub nii eitava kõne märgend *neg* kui ka jaatava kõne märgend *af*. Nt *ei* (ei+0 //_V_ aux neg //) *olnud* (ole+nud //_V_ aux indic impf ps neg//) *säilinud* (säili+nud //_V_ main partic past ps //).

mas-vorm ja *da*-infinitiiv võivad olla nii verbiahela osad kui esineda ka lauses adverbiaali funktsioonis. Kui need vormid on verbiahela osad, siis peaks nendega kokku kuuluv *olema*-verbi vorm saama abiverbi analüüsi (_V_aux), kuid seda, kas antud infiniitvorm, eriti *mas*-vorm, kuulub öeldisverbi koosseisu või mitte, saab otsustada põhiliselt siiski semantika põhjal. *mas*-vorm moodustab koos *olema*-verbiga kestvat protsessi või sündmuse eelfaasi väljendava perifrastilise verbivormi. Millal *mas*-vorm on finaal-lokaalse adverbiaalina talitleva sekundaartarindi peasõna, millal perifrastilise verbivormi osa, sõltub infiniitvormis verbi leksikaalsest tähendusest[23]. Selles osas on edaspidise automaatse ühestamise huvides märgendussüsteemi lihtsustatud ja *mas*-vormi ning *da*-infinitiiviga seostuv *olema*-verb on alati märgendatud põhiverbiks.

Konstruktsiooni *saama*+supiini illatiiv kasutatakse eesti keeles tuleviku väljendamiseks. (Näit. *Kirjutamine ise **saab olema** kerge*) "1984" morfoloogilisel analüüsil märgendati verb *saama* sellistes konstruktsioonides põhiverbiks, sest abiverbiks märgendamine tähendanuks tulevikukategooria toomist vormikirjeldusse. Tulevikukategooria loomata jätmist põhjendab ka asjaolu, et verbi *saama* kasutati tuleviku väljendamiseks (***saab** olema*) vaid kolmel juhul *saama*-verbi 380 esinemiskorrast kogu märgendatud tekstis.

Probleeme tekkis ka sõnadega, mille algvormi polnud tekstis esineva käände- või pöördevormi põhjal võimalik üheselt kindlaks määrata, st et kirjakeeles aktsepteeritakse mitut paralleelset algvormi varianti, mida ka analüsaatori sõnastik välja pakkus. Sellisel juhul valisime enam kasutatava sõna või õigekeelsussõnaraamatus antud õige vormi. Kokkuleppelised eelistused puudutasid järgmisi sõnu:
väike, väikene – valisime *väike*
päike, päikene – päike
poisikesepõlv, poisikesepõli – poisikesepõlv
neiu, neid – neiu
talv, tali – talv
manner, mander – manner
kaitsma, kaitsema – kaitsma.

[22] A. Voutilainen, J. Heikkilä, A. Anttila, Constraint Grammar of English. A Performance-Oriented Introduction. University of Helsinki, Departement of General Linguistics. Helsinki, 1992, s. 5.
[23] M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, Eesti keele grammatika II. Süntaks. Lisa: Kiri. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn, 1993, lk. 259.

Sõna *jah* märgenditeks pakkus morfoloogiline analüsaator algselt määrsõna ning hüüdsõna. Ühestades valisime sõnaliigiks alati määrsõna, ühtlustades sõnade *jah* ja *ei* märgendid, kuna *ei* ei olnud hüüdsõnana välja toodud.

Praeguses käsitsi ühestatud versioonis on püütud kõik eespool nimetatud küllalt keerukad süntaktilis-semantilised seigad sõnaliikide ja -vormide määratlemisel arvesse võtta. Kuivõrd põhjendatud selline vahetegemine on, peavad ilmselt otsustama analüsaatori kasutajad. Morfoloogiline märgendus on lõpptulemusena siiski küllalt detailne ja taotleb ennekõike sisulist adekvaatsust.

Kokkuvõtteks

On alust eeldada, et mistahes uue korpuse (uue kas keele või märgendussüsteemi poolest) märgendamisel on möödapääsmatu aeglane ebaefektiivne automatiseerimata, kuid kvaliteetne käsitsi märgendamine.

Artikkel kirjeldas G. Orwelli "1984" eestikeelse tõlke morfosüntaktilist märgendamist, mille puhul oli põhieesmärgiks kvaliteet. Seejuures püüti saavutada ka kolme alleesmärki:

1. Luua testmaterjal keeletehnoloogiliste programmide kontrollimiseks.
2. Kontrollida, missugune eesti keele morfoloogiasüsteemi kirjeldus vastab tegelikele tekstidele kõige paremini ja täiendada-kohandada kirjeldusi konkreetse teksti vajadustest lähtuvalt.
3. Luua eeldused edasiseks märgendamiseks.

Sama teksti ühestati erinevate inimeste poolt kaks korda. Filoloogidevahelised erimeelsused lahendati probleemide arutamise teel. Selgus, et kui ühestamisel piirduda ainult etteantud variantidest ühe valimisega, siis on töö umbes kaks korda kiirem kui juhul, kui esiteks kontrollitakse ka neid juhtumeid, kus morfoloogiaanalüsaator on andnud ainult ühe variandi ning teiseks ühtlustatakse eri filoloogide tehtud valikud.

Ilmnesid ka mitmed lingvistilised probleemid:

1. Vajadus valida mitmest alternatiivsest algvormist üks ja ainus (olukord, mida tavaliselt on tänu mitmetitõlgendatavuste aktsepteerimisele võimalik vältida).
2. Eesti keelele omased raskused substantiivide ja adpositsioonide ning adverbide ja konjunktsioonide eristamisel.
3.. Paljudele tüpoloogiliselt erinevatele keeltele omased raskused adjektiivide ja verbide partitsiipide eristamisel.