TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
Eesti ja soome-ugri keeleteaduse osakond

Helen Nigol

Voorusisesed parandused, kordused ja valestardid suulises eesti keeles: nende tuvastamine ja normaliseerimine

Magistritöö

Juhendaja M.Sc. Heli Uibo

Tartu 2006

Sisukord

Sissejuhatus

Süntaktiline analüüs on üks keele automaattöötluse olulisi ülesandeid. Eesti keele jaoks on Tartu Ülikoolis loodud kitsenduste grammatikal põhinev süntaksianalüsaator e parser (ESTKG parser), mis tuleb edukalt toime kirjaliku keele lausete analüüsimisega (Müürisep 2000). Käesolev töö annab panuse analüsaatori suulisele keelele kohandamiseks.

Süntaksianalüsaatorite arendamiseks vajatakse märgendatud korpusi. Oma bakalaureusetöös (Nigol 2004) käsitlesin süntaktiliselt analüüsitud korpusi e puude panku. Kirjaliku eesti keele puude panga loomist alustati 2002. aastal (Uibo, Nigol 2006). Seejärel sündis idee luua ka suulise eesti keele puude pank. Suuline keel on aga oma olemuselt teistsugune ja see toob kaasa uusi probleeme võrreldes kirjaliku keele analüüsiga. Nimelt on suuline keel lineaarne protsess, mille käigus jõuab kõneleja jutt kuulajani rohkem või vähem soravana. Võib välja tuua kaks suuremat probleemide hulka, millega peab arvestama kirjakeele süntaksianalüsaatori ümberkohandamisel suulisele keelele: 1) mittesoravused[1] (ingl *disfluencies*) ja 2) sõnaklassid (Gibbon jt 2000: 27). Käesolevas töös keskendutakse mittesoravuste analüüsiga seotud probleemidele.

Mittesoravusteks peetakse täidetud pause, partikleid ning selle töö uurimis-objektiks olevaid kordusi, parandusi ja valestarte. Mitte kõik suulise kõne süntaktilisel analüüsil ilmnevad probleemid ei ole reeglitega ega statistiliselt lahendatavad. Täidetud pause ja partikleid on automaatselt lihtsam analüüsida kui komplekssseid mittesoravusi, nagu seda on parandused, kordused ja valestardid. Üheks võimaluseks mittesoravustega seotud probleeme lahendada on eeltöötlemisetapil need märgendada, mille käigus tehakse ebagrammatilised lausungid süntaksianalüsaatori jaoks grammatilisteks, nö normaliseeritakse (ingl *normalization*). Eraldi analüüsietapina on mittesoravusi märgendatud näiteks suulise kõne korpuses Switchboard (Meteer jt 1995), Susanne allkorpuses Christine[2] ja ICE-GB (Meyer 2002: 96). Normaliseerimine põhineb eeldusel, et suulises keeles esinevad mittesoravused pole sugugi tavaline keeleline

[1] Artiklis Müürisep jt 2006 on kasutatud mõistet *mitteladusused*.
[2] http://www.grsampson.net/RChristine.html

müra, vaid vastupidi, mittesoravustel on kindel struktuur, mis aitab kuulajal kõneleja mittesoravast jutust aru saada. See struktuur on samuti rakendatav erinevates kõneanalüüsiga seotud programmides. Normaliseerimise puuduseks on see, et ta võtab kaua aega, mistõttu sel moel normaliseeritud keel pole online-rakendustes kasutatav. Küll aga annab normaliseerimine väga palju olulist infot, mida vajatakse enne automaatse süntaktilise analüüsi rakendamist: vaja on teada, mida otsida ja millises suunas süntaksianalüsaatorit täiustada.

Suulises inglise keeles esinevaid mittesoravusi on arvutilingvistilisest vaatepunktist juba palju uuritud. Teistest keeltest on käsitletud nt ka rootsi, taani, saksa ja jaapani keelt. Suulises eesti keeles on põhjalikult uuritud kõneakte ning teatud partiklite esinemist ja tähendust erinevates kontekstides. Samas tähenduses, kui parandamisi käesolevas töös käsitletakse, on neid vaadelnud konversatsioonianalüüsist lähtuvalt Tiit Hennoste. Krista Strandson on oma magistritöös (2002) samuti uurinud parandusi, kuid teise nurga alt, nimelt kuulaja algatatud eneseparandusi. Erinevate parandamisvõtete kasutust spordikommentaatorite kõnes uuris oma bakalaureusetöös Mare Viks (2001).

Esimene katse suulist eesti keelt ESTKG analüsaatoriga analüüsida tehti 2005. aastal. Suulisele keelele omaste konstruktsioonide nagu korduste, paranduste ja valestartide esinemist eraldi ei vaadatud. Rohkem huvituti, kuidas muudetud arvutigrammatika reeglid suulise keele peal töötavad. Käesoleva töö ühe praktilise osana on tähelepanu pööratud just sellele, kuidas süntaksianalüsaator eelpool mainitud konstruktsioonidega hakkama saab.

Käesolevas magistritöös analüüsitakse kõneleja voorusiseseid parandusi (ingl *self-initiated self-repair*), kordusi ja valestarte. Peatutakse ka täidetud pausidel ja partiklitel, aga seda sel määral, kui palju on nende tõlgendamist vaja korduste ja paranduste analüüsil. Uuritakse paranduste, korduste ja valestartide nn pindmist struktuuri (ingl *surface structure*; Shriberg 1994), st seda, kuidas nad on esitatud. Analüüsitavad dialoogid on võetud Tartu Ülikooli eesti dialoogikorpusest. Litereeringuid ei ole üle kontrollitud.

Praktiline töö on jaotatud kolmeks. Esimesena on läbi viidud paranduste, korduste ja valestartide märgendamine, mille jooksul töötati välja teatavad

märgendamise põhimõtted ja kasutatavad märgendid. Eeskuju võeti Switchboardi korpuse mittesoravuste märgendamisjuhendist. Teises osas analüüsitakse märgendatud lausungeid ning otsitakse parandustele, kordustele ja valestartidele viitavaid prosoodilisi ja leksikaalseid markereid. Praktiline osa lõpeb eksperimendiga, mille hüpoteesiks see, et analüsaator on normaliseeritud lausungite analüüsil edukam kui normaliseerimata lausungite puhul.

Käesolev uurimus on üks osa eesti keele süntaksianalüsaatori suulisele eesti keelele kohandamisest. Töö käigus märgendatud korpus on heaks testkorpuseks, mille alusel saab hinnata analüsaatori töö paranemist mittesoravusi sisaldavate lausungite analüüsil.

Käesolev magistritöö koosneb viiest peatükist.

Esimeses peatükis antakse ülevaade mittesoravustest ja nende käsitlemisel kasutatava terminoloogia mitmekesisusest. Tutvustatakse erinevaid mittesoravuste liigitusi. Samuti antakse laiem taust erinevaist mittesoravuste uurimustest, nii psühholingvistilisest kui ka arvutilingvistilisest vaatenurgast, st vaadeldakse uurimusi selle kohta, kuidas töötleb mittesoravusi inimesest kuulaja ja kuidas teeb seda arvuti. Käesoleva töö huviorbiidis on just see viimane. Peatüki lõpetab ülevaade mittesoravuste märgendamisest. Toodud inglisekeelsed näited on tõlgitud seal, kus see on sisuliselt vajalik olnud, ja muudel juhtudel jäetud tõlkimata.

Teises peatükis tutvustatakse analüüsi aluseks olevat korpust, analüüsiüksuseid ja märgendamist.

Kolmanda peatüki alguses tutvustatakse empiirilise materjali liigitamist, mis põhineb parandamiste struktuuril, s.t oleneb sellest, kas midagi asendatakse, lisatakse, korratakse või loobutakse millestki. Sellest tulenevalt vaadatakse eraldi parandusi, kordusi ja valestarte.

Neljandas peatükis tutvustatakse läbiviidud automaatse süntaksianalüüsi eksperimenti ja selle tulemusi.

Viiendas peatükis esitatakse kokkuvõte ja mõned edasised uurimissuunad.

Töö lõpus on lisadena toodud transkriptsioonimärkide loend (lisa 1), märgendatud parandused (lisa 2), kordused (lisa 3) ja valestardid (lisa 4) ning kõige lõpus morfoloogiliste ja süntaktiliste märgendite loend (lisa 5).

1. Mittesoravuste uurimine

Järgnev peatükk annab sissejuhatava ülevaate mittesoravustega seotud temaatikast. Tutvustatakse terminoloogiat ja antakse ülevaade mittesoravuste erinevatest käsitlustest. Kuna seda temaatikat käsitlevaid uurimusi on palju, siis põhirõhk on pandud arvutilingvistiliste lähenemiste kajastamisele, mis on olulised just selle magistritöö seisukohast.

1.1. Mittesoravused ja nende struktuur

Üldiselt defineeritakse mittesoravusi kui fenomene, mis pidurdavad kõne vaba voolavust ja mis ei lisa öeldusse propositsionaalset sisu. Kõnes tekkivad mittesoravused annavad märku kõne planeerimisega seotud probleemidest (Shegloff jt 1977, Levelt 1983, 1989).

Mittesoravustena käsitletakse täidetud pause, partikleid, venitamisi, poolikuks jäänud lausungeid, kordusi, parandusi, lausungisulameid e anakoluute. Täidetud pausid, pikemalt hääldatud silbid ja kordused on suulises kõnes kõige tavalisemad kõne tootmisel kasutatavad vahendid võitmaks aega mõtlemisel, mida öelda. Lausungi süntaksis ja semantikas ei juhtu midagi, lausung jõuab kuulajani lihtsalt väikeste takerdustega. Struktuurilt on kõige keerukamad parandused. Kõneleja alustab oma kõnevooru, katkestab siis selle ning kas asendab midagi juba öeldut lausungis, lisab sinna hoopis midagi või kustutab sootuks terve lausutud fraasi ja alustab uut.

Levelt (1989: 484-499) on välja toonud neli aspekti, kuidas kõneleja aitab kuulajal soovitud informatsiooni kätte saada. Esiteks, keelelisest küljest vaadatuna on parandusel eelneva lausungiga süntaktiline seos, mis teeb võimalikuks, et kuulaja saab algse ja parandatud lausungi peale kokku kätte siiski kõneleja soovitud interpretatsiooni. Teiseks, kõneleja võtab suuremal või vähemal määral algsest lausungist midagi selle parandatud varianti kaasa, mis annab kuulajale teada, kas paranduse põhjuseks oli viga või ebatäpsus. Kolmandaks, paranduse esimene sõna on kuulajale tavaliselt piisav selleks, et teada, kuhu parandus algses lausungis paigutada. Neljandaks, kõneleja esitab parandatud osa alati teatud rõhuga, nii et kuulaja saab aru, et tegu on parandamisega.

Parandustel on teatud standardne vorm. Levelt (1983) on paranduste struktuuri jaganud järgmiselt. Clark (1996: 258) on nimetanud seda ka katkestuse mudeliks (ingl *disruption schema*). Analüüsime samaaegselt ka näitelauset: et `sööke nende hin- + `selle hinna sees ei=`ole.

- **parandatav osa** (ingl *reparandum*)

See osa lausungist, mida kõneleja tahab parandada. Näide lausungist kuni katkestuseni:

et `sööke nende hin-

- **katkestuskoht** (ingl *interruption point, hiatus, interregnum*)

Katkestusosa märgib parandatava üksuse lõppu ja parandamisosa algust. Antud näite puhul viitab katkestusosale sõnakatke hin-. Teiste paranduste korral ei ole katkestuskoha tuvastamine nii ilmne, seega paranduste märgendamise käigus markeeritakse see koht paranduses alati eksplitsiitselt, nt plussmärgiga (vt ka ptk 1.2.3.).

et `sööke nende hin- +

- **parandamisosa algus** (ingl *editing term*)

Toimetamisfraasi algus. Kõneleja võib mitmete leksikaalsete ja prosoodiliste vahenditega kuulajale märku anda, et nüüd järgneb parandus. Leksikaalseteks vahenditeks võivad olla partiklid, täidetud pausid, teatavad toimetamisfraasid ja prosoodilisteks nt pausid ja rõhud. Meie praeguse näite puhul alustab kõneleja eneseparandamist kohe, ilma leksikaalse märguandeta, kuid näeme, et kõneleja on paranduse esimest sõna rõhutanud, mis on kõneleja üks võte kuulajale märku anda parandatavast osast, nagu Levelt seda väga täpselt neljanda aspektina on esitanud.

et `sööke nende hin- + `selle hinna

- **parandatud osa** (ingl *alteration, reparans*)

Kõneleja teeb paranduse ja lõpetab lausungi.

et `sööke nende hin- + `selle hinna sees ei=`ole.

Parandatud osa saab käsitleda kolmest vaatepunktist (Hennoste 2000: 2702):

- mida reformuleeritakse, kas nt grammatikat või sõnavara;
- kuidas reformuleeritakse, kas nt asendatakse või lisatakse;

- milline on tagasimineku ala, mis näitab seda osa, mida reformuleerimisel korratakse.

Vaatame nende kolme punkti taustal meie näidet: et `sööke nende hin- + `selle hinna sees ei=`ole. Kõneleja parandab grammatikat, pluuralist saab singular. Tegemist on asendamisega, kusjuures näeme, et kõneleja on poole sõna pealt otsustanud fraasi ümber sõnastada. Paranduse tulemuseks on enamasti algset osa välja jättes süntaktiliselt korrektne konstruktsioon. Antud juhul oli kõneleja soov edasi anda sellist informatsiooni: et `sööke `selle hinna sees ei=`ole.

Kõneleja läheb parandades alati mingi sõnani tagasi, kas ühe sõna võrra või rohkem. Levelt (1983: 75) leiab, et sõnani, milleni kõneleja lausungis tagasi läheb, on seotud sellega, et tulemuseks oleks korrektne süntaktiline koordinatsioon eelneva lausungiga. Kindlat reeglit aga siiski pole. Nooteboom (1980)[3] on tähele pannud, et foneetilised vead näitavad vähem tagasiminekut kui leksikaalsed vead. Antud näite puhul on kõneleja tagasi läinud fraasi algusesse ja alustanud kogu fraasi uuesti.

Mitmete eksperimentide saadud mittesoravuste esinemissagedused kõiguvad 2 sõnast kuni 26 sõnani 100 sõna kohta (Fox Tree 1995: 709). On leitud, et mida pikem lausung, seda suurem on tõenäosus, et selles sisaldub mittesoravust. Oviatt (1995) viis läbi uurimuse, kus ainuüksi teades lausungi pikkust, võis 77% juhtudel kindlalt oletada, et lausungis esineb mittesoravust. Nt Trains korpust analüüsides leiti, et 23% kõikidest kõnevoorudest sisaldas vähemalt ühte parandust. 54% kõnevoorudest, mis olid vähemalt 10 sõna pikad, ja 70% kõnevoorudest, mis olid vähemalt 20 sõna pikad, sisaldasid alati vähemalt ühte parandust. (Heeman 1997: 10)

1.1.1. Terminoloogia

Järgnevas peatükis antakse lühike ülevaade mittesoravustega seotud terminoloogilistest probleemidest. Samuti selgitatakse selle töö keskseid mõisteid.

Esimese detailselt väljatöötatud mittesoravuste (ingl *non-fluencies*) alase uurimuse ja esimese katse mittesoravusi klassifitseerida tegi Wendell Johnson Iowa ülikoolist 1940-ndatel. Järgnevatel aastatel toetusid kõik uurimused Johnsoni uurimusele. 1961. aastal kasutas Johnson esimest korda mõistet *disfluency*, mis esialgu oli mõeldud *stuttering* (e.k. kogelemine) vasteks, aga peagi saadi aru, et *disfluency* on

[3] Viidatud Levelt (1983) kaudu.

siiski laiem mõiste ja et takerdumisi võib kõnes tekkida mitmetel erinevatel põhjustel lisaks kogelemisele. (Eklund 2004: 56)

Erialast kirjandust lugedes võib tihti leida ka sõnakuju dy*sfluency*, mis lubab oletada, et *dysfluency* ja *disfluency* on sünonüümid ja et neid võib kasutada vaheldumisi. Kuid tegemist pole sünonüümidega. Prefiks *dis*- märgib millegi puudumist. Selles mõttes tähendab *disfluency* soravuse puudumist kõnes või teisisõnu keelt, mis pole sorav. Prefiks *dys*- tähendab halba, rasket, ebaloomulikku ja seda kasutatakse enim meditsiinilises kontekstis viitamaks patoloogilistele põhjustele. Seega sõna *dysfluency* viitab keelele, mis on ebaloomulik, ebanormaalne. (Eklund 2004: 158-160)

Inglise keeles on mõiste *disfluency* kõrval katusmõistetena kasutatud veel ka nt *repairs*, *corrections*, *reformulations*, *restarts*, *edits* ja *hesitations*. Clark (1996: 259) kasutab katusmõistena sõna *disruptions*, mille ühe liigina käsitleb mittesoravusi. Ta väidab, et kõneleja lausung võib mis iganes põhjusel peatuda. Lisaks mittesoravustele võib põhjuseks olla ka nt naermine, kohvi rüüpamine, köhatamine.

Lisaks võib ühe uurija katusmõiste olla teise uurija alaliigimõisteks. O'Shaughnessy (1994) kasutab katusmõistena *restart*, kuid Ermani käsitluses (1987) on *restart* üks parandusvõte koos kordamise (ingl *repetition*), lisamise (ingl *insertion*) ja parandamisega (ingl *correction*); *disfluency* on katusmõiste Lickley (1994), Fox Tree (1993) ja Shribergi töödes, kuid Postma, Kolk ja Povel (1990) on kasutanud seda ühe liigina kõnevigade (ingl *speech error*) ja eneseparanduste (ingl *self-repair*) kõrval. (Shriberg 1994: 11)

Nii nagu katusmõiste endaga, nii on ka erinevatel mittesoravustel mitmeid erinevaid nimetusi, kuigi sisu on neil väga kattuv. Nii on nt kõne edasilükkamisega seotud mitmed mõisted: *abridged repair, filler, filled hesitation, hesitation, filled pause, stutter*. Valestardi märkimiseks on kasutusel samuti mitmeid mõisteid: *full sentence restart, restart, fresh start, restart, sentence correction, sentence incompletion, sentence restart*. (Shriberg 1994: 10)

Seda mõistete virvarri võib seletada sellega, et diskursuseanalüüsil on omad mõisted, millega opereeritakse, konversatsioonianalüüsil omad, psühholingvistikal omad. Arvutilingvistikasse on erinevaid mõisteid iga autor vastavalt oma eelistustele lihtsalt üle kandnud. Erandiks on ehk ainult sõnakatked (ingl *word fragments*), neile

pole varemalt niivõrd suurt tähelepanu ei mittesoravuste analüüsil ega liigitamisel pööratud.

Eesti keeles tekitab segadust sõna parandamine, mis ühelt poolt väljendab parandusprotsessi ja teiselt poolt ühte mittesoravuse liiki. Et need kaks tähendust omavahel segamini ei läheks, siis lepime siin kokku, et *parandamine* tähendab parandamise protsessi ja *parandus* on üks mittesoravuse liik.

Järgnevalt esitatakse käesoleva töö keskmeks olevad mõisted.

- (**takerdus)pausid** (ingl *pauses*): parandussegmendi pausid, millega lükatakse teksti moodustamist edasi või tehakse parandusi. (Hennoste 2000: 2025)
- **täidetud pausid** e üneemid (ingl *filled pauses*): kõneleja kasutab neid teksti tootmisel planeerimisaja võitmiseks ning kõne sujuvuse tagamiseks (Hennoste 2000:1567) ;
- **(diskursuse)partiklid** (ingl *discourse markers*): partiklitel on palju funktioone, aga selle töö seisukohalt on oluline partiklite funktsioon parandamisel (Hennoste 2000: 1775);
- **(enese)parandused** (ingl *(self)-repair*): vastavalt paranduse struktuurile, kas lausungis midagi asendatakse või lisatakse, on parandamine jagatud asendamiseks (ingl *replacement*) ja lisamiseks (ingl *insertion*) (Hennoste 2000: 2704); käesolevas töös käsitlen paranduste all ka sõnakatkeid;
- **kordused** (ingl *repetition*): ühe sõna kordamine;
- **valestardid** (ingl *false starts*): üks lausung jääb lõpetamata, kui alustatakse poole pealt uut. Selgituseks olgu öeldud, et kasutusel on ka selline mõiste nagu *fresh start*, millega märgib valestardile järgnevat uut lausungit (Fox Tree 1995: 710).

1.1.2. Liigitamine

Erinevad mittesoravuste liigitused põhinevad vastavusel parandatava ja parandatud osa vahel. Shriberg (1994: 10) on leidnud, et iga liigitus eristab minimaalselt nelja klassi: 1) vastavus ilma muutusteta parandatava ja parandatud osa vahel, nn täpsed kordused; 2) vastavus teatud muutustega parandatava ja parandatud osa vahel, nn asendused, lisandused; 3) lausungi pooleli jätmine, vastavus parandatud osaga puudub, nn valestart;

4) parandatavas osas muutub materjal, pole midagi parandada, nn edasilükkamine, täidetud pausid.

Käesolevas töös tuuakse välja järgmiste seitsme uurija(rühma) jaotused: Hindle (1983), Levelt (1983), Bear jt (1993), Heeman ja Allen (1994), Giachin ja McGlashan (1997), McKelvie (1998), Kurdi (2002). Näeme, kuivõrd erinevad on iga uurija liigitused – kes on käsitlenud kõiki mittesoravusi, kes ainult parandusi.

Hindle (1983) on kõige sagedasemate mittesoravuste klassina välja toonud järgnevad neli:

1) Ebaharilikud konstruktsioonid: produktiivne lausungitüüp, mis esineb suulises keeles väga sageli, nt *That's the only thing he does is fight* (e.k. see on ainus asi millega ta tegeleb on kaklemine). Ei ole mittesoravus, aga vajalik siiski korpuses ja grammatikas ära märkida.
2) Selged ebagrammatilisused: ebaproduktiivne lausungitüüp, nt *I've seen it happen is two girls fight* (e.k. ma olen näinud seda juhtumas on kahe tüdruku kaklus). Ütleb ka ise, et raske vahet teha, millal lausung just siia liiki satub, aga seda võiks nimetada sõnamulinaks (ingl *gibberish*). Peab seda harvaks nähtuseks, seega analüüsil sellele suurt tähelepanu ei pööra. Hilisemate uurimuste kohaselt oleks tegu anakoluudiga.
3) Eneseparandused: kui parandatav osa asendada uue, parandatud osaga, peaks tulemuseks olema süntaktiliselt igati korralik lausung.
4) Mittesüntaktilised vead: nt vead fonoloogias, morfoloogias ja/või semantikas.

Levelt (1983) on parandused vastavalt kõneleja motiividele jaganud viieks:

1) D-parandused (ingl *different*): kõneleja katkestab pooleli oleva lausungi ja otsustab öelda midagi muud;
2) A-parandused (ingl *appropriate*): kõneleja leiab, et mingi osa öeldud lausungist oleks vaja sobivamaks ümber teha;
3) E-parandused (ingl *error*): kõneleja avastab, et öeldu sisaldab viga;
4) C-parandused (ingl *covert repair*): kõneleja lükkab lausungi moodustamist täidetud pauside ja sõnakordustega edasi. Levelt nimetab neid varjatud

parandusteks. Nende abil saab kõneleja enne lausungi väljaütlemist varjatult oma kõne planeerida ja selles parandusi teha;

5) R-parandused (ingl *rest*): näited, mis ei mahu esimesse nelja klassi, kuuluvad siia, nn ülejäägid.

Bear jt (1993) liigitus on mõeldud automaatsele tuvastusprogrammile, et see teeks vahet erinevat liiki mittesoravustel. Programm peab suutma vahet teha neljal liigil: 1) parandustel, mis sisaldavad asendust või lisandust, nt *Show me flights from Boston from Denver to Dallas* (e.k. näita mulle lende Bostonist Denverist Dallasesse); 2) ühe või enama sõna kordustel, nt *Show me show me the flights …* (e.k. näita mulle näita mulle lende …); 3) valestartidel, nt *Show me the What are the flights …* (e.k. näita mulle neid Millised lennud …); 4) sõnakatketel, nt *Show me the flights from Bos- Denver* (e.k. näita mulle lende Bos- Denverist).

Heeman ja Allen (1994) on parandused jaganud kolmeks:

1) Loobumine (ingl *fresh starts*): kõneleja jätab lause pooleli ja alustab selle asemel uut, nt *so it´ll take • um so you want to do what* (e.k. see võtab • ee nii sa tahad teha siis mida);
2) Ümbertegemine (ingl *modifications*): kõneleja vahetab sisuliselt või grammatiliselt väära sõna/lauseosa uue vastu välja või lisab sõna/lauseosa tagantjärele juurde, nt *so that will total • will take seven hours to do that* (e.k. see kokku • võtab seitse tundi, et seda teha);
3) Edasilükkamine (ingl *abridged*): kõneleja vajab aega, kasutades selleks pause, partikleid, üneeme, nt *we need to • um manage to get the bananas to Dansville* (e.k. me peame saama need banaanid Dansville'i)

Giachin ja McGlashan (1997: 114-115) on liigitanud nn restarte, mis on jagatud kolmeks: 1) täpne kordus, nt *I would like to go – to go to Milano* (e.k. ma tahaksin minna – minna Milanosse); 2) parafraseerimine, nt I *would like to leave – I leave from Torino* (e.k. ma tahaksin lahkuda – ma lahkun Torinost); 3) sõnakatke, nt *I leave from Tori- from Torino* (e.k. ma lahkun Tori- Torinost). Nende parafraseerimine on ilmselt võrdväärne Heemani ja Alleni ümbertegemisega.

McKelvie (1998: 10) on parandused jaganud neljaks, sealjuures tehes vahet hesitatsioonidel ja parandustel, kus esimesel juhul on tegu lihtsalt edasilükkamisega ja teisel juhul lausungis millegi parandamisega. Parandused on jagatud omakorda veel nn päris parandusteks ja reformuleeringuteks. Neljandana on käsitletud valestarte. McKelvie jaotus näeb seega välja järgmine:

1) Hesitatsioonid
2) Parandused
3) Reformuleeringud
4) Valestardid

McKelvie jaotuses on huvitav just paranduste ja reformuleeringute vahel vahetegemine. Mida näiteks Bear jt on pidanud sõnakatketeks, on tema pidanud parandusteks. Parandus on tema käsitluse järgi see, kus on lõpetamata moodustaja, millele järgneb lõpetatud ja parandatud variant sellest, nt I´ve no g- I´ve not got any tribal settlement at all. Reformuleeringuks peab ta neid juhtumeid, kus kaks moodustajat on eraldatud parandusele viitava fraasiga, seejuures teine moodustaja peaks parandama esimest, kuigi esimene moodustaja on samuti lõpetatud, nt turn to your right, well left, sorry.

Kurdi (2002) on mittesoravused jaganud 5 tüüpi: 1) täidetud pausid, sõnakatked; 2) kordused; 3) valestardid; 4) eneseparandused; 5) poolelijäänud lausungid. Kordused on omakorda jaotatud kordusteks koos lisatud sõnaga ja kordusteks ühe sõna kustutamisega. Parandused on jaotatud kolmeks: lisamine, asendamine, sõnajärje muutmine.

Käesoleva töö aluseks olevat mittesoravuste liigitust tutvustatakse peatükis 3.

1.2. Mittesoravuste analüüsimine

1.2.1. Psühholingvistilised lähenemised

Väga põhjaliku ülevaate on mittesoravuste erinevatest uurimustest andnud Robert Eklund oma doktoritöös (2004: 51-171), kus on juttu mittesoravustest nii kõnetootmisel, kirjakeeles, mitteemakeelerääkijate keeles kui ka viipekeeles. Järgnevas peatükis

antakse põgus ülevaade (psühho)lingvistilistest uurimustest: mittesoravuste tekke-põhjustest ja sellest, kuidas kuulaja mittesoravusi töötleb.

Mittesoravused võivad tekkida igal kõne tootmise tasandil: kõnelejal ei pruugigi üldse selge olla, mida ta just täpselt öelda tahab, aga ta alustab oma kõnevooru, mis tähendab, et tal on kohustus jätkata – Clark ja Wasow (1998) nimetavad seda siduvuse strateegiaks (ingl *commitment*). Sellega on omakorda seotud jätkuvuse strateegia (ingl *continuity*), st teatud moodustajate kordamisega taastab kõneleja enda kõnevooru soravuse ehk aktiveerib varem öeldu. Näilist soravust tekitatakse täidetud pauside ja pikaks venitatud silpidega. Brennan ja Williams (1995) on leidnud, et täidetud pausid tunduvad lühemana kui sama pikad vaiksed pausid. Nt kordusi tekitab ka moodustajate keerukus, see on nn keerukuse strateegia (ingl *complexity*) (Clark, Wasow 1998; Arnold jt 2000). Moodustajaid on seda raskem planeerida, mida suurem on nende grammatiline kaal (ingl *grammatical weight*). Clark ja Wasow leidsid oma katsetes, et kõnelejad kordavad artikleid *the* ja *a* sagedamini just grammatiliselt keerukamate noomenifraaside (NP) ees, nt *the, the time we were there at the warehouse*.

Arnoldi jt (2000) katsetest on välja tulnud, et mittsoravusi ilmneb rohkem just viitamisel asjadele, mis on diskursuse seisukohast uued. Eksperimendis osalejad olid paaridesse jaotatud ja jagasid vastastikku juhiseid, kuidas ümber paigutada teatud objekte, mis olid diskursuses juba kord mainitud või diskursuse seisukohast uued. Analüüsid näitasid, et kõikidest NP-dest (n = 5128) 21% moodustasid mittesoravad uued NP-d, juba eelnevalt esinenud mittesoravaid NP-sid oli ainult 16%.

Ühest vastust küsimusele, kuidas vaatamata keele kõrvalekaldele grammatilistest normidest on inimestel siiski vähe raskusi mõistmaks mittesoravat kõnet, pole. Omaette suureks küsimuseks jääb ka, kuidas suudavad lapsed omandada grammatika, kuuldes vaheldumisi nii grammatilisi kui ka ebagrammatilisi lausungeid. On arvatud, et kuulaja filtreerib mittesoravused ja interpreteerib ainult lõplikku soravat lausungit. Arvesse võttes aga, et suuline kõne on siin ja praegu toimuv suhtlusakt, nn ühistegevus (ingl *joint activity*; Clark 1996), kus sageli ehitavad kõnevooru kahepeale kokku kõneleja ja kuulaja, on raske seda uskuda. Filtreerimise ideele räägib vastu ka asjaolu, et kui sõna on juba kord lausutud ja seotud juba olemasoleva süntaktilise struktuuriga, püütakse seda semantiliselt ikka interpreteerida, sellele tähendust anda, mis iganes sõnu enne ka

ei öeldud. See tähendab, et kui öeldu võib olla ka süntaktiliselt vale, püüame sellest siiski aru saada. (Bailey, Ferreira 2003)

Üldine oletus on, et kõik need kõhklemised ja parandused aeglustavad kõnest arusaamist. Kuulaja ootab ühte lauset, aga peab siis selle kõrvale jätma ja hakkama uut ehitama. Süntaksipuu seisukohast vaadatuna peab kuulaja poole süntaksipuust uuesti ehitama, viskama vana osa välja ja püüdma uut infot sinna sobitada. Kui kuulaja kuuleb uut algust, peab ta suutma meenutada, mida öeldi enne seda, ja siis siduma uuesti öeldu sobivasse kohta lausungis. Raskusi võivad tekitada just lausungi keskel uuesti alustatavad fraasid. Lause alguse otsast alustamise korral pole kuulajal vaja midagi meelde jätta ega kuskilt midagi jätkata. (Fox Tree 1995)

Katsetest selgub, et vastupidiselt ootustele, ei takista/pidurda mittesoravused alati arusaamist. See sõltub sellest, millist liiki mittesoravustega on tegu ja kuhu nad kõnevoorus langevad. Vaatamata asjaolule, et ka kordused takistavad kõne vaba voolavust ja süntaktilist koherentsi ning nõuavad kuulajal kuuldud lausungi ümber-interpreteerimist, paistavad kordused üsna sageli arusaamisele hoopis kaasa aitavat, kindlasti nad ei takista arusaamist. Öeldut ainult korratakse, lausungi süntaks ja semantika jäävad samaks. (Fox Tree 1995)

On tehtud ka mitmeid eksperimente, kus on mõõdetud osalejate reaktsioonikiirust vastavalt sellele, mida neile öeldakse. Tavalisim selline eksperiment toimub nii: kõneleja annab edasi teatud juhiseid kord korrektselt, kord ennast parandades, korrates, pause tehes, nt *vajuta kollast rohelist nuppu, vajuta sinist ee punast nuppu*, ja seejärel mõõdetakse kuulajate arusaamise kiirust sellega, kui kiiresti nad vastavat nuppu vajutavad. (Detailsemalt sellistest eksperimentidest Brennan, Schober (2001).)

Marslen-Wilson jt (1988)[4] on uurinud lausungi süntaksit, semantikat ja pragmaatikat – kuidas need aitavad kaasa antud sõna antud kontekstis kiirele äratundmisele. Nad väidavad, et nt sõna *kitarr* tuntakse ära kiiremini nn normaalses lauses nagu *poiss hoidis kitarri* kui semantiliselt ebatõenäolises lauses *poiss jõi kitarri* või pragmaatiliselt ebatõenäolises lauses *poiss mattis kitarri*.

[4] Viidatud Fox Tree (1995) kaudu.

Tekstianalüüsi seisukohast on huvitav ka see, kuivõrd kuulaja saab hinnata kõneleja teadmisi mingil teemal tänu mittesoravuste esinemisele (Bailey, Ferreira 2003). Nt kui kõnelejalt küsitakse *Kes oli Salme Reek*? ja vastuseks tuleb *ee näitleja*, mitte lihtsalt *näitleja*, siis kuulaja võib kahelda kõneleja kindlas teadmises, et Salme Reek oli tõesti näitleja. Kui aga kõneleja vastaks *ee ma ei tea* konkreetse *ma ei tea* asemel, siis võib kuulajal tekkida hoopis vastupidine kahtlus, et kõneleja teab küll, aga lihtsalt ei taha öelda.

1.2.2. Arvutilingvistilised lähenemised

Kui lingvistid ja psühholingvistid on mittesoravusi uurinud lausungite tootmise seisukohast, siis arvutilingvistid on seda teinud just mittesoravuste tuvastamise seisukohast. Loomuliku keele töötlemisel on tuvastamine tunduvalt keerulisem ülesanne kui genereerimine. Nt dialoogsüsteem, mis suhtleb kasutajaga loomulikus keeles, peab ära tundma lausutud sõnad, võimalikud kordused ja parandused ning tabama lõpuks ka öeldu kavatsetud tähenduse. Määrata kindlaks kõneleja öeldu kavatsetud tähendus, on suulise kõne analüüsi üks väljakutsuvamaid ülesandeid. Isegi kui dialoogsüsteemides suudab kõnetuvastaja leida kõik sõnad, on öeldu kavatsetud tähendust sagedaste pauside, takerdumiste ja paranduste tõttu raske kindlaks määrata. Sõnad, mida asendatakse või korratakse, pole enam osa kavatsetud lausungist, ja need on vaja kindlaks määrata ja normaliseerida.

Süntaktilise analüüsi etapis tuleb leida lahendused, kuidas analüüsida täidetud pause, sõnakatkeid, süntaktilist ebatäielikkust, üksikuid fraase, parandatavate ja parandatud osade jadasid, kordusi, süntaktilisi ühtesulamisi e anakoluute (Leech jt 1998; LE-EAGLES WP4).

Võimalik on rakendada kahte sisuliselt erinevat strateegiat: normaliseerimist ja märgendamissüsteemi laiendamist, st ka mittesorav lausungiosa saab analüüsi. Kolmas võimalik alternatiivne lähenemine kahele eelnevale on osaline parsimine (ingl *partial parsing*).

Igasugune programm, mis analüüsib suulist keelt, peab hakkama saama mittesoravustega, nii nende tuvastamise kui ka normaliseerimisega. Üheks võimaluseks mõlemaga hakkama saada on mittesoravused enne süntaktilist analüüsi käsitsi

märgendada, misjärel pole parandatav osa järgmiste analüüsietappide jaoks enam kättesaadav (nt Charniak, Johnson 2001). (Reaalsetes rakendustes pole muidugi mõeldav, et mõni vahepealne osa tehakse käsitsi.) See pole aga alati parim lahendus. Nimelt, Core ja Schubert (1999) leiavad, et ka parandatav osa peab jõudma semantilisse analüüsi. Eriti oluline on see just anafooride määramisel. Nende kuulus näide: *have the engine take the oranges to Elmira, um, I mean, take them to Corning*. Siin viitab *them* sõnale *oranges*, kui viimane ära kustutada, siis jääb selgusetuks, mida tuleb Corningisse viia. Aga selliste näidete hulk on võrreldes ülejäänud juhtudega väga väike.

McKelvie (1998) on veendunud, et teatud tüüpi mittesoravused on parsitavad ja neid tuleb parsida, nt parandused. Mittesoravuste kaasamine süntaktilisse analüüsi ei ole lihtne ülesanne. Selleks tuleb muuta kas reegleid või parserit ennast. Core (1999) leiab, et ainult reeglite muutmine ei aita. Näiteks toob ta McKelvie (1998) uurimuse, kus muudeti vaid reegleid, mille tulemuseks oli see, et parser sai hakkama ühe moodustaja parandamisega, kuid mitte suurema arvu moodustajatega. See probleem esineb ka eesti keele kitsenduste grammatika kohandamisega suulisele keelele (vt ptk 4).

Järgnevalt vaatame, kuidas on mittesoravusi lause süntaktilisse analüüsi kaasatud. Katsetatud on nii sulundamist, moodustajatepuud kui ka sõltuvuspuud.

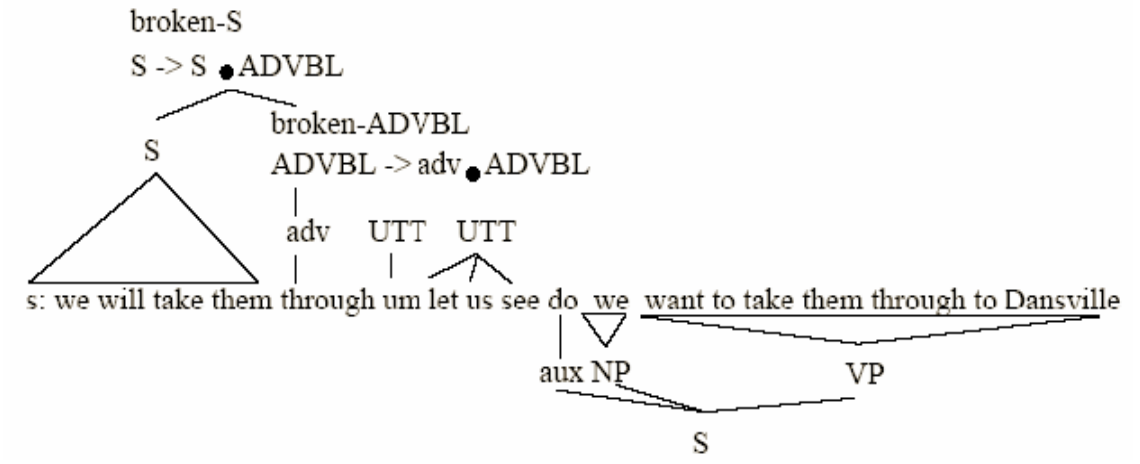
Näide sulundatud parandusega lausungist Christine korpusest (Sampson 1995)[5], kus mõlemal pool katkestuskohta (#) esinevaid relatiivlauseid (esimene lõpetamata) käsitletakse kui ühe nimisõnafraasi erinevaid laiendajaid, nt

and that [NPs any bonus [RELCL he] # money [RELCL he gets over that]] is a bonus

Fraasistruktuuripuusse on mittesoravusi kaasanud Core ja Schubert (1999). Kõneleja on alustanud oma lausungit *we will take them through*, kuid ei lõpeta seda kohe (joonis 1). Selle asemel võtab kõneleja endale aega, kasutades täidetud pausi *um*, ning seejärel kuulajale märku andes, et eelnev lausung jääbki lõpetamata *let us see*. Kõneleja formuleerib uue lausungi *do we want to take them through to Dansville*. Näeme, et pooleli jäetud lausung saab samuti väga põhjaliku analüüsi. Minimaalne analüüsitav

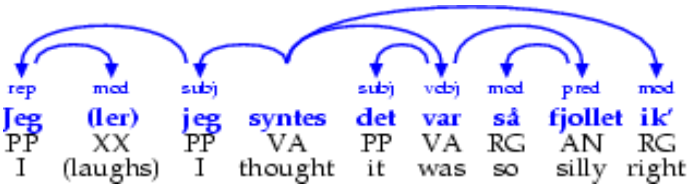
[5] Viidatud Leech jt 1998 kaudu

terviklik üksus on saanud lause analüüsi S, mille lõppu märgib ● enne poolikuks jäänud adverbiaali (broken-ADVBL).



Joonis 1. Näide paranduse esitamisest fraasistruktuuripuus

Sõltuvuspuu näide on pärit suulise taani keele sõltuvuspuudepangast[6] (joonis 2). Näeme, kuidas sõltuvuspuus on lahendatud korratud subjekti probleem. Korratud subjekt saab endale õige subjekti (subj) rolli ja esimene algne subjekt korduse (rep) rolli.



Joonis 2. Näide korduse esitamisest sõltuvuspuus

Mittesoravuste automaatseks tuvastamiseks ja normaliseerimiseks on rakendatud erinevaid meetodeid. Tuvastamiseks on nt rakendatud metareegleid (Hindle 1983; McKelvie 1998), statistilist lähenemist (Stolcke, Shriberg 1996), nn parandustele vihjajaid (ingl *triggers*) (Spilker jt 2000). Normaliseerimiseks on rakendatud

[6] http://www.id.cbs.dk/~mtk/ddt/spoken.html

prosoodilis-akustilisi vahendeid (Nakatani, Hirschberg 1994), mustrite sobitamist (Bear jt 1992; Heeman, Allen 1994, 1997; Kurdi 2002), käsitletud masintõlke ülesandena (Spilker jt 2000).

Järgnevalt tutvustatakse üheksa uurimuse tulemusi. Kuna nende uurimuste süstematiseerimine on nende erineva eesmärgi, vahendite, materjali tõttu väga keeruline, siis on uurimused esitatud ajalises järjekorras. Erinevate meetodite tõhususest saab aimu esitatud saagisest (ingl *recall*) ja täpsusest (ingl *precision*), kuigi alati on oluline ka vaadata, millistel tingimustel on tulemused saadud. Tuvastamise saagis näitab õigesti tuvastatud parandamiste osakaalu kõikide parandamiste suhtes ja täpsus tuvastatud parandamiste osakaalu kõikide leitud tuvastamiste suhtes. Parandamise saagis näitab õigesti normaliseeritud parandamiste osakaalu kõikide parandamiste suhtes ja täpsus õigesti normaliseeritud parandamiste osakaalu kõikide leitud normaliseerimiste suhtes.

Hindle (1983) uurimus oli üks esimesi arvutilingvistilisi lähenemisi mittesoravuste analüüsile. Hindle kasutas deterministlikku parserit Fidditch, mida täiendati reeglitega, mis vaatasid kattuvaid kategooriaid ja sõnade stringe. Mudel töötas tingimusel, et parandusele viitab kindel marker. (Hilisemad uurimused on selle väite ümber lükanud.) Hindle saavutas oma testkorpuse peal parandamise täpsuseks 97%. Nii kõrgele tulemusele aitas kindlasti kaasa oletus, et parandused algavad kindla parandamisosa algusega, mis olid enne korpuses käsitsi ära märgendatud.

Bear jt (1992) kasutasid 406 lausungi peal mustrite sobitamise (ingl *pattern-matching*) meetodit. Mustrite sobitamise komponent otsis identseid sõnade jadasid ja lihtsaid süntaktilisi anomaaliaid, nagu *a the* ja *to from*. 406 lausungist, mis kõik sisaldasid mingisugust parandust, leidis programm edukalt 309 ja nendest 177 suudeti ka korrektselt parandada.

Nakatani ja **Hirschberg** (1994) kasutasid paranduste leidmiseks prosoodilist informatsiooni ja sõnakatkete esinemist. Nad tegelesidki ainult tuvastamisega. 74% nende korpuses leiduvatest parandustest olid märgitud sõnakatketega. Kasutades käsitsi märgitud prosoodilist märgendamist, treenisid nad oma parserit treeningkorpusel, mis koosnes 172 lausungist, milles igaühes leidus vähemalt üks parandus. Nakatani ja Hirschberg leidsid, et olulised on sõnadevahelised pausipikkused, sõnakatkete

esinemine ja leksikaalne sõnade kattumine. Testkorpusel, mis koosnes 186 lausungist ja 223 parandusest, saavutasid nad paranduste leidmise saagiseks 83,4% ja täpsuseks 93,9%.

Heeman ja **Allen** (1994) rakendasid paranduste tuvastamiseks mustrite sobitamise meetodit. Mustrite ülegenereerimise vähendamiseks kasutasid nad morfoloogilist analüsaatorit statistilise filtrina, mis hindas iga korduse ja paranduse tõenäosust. Mittesoravuste testkorpusel saadi tuvastamise saagiseks 83% ja täpsuseks 89% ning normaliseerimise saagiseks 80% ja täpsuseks 86%.

Core ja **Schubert** (1999) analüüsivad ka parandatavat osa, mis saab endale fraasistruktuuri. Katkestuskohti käsitletakse eraldi lausungitena suurema lausungi sees (vt joonist 1; märgitud UTT). Nad käsitlevad oma mudelis ka pealeräkimisest tulenevaid probleeme, püüdes vahet teha erinevatel situatsioonidel: millistel juhtudel peab teist kõnelejat käsitlema iseseisvalt ning millal teevad kõneleja ja kuulaja lausungi kahepeale valmis ning millal üks parandab teist. Mõeldud on ka sellele, kui kuulaja tahab vahele segada ja proovib seda ka teha, lausudes paar sõna, kuid andes alla, kui kõneleja ei anna vooru üle. Sellised juhtumid on liigitatud lõpetamata lausungite alla ja need on samuti dialoogi interpretatsiooni kaasatud. Pealeräkimiste analüüsimine ei ole kerge ülesanne. Seda näitab ka nende tehtud eksperimendi tulemus: 106 pealeräkimisega lausungist sai parser hakkama 24-ga. Nende eksperimendi üldised tulemused: 284 parandust õigesti tabatud, 371 valealarmi ja 257 märkamatuks jäänud parandust ehk saagis 52,50% ja täpsus 43,36%.

Spilker jt (2000) on tegelenud Verbmobili kõnesüsteemi selle osaga, mis puudutab mittesoravuste analüüsi. Nende välja töötatud paranduste tuvastamise ja normaliseerimise algoritm sisaldab kolme sammu. Esiteks, kasutusel on parandustele viitajad (ingl *triggers*), mis viitavad potensiaalsetele katkestuskohtadele. Viitajatena kasutatakse akustilis-prosoodilisi märguandeid ja sõnakatkeid. Seejärel püüab stohhastiline mudel leida igale tuvastatud katkestuskohale sobiva paranduse, pakkudes selleks kõige tõenäolisemat varianti. Selle saavutamiseks käsitletakse paranduse töötlemist statistilise masintõlke probleemina, kus parandatav osa on parandatud osa tõlkeks. Testkorpus sisaldas 549 parandust. Tulemused on toodud tabelis 1. Testi 2 puhul oletati, et eksisteerib täiuslik sõnakatkete tuvastaja. Testimisel kasutati litereeritud

tekste, kuna soov oli mõõta paranduste normaliseerimise protsessi kvaliteeti, mitte kõnetuvastaja tööd.

	Tuvastamine		**Normaliseerimine**	
	Saagis	**Täpsus**	**Saagis**	**Täpsus**
Test 1	49%	70%	47%	70%
Test 2	71%	85%	62%	83%

Tabel 1. Suulise keele automaatse tuvastamise ja normaliseerimise tulemused kõnetõlke-programmis Verbmobil

Charniak ja **Johnson** (2001) on välja töötanud kahekordse läbimise mudeli (ingl *two-pass architecture*). Esimene läbimine aitab leida parandatava osa, nt *Why didn´t he, why didn´t she stay home*?. Teine läbimine eemaldab need sõnad stringist. Eksperimendi tulemused näitasid, et parsimine õnnestus palju paremini, kui kindlad katkestuskohad olid ette antud. Täpsuseks saadi 85,3% ja saagiseks 86,5%.

Kurdi (2002) on oma mudelis ühendanud mustrite sobitamise meetodi ja pindsüntaktilise parsimise. Paremaks süntaktiliseks normaliseerimiseks on rakendatud ka känk-parsimist (ingl *chunk parsing*), mis lubab sisendi paindlikumat analüüsi. Võrreldes teiste siin toodud uurimuste tulemustega, on Kurdi omad väga head: tuvastamise saagis 89,67% ja täpsus 92,76% ning normaliseerimise saagis 84,47% ja täpsus 86,61%.

Snover jt (2004) kasutasid mittesoravuste tuvastamisel leksikaalseid markereid, so sõnu endid ja sõnaliigi märgendeid. Nad jagasid mittesoravused vastavalt nende omadustele: parandajateks (ingl *edits*) ja edasilükkajateks (ingl *fillers*). Iga sõna varustatakse vastavalt kas parandaja, edasilükkaja või sorava sõna (ingl *fluent*) märgendiga. Tegemist on transformatsioonidel põhineva lähenemisega, kus õpitud reegleid on 106. Õpitavad reeglid on seotud mitmete tunnustega: kas sõnale järgneb paus ja kas sõna on sagedase esinevusega, st kas sõna on sagedasem ühe kõneleja keelepruugis kui ülejäänud korpuses. See on oluline selleks, et saaks vahet teha sõnadel, mis on tavaliselt soravad, aga vahel ka mitte, nt *like*. Kui kõneleja kasutab sõna *like* väga tihti, siis on tõenäoline, et ta kasutab seda täitena. *like* esineb korpuses näiteks 22% juhtudest mittesoravana. Eksperiment viidi läbi kahe korpuse peal, üks sisaldas transkribeeritud telefonivestluseid ja teine suulist kõne. Esimesel juhul oli parandajate

tuvastamise vigu 68,0% ja edasilükkajate omi 18,1%, teisel juhul vastavalt 87,9% ja 48,8%. Kõne analüüsi kõrged veaarvud on osalt tingitud kõne tekstiks ümbertegemisest. Parandajate suurem veaarv lubab oletada, et need pikad, mistõttu on neid keerukam üles leida. Raskemaks teeb nende analüüsi ka see, et kui edasilükkamise osa on sorav, siis ei oska programm midagi kahtlustada, nt [*and whenever they come out with a warning*] *you know they were coming out with a warning about trains*. Katkestuskoha võib küll prosoodiliste vahenditega tuvastada, aga seda, kustkohast lausungis parandamine algab, on vaja samuti tuvastada.

Ferreira jt (2004) on toonud üheks põhjuseks, miks praegused mittesoravuste automaatse tuvastamise uurimused pole andnud soovitud head tulemust, selle, et mittesoravused on väga heterogeenne klass keelelisi nähtusi. Kui nt täidetud pausid on suhteliselt fikseeritud klass, siis parandused ja kordused on statistiliselt tunduvalt keerulisemad käsitleda, kuna neil puudub kindel leksikaalne struktuur, st paranduses võib korraga olla nii kaks sõna kui ka terve lausung.

Praegu on rohkem siiski selliseid uurimusi, mis on tegelenud transkribeeritud dialoogide analüüsiga. Automaatset mittesoravuste tuvastamist ja normaliseerimist praegu veel väga vähe ja häid tulemusi on saadud üksnes sellega, kui katkestuskoht on käsitsi eelmärgendatud.

1.2.3. Mittesoravuste märgendamine

Mittesoravuste järjepidev märgendamine algas 1990. aastail. Tekkis suurem huvi mittesoravuste strukuuri vastu, st huvikese küsimuselt miks (kõnelejad takerduvad ja ennast parandavad) kandus üle küsimusele kuidas (süntaktilises analüüsis nendega hakkama saada).

Tuntumad suulise keele korpused, milles on mittesoravusi märgendatud, on Switchboard, TRAINS-93, Christine, ATIS, ICE-GB, Monroe ja Verbmobil. Nendest ainult Verbmobili korpus koosneb reaalsest suulisest keelest, teised sisaldavad juba transkribeeritud suulist keelt.

ICE-GB ja Monroe korpused on sellised, kus lausungisse mittesobiv lihtsalt eemaldatakse. ICE-GB korpuses (Meyer 2002: 96) on selleks kasutusel vastav TEI

kustutamisvahend[7], mis enne süntaktilist analüüsi eemaldab mittesoravused. Nt lausung *can I can I can we take that again* (ee võin ma võin ma võime me selle uuesti võtta) märgendatakse ICE-GB korpuses järgmiselt:

<sent><}><->can I can I </-><=>can we</=></}>take that again<$?>

Monroe korpuses pannakse mittesorav sõna või fraas nurksulgudesse, nt [*We're gonna send the digging crew*] *we're gonna send the road crew from RGE to Elmwood bridge*. Sama tehakse kordustega, nt [*Can you*] *can you go over the thing for me again*? Eraldi märgistuse saavad ka lõpetamata või ebagrammatilised lausungid. Lõpetamata lausungeid on iseloomustatud järgmiselt: 1) subjekt ja verb on olemas, kuid vähemalt üks lausemoodustaja on veel vajalik, et öeldu mõtet välja lugeda, nt *It and that is*; 2) lausung sisaldab ühte või mitut sõna, mis ei klassifitseeru eelmise lausungi elliptiliseks vastuseks; 3) lausung lõpeb poole sõna pealt. Ebagrammatiliste lausungite lisajoontena on välja toodud: 1) oluline moodustaja (nt artikkel või prepositsioon) on puudu; 2) sõnastus on vale; 3) morfoloogilised vead, nt ühildumisvead. (Swift jt 2004)

Lisaks ICE-GB ja Monroe tüüpi normaliseerimisele on sellist, kus väga täpselt püütakse erinevate mittesoravuste vahel vahet teha ja märgendamine tähendab enamat kui lihtsalt lausungist millegi väljaviskamist. Märgendatakse väga erinevat informatsiooni: sõnade hulka paranduses, mittesoravuste liiki, katkestuskohta, kõneleja sugu, vanust jne. Shriberg (1994: 50) on nt parandused esinevad sõnad jaganud sõnad lausungis ja nn tõhusad sõnad lausungis. Viimast võiks pidada normaliseerimise mõõduks. Näitena on ta toonud järgneva lausungi, mis kokku sisaldab 24 sõna, aga tõhusaid neist on 16. Lausungis on läbi kriipsutatud parandatav osa ja katkestuskoht: ~~um i would like uh~~ i would like to ~~book a flight~~ book a flight for sunday from miami florida to las vegas nevada. Süntaksianalüsaatori seisukohast vaadatuna on suur vahe, kas analüsaator peab analüüsima 24 sõna või 16 sõna.

Bear jt (1993) märgendavad katkestuskohta, toovad välja sõnade seosed parandatava ja parandatud osa vahel ning täidetud pausid ja sõnakatked. Katkestusosa märkimiseks kasutavad nad püstkriipsu (|), mis paigutatakse lausungi alla vastavasse kohta järgmisel real, nt

[7] http://www.tei-c.org/P4X/TS.html

List these in increasing in order of increasing fare

|

Valestarte märgendatakse kahte moodi. Kui parandatava ja parandatud osa vahel eksisteerib semantiline seos, märgendatakse .| katkestuskoht, kusjuures sellisel juhul tuleb märgendajal tuua välja suhted parandatava ja parandatud osa vahel. Kui kõneleja alustab poole lausungi pealt uue mõttega, siis märgitakse see || ning märgendaja ei pea paranduses olevate sõnade suhteid välja tooma. Aga on ka selliseid lausungeid, kus on märgendaja otsustada, kas kasutada .| või ||. Järgmises näites on nii lausungi semantilise plaani muutumine kui ka kattuvad sõnad olemas, kuid kuna sõnadevahelisi seoseid oleks väga raske välja tuua, siis on mõistlik lausung valestardiks märgendada:

What time does this flight arrive where does this flight make a stop

||

Kui katkestuskoht on lausungis määratud, märgendatakse lausung sõnatasandil. Selleks on kasutusel neli märgendit: M (ingl *matching*) kattumine/kordumine, R (ingl *replacement*) asendamine, X lisamine või kustutamine, C (ingl *cue word*) parandusele viitav sõna/fraas. Järgnevalt üks näide sõnade märgendamisest. Kattuvad sõnad saavad märgendi M, asendatavad R ja parandamisele viitav fraas C ning indeksid näitavad, mitu sõna kattuvad.

from Atlanta back to Pittsburgh I'm sorry back to Denver

M^1 M^2 R^1 C C | M^1 M^2 R^1

Järgnevalt vaatame, kuidas näeb normaliseerimine välja Switchboardi korpuses (Meteer jt 1995). Sellest sai alguse ka minu idee püüda suulises eesti keeles esinevaid mittesoravusi märgendada. Switchboardi korpuse mittesoravuste põhjalikust märgendamisest annab parima ülevaate ühe väikese dialoogi analüüs – üks dialoog on esitatud kolmel kujul: algsel, märgendatud ja normaliseeritud kujul.

Algsel kujul: dialoog sisaldab kordusi ja parandusi.

A: he's pretty good. He stays out of the street and, uh, if I catch him I call him and he comes back. So he, he's pretty good about taking to commands and --

B: Um.

A: --and things.
B: Did you bring him to a doggy obedience school or --
A: No --
B: -- just --
A: -- we never did.
B: --train him on your own and,
A: I, I trained him on my own and, uh, this is the first dog I've had all my own as an adult.
B: Uh-huh.

Märgendatud kujul: märgendatud on lausungipiirid, kordused, parandused, mittesoravad konjunktsioonid ja täidetud pausid.

1| A: he's pretty good. / He stays out of the street / {C and, } {F uh, } if I catch him I call him
2| {C and } he comes back. / {D So } [he, + he's] pretty good about taking to commands
3| [and + --
4| B: {F Um. }
5| A: --and] things. /
6| B: Did you bring him to a doggy obedience school or --
7| A: No -- /
8| B: -- just --
9| A: -- we never did. /
10| B: --train him on your own / {C and, } -/
11| A: [I, + I] trained him on my own / {C and, } {F uh, } this is the first dog I've had all my
12| own as an adult. /
13| B: Uh-huh. /

Normaliseeritud kujul: dialoogist on eemaldatud kõik mittesorav.

A: he's pretty good. He stays out of the street. if I catch him I call him. he comes back. he's pretty good about taking to commands and things.
B: Did you bring him to a doggy obedience school or just train him on your own
A: No we never did. I trained him on my own. this is the first dog I've had all my own as an adult.
B: Uh-huh.

Nagu näha võib, märgendatakse Switchboardi korpuses lisaks parandustele palju muudki. Lausungid (ingl *slash-units*) on jagatud kaheks: lõpetatud lausungid saavad märgendi / ja lõpetamata ehk pooleli jäänud lausungid -/. Lisaks on oma reeglid, millal

koordineeritud konjunktsioon *ja* kõrvaldatakse dialoogist ja millal mitte, nt vt märgendatud dialoogi 10. rida.

Parandused ja kordused pannakse nurksulgudesse, milles vasakule poole katkestuosa (+) jääb parandatav sõna ja paremale poole parandatud sõna, vt ridu 2 ja 11. Switchboardi korpuses ei vaadata dialoogi ühe kõnevooru kaupa, st kui kõneleja kõnevoor katkeb, siis vaadatakse, kas see jätkub ülejärgmisel real (--). Lõpetamata lausungiga kõnevooru saab lõpetada ainult sama kõneleja, vt ridu 7, 9, 11. Teise kõneleja vahele öeldu moodustab omaette lausungi. Suulisele kõnele omased täitesõnad on jagatud viieks: täidetud pausid (F), parandustele viitavad sõnad/fraasid (E), partiklid (D), koordineerivad konjunktsioonid (C), infolisad (A). Erinevate täitesõnade vahel vahetegemine on kohati ebaselge, eriti mis puudutab täidetud pauside, parandustele viitavate sõnade ja partiklite eristust. Üksteisest eristamist raskendab ka see, et ühte asja tuleb ühes kontekstis märgendada ühtmoodi ja teises teistmoodi, mistõttu ongi mõistetamatu nende kolme eristamine sellisel viisil.

Erinevaid mittesoravuste märgendamisi on veel mitmeid, aga lõppkokkuvõttes märgendatakse siiski samu nähtusi, lihtsalt märgendid on erinevad. Nägime nii siinsetes kui ka eespool toodud näidetes, et nt Sampson kasutas katkestusosa märkimiseks trelli, Bear jt püstkriipsu, Switchboardi korpuses on selleks kasutusel plussmärk, Shriberg (1994: 57) märgib katkestusosa punktiga (kuigi muidu on märgendamisalused samad mis Switchboardi korpuses), Heeman (1997) ülespoole suunatud noolega (↑). Heemani märgendamisskeemist võib välja tuua veel seda, et lisaks katkestuskoha ning parandatavate ja parandatud sõnade suhete väljatoomisele saab iga parandus endale ka liigi märgendi, mis näitab, kas tegemist on loobumise (:can), ümbertegemise (:mod) või edasilükkamisega (:abr) (vt ka ptk 1.1.2.). Kuna mõnikord on loobumise ja ümbertegemise vahel raske vahet teha, siis mitmese analüüsi korral lisatakse liigitähisele veel ka plussmärk, st valitakse küll üks liik, aga plussmärgiga näitab märgendaja oma kõhklust, nt

engine two from Elmi(ra)- or engine three from Elmira

m1 r2 m3 m4 ↑ et m1 r2 m3 m4

ip:mod+

Omaette teema on, kui laialt märgendada parandust, st kust algab ja lõpeb parandus. Sampson on öelnud, et tema kogemuste põhjal on paranduse alguse määramine tihti kunstlikult paika pandud, võib olla mitu alternatiivset analüüsi, sest sõnad, mis järgnevad katkestusosale, ei asenda alati täpselt öeldud sõnu. Alati on olemas selged juhud, kus pole kaksipidi mõtlemist, ja segased juhud, kus märgendaja on kohustatud siiski teatud valiku langetama. Et selliseid valikuid poleks vaja teha, peaks olema igasugune märgendamine detailselt ära seletatud ja kategooriate vahel selged piirid välja toodud. Kuna keeletehnoloogia tänapäevastes rakendustes on suulisel keelel põhinev inimese ja arvuti vaheline vestlus järjest levinum, leiab Sampson, et standardiseeritud skeemide olemasolu mittesoravuste märgendamiseks on väga soovitatav. Standardiseerimine aitaks vältida olukorda, kus kõik märgendavad samu keelenähtuseid, kuid nimetavad ja märgendavad neid erinevalt. (Sampson 1998) Iseasi on muidugi see, kas selline standardiseerimine väljaspool inglise keelt on kasulik ja efektiivne.

Järgmises peatükis tutvustatakse suulise eesti keele märgendamist: põhimõtteid ja kasutatavaid märgendeid.

2. Analüüsitav korpus ja märgendamine

Peatükiga 2, mis on aluseks järgnevale analüüsile (ptk 3) ja eksperimendile (ptk 4), algab käesoleva magistritöö praktiline osa.

2.1. Korpus

Materjal on võetud Tartu Ülikooli eesti dialoogikorpusest[8] suuliste dialoogide hulgast. Korpus sisaldab 2005. aasta detsembri seisuga 873 litereeritud teksti, neist 715 telefonikõnet ja 116 silmast silma vestlust, kogupikkusega umbes 150 000 tekstisõna. Litereerimisel kasutatud transkriptsioonimärgid on toodud Lisas 1. Transkriptsioone ei ole üle kontrollitud. Analüüsiti 35 infodialoogi (13 168 sõna, 1991 lausungit) (vt tabelit 2). Lühim dialoog sisaldas 31 sõna ja pikim 1962 sõna. Dialoogid valiti juhuslikult.

Dialoog	**Telefonikõne**	**Sõnade arv**	**Lausungite arv**
91_a3	arvutifirmasse	135	14
334_a3	bussijaama	178	32
452_a3	bussijaama	418	58
475_b14	elektroonikakauplusesse	372	53
97_a6	hambakliinikusse	90	17
364_a5	hambakliinikusse	131	34
427_b16	hotelli	328	69
338_a1_info	infotelefonile	378	74
347_a2_info	infotelefonile	146	34
455_a28_info	infotelefonile	58	18
456_b28_info	infotelefonile	46	14
460_a15_info	infotelefonile	138	44
475_b3_info	infotelefonile	442	84
63_a31_info	infotelefonile	39	12
354_a4	kinnisvarabüroosse	219	43
380_a7	kondiitriärisse	415	57
259_b7	lasteaeda	95	24
403_a1	lennujaama	107	14
96_b8	mööblikauplusesse	164	27
334_a5	panka	317	38
398_a6	panka	71	15
373_a5	postimüügifirmasse	303	49
428_a36	registratuuri	275	40
384_a2	reisibüroo	426	52
233_a4	reisibüroosse	297	32

[8] http://www.cs.ut.ee/~koit/Dialoog/EDiC

353_a3	reisibüroosse	1962	183
355_a7	reisibüroosse	75	18
357_a6	reisibüroosse	761	113
367_a4	reisibüroosse	274	25
355_a11	reisibüroosse	337	58
356_a4	reisibüroost kliendile	1000	171
359_a3	reisibüroosse	1797	307
354_a8	spordikeskusesse	400	72
338_a3	ülikooli	943	88
427_a1_info	ülikooli infotelefonile	31	8
KOKKU		13 168	1991

Tabel 2. Analüüsitud dialoogid

2.2. Analüüsiüksused

Kuna analüüsitakse kõneleja enda teksti siseseid parandusi, kordusi ja valestarte, siis on analüüsiüksusteks kõneleja lausungid, mis suuremal või vähemal määral vastavad kirjaliku keele lausetele. Vastavalt transkriptsioonile märgib lausungi lõppu punkt (intonatsiooni langus) ja osalausungeid koma (poollangev intonatsioon), nt

> (1) .hh kõrva kurgu arst on nüüd üheteistkümne- üksteist september tuleb tööle, ta on praegu kahenädalasel puhkusel. (428_a36)

Kuid on ka selliseid kõnevoore, mis koosnevad mitmest lausungist ja osalausungist, kuid mille piiridel pole intonatsioonimuutusi. Sel juhul vaadatakse grammatilist/süntaktilist lõpetatust. Järgnevas näites on kokku 3 grammatilist lausungit: esimene lausung lõpeb sõnaga *kaupa*, teine sõnaga *öelda* ja mitmest osalausungist koosnev viimane lausung sõnaga *teile*. *siis kui* on valestart.

> (2) kas see täitub nüüd `tänase päevaga ja kas ta täitup: mingisuguse väikse `grupi näol või ta tilgub üks (.) `ühe inimese kaupa (.) ma ei oska `öelda siis kui > tõenäoliselt on siis õige `aeg kui te otsustate et te tahate `sõita et teil on `aeg ja ja see marshuut `sobib teile. < (357_a6)

Intonatsioonilangus ei tähenda alati, et lausung on lõpetatud. Nii võib vahel ühe grammatilise terviku kokku anda alles kaks lausungit. Näites (3) näeme, kuidas parandatav osa paikneb eelmise lausungi lõpus ja parandatud osa uue lausungi alguses.

(3) see maaalune ekskursioon kestaks kuskil `täiendavalt tund aega siis jääks lihtsalt `Pihkva peal aega liiga **`vähe. (0.5) väheks**=et=noh=et `see peaks olema juba nagu `kahepäevane reis võipola siis. (357_a6)

Lausungid, mis sisaldavad pealerääkimist, analüüsiti samuti, juhul kui sõnad on välja kuuldud (pealerääkimised on transkriptsioonis märgitud []). Lausungid, millest mingid osad on kuulamisel ebaselgeks jäänud (transkriptsioonis märgitud {}), võetakse ka analüüsis arvesse ja neid analüüsitakse täpselt nii palju, kui on analüüsida, nt

(4) [ahhaa.] las- lastel [on nagu {-}] (359_a3)

Kui kõneleja vooru keskel annab kuulaja tagasisidet, kasutades selleks tagasisidepartikleid *mhmh*, *ahah*, siis sellest tulenevad kõneleja lausungi hakkimised erinevatele ridadele on paranduste esinemise korral analüüsi jaoks kokku liidetud. Nt

(5) V: loomu`likult. .hh `mina ei ole kahjuks `ise `isiklikult seal `veel `käinud,
H: ahhaa
V: sinna veel `sisse `saanud.

Lõpptulemus näeb sellisel juhul välja selline:

(6) loomu`likult. .hh `mina ei ole kahjuks `ise `isiklikult seal `veel `käinud, sinna veel `sisse `saanud. (356_a4)

2.3. Märgendamine

Analüüsi käigus märgendatakse vastavate märgenditega (vt tabelit 3) kõik voorusisesed kordused, parandused, valestardid, samuti täidetud pausid ja partiklid, kui need esinevad katkestuskohas. Märgendamise käigus määratakse paranduse ja korduse asukoht lausungis, parandatav osa ja selle uus variant ning katkestuskoht.

Eeskuju on võetud Switchboardi korpuse mittesoravuste märgendamisjuhendist (vt ka ptk 1.2.3.). Käesoleva töö autor on Switchboardi märgendamisskeemi suulise eesti keele normaliseerimiseks mõneti täiustanud. Nimelt, kasutusele on võetud lühendid *RP* ja *RE*, et mittesoravuste märgendid oleksid transkriptsioonimärgenditest lihtsamalt eristatavad ja ei kattuks pealerääkimist märkivate märgenditega. Kuna

Switchboardi märgendamisskeemist võeti üle inglisekeelsed lühendid *D* ja *F* märkimaks (diskursuse)partiklit ja täidetud pausi, siis ühtluse huvides võeti ka parandust ja kordust märkivad lühendid *RP* ja *RE* üle inglisekeelsetest sõnadest *repair* ja *repetition*. Lühendid *RP*, *RE*, *D*, *F*, *A* ja *X* määratlevad sulu sisu, st kas tegu on paranduse, korduse, partikli, täidetud pausi, lisamise või analüüsimatu üksusega.

Märgendamise tulemuseks peaks olema pärast parandatava osa eemaldamist süntaktiliselt korrektne lausung. Järgnevalt tutvustatakse lühidalt, kuidas ühte või teist mittesoravust märgendati.

Märgendid	**Seletus**	**Näide**
[RP…+…]	parandus	**[RP** selli- **+** sellist**]**
[RE…+…]	kordus	**[RE** nii **+** nii**]**
+/	valestart	siis saate sealt **+/** minu=arust ´dekanaat väljastab niisugused ´tõendid
{D…}	partikkel	**{D** nagu**}; {D** noh**}**
{F…}	täidetud paus	**{F** ee**}; {F** õ**}**
{A…}	lisamine	ja sis lapse hind on: **{A** ma vaatan siin vel**}** kaks tuhat öheksa`sada
{X…}	analüüsimatu	meil kül `präegu sin `kohapeal **{X** meil**}** `sellist vari`anti ei `paista.

Tabel 3. Normaliseerimisel kasutatavad märgendid

Parandamisse kaasatud lausungi osa, nii selle algus kui ka lõpp markeeritakse nurksulgudega ([]). Katkestuskoht, st see koht lausungis, kus kõneleja peatub ja hakkab oma lausungit ümber formuleerima, märgitakse plussmärgiga (+), nt

(7) aga mingit ´firmat kes seal **[RP** seda **+** (.) pügi´veoga**]** ´tegeleb.
(347_a2_info)

Paranduse algus ja lõpp pole alati üheselt määratav, st millise sõnani läheb kõneleja oma paranduses tagasi ja millise sõnaga lõpeb kogu parandus. Kui algust ja lõppu on olnud raske määrata, siis on valitud vastavalt süntaksile ja semantikale väikseim

võimalik parandatav ja parandatud osa. Näites (8) on parandatavaks osaks *ned ´ainete pe-* ja parandatud osaks *ned ´ained*, mitte nagu võiks konteksti põhjal arvata *ned ´ained nagu oma ´peaainete=ja kõrvalainete ´kaupa*.

(8) aga **[RP** ned ´ainete pe- **+ {D** tähendab selles mõttes**}** ned ´ained**]** nagu oma ´peaainete=ja kõrvalainete ´kaupa te peaksite nendes (0.5) ´õppetoolides ´ ära kontrollima (338_a3)

Sageli on raske ka vahet teha, kas eelnev üksus on pooleli jäetud ja alustatud uut või on tegu parandamisega, nt

(9) kas te k- **+** mille põhjal te nagu ´otsustate seda

Kõikide paranduste korral, mis leiavad aset lausungi algul, on raske vahet teha, kas tegu on parandamise või valestardiga. Lausealguse puhul on ka vähe informatsiooni, mille põhjal otsustada, kas mõte jäetakse pooleli või hoopis parandatakse seda. Seega märgendatakse ühtlase tulemuse huvides kõik sellised sõnakatked alati parandusteks. Nii saab eelmine lausung sellise analüüsi:

(10) **[RP** kas te k- **+** mille põhjal te**]** nagu ´otsustate seda (373_a5)

Kuid on ka erandeid. Näites (11) on poolikuks jäänud lausung pealeräägitud osas. Helistaja tunneb huvi, et kas `bussi peale peaks=ee (0.5) nagu: `oma sööki ka kaasa võtma või kui `bussiga sõit on siis saab nagu bussist ka midagi. (.) [no=ütleme=ku mingid `soojad `supid või]. Helistaja lõpetab oma küsimuse, järgneb mikropaus ((.)), misjärel hakkab helistaja oma küsimust täpsustama ja vastaja samal ajal eelmisele küsimusele vastama. Vastaja annab helistajale õiguse lausung lõpetada, loobudes enda omast. Seega saab vastaja lausungit pidada ainult valestardiks.

(11) [no=bussis=sa- **+/** (0.5) jaa, (.)] **[RE** bussis + **{F** ee**}** bussis**]** `kindlasti on (.) kuum `kohv, (.) `tee, (.) `puljong. (359_a3)

On ka väga selgeid näiteid sellest, kus üks lausung jääb pooleli ja alustatakse uut, nt küsimusele, millist torti soovitakse, vastab helistaja järgmiselt:

(12) ma=i=´oska nagu nimodi kohe täpselt ´öelda et noh selline: **+/** .hh m ´tegemist on ühe üliõpilasorganisatsiooni ´aastapäevaga. (380_a7)

Valestartideks on analüüsitud ka need juhud, kus kõneleja justkui plaanib jätkata loetelu või alustab osalauset teatud konjunktsiooniga, kuid järgnev (osa)lausung sugugi ei seostu sellega. Sellisel juhul markeeritakse vastav konjunktsioon märgenditega **/** ja **+/**, nagu on seda tehtud järgnevas näites.

(13) ja `millega te tahate minna kas te tahate seal .hh ee mingi `transpordiga **[RE[[**=või + või**]** + või**]** + või**]** =`jalgrattaga **/** =või **+/** (.) [noh] neid vari`ante võib ka `erinevaid olla. (353_a3)

Samamoodi märgendatakse ka sellised valestardid, kus kõneleja jätab lausungi keskel fraasi/osalausungi poolikus, hüpates teise peale üle, nt

(14) kui teil on tudengipakett `ka juba tehtud, (.) sis peale (.) esimest ok`toobrit, (.) tulete uuesti `panka, ja=s **/** saate ai noh **+/** (.) `kohe tehakse teiega laenu`leping. (334_a5)

Kordusteks on analüüsitud kõik nn puhtad kordused, st kus korratud on ühte ja sama sõna, nt

(15) .hh **[RE** aga **+** aga**]** noh siis ´seletate (338_a3)

Lisamised (kiilungid ja sabalisandused) on esile tõstetud märgendiga {A…}.

(16) mnjah, ma vaatan präegu=et sellist nagu ei `paistagi siin **{A** sellist vari`anti**}** etkel. (475_b14)

Paranduste ja korduste katkestusosas esinevad partiklid ja täidetud pausid on samuti analüüsitud, need on märgendatud vastavalt {D…} ja {F…}. Sagedasemad partiklid, mis viitavad lausungi ümbertegemisele, on teatud variatsioonidega *tähendab*, *selles mõttes*, *noh*, ning järgnevas näites esinevad *nagu* ja *ütleme*. Täidetud pausidest on kõige sagedasem *ee*.

(17) ma teen **[RP** sellisel **+ {D** =nagu=ütleme**}** ´suurele ´plaadile**]** (380_a7)

Leksikaalsete parandustele viitajate kõrval on olemas ka prosoodilised – pausid ja rõhud. Kui katkestuskohale ei viita leksikaalne informatsioon, siis võib sellele viidata paus, nt

(18) siis on äkki vaja se **[RE** `pistik **+** (.) pistik**]** seal `lihtsalt ära vahetada. (475_b14)

Eelpool toodud näitelausungeid võib nimetada hästi moodustatud parandusteks, st neil on selge parandatav osa ja kindel parandatud osa. Kuid on ka selliseid parandusi, mille kohta võib öelda, et nad on nö halvasti moodustatud (ingl *non-well formed self repairs*) (McKelvie 1998: 8), mistõttu on neid ka keerulisem märgendada/normaliseerida. Näites (19) on näha kõneleja pingutusi öelda sõna põhjapõdrasõit. Kõneleja katkestab kaks korda sõna, enne kui saab öeldud *põdrapõhjasõit*, mis pole samuti see, mida ta tegelikult öelda tahtis. Parandussekvents jätkub sissehingamise ja täidetud pausiga, pärast neljandat katset saab kõneleja öeldud, mida oli tahtnud – *põhjapõdrasõit* – lõpetades lausungi parandamisele viitava markeriga *vabandust*.

(19) see `mootorkelgusõit **[RP[**jõ- + `põdrapõhja**[RP** sõi- + sõit**]]** + .hh **{F** ee**}** $ `põhjapõdrasõit $ **{D** vabandust**}]** (367_a4)

Kui märgendaja pole suutnud parandusi sobivalt märgendada, on need kas üldse märgendamata jäetud või on lausungi soravuse huvides mõni lauseliige märgendatud {X…}, nt

(20) et ühesõnaga ma=san=aru **{X** =et=see aast- kuni ühe`teistkümnendast üheteistkümnenda `jaanuarini**}** =et=siss ühe`teistkümnendast `jaanuarist peaks sis olema `odavam. (384_a2)

Märgendatud parandused, kordused ja valestardid on esitatud töö lõpus, vastavalt Lisa 2, Lisa 3 ja Lisa 4.

3. Märgendatud materjali analüüs

Järgnevas peatükis analüüsitakse märgendatud parandusi, kordusi ja valestarte. Esmalt tutvustatakse, kuidas märgendatud nähtused on jaotatud, seejärel on vaadeldud iga rühma eraldi kahest vaatepunktist. Kõigepealt, kuidas kõneleja on ühte või teist parandusvõtet kasutanud, ja seejärel, kas ja kuidas on kõneleja järgnevast parandusest/kordusest kuulajale märku andnud, st kas on teatud kindlaid prosoodilisi ja leksikaalseid markereid, mis viitavad parandamisele. Selliste markerite olemasolu aitaks mittesoravused automaatselt tuvastada.

Hennoste (2000: 1143) on kõne planeerimisega seotud võtted jaganud vastavalt sellele, kas probleem on mõnes juba väljaöeldud tekstiosas või on see alles väljaütlemata tekstiosas, kaheks: edasilükkamiseks ja parandamiseks e reformuleerimiseks. Edasilükkamisvahenditena on ta välja toonud puhtad pausid, täidetud pausid e üneemid, partiklid, kordused ja venitused (Hennoste 2000: 2693–2697). Reformuleerimise tüüpidena on ta esitlenud lisamist ja asendamist. Lisaks kaks analoogset võtet: kiilumine lisamise alaliigina ja loobumine asendamise alaliigina. (Hennoste 2000: 2701–2708). Käesolevas töös vaadatakse mittesoravusi mõneti erinevalt. Vastavalt parandamise struktuurile, kas midagi asendatakse, lisatakse, korratakse või loobutakse millestki, vaatlen eraldi parandusi, kordusi ja valestarte (ning nende alaliike).

Vastavalt **paranduste** struktuurile olen need jaganud neljaks.

1) Sõnakatked: sõna jääb pooleli, millele transkriptsioonis viitab sidekriips (-), nt
   (1) see on nüd **[RP** mö- + möb´leerimata**]** korter. (354_a4)

2) Asendused: asendamise käigus kõneleja loobub millestki, asendades selle grammatiliselt/semantiliselt täpsema sõna/fraasiga, nt
   (2) ikkagi=noh `erinevatel `päevadel=on võimalik=sis **[RP** `mägi + mäge**]** valida. (355_a11)

3) Lisamised: lausungi keskele või lõppu pannakse midagi täpsustavat, selgitavat juurde, nt

(3) tahaks `teada kas: on teil `andmeid (0.5) laste: (.) **{A** mitte `beebikoolide aga noh selliste `kaheaastaste laste**}** (.) .hh mingite `laulu`ringide kohta. Tartus. (475_b3)

(4) Tartu `kristlik `perekeskus, siin on nüt `väikelaste laulu`ring > **{A** või=tähndab beebide `ja väikelaste kuus `kuud kuni neli `aastat**}**. (475_b3)

4) Muu: siia alla on liigitatud juhtumid, mida on raske üheks või teiseks pidada, nt

(5) jah, et kui päris niiöelda võ- `väikese reisi Ka`naari saartele ei `võta sis iga `maa pakub ikka väga palju $ [`huvitavat] ku `vaadata. $ (233_a4)

Kordused olen jaganud kaheks vastavalt sellele, kas korduse vahel esineb katkestuskoht või ei:

1) Kordused katkestuskohaga: katkestuskohas võivad esineda (täidetud) pausid või partiklid, nt

(6) ee talvel muidugi **[RE** saaks (.) + saaks**]** seal nagu (.) `suusatada (353_a3)

2) Kordused katkestuskohata, nt

(7) aga meil on: niimoodi=et meil .hh `bussi`juht (.) seisab `järje`korras [ja] `elavast järjekorrast **[RE** võtab + võtab**]** meile `piletid. (356_a4)

3) Muu: siia alla olen paigutanud ahelkordused, nt

(8) töö peate kaitsma eks ´õigeaegselt sis, .hh et te saaksite kevadel ´lõpetada **[RE[[[**=ja + .hh ja**]** + ja**]** + ja:**]** (.) + ja**]** =nii=´edasi. (338_a3)

Valestarte olen käsitlenud omaette rühmana, kuigi, nagu juba märgendamise peatükis (ptk 2.3.) tõdeti, on vahel raske vahet teha parandusel ja valestardil. Valestarte eristab parandustest see, et valestardil puudub parandatava ja parandatud osa vahel seos, nii süntaktiline struktuur kui ka semantika muutuvad (kuigi mitte alati). Nt järgnevas näites jääb küll lausung süntaktiliselt pooleli, kuid kõneleja jätkab siiski sama teemat:

(9) .hh < no kuskil:: ´kolmkümmen:d > oota se oli: **+/** (0.5) päev enne esimest detsembrit on vist (.) ´kolmkümmend. (.) no´vember. (380_a7)

Mõnikord on kõneleja kasutanud mitut parandamisvõtet samaaegselt. Järgnevas näites on tegu nii sõnakatke kui ka lisamisega. Sellistel juhtudel on näide alati liigitatud sõnakatkete alla, kuna poolikuks jäänud sõna on kõige kindlam marker, et toimub parandamine, nt

(10) ahah. .hh e tähendab siss=ee **[RP** te so- + kas te soovite**]** muidu seda lsik
(.) Maestro kaarti ka `taotleda või. (334_a5)

Tabelis 4 on arvuliselt välja toodud, kui palju ühte või teist mittesoravust analüüsitud 35 dialoogist leiti.

Liik			**Kokku**
Parandused	Sõnakatked	53	131
	Asendused	50	
	Lisamised	13	
	Muu	15	
Kordused	Katkestuskohaga	46	113
	Katkestuskohata	52	
	Muu	15	
Valestardid			33

Tabel 4. Erinevate mittesoravuste esinemus

Järgnevalt ei otsita niivõrd vastust küsimusele, miks kõnelejad end parandavad, kuivõrd sellele, kuidas nad seda teevad ja kas on mingi kindel marker, mis viitab parandamisele. Eeskätt pakuvad huvi prosoodilised ja leksikaalsed markerid. Praegusel kujul eesti keele kitsenduste grammatika transkriptsioonimärke ei arvesta (need eemaldatakse enne analüüs), mistõttu igasugune prosoodiline info läheb hetkel veel kaotsi. Seega saame praeguse süntaksianalüsaatori tööd parandada/mõjutada vaid teatud leksikaalsete indikaatoritega. Paljud uurimused on näidanud, et kui parandusele viitab kindel katkestuskoht, siis on süntaksianalüsaator võimeline parandamiste analüüsiga hakkama saama. Kuid samas on ka paljudes uurimustes tõdetud, et otsitud leksikaalseid markereid lihtsalt pole. Analüüsides suulist rootsi keelt märkis Eklund (2005: 255), et tema materjalis leiduvatest parandustest sisaldas vaid 1% katkestuskohas esinevaid viiteid. Suulist soome keelt analüüsides on leitud, et kõneleja annab parandustest teatud leksikaalsete vahenditega märku 14% juhtudest, ülejäänud 84% juhtudest sõnakatketega

(Sorjonen, Laakso 2005). Suulise eesti keele parandusi analüüsides on Hennoste (2006) leidnud, et 35% juhtudest kasutab kõneleja parandusele viitavaid leksikaalseid markereid ja 65% juhtudest sõnakatkeid. Vaatame, kui palju prosoodilisi ja leksikaalseid viiteid käesolevas materjalis leidus ning kuidas markeerivad kõnelejad kordusi ja valestarte.

3.1. Parandused

3.1.1. Sõnakatked

Sõnakatkestamist võib pidada ainsaks monofunktsionaalseks parandamisele osutajaks (Hennoste 2000: 2705). Kuna siia alla on koondatud kõik lausungid, mis sisaldavad sõnakatkeid, siis see klass on oma struktuurilt väga erinev.

Sagedasemad ja lihtsamad sõnakatked on need, kus sõnast on välja öeldud üks silp ja seejärel lausutakse parandamise käigus kogu sõna (näide 11). Clark ja Wasow (1998: 226) on leidnud, et sõnakatked on põhjustatud katkestusest fonoloogilisel tasandil. Nende väide kehtib suurel osal sellistel sõnakatketel, nagu näeme näidetes (11), (12), (13), kuid ei kehti näidete (14), (15), (16), (17), (18) ja (19) puhul, kus on selgelt näha, et katkestus on tekkinud semantilisel tasandil, st kõneleja on ühe sõna lausumise jooksul otsustanud seda kuidagi täpsustada, asendada parema sõnaga.

(11) `tagasijõudmine on no õhtul ütleme **[RP** tav- + tavaliselt**]** jõuavad nad kell kaheksa ka `tagasi. (356_a4)

Selliseid sõnakatkeid on 52 näite hulgas 32. Siia alla on arvestatud kõik need juhud, kus pooleli jäänud sõnad asendatakse ühe sõnaga, ükskõik, kas öeldud esitäht või silp kattub hiljem tervenisti välja öeldud sõnaga. Nii olen siia hulka lugenud ka sellised juhud, kus sõna algus ei ühti tervenisti lausutud sõnaga, nt

(12) saate (.) uue **[RP** pa- + püsipa`rooli**]** sis (398_a6)

Samuti olen siia alla liigitanud liitsõnade parandamise, kus esimest poolt üle ei korrata, korrigeeritakse ainult liitsõna teist poolt. Märgendamise seisukohast olen probleemi

lahendanud järgmiselt. Kuna kõneleja parandab ainult liitsõna teist poolt *varustus* (näide 13), siis on vastavalt ka märgendatud, st liitsõna esimene pool *suusa* on märgendamisest välja jäetud. Nii saame tulemuseks sõna, mida kõneleja esimesel katsel ei suutnud lõpuni viia – *suusavarustus*.

(13) suusa**[RP** varut- + varustus**]** kas=võtate kaasa või rendite `kohapealt (355_a11)

Võrdlusena võib tuua analoogse näite, aga kõneleja kordab ka liitsõna esimest poolt, võttes seejärel veel väikse mõttepausi ning seejärel lõpetades parandussekventsi, nt

(14) .hh ja siis **[RP** linnaval- + linna (.) hooldus´amet**]** on siin antud otseselt (347_a2_info)

Üks näide on ka sellest, kus transkribeerimise käigus pole poolik sõna saanud endale sidekriipsu, sest kõneleja on poole sõna pealt täidetud pausile üle läinud, mis on vastavalt ka transkribeeritud, nt

(15) ee suusareise pakume **{X** tõe=õõ**}** te **[RP** arvat + **{F** =õõ**}** mõtlete**]** nüüd `veebruarikuus=`jah? (359_a3)

Sõnakatkete kõrval, kus ühte poolikuks jäänud sõna parandatakse teisega, on sõnakatked lisamisega, st kõneleja tahab poole sõna peal fraasi täpsustada/ümber formuleerida, nt

(16) jah, (0.5) ästi, **[RP** õhe- + natuke õheksa**]** `läbi kolma`päeval. (259_b7)

(17) ja kindlasti `tagasiteel ee võib `ka **[RP** kusagil Pa- + teha kas siis `Paide lähedal või kusagil**]** .hh `peatus (356_a4)

Üksikuid näiteid on ka sellest, kus kõneleja jätab pooliku sõna/fraasi parandatud osas üldse välja, asendades selle hoopis teise sõnaga, nt

(18) [et] on `näha et **[RP** nende (.) mingi sisse- + midagi**]** laekub `panka. (334_a5)

(19) =nii=et sinna lähevad ilusti nii `suusad **[RE** kui + kui**]** =ee **[RP[**varus- + **{D** või**}** see**]** + **{F** ee**}** `kotid**]** (359_a3)

Prosoodilised ja leksikaalsed markerid. Nagu esitatud sõnakatkete näidete puhul näha, on nende ainsaks ühiseks jooneks see, et parandatavas osas on üks sõna poolikuks jäänud, mis on enamusel juhtudel transkriptsioonis märgitud sidekriipsuga, mis viitab kindlalt parandusele. Sõnakatkete parandamise mudeli võiks kirja panna järgmiselt:

[RP sõnakatke- **+** {F/D/(paus)/Ø} parandatud osa**]**

Näeme, et lisaks sõnakatkele võivad parandamisele katkestuskohas viidata ka (täidetud) pausid ja partiklid. Kuid 79% juhtudel viitab parandamisele üksnes poolikuks jäetud sõna.

Näidetest ilmnes, et paranduses olevate sõnade hulk kõigub kahest kuni üheksa sõnani. Sõnakatked on küll automaatse paranduste tuvastamise seisukohast parim marker märkimaks paranduse esinemist, kuid teisest küljest on nende automaatne parandamine jällegi sedavõrd raskem, kuna parandatud sõnade hulk ja kasutus on väga varieeruvad.

3.1.2. Asendused

Asendamise kõige lihtsama klassina võib välja tuua ühe sõna asendamise teisega, sellistel juhtudel tavaliselt katkestuskoht puudub, puudub nii edasilükkav paus kui ka parandamisele viitav leksikaalne marker:

- asendamine sisult täpsema sõnaga, nt

(20) see **[RP** tegeleb + tegutseb**]** nüd aadressil `Puusepa `kümme, siin on `laulumängu`ring. (475_b3)

- asendamine grammatiliselt korrektsema sõnaga, nt

(21) siin on `märgitud=et=ee (.) **[RP** kahe + kahest**]** kuni `viie tärni hotellini (384_a2)

Palju on selliseid parandusi, kus kasutatakse edasilükkamispartiklina pronoomenit *see*. Näites (22) alustab kõneleja fraasi *sellist*, mille ta pärast mitut edasilükkamist asendab

sõnaga *süvenemiseks*. Selliste asenduste puhul on väga sage just see, et parandatav ja parandatud osa on erinevates käänetes.

(22) ähh ma kardan jaa et meil=on `aega selleks üsna `vähe=et meil on siin kliendid `käivad **[RE** =et + (.) hh et**]** meil lihtsalt **[RP** sellist + **{D** nii-öelda**}** (0.5) .hh `süvenemiseks**]** pole eriti `aega. (475_b14)

Samas edasilükkavas funktsioonis on ka järgnevas näites esinev proadverb *sinna*. Konteksti sobiva käände on kõneleja paika pannud, edasi otsib ta sobiva semantilise sisuga sõna, nt

(23) [jah,] vot ´seda peaksite nüd minema **[RP** oma sinna: + oma ´õppetooli**]** kus te ´kaitsete. (338_a3)

On ka keerukamaid ja pikemaid asendusi, kus kõneleja on parandusprotsessi haaranud rohkem sõnu, mõnda korrates, mõnda parandades. Kõneleja tahab parandada ainult fraasi *homme hommikul*, kuid alustab kogu küsimust uuesti, nt

(24) `kui ma `homme `ostan `ära ühe (.) `voodi, (.) **[RP** kas see > homme `hommikul, **+** kas see `tuuakse mulle päeva=jooksul**]** `ära ka. (96_b8)

Asendamise 50-st näitest 22-s ei markeerinud kõneleja kuidagi parandamise järgnemist. Ülejäänud 28 juhul on kõneleja kasutanud ühte järgnevaist võtteist (mõnel juhul ka mitut koos):

- partiklid (7), nt *tähendab*, *või*, *või niimoodi et noh*, *nagu ütleme*, *nii-öelda*;
- parandatava osa lõpus osaline intonatsioonilangus (7);
- katkestuskohas sissehingamine (3);
- paus (4);
- täidetud paus (3);
- sõna viimase silbi venitamine (3).

Prosoodilised ja leksikaalsed markerid. Asenduste parandamise mudel näeb välja selline:

[RP asendatav **+** {F/D/(paus)/Ø} parandatud osa**]**

Nagu näha võib, on katkestusosas jällegi kõik variandid võimalikud: kui kõneleja annab kuulajale leksikaalsete vahenditega märku parandamisest, siis seda võib ta teha nii (täidetud) pauside kui ka partiklitega. Lisaks tuli huvitava asjana välja, et kõneleja võib parandamist väljendada ka parandatava osa lõpus poollangeva intonatsiooniga. Samuti parandatava osa viimase sõna viimase silbi venitamisega, mis on transkriptsioonis märgitud kooloniga (:).

3.1.3. Lisamised

Lisamist sisaldavates lausungites ei ole midagi üleliigset, st kõik öeldu kuulub lausungi juurde. Lisamiste alla olen paigutanud kiilungid ja sabalisandused. Kiilumise korral paigutatakse ühe süntaktilise üksuse keskele teine. Kõneleja katkestab pooleli oleva lausungi süntaktilise konstruktsiooni, alustab uut, viib selle lõpule ning seejärel lõpetab pooleli jäänud põhilausungi (Hennoste 2000: 2703). Kaks kiilungit, mis leidusid, on mõlemad pärit reisibüroodialoogidest ja seotud infootsimisega, nt

(25) ja sis lapse hind on: **{A** ma vaatan siin vel**}** kaks tuhat öheksa`sada (359_a3)

Leidub ka kiilungeid, mis sisaldavad endas teatavat parandamist/täpsustamist, nt

(26) .hh tere. (.) .hh tahaks `teada kas: on teil `andmeid (0.5) laste: (.) **{A** mitte `beebikoolide aga noh selliste `kaheaastaste laste**}** (.) .hh mingite `laulu`ringide kohta. Tartus. (475_b3)

Teiseks lisamisevõtteks on sabalisandused (vt ka Hennoste 2001: 197). Kirjalikus keeles leidub neid samuti. Ülejäänud lausest eraldatakse need koma(de)ga. Näites (27) näeme, kuidas kõneleja on suuliste vahenditega sama tulemuse saavutanud, st pealausungi lõpus poollangev intonatsioon, mille järel pealausungi objekti täiendav lisa, nt

(27) .hh võtaks selle `mängumaa `ka, **{A** Anni `mängumaa**}**. (475_b3)

On ka selliseid segasemaid lisamisi, kus täpsustav info öeldakse lausungis suvalise koha peal. Järgnevas näites on kõneleja vajalikuks pidanud veel laiendada fraasi *selle pistiku* sõnadega *selle präeguse oleva*, nt

(28) keegi peab selle `pistiku nagu otsast `ära: võtma **{A** selle `präeguse oleva**}** ja sis sinna teise pistiku nagu `asemele panema. (475_b14)

Prosoodilised ja leksikaalsed markerid. Kindlad markerid puuduvad. Kui kõrvale jätta kaks erandit, siis võib väita, et kui kiilungitele, sabalisandustele ja ühe lausungi sisestele lisamistele midagi viitab, siis on selleks prosoodiline info (pausid, venitamine, poollangev intonatsioon).

3.1.4. Muu

Siia alla olen koondanud lausungid, mida on ühel või teisel põhjusel olnud raske märgendada või milles üks liige on saanud märgendi {X…}. Näites (29) näeme, kuidas kõneleja on püüdnud mitu korda minna lausungiga edasi, neljandal katsel on see tal ka õnnestunud.

(29) H: et ühesõnaga ma=san=aru **{X** =et=see aast- kuni ühe`teistkümnendast üheteistkümnenda `jaanuarini**}**=et=siss ühe`teistkümnendast `jaanuarist peaks sis olema `odavam. (384_a2)

3.2. Kordused

Hennoste (2000: 2696) on teinud vahet parafraasil, katkestatud kordusel ja edasilükkaval e puhtal kordusel. Selle määratluse järgi käsitlen ma puhtaid kordusi, st ühe sõna kordamist. Vastavalt sellele, kas korduses esineb katkestuskohas mõni takerdus või mitte, olen jaganud kordused katkestuskohaga ja katkestuskohata kordusteks ning ahelkordusteks.

3.2.1. Katkestuskohaga kordused

Katkestuskohaga korduste alla olen lugenud kordused, mis sisaldavad
- pause, nii mikropause kui ka pikemaid (näide (30) ja (31)). Hennoste (2000: 2025) on nimetanud neid takerduspausideks. Need on parandussegmendi pausid, millega lükatakse teksti moodustamist edasi või tehakse parandusi.

(30) siis on äkki vaja se **[RE** `pistik + (.) pistik**]** seal `lihtsalt ära vahetada. (475_b14)

(31) vaatame ültse **[RE** kas + (0.8) kas**]** `on lende. (403_a1)

- sissehingamist, nt

(32) aga noh kui te tahate nüd **[RE** oma + .hh oma**]** ütleme `seltskonnaga `minna (353_a3)

- esimese sõna viimase silbi venitamist, nt

(33) et **[RE** kas: + kas**]** on võimalik septembris ära teha. (428_a36)

- täidetud pause, nt

(34) **[RE** bussis + **{F** ee**}** bussis**]** `kindlasti on (.) kuum `kohv, (.) `tee, (.) `puljong. (359_a3)

Valdav osa kordustest sisaldab ühe sõna kordamist. Üksikud näited on ka sellest, kus on korratud fraasi, nt

(35) kokku=on ültse **[RE** kaheksa päeva: + kaheksa päeva**]** seda `sõitu (355_a11)

Korratud on nii erinevaid sõnaliike kui ka lauseliikmeid. Enim on aga konjunktsioonide kordamist, lausa 63%. See toetab ka Hennoste (2000: 2696) väidet, et edasilükkamis-kordustest moodustavad konjunktsioonid pea 2/3.

Tabelis 5 võib näha, milliseid võtteid on kõneleja kordamisel kasutanud. Kui kõneleja on korraga kasutanud mitut võtet, on neid eraldi loetud. Näeme, et kordamisel kasutab kõneleja katkestuskohas võrreldes parandamisega tunduvalt rohkem pause ja venitamisi. Samuti on 16 juhul katkestuskoht markeeritud sissehingamisega.

Prosoodilised ja leksikaalsed markerid. Katkestuskohaga korduste mudel näeb välja selline:

[RE korratav **+** {F/D/(paus)} korratud**]**

Seega võiks teiseks kindlamaks formaalseks paranduste indikaatoriks sõnakatkete kõrval pidada venitamisi. Ka asenduste puhul nägime, et mõnel juhul viitas parandamisele just sõna viimase silbi venitamine. Käesolevas töös ei saa seda väidet ei ümber lükata ega toetada, sest selleks oleks vaja rohkem materjali nii mittesoravate kui ka nn valehäiret andvate venitamiste kohta.

Kordused katkestuskohaga			Kokku
Pausid	(.)	14	21
	(0.5)	4	
	(0.8)	2	
	(1.2)	1	
Täidetud pausid	ee	6	9
	ää	2	
	noh	1	
Sissehingamised			16
Venitamised			16

Tabel 5. Korduste katkestuskohas esinevad prosoodilised ja leksikaalsed markerid

3.2.2. Katkestuskohata kordused

Kõige rohkem on muidugi korratud konjunktsioone *et* ning *ja*, kuid ka nt eestäiendeid.

(36) `rõhk oli seal `käsitööle **[RE[** =ja mt + ja**]** + =ja**]** noh sellistele .hh ee **[RE** kodus + kodus**]** valmistatud `esemetele. (353_a3)

(37) `praegu on meil ainult see **[RE** `teine + teine**]** no`vember. (356_a4)

Prosoodilised ja leksikaalsed markerid. Katkestuskohata korduse mudel näeb välja väga lihtne, see sisaldab korratavat ja korratud elementi.

[RE korratav **+** korratud**]**

Leksikaalne viide katkestuskohas kordamisele puudub, kuid on olemas teine viide – korratud sõna ise.

3.2.3. Ahelkordused

Siia alla olen paigutanud kõik ahelkordused, kuna nende puhul oli raske otsustada, kumma eelneva alla neid paigutada, nt

(38)´töö peate kaitsma eks ´õigeaegselt sis, .hh et te saaksite kevadel ´lõpetada **[RE[[[**=ja + .hh ja**]** + ja**]** + ja:**]** (.) + ja**]** =nii=´edasi. (338_a3)

3.3. Valestardid

Valestartide kaht erinevat märgendamist tutvustati juba märgendamise osas (vt ka ptk 3.).

Leidub selgeid valestarte, nt

(39) sellistesse **+/** no üldiselt on nimodi=et meil on `õnnestunud (357_a6)

On selliseid, kus konjunktsioon näitab, et tulema peaks teine lausungi jätk, kui tegelikult tuleb, nt

(40) no muidugi tehakse `peatused **/** sest no **+/** (0.5) esimese peatuse kindlasti teete juba `Lätis. (359_a3)

Lisaks on olemas sellised valestardid, kus kõneleja alustab lausungit, siis jätab selle pooleli, sest kõneleja arvates tuleks kuulajale veel lisainformatsiooni anda, ning seejärel alustab uuesti pooleli jäänud lausungit, nt

(41) ja siis nüd=ee **+/** (.) `**kaua see +/** (.) tändab=et see `aeg seal `veepargis. **kaua `see** on. mitu [`tundi.] (356_a4)

Prosoodilised ja leksikaalsed markerid. Valestardi mudel on järgmine:

parandatav lausung **+/** {(paus)/venitamine/Ø} uus lausung

Materjali põhjal võib teha samad järeldused nagu ka lisamiste puhul, et kui teatavad parandamisele viitavad markerid eksisteerivad, siis on need samuti prosoodilised. 32 näitest 17-s esines valestardi lõpus ja enne uue lausungi algust paus või sissehingamine. Ülejäänud juhtudel oli üleminek uuele lausungile sujuv.

3.4. Kokkuvõte

Kokku tuvastati 13 168-sõnalises korpuses 131 parandust, 113 kordust ja 33 valestarti. Seejärel otsiti leksikaalseid ja prosoodilisi markereid, millele oleks võimalik mitte-soravuste automaatsel analüüsil toetuda. Sõnakatkete puhul saab rääkida kindlast parandamise markerist. Selleks on sidekriips, mis transkriptsioonis viitab pooleli jäänud sõnale. Ülejäänud uuritud mittesoravuste puhul saab rääkida vaid võimalikest

markeritest. Asenduste näidetes oli katkestuskoht markeeritud 56% juhtudest. Selle markeerimiseks olid kõnelejad kasutanud erinevaid võtteid (mõnel juhul ka mitut koos): leksikaalsetest võtetest partikleid ja täidetud pause, prosoodilistest võtetest parandatava osa lõpus poollangevat intonatsiooni, sissehingamist, pause ja sõna viimase silbi venitamist. 98 ühesõnakordamistest 47% juhtudel markeeris kõneleja katkestuskoha, kasutades selleks leksikaalsetest markeritest täidetud pause ja prosoodilistest pause, sissehingamist ja esimese sõna viimase silbi venitamist. Lisamiste ja valestartide näiteid analüüsides ilmnes, et kui lausungis neile kahele üldse midagi viitab, siis on nendeks prosoodilised markerid, nagu pausid ja venitamised.

4. Eksperiment kitsenduste grammatikal põhineva süntaksianalüsaatoriga

Käesolevas peatükis tutvustatakse eesti keele kitsenduste grammatikal põhineva analüsaatoriga läbi viidud eksperimenti. Enne eksperimendi tulemuste esitamist ja analüüsi antakse põgus ülevaade süntaksianalüsaatori tööst ja selle kohandamisest suulise eesti keele analüüsiks ning sellest, kuidas materjal eksperimendiks ette valmistati.

4.1. Analüsaator ja suuline keel

Kõigepealt tutvustatakse lühidalt eesti keele kitsenduste grammatika süntaksianalüsaatori (edaspidi ESTKG analüsaator) olemust ja tööpõhimõtteid ning seejärel tutvustatakse esimest ESTKG analüsaatoriga tehtud suulise eesti keele analüüsi katset ja selle tulemusi. Peatükk põhineb Müürisepa ja Uibo artiklil (2006).

Eesti kirjakeele kitsenduste grammatika analüsaatori töötasid aastail 1996-2000 välja Tiina Puolakainen ja Kaili Müürisep. Süntaktiline analüüs koosneb kolmest etapist: morfoloogilisele analüüsile järgnevast morfoloogilisest ühestamisest, osalausungipiiride määramisest ja sõnade süntaktiliste funktsioonide määramisest. Lause iga sõna varustatakse teatud hulga märgenditega, mis näitavad selle sõna inflektsiooni- ja derivatsiooniomadusi, sõnaliiki ja süntaktilist funktsiooni. Kitsenduste grammatika süntaks on sõnapõhine, st analüüsi käigus ei leita lause puukujulist fraasistruktuuri, vaid ainult sõnade süntaktiline funktsioon lauses. Fraasipõhjaks võivad olla subjekt, objekt, adverbiaal või predikatiiv. Fraasi laiendid saavad endale samuti märgendid, mis näitavad fraasi põhja leidumise suunda. Põhjad ja laiendid ei ole formaalselt seotud, st laiendid ei viita ühelegi sõnale konkreetselt. Predikaadi märgendeid on viis: finiitne või infiniitne põhiverb või abiverb ja verbi eitus. Süntaktiliste funktsioonide määramine käib nii, et analüüsi alguses lisatakse igale sõnale kõik võimalikud analüüsivariandid ja seejärel hakatakse konteksti mittesobivaid eemaldama. Eesti keele süntaksi arvutigrammatika sisaldab 1240 morfoloogilise

ühestamise reeglit, 47 osalausepiiride määramise reeglit, 180 märgendite lisamise reeglit ja 1118 süntaktilist kitsendust.

Esimese katse analüüsida ESTKG analüsaatoriga suulist eesti keelt viisid läbi Müürisep ja Uibo 2005. aastal (vt ka Müürisep jt 2006). Süntaksianalüsaatori sisendina kasutati juba eelnevalt morfoloogiliselt analüüsitud dialooge (vt ka Hennoste jt 2002). Paralleelselt morfoloogilise analüsaatoriga ESTMORF töötas spetsiaalne oletaja, mille ülesanne on tuvastada sõnu, mis transkribeerimisel on kirja pandud teisiti, kui nende ortograafiline kirjapilt muidu ette näeb, nii saab nt *kolmkend* sama analüüsi kui *kolmkümmend*.

Algselt kirjakeele analüüsimiseks koostatud kitsenduste grammatika kohandati ümber suulisele keelele. Selleks muudeti lausungipiiride tuvastamise reegleid ja mitmeid süntaktilisi kitsendusi. Algselt kirjaliku keele osalausepiiride määramise reeglid vaadati üle, sest kirjavahemärkide tähendus on suulise keele tekstides erinev. Uutes reeglites kasutati intonatsioonimärke, partikleid ja täidetud pause. Kui süntaksianalüsaator kahtleb, kas tegemist on ikka kindla osalausepiiriga, siis lisab ta sõnale oletatava osalause tunnuse CLB-C. Punkti loeb süntaksianalüsaator kindlaks lause lõpu tunnuseks. Partikleid ja täidetud pause käsitletakse lause lõpus tunnustena, kui kummalgi pool kontekstis leidub finiitseid verbivorme. Sisse toodi ka kaks uut märgendit: @B partikli ja @T tundmatu süntaktilise funktsiooniga sõna märkimiseks. Käesoleva töö raames tehtud eksperiment viidi läbi samades tingimustes.

Müürisep ja Uibo testisid kohandatud süntaksianalüsaatorit 2543-sõnalisel argivestlustest koosneval suulise eesti keele korpusel. Süntaksianalüsaatori väljundit võrreldi käsitsi märgendatud korpusega ning tulemused, mis saadi, olid järgmised: saagis 97.3%, täpsus 89.2% ja ühesus 91.5%, kusjuures võrreldes kirjaliku keele analüüsi tulemustega, on saagiste vahe üllatavalt väike, ainult 1,2%. Veatüübid olid järgmised: valest osalausungipiiri määramisest tingitud vead, tundmatu süntaktiline funktsioon, adjektiivi käitumine substantiivina, varasem vale analüüs, kordused. Parandustele, kordustele ja valestartidele eraldi tähelepanu ei pööratud. Peatükis 4.3. tutvustatakse käesoleva eksperimendi tulemusi ja vaadatakse, kas ja kuidas on süntaksianalüsaator selliste suulisele kõnele omaste nähtustega hakkama saanud.

Käesoleva eksperimendi materjali valmistas ette käesoleva töö autor, tehniliselt viis selle läbi Kaili Müürisep.

4.2. Materjali ettevalmistamine

Eksperimendi tarbeks valmistati peatükis 3 esitletud materjal ette järgmiselt. Süntaksianalüsaatorile anti kaks korda analüüsida ühtesid ja samu lausungeid, ainult selle erinevusega, et ühel juhul oli tegu algsete (vt näidet (1)) ja teisel juhul normaliseeritud lausungitega (vt näidet (2)). Kuna mittesoravuste süntaktiline analüüsimine on algjärgus, siis süntaksianalüsaator ei oska arvestada veel mittesoravuste märgenditega (vt näidet (3)), seega kustutati normaliseeritud lausungi saamiseks parandamist vajav osa, samuti katkestuskohas esinenud pausid, partiklid ja täidetud pausid käsitsi. Märgendamisel analüüsimatuks osutunud sõnad (saanud märgendi X) eemaldati samuti.

(1) .hh ta on kuskil: .h tändap viie`teistkümnendal sajandil on ta `asutatud.

(2) .hh viie`teistkümnendal sajandil on ta `asutatud.

(3) .hh **[RP** ta on kuskil: + .h **{D** tändap**}** viie`teistkümnendal sajandil on ta**]** `asutatud

Analüüsitavaid sõnu algsetes lausungites oli

Normaliseeritud lausungites tehti sõltuvalt neis leiduvatest probleemidest enne analüüsi väikeseid muudatusi. Asenduste, korduste ja mõnede lisamistega polnud vaja teha muud, kui kustutada parandamist vajav osa ning katkestuskohas esinenud pausid, partiklid ja täidetud pausid.

Teatud liiki valestartide ja lisamiste puhul tuli aga (kustutamise kõrval) lisada ka kirjavahemärke. Valestartides, mis olid märgendatud nii nagu näites (4), kustutati esmalt kaldkriipsude vahele jääv poolik lausung ja seejärel lisati alles jäänud esimese lausungi lõppu punkt, et tekiks kaks analüüsitavat lausungit (vt näidet (5)).

(4) no muidugi tehakse `peatused **/** sest no **+/** (0.5) esimese peatuse kindlasti teete juba `Lätis. (359_a3)

(5) no muidugi tehakse `peatused. (0.5) esimese peatuse kindlasti teete juba `Lätis. (359_a3)

Kiilude puhul eraldati kiil mõlemalt poolt komadega, nt

(6) ja sis lapse hind on:, ma vaatan siin vel, kaks tuhat öheksa`sada (359_a3)

Lisamised, mis leidsid aset kaugemal lausungis, kustutati, sest ei punkti panek ega komadega eraldamine ei teeks lisamist lausungi seisukohalt selgemaks (vt näiteid (7) ja (8)).

(7) keegi peab selle `pistiku nagu otsast `ära: võtma **selle `präeguse oleva** ja sis sinna teise pistiku nagu `asemele panema. (475_b14)

(8) keegi peab selle `pistiku nagu otsast `ära: võtma ja sis sinna teise pistiku nagu `asemele panema. (475_b14)

Algsete lausungite korpus sisaldas 4701 sõna ja normaliseeritud lausungite korpus 3864 sõna, seega eemaldati normaliseerimise käigus 837 sõna, mis olid nö mittetõhusad (vt ka ptk 1.2.3.).

4.3. Eksperimendi tulemused

Eksperimendi eesmärk oli näha, kas normaliseeritud lausungitega saab analüsaator paremini hakkama kui normaliseerimata e algsete lausungitega. Algsete lausungite puhul eeldati, et ilmnevad analüüsivead on seotud enamjaolt parandamiste, kordamiste ja valestartidega, kuna kõik lausungid sisaldasid ühte neist mittesoravustest. Normaliseeritud lausungite puhul eeldati, et kuna ilmsed veakohad on eemaldatud, siis esinevad vead on seotud muude nähtustega, nt osalausungipiiride määramise vead olid mõnel juhul põhjustanud kogu lausungi vale analüüsi. Seega ei ole eraldi vaadatud algsetes lausungites ja normaliseeritud lausungites ilmnenud vigu, vaid neid on vaadeldud võrdlevalt, st millistel juhtudel oli tulemus parem algses ja millistel juhtudel normaliseeritud lausungites ning millistel juhtudel ei muutunud analüüsis midagi. Võrreldakse lausungite paare igast probleemi liigist.

Eksperimendi tulemused on toodud tabelis 6. Võrdlevalt on toodud saadud tulemused nii normaliseerimata kui ka normaliseeritud lausungite korpusel[9]. Kuna morfoloogiline ühestamine tehti käsitsi, siis statistika näitab üksnes süntaksiprobleeme. Näeme, et erinevate probleemihulkade tulemused on erinevad. Normaliseerimine on kõige vähem aidanud tulemuste paranemisele kaasa korduste puhul, paranemine saagise osas 98.24%-lt 98.57%-le ja täpsuse osas 90.66%-lt 91.76%-le. Kuna võrreldes muude kordustega korrati just palju konjunktsioone, siis see seletab ka tulemuste nii vähest paranemist, sest problemaatilisi kordusi oli tunduvalt vähem (vt ka ptk 3.2.). Paljud eemaldatud kordused (nt *et et et*) olid triviaalse süntaktilise funktsiooniga, nende olemasolu algses korpuses tõstis korrektsuse näitajaid.

	Parandused		**Kordused**		**Valestardid**	
	Norm-mata	Norm-tud	Norm-mata	Norm-tud	Norm-mata	Norm-tud
Saagis	94.38	96.17	98.24	98.57	97.44	98.86
Täpsus	84.56	87.33	90.66	91.76	89.96	93.80
Ühesus	91.20	92.45	93.40	94.01	93.39	94.89

Tabel 6. Analüüsi tulemused (%)

Paranduste ja valestartide puhul on tulemuste paranemine juba märgatavam. Paranduste saagis tõusis 94.38%-lt 96.17%-le ja täpsus 84.56%-lt 87.33%-le. Valestartide puhul vastavalt 97.44%-lt 98.86%-le ja 89.96%-lt 93.80%-le. Vaatame järgnevalt iga mittesoravuse liiki eraldi.

Sõnakatked. Kui morfoloogilise analüüsi käigus on sõnakatke saanud õige märgendi (T), siis süntaktilises analüüsis see probleeme ei valmista. Kuid seda ainult juhul, kui sõnakatkestamine puudutab ühte sõna.

Vaatame näidet[10] 9, kus algses lausungis poolikus jäänud sõna on morfoloogilise analüüsi käigus õigesti saanud märgendi T (morfoloogiliselt analüüsimatu sõna), nii on

[9] Süntaktiliselt märgendatud korpused leiab aadressilt http://math.ut.ee/~kaili/Korpus/Spoken/.
[10] Näidetes esinevate morfoloogiliste ja süntaktiliste märgendite seletused on toodud Lisas 5. Olgu veel öeldud, et näidete parema esitamise huvides on morfoloogilisest infost alles jäetud vaid sõnaliik (märgitud topelt kaldkriipsude vahel).

süntaktilise analüüsi käigus saanud see märgendi @T ja parandatud sõna *palk* märgendi @SUBJ. Normaliseeritud lausungis pole analüüsis samuti probleeme tekkinud, kõik sõnad on saanud endale õige analüüsi. Seega võib väita, et kui morfoloogiline analüsaator suudab poolikuks jäänud sõnale anda õige märgendi, siis saab selle analüüsimisega hakkama ka süntaksianalüsaator.

Algne lausung	**Normaliseeritud lausung**
et et+0 //_J_// **CLB @J nendel tema+del //_P_// @ADVL peab pida+b //_V_// @+FCV pa- pa+0 //_T_// @T hh hh+0 //_B_// @B palk palk+0 //_S_// @SUBJ laekuma laeku+ma //_V_// @-FMV panka panka+0 //_S_// @ADVL	et et+0 //_J_// **CLB @J nendel tema+del //_P_// @ADVL peab pida+b //_V_// @+FCV palk palk+0 //_S_// @SUBJ laekuma laeku+ma //_V_// @-FMV panka panka+0 //_D_// @ADVL

Näide 9. et nendel peab **[RP** pa- + hh (.) `palk**]** laekuma `panka? (334_a5)

Sõnakatked ei tähenda aga alati seda, et üks poolikuks jäänud sõna asendatakse teisega. On ka terveid fraase, mille parandamise algusele võib sõnakatke viidata. Kui katkestatakse kogu fraas, siis vastavalt katkestatud struktuurile, võib see jällegi terve lausungi analüüsi segi paisata. Sel juhul aitab ainult see, kui analüsaator suudaks lisaks katkestatud sõnale varustada ka eelneva fraasiosa, mis kuulub ümbertegemisele, (kitsenduste grammatika formalismi kohaselt) märgendiga @T. Aga ühtegi reeglit pole võimalik sellisteks juhtudeks kirjutada. Vaatame näidet 10, mille analüüsiga on süntaksianalüsaator nii ühel kui ka teisel juhul hästi hakkama saanud, kuid algse lausungi analüüs on siiski mõneti segadusse ajav. Analüüsist võib välja lugeda, nagu oleks sõnal *hinna* kaks laiendit – *nende* ja *selle*. Tegelikult see nii aga pole, adekvaatne oleks analüüs siis, kui ka *nende* saaks endale kas märgendi @T või märgitaks ta ära kuidagi teisiti, et tegemist on parandamises oleva fraasi sõnaga.

Algne lausung	Normaliseeritud lausung

sööke
  söök+e //_S_// @SUBJ
nende
  see+de //_P_// **@NN>**
hin-
  hin+0 //_T_// **@T**
selle
  see+0 //_P_// @NN>
hinna
  hind+0 //_S_// @P>
sees
  sees+0 //_K_// @ADVL
ei
  ei+0 //_V_// @NEG
ole
  ole+0 //_V_// @+FMV
	sööke
	  söök+e //_S_// @SUBJ
	selle
	  see+0 //_P_// @NN>
	hinna
	  hind+0 //_S_// @P>
	sees
	  sees+0 //_K_// @ADVL
	ei
	  ei+0 //_V_// @NEG
	ole
	  ole+0 //_V_// @+FMV

Näide 10. `sööke **[RP** nende hin- + `selle hinna**]** sees ei=`ole. (356_a4)

Vaatame veel ka juba teisest peatükist tuttavat lausungit (näide 11), kus kõneleja saab alles neljandal katsel öeldud sõna *põhjapõdrasõit*. Kumbki analüüs pole saanud päris õiget analüüsi. Sisuliselt on tegemist loeteluga, kuid kuna sellele midagi ei viita, siis on algsest lausungist saanud predikatiivlause ja normaliseeritud lausungis on *põhjapõdrasõit* saanud endale kaks eestäiendit. Suulises keeles on mittesoravuste kõrval muudki, mis vajab normaliseerimist.

Algne lausung	Normaliseeritud lausung

see
  see+0 //_P_ // **CLB @NN>
mootorkelgusõit
  mootor_kelgu_sõit+0 //_S_// @SUBJ
jõ-
  jõ+0 //_T_// @T
põdrapõhjasõi-
  põdrapõhjasõi+0 //_T_// @T
sõit
  sõit+0 //_S_// **@NN>**
ee
  ee+0 //_B_// @B
põhjapõdrasõit
  põhja_põdra_sõit+0 //_S_// **@PRD**
vabandust
  vabandus+t //_B_// @B
	see
	  see+0 //_P_// @NN>
	mootorkelgusõit
	  mootor_kelgu_sõit+0 //_S_// **@NN>**
	põhjapõdrasõit
	  põhja_põdra_sõit+0 //_S_// @SUBJ

Näide 11. see `mootorkelgusõit **[RP[**jõ- + `põdrapõhja**[RP** sõi- + sõit**]]** + .hh **{F** ee**}** $ `põhjapõdrasõit $ **{D** vabandust**}]** (367_a4)

Asendused. Kui sõnakatked tänu sellele, et need on transkriptsioonis spetsiaalselt markeeritud, on kergema vaevaga tuvastatavad ja normaliseeritavad, siis asendused on märksa keerukam probleemide klass. Parandamist vajav osa võtab ülejäänud lausungilt õige analüüsi. Süntaksianalüsaator vaatab lauset vasakult paremale, st lausungi algusest, kuid parandatavates lausungites on õige osa alati just lausungi lõpus. Süntaksianalüsaatorile ei saa kirjutada ühtegi lisareeglit, kuidas asendusi ära tunda, kui pole ühtegi kindlat viidet neile. Abiks on asenduste puhul ainult katkestuskohas esinevad parandusele viitajad (näide (12)), mis saavad endale oletava lausungipiiri märgendi CLB-C. Kui asendustele on nii viidatud ja süntaksianalüsaator suudab konteksti põhjal vastava märgendi panna, siis on ainsaks probleemiks nagu eelmise näite puhul see, kuidas analüüsis asendatav fraas varustada märgendiga @T.

Algne lausung	**Normaliseeritud lausung**

et
  et+0 //_J_// **CLB @J
ja
  ja+0 //_J_// @J
noh
  noh+0 //_B_// @B
mäed
  mägi+d //_S_// @SUBJ
on
  ole+0 //_V_// @+FMV
lahti
  lahti+0 //_D_// @ADVL
tähendab
  tähenda+b //_B_// **CLB-C @B
tõstuk
  tõstuk+0 //_S_// @SUBJ
töötab
  tööta+b //_V_// @+FMV
kella
  kell+0 //_S_// @NN>
neljani
  neli+ni //_N_// @ADVL
	et
	  et+0 //_J_// **CLB @J
	ja
	  ja+0 //_J_// @J
	noh
	  noh+0 //_B_// @B
	tõstuk
	  tõstuk+0 //_S_// @SUBJ
	töötab
	  tööta+b //_V_// @+FMV
	kella
	  kell+0 //_S_// @NN>
	neljani
	  neli+ni //_N_// @ADVL

Näide 12. et ja=noh **[RP** `mäed=on `lahti: + **{D** tähendab**}** `tõstuk töötab**]** kella `neljani (355_a11)

Näites 13 näeme, mis juhtub analüüsiga, kui parandatav osa jääb analüüsi. Nii on sõna *mägi* tinginud selle, et *võimalik* on saanud järeltäiendi märgendi @AN> ja *valida* lausa kolm märgendit – infiniitse verbi, verbi infinitiivse vormi järeltäiendina ja määruse

märgendi. Normaliseeritud lausung on saanud korrektse analüüsi: *võimalik* on lauses predikatiiviks ja *valida* subjektiks.

Algne lausung

```
ikkagi
    ikkagi+0 //_D_// **CLB  @ADVL
noh
    noh+0 //_B_//  @B
erinevatel
    erinev+tel //_A_//  @AN>
päevadel
    päev+del //_S_//  @ADVL
on
    ole+0 //_V_//  @+FMV
võimalik
    võimalik+0 //_A_//  @AN>
sis
    sis+0 //_D_//  @ADVL
mägi
    mägi+0 //_S_//  @SUBJ
mäge
    mägi+0 //_S_//  @OBJ
valida
    vali+da  //_V_//     @-FMV  @<INF_N
@ADVL
```

Normaliseeritud lausung

```
ikkagi
    ikkagi+0 //_D_//   @ADVL
noh
    noh+0 //_B_//  @B
erinevatel
    erinev+tel //_A_//  @AN>
päevadel
    päev+del //_S_//  @ADVL
on
    ole+0 //_V_//  @+FMV
võimalik
    võimalik+0 //_A_//  @PRD
sis
    sis+0 //_D_//  @ADVL
mäge
    mägi+0 //_S_//  @OBJ
valida
    vali+da //_V_//  @SUBJ
```

Näide 13. ikkagi=noh `erinevatel `päevadel=on võimalik=sis **[RP** `mägi + mäge**]** valida. (355_a11)

Algne lausung

```
meil
    mina+l //_P_//  @ADVL
lihtsalt
    lihtsalt+0 //_D_//  @ADVL
sellist
    selline+t //_P_//  @SUBJ
nii-öelda
    nii-ütle+da //_D_//  @ADVL
süvenemiseks
    süvene=mine+ks //_S_//  @ADVL
pole
    ole+0 //_V_//  @+FMV
eriti
    eriti+0 //_D_//  @ADVL
aega
    aeg+0 //_S_//  @ADVL
$.
. //_Z_ Fst //
```

Normaliseeritud lausung

```
meil
    mina+l //_P_//  @ADVL
lihtsalt
    lihtsalt+0 //_D_//  @ADVL
süvenemiseks
    süvene=mine+ks //_S_//  @ADVL
pole
    ole+0 //_V_//  @+FMV
eriti
    eriti+0 //_D_//  @ADVL
aega
    aeg+0 //_S_//  @SUBJ @ADVL
$.
. //_Z_ Fst //
```

Näide 14. meil lihtsalt **[RP** sellist + **{D** nii-öelda**}** (0.5) .hh `süvenemiseks**]** pole eriti `aega. (475_b14)

Vaatame edasi näidet 14, kus algses lausungis on kõneleja alustanud fraasi sõnaga *sellist*, seejärel kasutanud veel nn venitamistaktikat, öeldes *nii-öelda* ning siis lõpetades fraasi sõnaga *süvenemiseks*, mis grammatiliselt ei ühildu fraasi algusega. Nii ongi *sellist* saanud subjekti märgendi ja *aega* vale adverbiaali märgendi. Ka normaliseeritud lausung pole saanud ühest analüüsi, nimelt *aega* on saanud kaks märgendit, subjekti ja adverbiaali märgendi, kuid vähemalt subjekti märgend on õige. Seega võib öelda, et normaliseerimisest on kasu olnud.

Lisamised. Lisamiste analüüsiga probleeme ei tekkinud, sest lisamist sisaldavates lausungites ei ole midagi üleliigset, st midagi pole vaja kustutada, kõik öeldu kuulub lause juurde. Nagu nägime peatükis 3.1.3, ei paikne kõik lisatav lausungi ülesehituse seisukohast vaadatuna just kõige õigemas kohas, kuid süntaksianalüsaatori tööd pole see kuidagi mõjutanud (näide 15).

Algne lausung	**Normaliseeritud lausung**

keegi
  keegi+0 //_P_ // @SUBJ
peab
  pida+b //_V_ // @+FCV
selle
  see+0 //_P_ // @NN>
pistiku
  pistik+0 //_S_ // @OBJ
nagu
  nagu+0 //_D_// @ADVL
otsast
  ots+st //_S_ // @ADVL
ära
  ära+0 //_D_// @ADVL
võtma
  võt+ma //_V_// @-FMV
selle
  see+0 //_P_ // @NN>
präeguse
  präeguse+0 //_A_// @AN>
oleva
  olev+0 //_S_// @<NN
	keegi
	  keegi+0 //_P_ // @SUBJ
	peab
	  pida+b //_V_ // @+FCV
	selle
	  see+0 //_P_ // @NN>
	pistiku
	  pistik+0 //_S_ // @OBJ
	nagu
	  nagu+0 //_J_// @J
	otsast
	  otsast+0 //_D_// @ADVL
	ära
	  ära+0 //_D_// @ADVL
	võtma
	  võt+ma //_V_ // @-FMV

Näide 15. keegi peab selle `pistiku nagu otsast `ära: võtma **{A** selle `präeguse oleva**}** (475_b14)

Kordused. Kuna enne lausungite automaatset analüüsi eemaldati kõik transkriptsioonimärgendid peale punktide ja komade, siis pärast morfoloogilist ja

süntaktilist analüüsi ei saa enam vahet teha katkestuskohaga ja katkestuskohata kordustel. Korduste analüüsil tekitavad probleeme just subjektide ja objektide kordused. Näites 16 on näha, kuidas korduse esimene sõna *pistik* laiendab korratud teist sõna *pistik*, mis on objekti rollis.

Algne lausung

```
siis
    siis+0 //_D_//   **CLB @ADVL
on
    ole+0 //_V_//  @+FMV
äkki
    äkki+0 //_D_//  @ADVL
vaja
    vaja+0 //_D_//  @ADVL
se
    se+0 //_P_//  @NN>
pistik
    pistik+0 //_S_//  @NN>
pistik
    pistik+0 //_S_//  @OBJ
seal
    seal+0 //_D_//  @ADVL
lihtsalt
    lihtsalt+0 //_D_//  @ADVL
ära
    ära+0 //_D_//  @ADVL
vahetada
    vaheta+da //_V_//  @SUBJ
```

Normaliseeritud lausung

```
siis
    siis+0 //_D_//   **CLB @ADVL
on
    ole+0 //_V_//  @+FMV
äkki
    äkki+0 //_D_//  @ADVL
vaja
    vaja+0 //_D_//  @ADVL
se
    se+0 //_P_//  @NN>
pistik
    pistik+0 //_S_//  @OBJ
seal
    seal+0 //_D_//  @ADVL
lihtsalt
    lihtsalt+0 //_D_//  @ADVL
ära
    ära+0 //_D_//  @ADVL
vahetada
    vaheta+da //_V_//  @SUBJ
```

Näide 16. siis on äkki vaja se **[RE** `pistik + (.) pistik**]** seal `lihtsalt ära vahetada. (475_b14)

Algne lausung

```
mul
    mina+l //_P_//  **CLB-C @ADVL
see
    see+0 //_P_//   @NN>
nimi
    nimi+0 //_S_//  @SUBJ
ei
    ei+0 //_V_//  @NEG
tulnud
    tule+nud //_V_//  @+FMV
ei
    ei+0 //_V_//  @NEG
tulnud
    tule+nud //_V_//  **CLB-C @+FMV
nagu
    nagu+0 //_D_//  @ADVL
hästi
    hästi+0 //_D_//  @ADVL
ette
    ette+0 //_D_//  @ADVL
```

Normaliseeritud lausung

```
vabandage
    vabanda+ge //_V_//  @+FMV
mul
    mina+l //_P_//  **CLB-C @ADVL
see
    see+0 //_P_//   @NN>
nimi
    nimi+0 //_S_//  @SUBJ
ei
    ei+0 //_V_//  @NEG
tulnud
    tule+nud //_V_//  @+FMV
nagu
    nagu+0 //_D_//  @ADVL
hästi
    hästi+0 //_D_//  @ADVL
ette
    ette+0 //_D_//  @ADVL
```

Näide 17. vabandage mul see `nimi **[RE** ei [tul]nud + .hh ei tulnud**]** nagu hästi `ette. (259_b7)

Laiendite ja verbidega saab analüsaator kenasti hakkama, määrates esimesel juhul põhjale lihtsalt kaks ühesugust laiendit ja teisel juhul teise verbi uue osalausungi algusesse (näide 17). Korduste analüüs võiks olla sedavõrd lihtsam, kui pärast morfoloogilist analüüsi oleks teatav vahepeale programm, mis markeerib ühe kahest kõrvuti asetsevast sama morfoloogilise analüüsiga sõnast, nii et süntaktilisse analüüsi jõuab edasi ainult üks sõna.

Valestardid. Müürisep ja Uibo (2006) on leidnud, et valestarte saab reeglitega tuvastada, märgendades neid osalausepiiri märgenditega. Kuid see õnnestub vaid juhtudel, kui valestart sisaldab verbi. Vaatame näidet 18, kus valestart ei sisalda verbi.

Algne lausung	**Normaliseeritud lausung**

ma
  mina+0 //_P_// @SUBJ
i
  i+0 //_V_// @+FCV
oska
  oska+0 //_V_// @+FMV
nagu
  nagu+0 //_J_// @J
nimodi
  nimodi+0 //_D_// @ADVL
kohe
  kohe+0 //_D_// @ADVL
täpselt
  täpselt+0 //_D_// @ADVL
öelda
  ütle+da //_V_// @OBJ
et
  et+0 //_J_// **CLB @J
noh
  noh+0 //_B_// @B
selline
  selline+0 //_P_// **@SUBJ**
m
  m+0 //_T_// @T
tegemist
  tege=mine+t //_S_// **@ADVL**
on
  ole+0 //_V_// @+FMV
ühe
  üks+0 //_N_// @NN>
üliõpilasorganisatsiooni
  üli_õpilas_organisatsioon+0 //_S_// @NN>
aastapäevaga
  aasta_päev+ga //_S_// @ADVL
	ma
	  mina+0 //_P_// @SUBJ
	i
	  i+0 //_V_// @+FCV
	oska
	  oska+0 //_V_// @+FMV
	nagu
	  nagu+0 //_D_// @ADVL
	nimodi
	  nimodi+0 //_D_// @ADVL
	kohe
	  kohe+0 //_D_// @ADVL
	täpselt
	  täpselt+0 //_D_// @ADVL
	öelda
	  ütle+da //_V_// @OBJ
	$.
	  . //_Z_ Fst //
	LL
	  ####
	LA
	  #### **CLB
	tegemist
	  tege=mine+t //_S_// @SUBJ
	on
	  ole+0 //_V_// @+FMV
	ühe
	  üks+0 //_P_// @NN>
	üliõpilasorganisatsiooni
	  üli_õpilas_organisatsioon+0 //_S_// @NN>
	aastapäevaga
	  aasta_päev+ga //_S_// @ADVL

Näide 18. ma=i=´oska nagu nimodi kohe täpselt ´öelda et noh selline: **+/** .hh m ´tegemist on ühe üliõpilasorganisatsiooni ´aastapäevaga. (380_a7)

Algses lausungis jääb kõrvallausung *et noh selline* pooleli. Kuna osalausungipiire samuti pole, siis on *selline* saanud endale subjekti märgendi ja *tegemist*, mis on tegelikult uue lausungi subjekt, hoopis adverbiaali märgendi. Normaliseeritud lausungi analüüs on igati korrektne.

Näites 19 näeme, kuidas normaliseeritud lausung pole saanud ühest analüüsi, aga vähemalt algse lausungi probleemse koha eemaldamine on analüüsi natuke paremaks teinud. Lisaks illustreerib see näide hästi seda, et punkt ei tähenda alati lausungi lõppu.

Algne lausung

```
no
    no+0 //_B_//   @B
kuskil
    kuskil+0 //_D_//  @ADVL
kolmkümmend
    kolm_kümmend+0 //_N_//  @OBJ @ADVL
oota
    oota+0 //_V_//  @+FMV
se
    se+0 //_P_//  **CLB-C @SUBJ
oli
    ole+i //_V_//  @+FMV
päev
    päev+0 //_S_//  @PRD @OBJ @NN>
@ADVL
enne
    enne+0 //_K_//  @ADVL
esimest
    esimene+t //_N_//  **CLB-C @AN>
detsembrit
    detsember+t //_S_//  @SUBJ @OBJ
@ADVL @<P
on
    ole+0 //_V_//  @+FMV
vist
    vist+0 //_D_//  @ADVL
kolmkümmend
    kolm_kümmend+0 //_N_//  @SUBJ @OBJ
@PRD @ADVL
$.
    . //_Z_//
$LL$
    ####
$LA$
    #### **CLB
november
    november+0 //_S_//   @SUBJ
$.
    . //_Z_//
```

Normaliseeritud lausung

```
no
    no+0 //_B_//   @B
kuskil
    kuskil+0 //_D_//  @ADVL
kolmkümmend
    kolm_kümmend+0 //_N_//  @SUBJ
@ADVL
$.
    . //_Z_ Fst //
$LL$
    ####
$LA$
    #### **CLB
päev
    päev+0 //_S_//   @SUBJ @ADVL
enne
    enne+0 //_K_//  @ADVL
esimest
    esimene+t //_N_//  @AN>
detsembrit
    detsember+t //_S_//  @<P
on
    ole+0 //_V_//  @+FMV
vist
    vist+0 //_D_//  @ADVL
kolmkümmend
    kolm_kümmend+0 //_N_//  @SUBJ @PRD
@ADVL
$.
    . //_Z_//
$LL$
    ####
$LA$
    #### **CLB
november
    november+0 //_S_//   @SUBJ
$.
    . //_Z_//
```

Näide 19. no kuskil:: ´kolmkümmen:d > **/** oota se oli: **+/** (0.5) päev enne esimest detsembrit on vist (.) ´kolmkümmend. (.) no´vember. (380_a7)

Kuna algses lausungis pole eraldi oletatavat osalausungipiiri saanud *se* kõrval ka *oota*, siis see on tinginud selle, et *kolmkümmend* on saanud endale lisaks valele määruse märgendile ka vale objekti märgendi. Normaliseeritud lausungis on vale määruse märgendi kõrval ka õige subjekti märgend. Kuna valestart *see oli* jäi analüüsi, siis hakkas süntaksianalüsaator edasi analüüsima lausungit *see oli päev enne esimest detsembrist on vist kolmkümmend*. Süntaksianalüsaatori jaoks on selline lausung tõeline pähkel, millest annab ka tunnistust see, et *päev*, *detsembrit* ja *kolmkümmend* on jäänud algse lausungi analüüsil nelja märgendiga. Normaliseeritud lausungis on neil sõnadel vastavalt kaks märgendit, üks ja õige märgend ning kolm märgendit. Tegemist on muidu tavalise predikatiivlausega, kus *päev* on aluseks ja *kolmkümmend* predikatiiviks.

Algne lausung	**Normaliseeritud lausung**

kui
  kui+0 //_J_// **CLB @J
kui
  kui+0 //_J_// @J
kui
  kui+0 //_J_// @J
ta
  tema+0 //_P_// @SUBJ
seda
  see+da //_P_// **@ADVL @NN>**
seda
  see+da //_P_// **@PRD @ADVL**
tükina
  tükk+na //_S_// @ADVL
siin
  siin+0 //_D_// @ADVL
ei
  ei+0 //_V_// @NEG
ole
  ole+0 //_V_// @+FMV
	seda
	  see+da //_P_// @SUBJ
	tükina
	  tükk+na //_S_// @ADVL
	siin
	  siin+0 //_D_// @ADVL
	ei
	  ei+0 //_V_// @NEG
	ole
	  ole+0 //_V_// @+FMV

Näide 20. **[[RE** =kui + kui**]** + kui**]** ta **+/** (0.5) **[RE** {seda} + =seda**]** `tükina siin ei `ole. (475_b14)

Veel üks näide (20) valestardi analüüsist, kus lisaks valestardile esineb ka kordus. Siin on sarnaselt eelmise näitega probleemiks see, et parandatava lausungi *ta* dikteerib ülejäänud lausungi analüüsi, olles ise subjekt, kuigi tegelik subjekt on *seda*. Normaliseeritud lausungi analüüs on korrektne.

4.4. Kokkuvõte

Eksperimendi tulemuste analüüsi käigus selgus, et püstitatud hüpotees osutus õigeks. Normaliseerimine aitas paljudel juhtudel süntaksianalüsaatori tööle kaasa. Kõige vähem aitas tulemuste paranemisele kaasa korduste normaliseerimine, saagis paranes 98,24%-lt 98,57%-le ja täpsus 90,66%-lt 91,76%-le. Kuna võrreldes muude kordustega korrati just palju konjunktsioone, siis see seletab ka tulemuste nii vähest paranemist, sest problemaatilisi kordusi oli tunduvalt vähem. Paranduste ja valestartide puhul oli tulemuste paranemine juba märgatavam. Paranduste saagis tõusis 94,38%-lt 96,17%-le ja täpsus 84,56%-lt 87,33%-le. Valestartide puhul paranes saagis 97,44%-lt 98,86%-le ja täpsus 89,96%-lt 93,80%-le.

5. Kokkuvõte

Käesolev töö on eeltöö mittesoravuste automaatseks normaliseerimiseks. Töö käigus analüüsiti kõneleja voorusiseseid parandusi, kordusi ja valestarte ning nende pindmist struktuuri, st kuidas kõneleja neid esitab. Samuti vaadeldi, kuidas need suulisele keelele omased nähtused mõjutavad eesti keele kitsenduste grammatikal põhineva analüsaatori tööd. Suuline keel on võrreldes kirjaliku keelega oma olemuselt teistsugune ja see toob automaatsel süntaktilisel analüüsil kaasa uusi probleeme. Üheks võimalikuks lahenduseks on parandused, kordused ja valestardid enne süntaktilist analüüsi nö normaliseerida, st ebagrammatilised lausungid tehakse grammatilisteks, nt lausungist *ma ma ei taha tähendab ei soovi seda* eemaldatakse kordus (*ma*) ja asendus koos leksikaalse parandamisele viitava partikliga (*ei taha tähendab*).

Praktiline töö oli jaotatud kolmeks. Esimesena viidi läbi paranduste, korduste ja valestartide märgendamine (normaliseerimine), mis põhineb eeldusel, et suulises keeles esinevatel parandamisvõtetel on kindel struktuur, mis aitab kuulajal kõneleja mittesoravast jutust aru saada. Töötati välja teatavad märgendamise põhimõtted ja kasutatavad märgendid. Eeskuju võeti Switchboardi korpuse mittesoravuste märgendamisjuhendist. Analüüsitavad dialoogid võeti Tartu Ülikooli eesti dialoogikorpusest. Analüüsiti 35 infodialoogi (13 168 sõna). Kõiki lausungeid ei olnud ega olegi võimalik normaliseerida. Normaliseerimine aitab suuremal osal juhtudel, kuid on ka lausungeid, kus see ei aita, nt kus parandamine on lihtsalt „sõnamulin“ või kus kõneleja kordab juba öeldud sõna veelkord suvalise koha peal lausungis.

Teises osas analüüsiti märgendatud lausungeid. Vastavalt struktuurilistele omadustele jaotati parandused sõnakatketeks, asendusteks ja lisamisteks ning kordused katkestuskohaga ja katkestuskohata ning ahelkordusteks. Valestarte ei liigitatud. Kokku tuvastati 13 168-sõnalise korpuses 131 parandust, 113 kordust ja 33 valestarti. Seejärel püüti leida leksikaalseid ja prosoodilisi markereid, millele saaks toetuda mittesoravuste automaatsel analüüsil. Sõnakatkete puhul saab rääkida kindlast parandamise markerist. Selleks on sidekriips, mis transkriptsioonis viitab pooleli jäänud sõnale. Ülejäänute puhul saab rääkida vaid võimalikest markeritest. Asenduste puhul oli 56% juhtudest katkestuskoht markeeritud. Selle markeerimiseks olid kõnelejad kasutanud erinevaid

võtteid (mõnel juhul ka mitut koos): leksikaalsetest võtetest partikleid ja täidetud pause, prosoodilistest võtetest parandatava osa lõpus poollangevat intonatsiooni, sissehingamist, pause ja sõna viimase silbi venitamist. 98 ühesõnakordamistest 47% juhtudel markeeris kõneleja katkestuskoha, kasutades selleks leksikaalsetest markeritest täidetud pause ja prosoodilistest pause, sissehingamist ja esimese sõna viimase silbi venitamist. Lisamiste ja valestartide kohta on esialgu raske midagi öelda, kuna neid esineb võrreldes teiste parandamisvõtetega vähem, siis esiteks polnud materjali hulk järelduste tegemiseks piisav ja teiseks, mõlemad mittesoravused on oma struktuurilt suhteliselt ebamäärased, mistõttu on ka mõlema tuvastamine keerulisem, seega ka normaliseerimine. Kuid olemasolevate näidete põhjal ilmnes, et kui lausungis lisamisele ja valestardile üldse midagi viitab, siis on nendeks prosoodilised markerid, nagu pausid ja venitamised.

Praktiline osa lõppes eksperimendiga, mille käigus anti süntaksianalüsaatorile kaks korda analüüsida samu lausungeid, ainult selle erinevusega, et ühel juhul oli tegu algsete ja teisel juhul normaliseeritud lausungitega. Eksperimendi hüpoteesiks oli, et analüsaator on normaliseeritud lausungite analüüsil edukam kui normaliseerimata lausungite puhul. Eksperimendi tulemuste analüüsi käigus selgus, et püstitatud hüpotees pidas paika. Normaliseerimine aitas mitmetel juhtudel süntaksianalüsaatori tööle kaasa. Kõige vähem aitas tulemuste paranemisele kaasa korduste normaliseerimine, saagis paranes 98,24%-lt 98,57%-le ja täpsus 90,66%-lt 91,76%-le, paranemine oli seega vastavalt 0,33% ja 1,1%. Kuna võrreldes muude kordustega korrati just palju konjunktsioone, siis see seletab ka tulemuste nii vähest paranemist, sest problemaatilisi kordusi oli tunduvalt vähem. Paranduste ja valestartide puhul oli tulemuste paranemine juba märgatavam. Paranduste saagis tõusis 94,38%-lt 96,17%-le ja täpsus 84,56%-lt 87,33%-le, paranemine oli vastavalt 1,79% ja 2,77%. Valestartide puhul paranes saagis 97,44%-lt 98,86%-le ja täpsus 89,96%-lt 93,80%-le, paranemisprotsent oli seega vastavalt 1,42% ja 3,84%. Ilmnes, et ühe sõna katkete, lisamiste ja lihtsate kordustega saab praegune suulise keele analüüsiks kohandatud süntaksianalüsaator hakkama, probleeme põhjustavad just asendused, subjekti-objekti kordused ja valestardid.

Suulise keele automaatse süntaktilise analüüsi juures on väga olulised lausungipiiride määramise reeglid, samuti enne süntaktilist analüüsi olev morfoloogiline

analüüs, mille käigus tuntaks õigesti ära partiklid, millel sageli mitu funktsiooni ja sõnaklassi (nt *tähendab*, *või*). Nende õigesti määramine mõjutab väga suurel määral edasist analüüsi. Lisaks, ESTKG süntaksianalüsaator vaatab lauset vasakult paremale, st alustab lause süntaktilist analüüsi lausealgusest. Samas suulise keele puhul on korrektne osa just lausungilõpus. Seega on normaliseerimine analüüsi eeletapina igati vajalik.

Edasised uurimissuunad. Edasi on plaanis täiustada ja testida käesoleva testmaterjali peal ESTKG süntaksianalüsaatori suulise keele analüüsi reegleid ning rakendada märgendamist suuremal hulgal suulise keele materjalil. Märgendamisse peaks lisama ka lausungipiiride märkimise. Neljandas peatükis nägime, kuidas mõnel juhul vale või puudulik osalausungipiiri määramine mõjutas terve lausungi analüüsi.

Samuti tuleb edasi mõelda, milline peaks välja nägema morfoloogilise ja süntaktilise analüüsi vahel olev etapp mittesoravuste automaatseks analüüsiks.

Kirjandus

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco, Ryan Ginstrom 2000. Heaviness vs. Newness: the effects of structural complexity and discourse status on constituent ordering. – Language, 76(1), pp. 28–55.

Bailey, Karl G. B., Fernanda Ferreira 2003. Disfluencies influence syntactic parsing. – Journal of Memory and Language, 49, pp. 183–200.

Bear, John, John Dowding, Elizabeth Shriberg 1992. Automatic Detection and Correction of Repairs in Human-Computer Dialog. Proceedings of the DARPA Speech and Natural Language Workshop.

Brennan, Susan, Michael F. Schober 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. Journal of Memory and Language, 44, pp. 274–296.

Brennan, Susan E., Maurice Williams 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. – Journal of Memory and Language, 34, pp. 383–398.

Charniak, Eugene, Mark Johnson 2001. Edit detection and parsing for transcribed speech. – Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting, pp. 118–126.

Clark, Herbert H. 1996. Using language. Cambridge. CUP.

Clark, Herbert H., Thomas Wasow 1998. Repeating Words in Spontaneous Speech. – Cognitive Psychology, 37, pp. 201–242.

Core, Mark G. 1999. Dialog Parsing: From Speech Repairs to Speech Acts. PhD thesis. University of Rochester, Rochester New York.

Core, Mark G., Lenhart K. Schubert 1999. A Syntactic Framework for Speech Repairs and Other Disruptions. 37nd Annual Meeting of the Association for Computational Linguistics.

Eklund, Robert 2004. Disfluency in Swedish human–human and human–machine travel booking dialogues. PhD thesis, Linköping University.

Ferreira, Fernanda, Ellen F. Lau, Karl G. D. Bailey 2004. Disfluencies, language comprehension, and tree adjoining grammars. – Cognitive Science, 28, pp. 721–749.

Fox Tree, Jean E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. – Journal of Memory and Language, 34, pp. 709–738.

Giachin, Egidio, Scott McGlashan 1997. Spoken Language Dialogue Systems. S. Young, G. Bloothooft (eds.), Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers, pp. 69–117.

Gibbon, Dafydd, Inge Mertins, Roger K. Moore 2000. Handbook of Multimodal and Spoken Dialogue Systems. Dordrecht: Kluwer Academic Publishers.

Heeman, Peter 1997. Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. PhD thesis, University of Rochester, Rochester New York.

Heeman, Peter, James Allen 1994. Tagging Speech Repairs. ARPA Workshop on Human Language Technology, pp. 187–192.

Hennoste, Tiit 2000-2001. Sissejuhatus suulisesse eesti keelde. – Akadeemia 2000, 5, lk 1117–1150; 7, lk 1553–1582; 8, lk 1773–1806; 9, lk 2011–2038; 12, lk 2689–2710; 2001, 1, lk 179–206.

Hennoste, Tiit 2006. Self-repair initiators in Estonian conversation compared with Finnish. Presentation on ICCA.

Hennoste, Tiit, Liina Lindström, Olga Gerassimenko, Airi Jansons, Andriela Rääbis, Krista Strandson, Piret Toomet, Riina Vellerind. Suuline kõne ja morfoloogiaanalüsaator. – Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim R. Pajusalu, T. Hennoste. Tartu, lk 161–171.

Hindle, Donald 1983. Deterministic Parsing of Syntactic Nonfluencies. Proceedings of the 21st Meeting of the Association of Computational Linguistics.

Kurdi, Mohamed-Zakaria 2002. Combining pattern matching and shallow parsing techniques for detecting and correcting spoken language extragrammaticalities. – 2nd Workshop on Robust Methods in Analysis of Natural Language Data, Italy, pp. 1–9.

Leech, Geoffrey, Martin Weisser, Andrew Wilson, Martine Grice 1998. Survey and Guidelines for the Representation and Annotation of Dialogue. LE-EAGLES-WP4-4, Integrated Resources Working Group.

Levelt, Willem J. M. 1983. Monitoring and self-repair in speech. – Cognition, 14, pp. 41–104.

Levelt, Willem J. M. 1989. Speaking: From intention to articulation. Cambridge, MA: MIT Press.

Lickley, Robin J. 1994. Detecting Disfluency in Spontaneous Speech. PhD thesis, University of Edinburgh.

Mary Swift, Myroslava Dzikovska, Joel Tetreault, James Allen. 2004. Semi-automatic syntactic and semantic corpus annotation with a deep parser. LREC.

McKelvie, David 1998. The syntax of disfluency in spontaneous spoken language. Technical Report HCRC/RP-95, Edinburgh University, Edinburgh, Scotland.

Meteer, M., A. Taylor, R. MacIntyre, R. Iver 1995. Dysfluency annotation stylebook for the Switchboard corpus. Distributed by LDC.

Meyer, Charles F. 2002. English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press.

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.

Müürisep, Kaili, Helen Nigol, Heli Uibo 2006. Eesti suulise keele korpuse automaatne pindsüntaktiline analüüs. – Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Toim M. Koit, R. Pajusalu, H. Õim. Tartu, lk 72–84.

Müürisep, Kaili, Heli Uibo 2005. Shallow Parsing of Spoken Estonian Using Constraint Grammar. Proceedings of NODALIDA special session on. Copenhagen Studies in Language #33/2006.

Nakatani, Christine, Julia Hirschberg 1993. A Speech-First Model for Repair Detection and Correction. Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, pp. 46–53.

Nigol, Helen 2004. Puude pankade teooriast ja praktikast eesti keele puude pangale mõeldes. Bakalaureusetöö arvutilingvistika erialal. Tartu Ülikool, eesti ja soome-ugri keeleteadus.

Oviatt, Sharon 1995. Predicting spoken disfluencies during human–computer interaction. – Computer Speech and Languages, 9(1), pp. 19–35.

Sampson, Geoffrey 1998. Consistent Annotation of Speech-Repair Structures. – A. Rubio et al. (eds.), Proceedings of the First International Conference on Language Resources and Evaluation, pp. 1279–82.

Schegloff, Emanuel A., Gail Jefferson, Harvey Sacks 1977. The preference for self-correction in the organization of repair in conversation. – Language, 53, pp. 361–382.

Shriberg, Elizabeth E. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California, Berkeley.

Snover, Matthew, Bonnie Dorr, Richard Schwartz 2004. A Lexically-Driven Algorithm for Disfluency Detection. Short Paper Proceedings of North American Association for Computational Linguistics and Human Language Technology Conference.

Sorjonen, Marja-Leena, Minna Laakso 2005. Cut-off, the particle ‘eiku’ and other practices for initiating self-repair, and the interactional functions of self-repair. – Virittäjä, 2, s. 244–271.

Spilker, Jörg, Martin Klarner, Günther Görz 2000. Processing Self-Corrections in a Speech-to-Speech System. W. Wahlster (ed.) Verbmobil: Foundations of Speech-to-Speech Translation. Springer, pp. 131–140.

Stolcke, Andreas, Elizabeth Shriberg 1996. Statistical Language Modeling for Speech Disfluencies. Proceedings of the International Conference on Audio, Speech and Signal Processing.

Strandson, Krista 2002. Vestluskaaslase algatatud reformuleeringud eesti vestlustes: reformuleeringualgatuse vahendeid ja põhjuseid. Magistritöö. Tartu Ülikool, üldkeeleteaduse õppetool.

Viks, Mare 2001. Suulise keele erijooni spordireportaažide keeles. Tartu Ülikool, eesti keele õppetool.

Uibo, Heli, Helen Nigol 2006. Puude pangad meil ja mujal. – Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Toim M. Koit, R. Pajusalu, H. Õim. Tartu, lk 36–51.

Same turn repairs, repetitions and false starts in spoken Estonian: detection and normalization

Summary

Spoken language contains disfluencies, which include repairs, repetitions and false starts. For automatized syntactic analysis these features have to be dealt with. One possible solution is to annotate disfluencies by normalizing ungrammatical utterances so that they can be parsed correctly. So far disfluencies were annotated in several corpora of spoken English language such as Switchboard, the Susanne corpus as part of the Christine corpus and ICE-GB.

The thesis contains three major parts. In the first part, a set of annotations and annotation rules for disfluencies are defined which partly incorporates ideas of the *Disfluency annotation stylebook for the Switchboard corpus*. Then, the resulting annotation scheme was applied on an information dialogue corpus of Estonian consisting of 13 168 words in which 131 repairs, 113 repetitions and 33 false starts were identified. Special attention was paid on how speakers themselves present and normalize repairs, repetitions and false starts within their utterances. Based on these observations a preliminary approach on how to identify patterns of repairs and how one could formalize them for automatic detection is demonstrated.

In the third part the results of a test run with the Estonian constraint-based parser are presented. All utterances containing repairs, repetitions and false starts were analyzed two times. The first run parsed the corpus in its original form; the second run parsed the same corpus after its normalization. For repetitions the recall rose slightly from 98.24% to 98.57% and precision from 90.66% to 91.67%. For repairs and false starts results improved more significantly. For repairs the recall rose from 94.38% to 96.17% and precision from 84.56% to 87.33%. For false starts recall rose from 97.44% to 98.86% and precision from 89.96% to 93.80%.

Lisa 1 – Transkriptsioonimärgid

. langev intonatsioon

, poollangev intonatsioon

? tõusev intonatsioon

(.) mikropaus (0,2 sekundit või lühem)

(…) mikropausist pikem paus

(1.2) pausi pikkus sekundites

` rõhk

>…< kiirendatud lõik

<…> aeglustatud lõik

… muust kõnest vaiksem lõik

: hääliku venitamine

$...$ naerva häälega öeldud sõna või pikem lõik

.hh häälekas sissehingamine

=h sõna lõpul olev väljahingamine

- sõna on poolikuks jäänud

= kaks iseseisvat üksust on hääldatud kokku

[pealerääkimise algus

] pealerääkimise lõpp

{või} halvasti kuuldud tekstilõik

{--} ebaselgeks jäänud sõna

{---} pikem ebaselgeks jäänud lõik

Lisa 2 – Märgendatud parandused

Sõnakatked

noh **[RP** pärisk- + päris kella**]** `kuueni **[RE** ärge $ + ärge**]** vast `jätke. (96_b8)

jah, (0.5) ästi, **[RP** õhe- + natuke õheksa**]** `läbi kolma`päeval. (259_b7)

see on `Paide liin, ta küll **[RP** p- (.) + `peatub**]** kõigis `peatustes aga see .hh `tempo on nagu `veidi kiirem kui > tavaliselt < .hh `kiirliin (334_a3)

e e j:ah, **[[RP** mi- + mi-**]** + `mida**]** nimelt. (334_a5)

.hh käendajatel on ainult see `nõue, et nendel peab **[RP** pa- + hh (.) `palk**]** laekuma `panka? (334_a5)

[et] on `näha et **[RP** nende (.) mingi sisse- + midagi**]** laekub `panka. (334_a5)

e tähendab siss=ee **[RP** te so- + kas te soovite**]** muidu seda Isik (.) Maestro kaarti ka `taotleda või. (334_a5)

aga **[RP** ned ´ainete pe- + **{D** tähendab selles mõttes**}** ned ´ained**]** nagu oma ´peaainete=ja kõrvalainete ´kaupa te peaksite nendes (0.5) ´õppetoolides ´ära kontrollima mis (.) **[RP** nendes õpetatavates õp- + neid ´aineid õpetavates õppe´toolides**]**. (338_a3)

et teil ei ole midagi **[RP** ma-, + ´maksmata**]** (338_a3)

[.hh ee] (.) ja kui ´kaua **+/** (.) .h **[RP** mi + ´mis**]** ajal peab puh **+/** noh enne kaitsmist ütleme mul on viie´teistkümnes millal ma pean selle bakalaureusetöö [´esitama.] (338_a3)

aga no **[RP** =nad põhimõtselt vist võimald- + kui ´võimalik s=nad korraldavad**]** noh kui **[RP** ika tahate gi- + ikka ´tahate**]** ´lõpetada ´kindlasti eksju, s=nad korraldavad selle ka võibla ´ühe inimese jaoks näiteks. (338_a3)

.hh ee kas ainult=ee **[RE** ´üks + ´üks**]** variant **[RE** või + või**]** **[RP** kak- + **{D** või=tändab**}** kahe**]** (338_a3)

see=on nimodi **[RP** läbi- + **{D** noh**}**, isikliku $ ´läbirääkimise $] .hh läbirääkimise tulemusel pannakse ´paika siis=et noh **[RP** kee- + kes**]** ´arvab mis kes ´on. (338_a3)

.hh ja siis **[RP** linnaval- + linna (.) hooldus´amet**]** on siin antud otseselt, .hh te soovite **[RP** üld- + (.) üld´numbrit**]** või ´info (347_a2_info)

siis ta peab ikka päris **[RP** ha- + `heal**]** tasemel **[RP** ole- + olema**]** **[RE** nagu + .hh nagu:**]** juba reno`veeritud see `maja. (353_a3)

a mis need **[RP** tus- + tu`rismitalude**]** hinnad muidu on. (353_a3)

see on nüd **[RP** mö- + möb´leerimata**]** korter. (354_a4)

aga ´kahetoalise puhul: peate arvestama ´sellega et siin nagu ´omanik tahab ´ette ´maksu saad a et siis läheks nagu .hh ee (.) noh (.) ´kuus=tuhat ´krooni nagu **[RP** e- + esialgu**]** kõik see asi ´maksma. (354_a4)

[RP suusavarut- + varustus**]** kas=võtate kaasa või rendite `kohapealt (355_a11)

`tagasijõudmine on no õhtul ütleme **[RP** tav- + tavaliselt**]** jõuavad nad kell kaheksa ka `tagasi. (356_a4)

ja kindlasti `tagasiteel ee võib `ka **[RP** kusagil Pa- + teha kas siis `Paide lähedal või kusagil**]** .hh `peatus (356_a4)

.hh et `sööke **[RP** nende hin- + `selle hinna**]** sees ei=`ole. (356_a4)

on küll jah [**[RP** natu- + natukene**]** `kurb, aga {mis seal} `ikka.] (356_a4)

ee suusareise pakume **{X** tõe=õõ**}** te **[RP** arvat + **{F** =õõ**}** mõtlete**]** nüüd `veebruarikuus=`jah? (359_a3)

mhmh (1.5) ja: **[RP** s- + `minek oleks sis**]** `bussiga siit. (359_a3)

aga seda: (.) mingit=e `interneti aadressi nagu teil ei ole **[RE** et + =et**]** kust **[RP** k- + **{F** ee**}** (.) kohast**]** [võiks infot leida.] (359_a3)

[ahhaa.] **[RP** las- + lastel**]** [on nagu {-}] (359_a3)

nii=et sinna lähevad ilusti nii `suusad **[RE** kui + kui**]** =ee **[RP[**varus- + või see**]** + **{F** ee**}** `kotid**]** (359_a3)

inimene ronis ee: vannist `välja ja `leidis kuskilt mingi `orgi=ja **[RP** vi- + vigastas**]** endale sääremarja (359_a3)

[RP kolmapäe- + järgmine kolmapäev**]** kõik ajad on `kinni. (364_a5)

[mhmh] .hh aga `küsiks natuke selle kohta **[RE** et=ee**}** + (0.5) et=ee**]** mis `hinnaklassis **[RP** se- + see**]** `on. (367_a4)

see `mootorkelgusõit **[RP[** jõ- + `põdrapõhja**[RP** sõi- + sõit]] + .hh {F ee} $ `põhjapõdrasõit $**]** vabandust, .hh siis on: `veelõbustuspark Kuusamos Kuusamo `troopika (367_a4)

[RE[et] + et**]** siin on ki- /+ et noh **[RP** ´pea- + kui ´peaks**]** juhtuma et ma=i=´ole äkki ´rahul, (.) [et siin on kir-] (373_a5)

[RE et + et**]** **[RP** kas te k- + mille põhjal te**]** nagu ´otsustate seda. (373_a5)

meil on olnud see .hh ee {-} kapslid=ee **[RP** äädi- + ´õunaäädikakapslid**]**. (373_a5)

on **[RP** sel- + sellega**] +/** vot ´see oli teine toode mis mind ´ka huvitas just. (373_a5)

`teie **[RP** selli- + sellist**]** `tööd ei `tee. (475_b14)

et=kui=teil=nüd=see X parool on `päris **[RP** är- + meelest `ära**]** läinud (398_a6)

[ee üks]`kõik millisesse pangakontorisse sobib `minna, saate (.) uue **[RP** pa- + püsipa`rooli**]** sis (398_a6)

.hh tere ma paluks (.) **[RP** rahandusülikon- + **{F** õõ**}** rahandusosakonnast**]** =õ (.) `Siili telefoninumbrit. (427_a1)

peale `seda on siis `õhtusöök, mis kuulub sis ka teie majutus**[RP** paki=si- + pa`keti**]** `sisse. (427_b16)

.hh kõrva kurgu arst **[RP** on nüüd üheteistkümne- + üksteist september tuleb tööle**]**, ta on praegu kahenädalasel puhkusel. (428_a36)

ma ei oskagi öelda, ja **[RP** siis edasi perearsti juu- + **{D** tändab**}** siis edasi kõrva kurgu arsti juurde**]**, (.) järjekorda panema (428_a36)

[RP [pa-] + nägemist**]**? (428_a36)

ee:=hh on see **[RP** `nädalavahe- + **{D** või noh [tähendab]**}** sin `pühade ajal**]**. (452_a3)

ega teil **[RP** `neid=i- + informatsiooni**]** ei=`ole. (452_a3)

[RP tääh + (1.5) tändab**]** =siin `viis või `kümme krooni ehk `on: (1.0) ja= **[RP** m- + mitte**]** `iga kord. mitte `iga liin see `oleneb `liinist=ja. .hhhh (1.8) et se=on `Põlva `kaudu. (1.0) mingi (.) `mikrobussiliin isegi läheb nagu `Kuressaarde aga noh see on `era`liin. (452_a3)

[RR sei- + seitse**]** `viiskend=ja (0.8) `kaheksa kakskend=ja no `hästi `tihedalt (452_a3)

tere. palun **[RP** L- + `Lunini**]** lastehaigla. seal mingi informatsi`oon `ka on. (460_a15)

järgmine on **[RP** kuusteist nul- + kuusteist null**]** null kuusteist nelikend=viis=[jah.] (460_a15)

Asendused

ee **[RP** `Kemeri poolt > + Kemerist**]** on see mul `ostetud Tartust (91_a3)

ja noh (.) nädala pärast tuli **[RP** see `monitor, + sama monitor**]** `tagasi, (.) täpselt sama `targalt. (91_a3)

`kui ma `homme `ostan `ära ühe (.) `voodi, (.) **[RP** kas see > homme `hommikul, **+** kas see `tuuakse mulle päeva=jooksul**]** `ära ka. (96_b8)

no **[RP** `meile, + `meie**]** präegu siis noh pa`ketis ei `ole `küll. (233_a4)

[tähendab]=et siis on niimodi=et=ee sis `teie võite tulla nüüd enne ok`toobrit, teha pangas tudengi`paketi sõlmida, siis te saate nagu **[RP** laenu{lt} + laenu**]** ja kõik `tasuta, muidu (.) on sellest aastast meil laenu taotlemine kolgend `krooni. (334_a5)

aa, a kas **[RP** mingi + `käendajatel peab mingi**]** kindel: .hh noh palga `suurus ka [`olema.] (334_a5)

ei=ole. (.) ja `kui neil palk (0.5) panka ei `laeku siis nad võtavad töö=juurest omale selle `tõendi **[RP** kolme kuu + (.) `kolme `viimase `kuu**]** `kohta. (334_a5)

[.hh] e mind uvitaksid Tallinnas asuvad **[RP** `kirjastused + raamatukirjastused**]**. (338_a1)

ee näiteks (.) **[RP** kas=teil + kas `Sinisukk**]** on ka `Tallinna (0.5) [oma]. (338_a1)

teil **[RP** peab + =hh kõik aine´punktid peavad**]** olema **[RP** nende: eu + hh ´erialati**]** ´koos tähendab oma ´peaaine omad peavad koos olema ja ´kõrvalained mis teil sis ´võetud on ja noh ´üldse ka ütleme kokku see sada=´kuuskend ainepunkti [noh,] (338_a3)

.hh no: tändab eks te sis ´seletage seal ´õppetoolis=et vot need ja need **[RP** ma need, **+** ma veel**]** teen ´ära eks. (338_a3)

ahaa. .hh ee ma kuulsin=et tuleb võtta: **[RP[** ühiselamutest + või sealt] + **{F** =ee**}** üliõpilas´külast**]** ka mingisugune ´tõend (338_a3)

no see on ´päris ´lõpus kui te juba tõesti **[RP** ´lõpetanud ´olete + **{D** või nimodi=et=noh**}** ´kaitsnud ´olete**]** ikka ´selge on et [te] tõesti ´lõpetate (338_a3)

selle te võiksite suhteliselt ´kiiresti ära teha **[RP** nii=palju +´nii kiiresti**]** kui te nüd ´jõuate sellepärast=et=ku=äkki ´selgub et teil on järsku ´mõni aine ´veel teha või ´midagi eks (338_a3)

[jah,] vot ´seda peaksite nüd minema **[RP** oma sinna: + oma ´õppetooli**]** kus te ´kaitsete. (338_a3)

mhmh (0.8) aga mingit ´firmat kes seal **[RP** seda + (.) pügi´veoga**]** ´tegeleb. (347_a2_info)

siis seal (.) `lähikonnas **[RP** on nüüd + (.) on `veel**]** (.) üks ho`tell see on `Bernhard. (353_a3)

see on sin (.) **[RP** selle (.) kauba`halli + uue kauba`halli**]** nagu (.) noh `kõrval {või} õigemini too otse `maja kohe jõe `ääres. (353_a3)

minu teada pole **[RP** sin: + midagi seal**]** `alustatud. (353_a3)

jah, **[RE** pigem + pigem**]** siis **[RP** `mujale + juba .hh `lõuna poole**]** nagu minna. (353_a3)

kas te .hh nüüd tahate `ringi sõita või tahate `paikselt **[RP** kusksi + kuskil**]** noh `pikemalt olla (353_a3)

[RP kuidas + kas**]** ´trepikojad on alt lukustatavad. (354_a4)

ja **[RP** =ta=on {muide} + tal on**]** vist süsteem nimodi=et nädal `Põlvas ja nädal `Võrus tööl. (354_a8)

et niimodi: `kaugemale välismaale **[RP** me jah, + `meie büroo**]** nagu=ei=ole orien`teeritud. (355_a7)

et ja=noh **[RP** `mäed=on `lahti: + **{D** tähendab**}** `tõstuk töötab**]** kella `neljani (355_a11)

ikkagi=noh `erinevatel `päevadel=on võimalik=sis **[RP** `mägi + mäge**]** valida. (355_a11)

see **[RP** =on + jääb**]** tee `äärde. (356_a4)

loomu`likult. .hh `mina ei ole kahjuks `ise `isiklikult **[RP** seal `veel `käinud, + sinna veel `sisse `saanud**]**. (356_a4)

siis veel `see küsimus **[RE** =et + et**]** kas se=on ainult see `veepark või **[RP** seal + on seal**]** midagi `juures ka (356_a4)

see maaalune ekskursioon kestaks kuskil `täiendavalt tund aega siis jääks lihtsalt `Pihkva peal aega liiga **[RP** `vähe. + (0.5) väheks**]** (357_a6)

.hh **[RP** ta on kuskil: + .h **{D** tändap**}** viie`teistkümnendal sajandil on ta**]** `asutatud (357_a6)

kas see täitub nüüd `tänase päevaga ja kas ta täitup: mingisuguse väikse `grupi näol või ta tilgub **[RP** üks + (.) `ühe**]** inimese kaupa (.) ma ei oska (357_a6)

sis ma küsiks veel `reisi kohta=et (.) **[RP[**`kas + **{F** =ee**}** **[RE** `kui + (0.8) **{F** ee**}** kui**]** `täis see grupp juba on (359_a3)

ja muidugi kui te saate Slovakkia **[RP** neid + **{F** =ee ee**}** `krooni**]** võtta **[RE[** sis + sis**]** + sis**]** se on `veel parem variant. (359_a3)

nii=et=see=on kõige (.) rohkem nagu (.) levinud **[RP** valuutakurss + [või] valuuta**]** jah. (359_a3)

sest meil oli sellised **{X** näiteks**}** .h juhtumid kui Ungaris **[RP** ho`tellis + kolmetärnilises hotellis**]** .h {-} inimene ronis ee: vannist `välja ja `leidis kuskilt mingi `orgi (359_a3)

mt ja et **[RP** `mis see=on + mis=seal=`on**]** (367_a4)

.hh no ´mina ei oska ütelda see {-} **[RP** mis me siin ´teeme + mis me ´välja niimodi ´saadame**]**, .hh sis see on meil kuskil ´viiekümne ´viie: kuuekümne ´krooni ümber on ´kilo ´hind (380_a7)

ma teen **[RP** sellisel + **{D** =nagu=ütleme**}** ´suurele ´plaadile**]** **[RE[**või .hh + või**]** + =või**]** vi´neeralusele **{X** ütleme teen**}** .hh sellistele suurtele ´üritustele oleme ´teinud (380_a7)

siin on `märgitud=et=ee (.) **[RP** kahe + kahest**]** kuni `viie tärni hotellini (384_a2)

[RP `oma: + `enda**]** elu kohajärgsest pan[ga`kontorist.] (398_a6)

et=ee **[RP** meil on nagu nii `laiad `katted, + **{D** tändab**}** =ned `taldrikud **[RE** ja + ja**]** `nõud asetatakse nii `laialt] .hh et põhimõt=siukest `küünarnukitunnet ei: tohiks `tekkida. (427_b16)

a **[RP** et ma peaks + **{F** m**}** (.) et nüüd peaks**]** nigu perearstiga rääkima (428_a36)

tere? (0.5) oskate mulle `öelda: (0.8) `Tartumaal Mäksa **{X** `valla seda**}** (0.8) `valla **[RP** seda + `maja**]** numbrit või=mis=se=`on=se. (456_b28)

< siis on > (1.2) õõ `Tartu `Maarjamõisa huvi`keskus, (0.5) see **[RP** tegeleb + tegutseb**]** nüd aadressil `Puusepa `kümme, siin on `laulumängu`ring. (475_b3)

[ja] sama`moodi on ka se `Lelula, (0.8) **[RP** et `Lelulan on `ka + et Lelulas on `ka**]** sin laste `mängutuba (475_b3)

.hh < käisin eile küsimas ühte: pistiku: `otsa mis käiks `mikrofoni teeks **[RP** stereost `monoks. > + **{D** või `vastupidi tähendab**}**, [monost `stereoks**]**. .hh s::e]=on `vale `asi. (475_b14)

ähh ma kardan jaa et meil=on `aega selleks üsna `vähe=et meil on siin kliendid `käivad **[RE** =et + (.) hh et**]** meil lihtsalt **[RP** sellist + **{D** nii-öelda**}** (0.5) .hh `süvenemiseks**]** pole eriti `aega. (475_b14)

.hh ää no=tähendab oõ `pistikud meil `on, aga meil ei=`ole sellist ee noh, (.) `üleminekutükki **[RE** nagu=te + .hh nagu te**]** `saite **{A** noh sellist**}**. (.) .hh < et > meil=on=noh (.) niiöelda **[RP** `tinutatavad `pistikud + `otsatinutatavad pistikud**]**. (475_b14)

Lisamised

´meile on täiesti piisanud ´sellest kui sinnasama väljavõtte peale kirjutatakse ´üles **[RE** et + **{D** noh**}** =et**]** **{A** ´ülevale kuskile nimodi**}** **[RE** =et + et**]** =noh aja´kirjandus kõrval´aine näiteks seal keskastmes et noh keskastmeni on ´korras ütleme **[RE** =et .hh +[et**]** võib] diplomi peale ´kirjutada=et teil on see (.) ´aine kesk´astmes. (338_a3)

aga ned ´ained nagu oma ´peaainete=ja kõrvalainete ´kaupa te peaksite nendes (0.5) ´õppetoolides ´ära kontrollima **{X** mis**}** (.) **{A** neid ´aineid õpetavates õppe´toolides**}**. (338_a3)

nojah, ee võibolla=et kui mingi: `asutuste sellised `üritused **{A** ühe=kahepäevased**}** (353_a3)

no sääl `lähikonnas on nüüd `päris `mitu (.) sellist=ee peatus (.) `kohta (.) kus `võiks nagu ütleme `olla. (0.5) (.) kohe Pühajärvel `endal on nüüd ho`tell, (.) **{A** `päris `korralik ho`tell**}**. (353_a3)

.hh nii **[RE** =et + =et**]** süia `saab **[RE** =ja + =ja**]** (.) minu teada on ka seal `Aura=keskuses **{A** ma võtan kohe selle präägu `lahti**}** on ka `seal ju kohvik `olemas. (356_a4)

ja sis lapse hind on: **{A** ma vaatan siin vel**}** kaks tuhat öheksa`sada (359_a3)

see=on nimodi=et `erinevad hinnad vastavalt `vanusele. `täiskasvanule on `kolm=tuhat=`nelisada üheksakümend, .hh `lapsele **{A** `kuue kuni `kaheteistaastane**}** on `kaks=tuhat=`nelisada üheksakümend, .hh ja: `kõige väiksem laps on `tuhat seitsesada üheksakümend. (367_a4)

Tartu `kristlik `perekeskus, siin on nüt `väikelaste laulu`ring > **{A** või=tähndab beebide `ja väikelaste kuus `kuud kuni neli `aastat**}**. (475_b3)

.hh tere. (.) .hh tahaks `teada kas: on teil `andmeid (0.5) laste: (.) **{A** mitte `beebikoolide aga noh selliste `kaheaastaste laste**}** (.) .hh mingite `laulu`ringide kohta. Tartus. (475_b3)

.hh võtaks selle `mängumaa `ka, **{A** Anni `mängumaa**}**. (475_b3)

mnjah, ma vaatan präegu=et sellist nagu ei `paistagi siin **{A** sellist vari`anti**}** etkel. (475_b14)

keegi peab selle `pistiku nagu otsast `ära: võtma **{A** selle `präeguse oleva**}** ja sis sinna teise pistiku nagu `asemele panema. (475_b14)

.hh ää no=tähendab oõ `pistikud meil `on, aga meil ei=`ole sellist ee noh, (.) `üleminekutükki **[RE** nagu=te + .hh nagu te**]** `saite **{A** noh sellist**}**.  (.) .hh < et > meil=on=noh (.) niiöelda **[RP** `tinutatavad `pistikud + `otsatinutatavad pistikud**]**. (475_b14)

Muu

jah, et kui päris niiöelda **{X** võ-**}** `väikese reisi Ka`naari saartele ei `võta sis iga `maa pakub ikka väga palju $ [`huvitavat] ku `vaadata. $ (233_a4)

aga ned ´ained nagu oma ´peaainete=ja kõrvalainete ´kaupa te peaksite nendes (0.5) ´õppetoolides ´ära kontrollima **{X** mis**}** (.) **{A** neid ´aineid õpetavates õppe´toolides**}**. (338_a3)

seal on võimalik ju jalgrattaga minna `Käärikule on võimalik sõita (353_a3)

kas te näiteks `kaarte või: mingit selliseid asju teil `ka on saate näiteks `kaarte välja laenutada või midagi taolist. (353_a3)

siis=ee (.) kas: `on ka nagu mingeid ma=i=tea `toidu kuidas öelda `pause või selliseid. (356_a4)

ee ma `kohe võtan selle `välja=hh. .hh niimodi `Aura `keskus. tähendab `veepark ja tähendap (0.8) seal=on `ujula koos õppebas[`seiniga,] (356_a4)

.hh ja seal on ka `terviseklubi noh ee .hh `mullivannid auru aroomi ja `leilisaunad (356_a4)

aga veel `üks asi **[RE** =et + =et**]** + =et**]** see on siis aint nagu **{X** =se**}** kooli`vaheajal=et nagu `hiljem te neid ei=`korralda **[RE** või + =või**]** kuidas see (356_a4)

nii=et=noh `demm `on **[RE** ja + =ja**]** jääb **{X** veel**}** pikaks ajaks veel käivesse (359_a3)

ma teen **[RP** sellisel + **{D** =nagu=ütleme**}** ´suurele ´plaadile**]** **[RE[**või .hh + või**]** + =või**]** vi´neeralusele **{X** ütleme teen**}** .hh sellistele suurtele ´üritustele oleme ´teinud (380_a7)

et ühesõnaga ma=san=aru **{X** =et=see aast- kuni ühe`teistkümnendast üheteistkümnenda `jaanuarini**}**=et=siss ühe`teistkümnendast `jaanuarist peaks sis olema `odavam. (384_a2)

ahah .hh aga kas see `aasta`vahetus **{X** näiteks aastavahetus**}** kas=seal=on aastavahetuse prog`ramm **{X** seal**}** kohe või **{X** mis=sis**}** on iga inimese enda `teha=et mida ta `teeb seal. .hh või: on teil kohe selline (384_a2)

siis `algab etendus kell `seitse? .hh ja kestab ligi `kolm `tundi ja:: `vaheaeg on: `kahekümne`minutiline. .hh et siis me pakume `kohvi **{X** nagu kohvi`paus on siis vahe`ajal**}**. (427_b16)

tere? (0.5) oskate mulle `öelda: (0.8) `Tartumaal Mäksa **{X** `valla seda**}** (0.8) `valla **[RP** seda + `maja**]** numbrit või=mis=se=`on=se. (456_b28)

meil kül `präegu sin `kohapeal **{X** meil**}** **[RE** sellist + `sellist**]** vari`anti ei `paista. (475_b14)

Lisa 3 – Märgendatud kordused

Katkestuskohaga

ma nagu `ise **[RE** seda + (0.5) seda**]** (.) asja konkreetselt ei tea `üldse (91_a3)

ahah, ahah vabandage mul see `nimi **[RE** ei [tul]nud + .hh ei tulnud**]** nagu hästi `ette. (259_b7)

ee jah võiks `küll **[RE** aga + (.) aga**]** äkki oleks ikkagi juba=sis `järgmine `nädal. (259_b7)

.hh võin aint `öelda selle **[RE** =mis + (0.5) mis**]** väljub kaksteist `viisteist. (334_a3)

[ee] `info **[RE** ja: + =ja**]** võibolla (.) kas teil neid `õppetoolide numbreid `ka on. (338_a1)

.hh **[RE** aga **+** (.) aga**]** nüd mis sellesse ´lõputöösse puutub sis et kui te tahate nüd ´aktusel saada: (.) ´diplomit siss **[RE** peate + (.) peate**]** {-} (338_a3)

´meile on täiesti piisanud ´sellest kui sinnasama väljavõtte peale kirjutatakse ´üles **[RE** et + **{D** noh**}** =et**]** **{X** ´ülevale kuskile nimodi**}** **[RE** =et + et**]** =noh aja´kirjandus kõrval´aine näiteks seal keskastmes et noh keskastmeni on ´korras ütleme **[RE** =et + .hh [et**]** võib] diplomi peale ´kirjutada=et teil on see (.) ´aine kesk´astmes. (338_a3)

ma=ei=kujuta=ette=et ise olen n:agu: `Lääne=Eestist pärit **[RE** =et + (.) et**]** meeldib see `kuppelmaastik ja=nii=edasi **[RE** =et + .hh et**]** just `mõtsime=et ei ole nagu matkamas `käinud (.) sellise `kümne=viieteist (.) ee `inimesega (353_a3)

nojah, ee võibolla=et kui mingi: `asutuste sellised `üritused ühe=kahepäevased otsitakse kohta kuskohal **[RE** võiks + .hh võiks**]** niiöelda: nädalavahetusel `aega nagu `veeta (353_a3)

ise oleme mõtelnud kõige rohkem nagu `rattamatka vari`anti **[RE** =et + .hh et**]** ei ole seal kandis `käinud, et (.) noh tõesti ei tea kus oleks võimalik `ööbida ja mis see hea `trajektoor oleks kus seal `sõita **[RE** =ja + (.) ja**]** nii `edasi. (353_a3)

no sääl `lähikonnas on nüüd `päris `mitu (.) sellist=ee peatus (.) `kohta **[RE** kus + (.) kus**]** `võiks nagu ütleme `olla. (0.5) / ee `üks **[RE** on + on**]** **+/** (.) kohe Pühajärvel `endal on nüüd ho`tell, (.) `päris `korralik ho`tell. (353_a3)

[RE[selle] + (.) selle**]** kokkuleppe ma nendega nagu `saaksin (353_a3)

ma ei ole `ise nüüd seda uuesti `näinud sest nad **[RE** alles + .hh alles**]** reno`veerisid seda `keskust (353_a3)

siis ta peab ikka päris **[RP** ha- + `heal**]** tasemel **[RP** ole- + olema**]** **[RE** nagu + .hh nagu:**]** juba reno`veeritud see `maja. (353_a3)

praegu kohe ei karga `pähe mis selle motelli (0.5) nimi `oligi ma pean kuskilt mõnest materjalist **[RE** nagu + (.) nagu**]** `vaatama. (353_a3)

ee talvel muidugi **[RE** saaks + (.) saaks**]** seal nagu (.) `suusatada (353_a3)

mõned kohad võibla: võib `tunde järgi $ **[RE** nagu $ + (.) nagu**]** `vaadata (353_a3)

`kala on seal `ka seda võib sealt ka saada seal saab isegi `suitsetada ja saab `kokku leppida=ja=noh selles mõttes **[RE** see + .hh `see**]** pool on seal küll `olemas. (353_a3)

aga noh kui te tahate nüd **[RE** oma + .hh oma**]** ütleme `seltskonnaga `minna (353_a3)

ma: saan teile siin ka `kaardi päl juba `näidata=et **[RE** kuhu + .hh `kuhu**]** nagu võiks minna **[RE** =ja + =ja**]** mingit `orientiire ja nende majutus`kohtade kohta. (353_a3)

et ee noh seal **[RE** võiks: + võiks**] [RE** päris + päris**]** vabalt nagu .hh `ööbida (353_a3)

kuna mul on saksa `rühmad seal `peatunud just **[RE** selles: + selles**]** mo`tellis, (.) siss nad on seal: `jalgratastega natukene `ringi `sõitnud eelkõige just=ee ümber `Pühajärve. (353_a3)

jah, sest kui teil=on=ikkagi `rühm **[RE** sis: + sis:**]** `eelkõige (0.5) muidugi soovitaks=siit `läbi tulla (353_a3)

[RE meie: + meie**]** niiöelda: saame juba: nendega `läbi rääkida ja mingisuguseid `boonuseid seal `välja kaubelda. (353_a3)

.h **[RE** aga: + aga**]** ma saan teile anda `Hermann reiside telefoninumbri. (355_a7)

kokku=on ültse **[RE** kaheksa päeva: + kaheksa päeva**]** seda `sõitu (355_a11)

[RE =ja: + =ja**]** siis=ee veepargi `pilet. (356_a4)

sest=`kui tahetakse `ainult ujulat sis=see .hh `ujula osa on **[RE** ju: + ju**]** noh `tavaline `ujula (356_a4)

niimodi et me `oleme saanud vaadata aga `seda me ei `luba. tähendab **[RE** =sinna: + sinna**]** nagu mingisse kirikusse `sisse. (357_a6)

[RE ja: + **{F** ee**}** ja**]** (.) `pükse nad üldiselt ei taha `ka. (357_a6)

see maaalune ekskursioon kestaks kuskil `täiendavalt tund aega siis jääks lihtsalt `Pihkva peal aega liiga **[RP** `vähe. + (0.5) väheks**] [RE** =et + **{D** =noh**}** =et**]** `see peaks olema juba nagu `kahepäevane reis võipola siis. (357_a6)

aga ned olid midagi (.) ehh ((ohkab)) **[RE** vist + **{F** =ee**}** vist**]** < `nelisada Eesti > krooni se nädalane `pilet või midagi `nii. (359_a3)

sis ma küsiks veel `reisi kohta=et (.) **[RP[**`kas + **{F** =ee**} [RE** `kui + (0.8) **{F** ee**}** kui**]** `täis see grupp juba on (359_a3)

nii **[RE** =et + .hh et**]** pidid=ee selle kiirabi `välja kutsuma. (359_a3)

kui se reisikindlustus on `olemas **[RE** sis + .hh siis**]** noh osutakse `esma**[RE** abi: + abi**]** (359_a3)

[RE bussis + **{F** ee**}** bussis**]** `kindlasti on (.) kuum `kohv, (.) `tee, (.) `puljong. (359_a3)

`kes närib `porgandit, kes õuna kes {-} `võileiva kes `mida **[RE** aga + .hh aga**]** millegipärast tõesti s(h)öögiisu tekib praktiliselt `kohe. (359_a3)

[mhmh] .hh aga `küsiks natuke selle kohta **[RE** et=ee**}** + (0.5) et=ee**]** mis `hinnaklassis **[RP** se- + see**]** `on. (367_a4)

ma: (.) uurin ´innakirju: **[RE** et: + (0.8) et**]** oleks vaja ´torti: ühele ´suurele üritusele et ma=i=tea

(0.5) kas=see: (.) ee (.) ´hind [sõltub ´kilost.] (.) **[RE** või + =[või**]** .hh oleneb kus´kohta ja **[RE** mis: + **{F** =ee**}** mis**]** ´torti. (380_a7)

vaatame ültse **[RE** kas + (0.8) kas**]** `on lende. (403_a1)

et **[RE** =`täna: + täna**]** öeldi `ära=et muidu oli `kinni aga:=on `olemas `küll. (427_b16)

et **[RE** kas: + kas**]** on võimalik septembris ära teha. (428_a36)

vot seda ma küll ei oska öelda, **/ [RE** see + (.) see**]** on jah kõrva kurguarsti **+/** (.) kõrva kurguarst oskab sellele vastata. (428_a36)

et lastekeskus Anni `mängumaa näiteks et seal on **[RE** nagu + (1.2) n::agu**]** =et `mängutoa kasutamisvõimalus (475_b3)

[RE et + **{F** =ää**}** et**]** ühesõnaga **[RE** =siss + .hh < siss:**]** =e > **[RE** täpselt + `täpselt**]** üsõnaga seal kus on `auk peaks olema niöelda `pistik ja `vastupidi jah (475_b14)

siis on äkki vaja se **[RE** `pistik + (.) pistik**]** seal `lihtsalt ära vahetada. (475_b14)

.hh ää no=tähendab oõ `pistikud meil `on, aga meil ei=`ole sellist ee noh, (.) `üleminekutükki **[RE** nagu=te + .hh nagu te**]** `saite **{X** noh sellist**}**. (475_b14)

ähh ma kardan jaa et meil=on `aega selleks üsna `vähe=et meil on siin kliendid `käivad **[RE** =et + (.) hh et**]** meil lihtsalt **[RP** sellist + **{D** nii-öelda**}** (0.5) .hh `süvenemiseks**]** pole eriti `aega. (475_b14)

Katkestuskohata

noh **[RP** pärisk- + päris kella**]** `kuueni **[RE** ärge $ + ärge**]** vast `jätke. (96_b8)

`kuna ta ära tuuakse, `peale kuut **[RE** või + või**]** =kusagil (96_b8)

[RE aga + aga**]** no=näiteks kui ka `kevade poole `on, et `on midagi. (233_a4)

[RE et + =et**]** `selles osas noh saate nagu `valiku teha (233_a4)

olenevalt tõesti ku- [reisi] `pikkusest **[RE** ja + ja**]** `rahast mis te selle eest ta[hate] `maksta. (233_a4)

.hh **[RE** aga + aga**]** noh siis ´seletate=et=näete=et selle ma ´teen veel, selle ma ´teen veel ja selle ma ´teen veel. (338_a3)

ee mis ´ajaks mul peab olema noh need hh vaadatud ned ee=hh [(.) et **[RE** ainete + ainete**]** kaupa et] (338_a3)

V: [RE minge + minge**]** ´aegsasti ´rääkige, ´öelge=et noh=et te arvate et te jõuate valmis umbes **[RE** sellel + sellel**]** ´ajal, [et] millal nad neil ´on. (338_a3)

.hh ee kas ainult=ee **[RE** ´üks + ´üks**]** variant **[RE** või + või**] [RP** kak- + **{D** või=tändab**}** kahe**]** (338_a3)

´meile on täiesti piisanud ´sellest kui sinnasama väljavõtte peale kirjutatakse ´üles et noh=et **{X** ´ülevale kuskile nimodi**} [RE** =et + et**]** =noh aja´kirjandus kõrval´aine näiteks seal

keskastmes et noh keskastmeni on ´korras ütleme **[RE** =et + .hh [et**]** võib] diplomi peale ´kirjutada=et teil on see (.) ´aine kesk´astmes. (338_a3)

et ee (.) `kuidas see nagu `välja näeks ise mõtlesime küll=et kas `Otepää kandis sis **[RE** või + või**]** `seal kuskil näiteks. (353_a3)

muidugi vari`ante on seal `palju tegelikult sest **+/** (.) **[RE** ja + =ja**]** noh (0.5) muidugi sõltub sellest `ka kas te .hh nüüd tahate `ringi sõita (353_a3)

no sääl `lähikonnas on nüüd `päris `mitu (.) sellist=ee peatus (.) `kohta **[RE** kus + (.) kus**]** `võiks nagu ütleme `olla. (0.5) **/** ee `üks **[RE** on + on**] +/** (.) kohe Pühajärvel `endal on nüüd ho`tell, (.) `päris `korralik ho`tell. (353_a3)

et ee noh seal **[RE** võiks: + võiks**] [RE** päris + päris**]** vabalt nagu .hh `ööbida (353_a3)

et `selle peale on ka nagu `mõeldud. .hh **[RE** ja + =ja**]** `siis on: veel mm: umbes kuus kilomeetrit `Pühajärvelt (.) on üks mo`tell. (353_a3)

nende `suurte maantedega ültse `arvestada ei `saa. .hh {peap} mingist `väiksemad nagu: `teed olema **[RE** kus: + kus**]** on: vähe `turvalisem selle `jalgrattaga. (353_a3)

see `järv on ikkagi `suur ja seal `järve ääres leiate `kindlasti koha **[RE** kus + kus**]** `telkida. (353_a3)

[jah,] (.) seal `suvel oli nüd küll mingid Kallaste `päevad toimusid, .hh aga noh **{X** `see oli `ka eelkõige**}** `rõhk oli seal `käsitööle **[RE[** =ja mt + ja**]** + =ja**]** noh sellistele .hh ee **[RE** kodus + kodus**]** valmistatud `esemetele. (353_a3)

[RE [aga] + =aga**]** =vat `majutamise pool on säl `vilets. (353_a3)

jah, **[RE** pigem + pigem**]** siis **[RP** `mujale + juba .hh `lõuna poole**]** nagu minna. (353_a3)

ma: saan teile siin ka `kaardi päl juba `näidata=et **[RE** kuhu + .hh `kuhu**]** nagu võiks minna **[RE** =ja + =ja**]** mingit `orientiire ja nende majutus`kohtade kohta. (353_a3)

`sinna läheb sis sisse meil üks `tunniajane ekskurs`joon mööda Tartu `linna see on [`Toome]mägi **[RE** =ja + =ja**]** (.) kõik=se `ülikool=ja [niuke] kõik. (356_a4)

see `plaan küll `on **[RE** et + et**]** neljakesi `minna. (356_a4)

.hh nii **[RE** =et + =et**]** süia `saab **[RE** =ja + =ja**]** (.) minu teada on ka seal `Aura=keskuses **{X** ma võtan kohe selle präägu `lahti**}** on ka `seal ju kohvik `olemas. (356_a4)

siis veel `see küsimus **[RE** =et + et**]** kas se=on ainult see `veepark või **[RP** seal + on seal**]** midagi `juures ka (356_a4)

aga meil on: niimoodi=et meil .hh `bussi`juht (.) seisab `järje`korras [ja] `elavast järjekorrast **[RE** võtab + võtab**]** meile `piletid. (356_a4)

aga veel `üks asi **[RE** =et + =et**]** + =et**]** see on siis aint **{X** nagu=se**}** kooli`vaheajal=et nagu `hiljem te neid ei=`korralda **[RE** või + =või**]** kuidas see (356_a4)

[RE[enne] + enne**]** `jõulu mingisuguse `reisi. `praegu on meil ainult see **[RE** `teine + teine**]** no`vember. (356_a4)

[meil] on olemas küll .hh ee: reisi `kirjeldus, täendap selle `päeva `kohta, ja niukene `üldsõnaline noh .hh (0.5) milliseid objekte te `näete ja selliseid asju **[RE** =ja + ja**]** seal `vahemaad ja `sellised. (357_a6)

.hh no: m:a ei `tea=h. **[RE** ma=i + ma=i**]** oska ju `öelda, `praegu on meil kuskil `kümmekond `kohta veel. (357_a6)

et teil on `aeg **[RE** ja + ja**]** see marshuut `sobib teile. < (357_a6)

see on väga ilus `koht, **[RE** ja + ja**]** väga ea o`tell on meil (359_a3)

seal on .hh ee: ned (0.5) `vabaõhumuuseumid **[RE** ja + ja**]** `koopad ja seal on (.) **[RE** väga + väga**]** nisugune uvitav programm `kah mis me pakume sinna `juurde. (359_a3)

kui te jätate oma telefoni`numbri=siss **[RE** või + =või**]** `faksi ((3. välde)) või `meili ((3. välde)) / või **+/** (.) sis saab natuke `aja pärast teile juba `vastata. (359_a3)

Tatranska on `üks sõna, **[RE** ja + =ja**]** Lomnitsa on `teine. (359_a3)

aga seda: (.) mingit=e `interneti aadressi nagu teil ei ole **[RE** et + =et**]** kust **[RP** k- + **{F** ee**}** (.) kohast**]** [võiks infot leida.] (359_a3)

ja se on kokku öheksa `päevane `reis, (.) **[RE** ja + ja**]** maksab `tõesti nii nagu ma ütsin kolm=tuhat kakssada `seitsekümend `krooni. (359_a3)

ei `ole, (.) **[RE** kõik + kõik**]** ööd on otellides (359_a3)

ei **[RE** täitsa + täitsa**]** vii:sakas o`tell on. (359_a3)

[RE [muidugi] + muidugi**]** [sest] `all on ju suured pagaashiruumid (359_a3)

nii=et sinna lähevad ilusti nii `suusad **[RE** kui + kui**]** =ee **[RP[**varus- + või see**]** + **{F** ee**}** `kotid**]** (359_a3)

sellist juhtumit ei `mäleta **[RE** aga + aga**]** `Ungaris näed oli `küll (359_a3)

[RE nii=et + nii=et**]** `seal saab natuke ee-ee keha kinnitada (359_a3)

[RE ja: ja=`siis ja=siis=ja=siis=ja=siis**]** vist (.) Slo`veenia. (0.5) nende {korundidega} **[RE** oli + oli**]** probleem aga ülejäänd va`luuta on ikka täitsa `olemas. (359_a3)

nii=et=noh `demm `on **[RE** ja + =ja**]** jääb veel pikaks ajaks **{X** veel**}** käivesse (359_a3)

ea=`küll. ai`täh teile info eest **[RE** ja + ja:**]** ma räägin sin oma perega `läbi (359_a3)

mm **[RE** siis + siss**]** sealt veel kõik tuleb (367_a4)

[RE[et] + et**]** siin on ki- /+ et noh **[RP** ´pea- + kui ´peaks**]** juhtuma et ma=i=´ole äkki ´rahul, (.) [et siin on kir-] (373_a5)

[RE et + et**] [RP** kas te k- + mille põhjal te**]** nagu ´otsustate seda. (373_a5)

ma: (.) uurin ´innakirju: **[RE** et: (0.8) + et**]** oleks vaja ´torti: ühele ´suurele üritusele et ma=i=tea (0.5) kas=see: (.) ee (.) ´hind [sõltub ´kilost.] (.) **[RE** või + =[või]**]** .hh oleneb

kus´kohta ja **[RE** mis: + **{F** =ee**}** mis**]** ´torti. (380_a7)

et kuskil **[RE** ´nii + nii**]** ´kaugele. (380_a7)

[RE et + **{F** =ää**}** et**]** ühesõnaga **[RE** =siss + .hh + < siss:**]** =e > **[RE** täpselt + `täpselt**]** üsõnaga seal kus on `auk peaks olema niöelda `pistik ja `vastupidi jah (475_b14)

mhmh no `sisend on=se `auk **[RE** =mis + mis**]** seal `sees on. (475_b14)

[[RE =kui + kui**]** + kui**]** ta /+ (0.5) **[RE** {seda} + =seda**]** `tükina siin ei `ole. (475_b14)

meil kül `präegu sin `kohapeal **{X** meil**}** **[RE** sellist + `sellist**]** vari`anti ei `paista. (475_b14)

`jah, nii nagu kokkulepe `oli=et kui ei=`sobi=et **[RE** =siis + [siis]**]** saab `tagasi=anda. (475_b14)

aa. .hh siin=on=nüd nii **[RE** =et + =et**]** `aastavahetuse hinnad ei=`olnud `esime=nädal on natukene `kallimad **[RE** =et + =et**]** `üheteistkümnendast jaanuarist (.) lähvad innad `odavamaks, ma=mõtlen sis `neid hindasid sin präegu. (384_a2)

et=ee **[RP** meil on nagu nii `laiad `katted, + **{D** tändab**}** =ned `taldrikud **[RE** ja + ja**]** `nõud asetatakse nii `laialt] .hh et põhimõt=siukest `küünarnukitunnet ei: tohiks `tekkida. (427_b16)

`Tartusse või `Tallina et noh `neid variante ma saan `ikka öelda=ja (1.2) **[RE** et + et**]** mis `kella `paiku teil vaja `oleks. (452_a3)

meil=ei=ole=jah `infot kahjuks **[RE** et + =et**]** (.) `Tartust nagu `jätkuvat **[RE** ja + =ja**]** .hh Tartu `bussijaamast ehk jah, saate seda `täpsemat `infot. (452_a3)

kirjutate neid `üles praegu **[RE** või + =või**]** (452_a3)

midagi e (0.8) `pool `kolm **[RE** või + =või**]** peale `kahte. (460_a15)

`poole kolme ajal sobis **[RE** =või. + või**]** mis. (460_a15)

Ahelkordused

e siin on muidugi neid=ee (.) **[RE[**`v:äga + =väga**]** + =väga**]** `palju. (338_a1)

´töö peate kaitsma eks ´õigeaegselt sis, .hh et te saaksite kevadel ´lõpetada
[RE[[[=ja + .hh ja**]** + ja**]** + ja:**]** (.) + ja**]** =nii=´edasi. (338_a3)

[jah,] (.) seal `suvel oli nüd küll mingid Kallaste `päevad toimusid, .hh aga noh **{X** `see oli `ka eelkõige**}** `rõhk oli seal `käsitööle **[RE[** =ja mt + ja**]** + =ja**]** noh sellistele .hh ee **[RE** kodus + kodus**]** valmistatud `esemetele. (353_a3)

.hh **[RE** et nii + =et=ee**]** + et**]** =ma ei `oska praegu hetkel teile seda `öelda (356_a4)

ma teen **[RP** sellisel **{D** =nagu**}** **{D** =ütleme**}** + ´suurele ´plaadile**]** **[RE[**või .hh + või**]** + =või**]** vi´neeralusele {X ütleme teen} .hh sellistele suurtele ´üritustele oleme ´teinud (380_a7)

.hh ja `millega te tahate minna kas te tahate seal .hh ee mingi `transpordiga **[RE[[**=või + või**]** + või**]** + või**]** =`jalgrattaga **/** =või **+/** (.) [noh] neid vari`ante võib ka `erinevaid olla. (353_a3)

aga: mis siis nagu `teie=poolt oleks sis, (.) et kas: `trajektoor on teie poolt / **[RE[**või + =või**]** + =või**]** mingit m- .hh noh, (.) et=ee (.) /+ kui `omal käel me peaksime seda tegema ja mida `teie omalt poolt saaks näiteks pakkuda oletame kui me läheks rattamatkale näiteks. (353_a3)

aga mn=ee ee tändab `Tartus **[RE[** =on + on**]** + on**]** siiski `piisavalt ka vabat `aega (356_a4)

aga veel `üks asi **[RE** =et + =et**]** + =et**]** see on siis aint **{X** nagu=se**}** kooli`vaheajal=et nagu `hiljem te neid ei=`korralda **[RE** või + =või**]** kuidas see (356_a4)

ja muidugi kui te saate Slovakkia **[RP** neid + **{F** =ee ee**}** `krooni**]** võtta **[RE** sis + **[**sis + sis**]]** se on `veel parem variant. (359_a3)

[RE[[ja: ja=`siis + ja=siis] + =ja=siis] + =ja=siis**]** vist (.) Slo`veenia. (0.5) nende {korundidega} **[RE** oli + oli**]** probleem aga ülejäänd va`luuta on ikka täitsa `olemas. (359_a3)

[[RE =kui + kui**]** + kui**]** ta **+/** (0.5) **[RE** {seda} + =seda**]** `tükina siin ei `ole. (475_b14)

[RE < [et] + =et**]** + et**]** noh > see tändab `seda=et noh keegi peab selle `pistiku nagu otsast `ära: võtma (475_b14)

[`kahe]tärnised akkavad **[RE[[**nüüd=hh (.) + nüd**]** + =nüd**]** + =nüüd**]** **+/** (.) odot=a=mis `kuupäeval te=tahtsite (384_a2)

Lisa 4 – Märgendatud valestardid

ja `nüüd on **+/** (0.5) ma=ei=`tea, (0.8) lubati: see monitor=siss (0.5) > `ära vahetada nüüd on mingisugune < (0.5) `venitamine hakanud `pihta. (91_a3)

kui teil on tudengipakett `ka juba tehtud, (.) sis peale (.) esimest ok`toobrit, (.) tulete uuesti `panka, ja=s / saate ai noh **+/** (.) `kohe tehakse teiega laenu`leping. (334_a5)

siis saate sealt **+/** minu=arust ´dekanaat väljastab niisugused ´tõendid (338_a3)

[.hh ee] (.) ja kui ´kaua **+/** (.) .h **[RP** mi + ´mis**]** ajal peab puh **+/** noh enne kaitsmist ütleme mul on viie´teistkümnes millal ma pean selle bakalaureusetöö [´esitama.] (338_a3)

et eeh ((naervalt)) (.) endal sõprade ga tekkis idee=et tahaks küll `matkama minna **/** aga: **+/** .hh mõtlesingi=et **+/** praegu `vaatasin telefoniraamatust (.) pakute sellist teenust. (353_a3)

muidugi vari`ante on seal `palju tegelikult **/** sest **+/** (.) **[RE** ja + =ja**]** noh (0.5) muidugi sõltub sellest `ka kas te .hh nüüd tahate `ringi sõita (353_a3)

ja `millega te tahate minna kas te tahate seal .hh ee mingi `transpordiga **[RE[[**=või + või**]** + või**]** + või**]** =`jalgrattaga **/** =või **+/** (.) [noh] neid vari`ante võib ka `erinevaid olla. (353_a3)

aga: mis siis nagu `teie=poolt oleks sis, (.) et kas: `trajektoor on teie poolt **/** **[RE[**või + =või**]** + =või**]** mingit m- .hh noh, (.) et=ee **+/** (.) kui `omal käel me peaksime seda tegema ja mida `teie omalt poolt saaks näiteks pakkuda oletame kui me läheks rattamatkale näiteks. (353_a3)

no sääl `lähikonnas on nüüd `päris `mitu (.) sellist=ee peatus (.) `kohta **[RE** kus (.) + kus**]** `võiks nagu ütleme `olla. (0.5) **/** ee `üks **[RE** on + on**]** **+/** (.) kohe Pühajärvel `endal on nüüd ho`tell, (.) **{A** `päris `korralik ho`tell**}**. (353_a3)

.hh et kas `teil **+/** oleks nagu huvitud **+/** see `reis maksab meil `kakssada=viiskend `krooni. (356_a4)

no=selles **+/** .hh see käib nüd selle asjaga `kokku kõik. (356_a4)

ja see on nüd **+/** `täis`pilet on siin ee (.) oopistükkis sada=kakskend=viis `krooni. (356_a4)

ja siis nüd=ee **+/** (.) `kaua see **+/** (.) tändab=et see `aeg seal `veepargis. kaua `see on. mitu [`tundi.] (356_a4)

sellistesse **+/** no üldiselt on nimodi=et meil on `õnnestunud (357_a6)

seal on: **+/** (.) **[RE[**ma=e + ma=i**]** + ma=i**]** oska tõesti öelda kui suur see terri`toorium seal on. (357_a6)

ma ei oska `öelda siis kui **+/** > tõenäoliselt on siis õige `aeg kui te otsustate et te tahate `sõita et teil on `aeg **[RE** ja + ja**]** see marshuut `sobib teile. < (357_a6)

[no=bussis=sa- **+/** (0.5) jaa, (.)] **[RE** bussis + **{F** ee**}** bussis**]** `kindlasti on (.) kuum `kohv, (.) `tee, (.) `puljong. (359_a3)

no muidugi tehakse `peatused / sest no **+/** (0.5) esimese peatuse kindlasti teete juba `Lätis. (359_a3)

kui te jätate oma telefoni`numbri=siss **[RE** või + =või**]** `faksi ((3. välde)) või `meili ((3. välde)) / või **+/** (.) sis saab natuke `aja pärast teile juba `vastata. (359_a3)

[RE[et] + et**]** siin on ki- **+/** et noh **[RP** ´pea- + kui ´peaks**]** juhtuma et ma=i=´ole äkki ´rahul, (.) [et siin on kir-] (373_a5)

on **[RP** sel- + sellega**] +/** vot ´see oli teine toode mis mind ´ka huvitas just. (373_a5)

.hh mhh ma=i=´oska nagu nimodi kohe täpselt ´öelda / et noh selline: **+/** .hh m ´tegemist on ühe üliõpilasorganisatsiooni ´aastapäevaga. (380_a7)

.hh < no kuskil:: ´kolmkümmen:d > **/** oota se oli: **+/** (0.5) päev enne esimest detsembrit on vist (.) ´kolmkümmend. (.) no´vember. (380_a7)

sest=ee ma ´ise valmistan ´siin ´laudasid ja näiteks ku ma=ei=ole ´võtnud seda .hh ee spetsiaalselt **+/ [RP** mõni tahab + siin pulma´torti nagu on [tahe]tud**]** ´aluste peal (380_a7)

ei ma: tahakski: umbes **+/** ma uurin praegu ´hinda (380_a7)

[[RE =kui + kui**]** + kui**]** ta **+/** (0.5) **[RE** {seda} + =seda**]** `tükina siin ei `ole. (475_b14)

[`kahe]tärnised akkavad **[RE[[**nüüd=hh (.) + nüd**]** + =nüd**]** + =nüüd**] +/** (.) odot=a=mis `kuupäeval te=tahtsite (384_a2)

ee ma `kuupäeva veel **+/** ma `vaatan siin et teil on alates öeksateistkümnendast (384_a2)

vot seda ma küll ei oska öelda, **/ [RE** see + (.) see**]** on jah kõrva kurguarsti **+/** (.) kõrva kurguarst oskab sellele vastata. (428_a36)

Lisa 5 – Morfoloogilised ja süntaktilised märgendid

Morfoloogilised märgendid – sõnaklassid

S – substantiiv

A – adjektiiv

V – verb

P – pronoomen

D - adverb

K – adpositsioon

J – konjunktsioon

N – numeraal

I – interjektsioon

X - muu

Y – lühendid

Z – kirjavahemärgid

Süntaktilised märgendid

- Öeldise märgendid

@+FMV - finiitne öeldis

@-FMV - infiniitne öeldis

@+FCV - *olema* liitaegades ning modaalverbid ahelverbides, finiitne vorm

@-FCV - *olema* liitaegades ning modaalverbid ahelverbides, infiniitne vorm

@NEG - verbi eitus

- Põhja märgendid

@SUBJ - alus ehk subjekt

@OBJ - sihitis ehk objekt

@PRD - öeldistäide ehk predikatiiv

@ADVL - määrus ehk adverbiaal, ka fraasiadverbiaal

- Laiendite märgendid

@AN> - omadus- ja järgarvsõna eestäiendina

@<AN - omadus- ja järgarvsõna järeltäiendina

@AD> - määrsõna eestäiendina

@<AD - määrsõna järeltäiendina

@PN> - kaassõna eestäiendina

@<PN - kaassõna järeltäiendina

@NN> - nimi-, ase- ja põhiarvsõna eestäiendina

@<NN - nimi-, ase- ja põhiarvsõna järeltäiendina

@VN> - partitsiip eestäiendina

@<VN - partitsiip järeltäiendina

@INF_N> - verbi infinitiitne vorm eestäiendina

@<INF_N - verbi infinitiitne vorm järeltäiendina

@<P - eessõna laiend,

@P> - tagasõna laiend

@<Q - kvantori järellaiend

@Q> - kvantori eeslaiend

- Muu

@J – sidend

@I - hüüatus