Õpikute keerukuse analüüs arvutitel

Hiie Asser, Heiki-Jaan Kaalep, Siret Linnas, Jaan Mikk, Kadri Muischnek, Merje Songe, Heli Uibo
Tartu Ülikool

1. Milleks õpikuid analüüsida?

Tänapäeval kulutab inimene enne tööle asumist üle veerandi oma elust selleks, et läbida kool(id) ja saada haridus. Oleks vaja, et see aeg kuluks tulemuslikult, s.t. et inimene omandaks selle aja jooksul võimalikult palju oskusi ja teadmisi. Antiikajal võrreldi õppimist veini valamisega vaadist karahvini. Tuleb valada paraja joana, mitte läigatada korraga liiga palju ega niristada liiga aeglaselt: mõlemal juhul jõuab karahvini vähem kui paraja tempoga valades.

Õppimine on protsess, mille tulemusena inimese teadmised ja oskused muutuvad. Õppimine tugineb varemomandatule, sisaldades nii teadaoleva kordamist kui uue lisamist. Õppeülesanded muutuvad raskemaks samm-sammult.

Viimasel ajal on üha sagedamini kõlanud mõte, et õppetöö koolis ei ole paljudele õpilastele jõukohane. Sellele juhiti tähelepanu ka OECD raportis (Juurak 2001). Mittejõukohane võib tähendada nii seda, et õppetöö on liiga raske, kui ka seda, et ta on hoopis liiga lihtne. Liiga keeruline ja mahukas õppetöö takistab õpilaste arengut, sealhulgas nende mõtlemise arengut. Samas, ka liiga lihtne ei ole hea: liiga väikeste sammudena uue omandamine tähendab aja raiskamist.

Õppekirjandus on õppekava kõige detailsem esitus. Sellest lähtuvad tavaliselt õpetajad oma tunde planeerides, õppekirjandusele toetuvad riigieksamite koostajad ja loomulikult on õppekirjandus õpilaste põhiliseks töövahendiks. Seega võib öelda, et jõukohase õppetöö aluseks on jõukohane õpik.

Õppeteksti jõukohasus sõltub selles esitatava materjali hulgast (teadusterminid, tähistused jne), selle abstraktsusest ja esituse struktuurist, sealhulgas lausete keerukusest ja materjali selgitamise detailsusest. Pahatihti selgitatakse üht mõistet erinevates ainetes erinevalt. Õpikute autoritel puudub selge ülevaade sellest, mida on õpilased õppinud eelnevates klassides ja nii võidakse eeldada õpilastel teadmiste olemasolu, mida neil tegelikult ei ole.

Õppetekstide erinevate karakteristikute mõju õppetöö tulemuslikkusele on korduvalt tõestatud (Mikk 2000). See teadmine on avaldanud õppetekstide kirjutamisele küll mõju, kuid mitte piisavalt. Kui õpikute autorid ja toimetajad saavad oma töö käigus varakult konkreetset

informatsiooni selle kohta, mis võib muuta õppeteksti õpilastele liiga raskeks või liiga lihtsaks, siis leiavad autorid ja toimetajad enamasti teisi võimalusi materjali esitamiseks. Peale muu tähendab see suurt kokkuhoidu õpikute koostamisel ja kirjastamisel: kõige kallim töö on halvasti tehtud töö ümbertegemine. Õpikute kasutajate poolel on kokkuhoid aga rahas mittemõõdetav: õppimiseks mittetulemuslikult kasutatud aeg on raisatud elu…

2. Kuidas mõõta teksti keerukust?

Teksti keerukus ehk loetavus, nagu ka arusaadavus, on intuitiivselt mõistetavad, kuid raskesti defineeritavad mõisted. Sellest hoolimata on kahe sama asja kirjeldava teksti keerukust võimalik võrrelda. Selleks esitatakse katsealustele küsimusi asjade kohta, mida tekstid kirjeldavad. Seejärel antakse tekstid katsealustele lugeda ja hiljem esitatakse neile jälle küsimusi, millele vastamine eeldab loetud teksti mõistmist. Tekst oli ilmselt paremini mõistetav ehk loetav, kui seda lugenud õpilaste õigete vastuste protsent on rohkem paranenud.

Teksti keerukust ei ole otse võimalik mõõta, kuid on mitmeid formaalseid tunnuseid, mille muutumine on seotud teksti keerukuse muutumisega. Nt. keerulisemates tekstides on tavaliselt pikemad laused kui lihtsates; seal on tavaliselt ka rohkem harvaesinevaid sõnu; sõnad on pikemad ja abstraktsemad.

Tüüpiline keerukuse uuringu skeem on järgmine:

1. Võetakse mõnikümmend teksti valdkonnast, mis uurijaid huvitab (näiteks loodusteadused koolis),
2. Õpilased õpivad neid tekste (iseseisvalt) ja siis vastavad küsimustele,
3. Arvutatakse tekstide formaalsed tunnused (nt. lause keskmine pikkus, komade arv tekstis, sõna keskmine pikkus),
4. Arvutatakse korrelatsioonid õigete vastuste protsendi ja tekstide formaalsete tunnuste vahel.

Statistiliselt oluliste korrelatsioonidega tunnused on seotud teksti keerulisusega. Nende väärtuste järgi on võimalik prognoosida ka uute tekstide jõukohasust õpilastele ja nende tunnuste oskusliku muutmise teel on võimalik teatud määral muuta teksti jõukohasust.

Arvestades teksti keerulisuse e. loetavuse suurt tähtsus kirjalikus kommunikatsioonis üldse, on inglise jm keelte jaoks koostatud arvutiprogramme, mis võimaldavad teksti loetavust (Readability index) määrata; nad on lülitatud ka nt. Microsoft Wordi grammatikakontrollija koosseisu.

Kui tegemist on võõrkeelse tekstiga, siis tema keerukust ei saa muidugi samade parameetrite alusel hinnata kui emakeelse teksti (mida me eelnenud käsitluses vaikimisi eeldasime) keerukust. Kuid saab hinnata nt. seda, kui haruldasi ja keerulisi sõnu ning lausekonstruktsioone kasutatakse ja kuivõrd sõnavara eri klassides kordub.

3. PHARE projekt

Euroopa Liidu PHARE Eesti keele õppe programmi (www.meis.ee/phare/new/index.php) raames on ette nähtud kirjastada uued eesti keele õpikud muukeelse kooli 1-9. klassidele. Et õpikute kvaliteeti kontrollida ja ehk ka parandada, on vaja neid võimalikult varases tegemise etapis analüüsida. Meie ülesanne oligi õpikute tekstianalüüs nii nende keerukuse kui ka õpikute omavahelise kooskõla osas:

1. Kas ja mil määral kordub sõnavara klassist klassi?
2. Kas õppetükid muutuvad järk-järgult raskemaks?
3. Kas õpikutes kasutatav sõnavara on sobiv ja vajalik (sagedussõnastik jm aspektid)?

Võõrkeele õppimisel on oluline harjutamine ja arusaamine. Arusaamine tähendab seoste loomist sõna ja selle tähenduse vahel. Seose loomiseks on omakorda tarvis seda seost näidata ja korrata. Eesti keele sagedussõnastiku ettelugemisest ei piisa eesti keele omandamiseks. Õppimine on edukam, kui võetakse üks väike osa (näiteks salm luuletusest), õpitakse see ära ja siis minnakse uue osa juurde. Selle tõttu tuleks uusi sõnu sisse tuua järk-järgult. Teiseks on otstarbekas õppida neid sõnu, mida sageli vaja läheb, see tähendab sagedussõnastiku sagedasemaid sõnu.

4. PAEL (Programm Analüüsiks, Eriti Lihtsustamiseks)

Eesti keele õpikute analüüsiks kasutasime programmi PAEL. Tema ülesehituse põhimõte on „tarkvara-Lego“, s.t. et väikestest programmidest saab ehitada küllalt keerulisi süsteeme, neid programme omavahel kombineerides.

PAELa töökeskkond on vabavara cygwin, s.o. Unix Windowsis. Selle põhjuseks on see, et esiteks on UNIXis lai valik tekstide töötlemiseks sobivaid käske, teiseks see, et neid käske saab väga lihtsalt omavahel kombineerida, ning lõpuks see, et UNIXi keskkond on stabiilne: inimene, kes õppis UNIXit kasutama aastal 1980, saab oma oskusi kasutada ka aastal 2000 ja ilmselt edaspidigi; kes õppis aga DOSi või Windowsi kasutama, peab iga paari aasta tagant ümber õppima. On selge, et pidev ümberõppimine takistab süvenemist.

Eestikeelsete tekstide analüüsil tuleb arvestada, et sõnad esinevad tekstis mitte algvormi kujul (s.o. millena sõnad on näiteks sõnastike märksõnadeks), vaid enamasti mingis muus vormis. Seega analüüsiprogrammi oluliseks osaks peab olema programm, mis teisendab sõnavormid algvormideks. Seejuures tuleb arvestada ka konteksti: lauses „Mees peeti kinni“ on „peeti“ algvormiks „pidama“, mitte „peet“. Kui konteksti mitte arvestada, siis oleks üle 40% sõnadest eestikeelses tekstis mitte-ühese morfoloogilise tõlgendusega (Kaalep, Vaino 1998), s.t. sõnadel on kas mitu võimalikku algvormi, nagu eeltoodud „peeti“ puhul, mitu võimalikku käände- või pöördevormi (nt „ema“ võib olla nii ainsuse nimetava, omastava kui osastava vorm) või mitu võimalikku sõnaliiki (nt. „alla“ võib olla nii kaassõna, „kapi alla“, kui ka määrsõna, „alla käima“).

Sõnade algvormide leidmiseks tuleb teha teksti täielik morfoloogiline analüüs, mis koosneb kahest etapist: üksiksõnade morfoloogiline analüüs ning ühestamine. Üksiksõnade morfoloogiline analüüs on eesti keele puhul teksti täieliku morfoloogilise analüüsi tingimata vajalik osa (morfoloogiliselt lihtsama keele, nt inglise keele puhul, võib ta ka puududa). Ta annab igale sõnale hulga analüüsivariante. Seejärel toimub mitmest variandist ühe, antud konteksti sobiva valimine e. ühestamine.

PAELa komponendid e klotsid on:

1. Morfoloogiline analüsaator ja ühestaja estyhmm (Filosoft OÜ) (Kaalep, Vaino 2000), mis teisendab sõnavormid algvormideks. Estyhmm omakorda koosneb järgmistest osadest:

1.1. Sõnastikule tuginev morfoloogiline analüüs. Ligikaudu 98±1% eestikeelse sisendteksti sõnadest on analüüsitav sel moel, et kasutatakse sõnastikust järelevaatamist, mitmesuguste morfeemide loendeid ja nende kombineerimise eeskirju.

1.2. Morfoloogiline oletamine nende sõnade jaoks, mida sõnastiku abil analüüsida ei saa. Selliseid sõnu on kuni 3% eesti kirjakeelsest tekstist. Tegemist on loomulikule keelele olemuslikult omase asjaga. On teada, et u. 3% eestikeelsetes tekstides olevatest sõnadest esineb seal ainult üks kord, kusjuures nad moodustavad teksti sõnavarast ligikaudu 50% (täpsed arvud sõltuvad teksti suurusest, tüübist ja sellest, kuidas loendame sõnu). Samuti on teada, et sõnastiku abil mitteanalüüsitavate sõnade hulk mistahes keeles on ligikaudu sama suur kui üks kord esinevate sõnade hulk. Need seaduspärasused kehtivad nii mõnekümne tuhande sõnaliste raamatute kui ka sadade miljonite sõnaliste tekstikorpuste korral (Baayen 2001). Et neid 3% analüüsida, oletame sõna välise kuju (suurtähelisuse, silpide arvu, lõpuhäälikute) alusel tema vormi, sõnaliiki ja algvormi. Vigu ja mitmesust tekib seejuures paratamatult rohkem kui sõnastikupõhisel analüüsil.

1.3. Morfoloogiline ühestamine. Kasutame Markovi Varjatud Mudeli nimelist statistilist meetodit (Kaalep, Vaino 1998), et mitmest võimalikust morfoloogilisest analüüsist valida antud konteksti sobiv.

2. Eesti kirjakeele sagedussõnastik (Kaalep, Muischnek 2002), mille aluseks on 1 miljon sõna ajalehe- ja ilukirjanduse tekste vahekorras 50:50.

3. Nimisõnade abstraktsuste sõnastik. Sõnastikus on 11516 nimisõna, millele abstraktsused on käsitsi määratud. Eesti keele õpikute analüüsi jaoks täiendati sõnastikku nii, et ta kataks kogu neis õpikutes olevat nimisõnade hulka. Abstraktsuse poolest on sõnad jagatud 3 astmesse: aste 1 on kõige konkreetsemad sõnad, s.t. nimisõnad, mis tähistavad meeltega vahetult tajutavaid esemeid ja olendeid, nt. „maja“; aste 2 on nimisõnad, mis tähistavad meeltega vahetult tajutavaid nähtusi ja protsesse, nt. „tuul“; aste 3 on nimisõnad, mis tähistavad meeltega vahetult mittetajutavaid objekte, nt. „saladus“.

4. Shelli skriptid (programmid) – „sideaine ja välisviimistlus“. Nende abil tehakse teksti ettevalmistust enne morfoloogilist analüüsi, tulemuste järjestamist ja tabelitesse paigutamist jms sellist, mis muudab analüüsitud materjali inimesele ülevaatlikumaks.

PAELa ülesanded antud projekti raames olid:

1. Teha teksti (sagedus)sõnastik
2. Võrrelda eri tekstide sõnavara nii omavahel kui ka sõnastike või loenditega
3. Leida teatud laused (nt. pikad)
4. Leida teatud sõnad (nt. lühendid, pikad)

PAELa väljund on mõeldud inimesele tõlgendamiseks; PAEL ei ütle, kas tekst on liiga keeruline või lihtne.

Mis siis tegelikult toimub, kui me PAELa käivitame? Algoritm sõnavara uurimisel on järgmine:

1. Teha teksti morfoloogiline analüüs ja ühestamine.
2. Selle alusel teha teksti sõnaloend, vajadusel sealt teatud tunnusega sõnu, nt lühendeid, välja filtreerides.
3. Panna sõnaloend kui tabeli veerg mingite teiste loenditega (nt. abstraktsete sõnade loend, eelmises klassis kasutatud sõnade loend) kokku ühte suurde tabelisse. Teisendada tabelit nii, et järele jääks üks veerg sõnade jaoks ja muud veerud muu info jaoks, nt. et abstraktsuse veerus on kirjas sõna abstraktsus (1, 2, 3 või -); 5. klassi veerus sõna esinemiste arv 5. klassi õpikus jne.
4. Valida tabelist huvitavad veerud või read, mis esitatakse inimesele tõlgendamiseks (liiga suurest tabelist on raske vajalikku infot välja lugeda).

5. Analüüsi tulemused

Eesti keele õpikute analüüsi eesmärgiks oli aidata autoreid ja toimetajaid. Analüüsisime 1.–3. klassi õpikuid koos ja 4.–9. klassi õpikuid (käsikirju) nii üksikuna kui koos, samuti õppetükkide kaupa. Põhirõhk oli sõnavara võrdlemisel; võõrkeele õpiku puhul ei saa kasutada samasuguseid keerukuse näitajad nagu nt keemia õpiku hindamisel.

Analüüsi eesmärgiks oli leida vastused järgmistele küsimustele:

1. Kas ja mil määral kordub 1.-3. klassi sõnavara 4. klassis, 4. klassi oma 5. klassis jne?
2. Kas õppetükid muutuvad järk-järgult raskemaks?
3. Kas õpikutes kasutatav sõnavara on sobiv ja vajalik (eesti keele sagedussõnastikuga võrreldes jm aspektid)?

Tabel 1 esitab olulisemaid keerukuse koondnäitajaid klasside kaupa.

Tabel 1. Keerukuse koondnäitajad

	1-3	4	5	6	7	8	9	Kokku
Sõnade üldarv	44713	22025	17599	18472	13225	16009	23819	155862
Erinevate sõnade arv	3050	2526	2731	3114	2838	3072	4239	10632
Klassi optimum (igas klassis 800 uut sõna)	2400	3200	4000	4800	5600	6400	7200	7200
1 kord esinevaid sõnu	933	858	1288	1484	1425	1463	2082	4423
Õpiku sõnade %, mis pole keele 10 000 sagedasema sõna seas	37	28	25	31	28	28	32	51
Eelmistes klassides mitte esinenud sõnade % õpikus		53	57	58	61	62	62	
Nimisõnade keskmine abstraktsus	1,49	1,64	1,69	1,67	1,83	1,88	1,83	1,73

„Erinevate sõnade arv“ näitab, kui suurt sõnavara on kasutatud antud õpikus. Rida „Klassi optimum“ näitab erinevate sõnade arvu, mis õpikus peaks olema, kui lähtuda teatud teoreetilistest põhimõtetest: et sõna meelde jääks, tuleks teda korrata rohkem kui üks kord

ning et just 800 on see sõnade arv, mida keskmine õpilane on suuteline antud õppetundide arvu puhul aasta jooksul omandama. Sel juhul omandaks õpilane 9 aasta jooksul 7200 sõna. Kui valida sagedussõnastikust (Kaalep, Muischnek 2002) kõik sõnad, mille esinemissagedus on suurem kui 7, siis neid sõnu on 7400 ja nad katavad tekstist 89% (kui seaksime sageduse läveks 8, siis saaksime sõnastiku suuruseks 6900). Seega kui õpikutes olevad sõnad oleksid valitud ainult sagedussõnastiku sagedusi arvestades, siis pärast 9. klassi lõpetamist peaks vene kooli õpilane saama aru u 89% eesti aja- või ilukirjanduse tekstis olevatest sõnadest. Sõnad, mille sagedus ei ole suurem kui 7, oleksid siis põhikooli õpikute seisukohalt vaadates keeles harvaesinevad sõnad. (Seda mõistet kasutame hiljem sõnade iseloomustamisel.)

Tabel illustreerib m. h. vastuolulisi nõudmisi, mida õpikutele esitatakse. Ühelt poolt nõutakse, et õpikutes kasutataks autentseid tekste, mitte spetsiaalselt õpikute jaoks kirjutatuid. Autentseid tekste iseloomustab aga see, et umbes pooled tekstide sõnavaras kasutatud sõnad esinevad seal ainult üks kord (hapax legomena). Teiseks on õpikute materjal grupeeritud temaatilistesse tsüklitesse. See tähendab samuti, et sõnad eriti ei kordu. Teiselt poolt teame, et „kordamine on tarkuse ema“ ja võib kahelda, kas üks kord esinenud sõnadel on õppimise seisukohalt üldse mõtet – „üks ei ole ühtegi“.

Varem on (Mikk jt. 1991) analüüsinud aastatel 1981 ja 1982 Riina Reneli ja Ingrid Sotteri poolt avaldatud inglise keele õpikuid. Väljavõtte tulemustest esitab tabel 2.

Tabel 2. Inglise keele algõpetuse õpikud

	4. klass	5. klass	6. klass
Sõnade koguarv õpikus	7132	11355	8853
Erinevate sõnade arv	302	555	636
Erinevate uute sõnade arv	302	352	267
Sõnade arv, mis kordusid õpikus alla 6 korra	118	283	420

Nendes õpikutes oli kolme klassi kohta 420 sõna, mis kordusid õpikuis vähem kui 6 korda. Analüüsitavais eesti keele õpikuis on üle 6000 sõna, mis korduvad alla 3 korra. Nüüd on õppetunde tunduvalt rohkem, kuid siiski tundub olevat vastuolu kahe seisukoha vahel: ühelt poolt soov kasutada autentseid tekste ja teiselt poolt soovitus korrata uut sõna õpikus umbes 7 korda selle kindlaks omandamiseks.

Silma hakkab ka tekstide suur abstraktsus, samuti suur erinevus sagedussõnastikust - kattuvus on ligikaudu 50%.

Kokkuvõttes võib öelda, et analüüsitud käsikirjad on pigem ebaühtlased ja keerulised kui lihtsad.

Ülaltoodud üldisest järeldusest autoritele küll oma tekstide muutmisel eriti palju kasu ei ole. Seetõttu koostasime me detailsemad sõnavara tabelid üksikute õpikute ja õppetükkide kaupa, et autorid saaksid ise otsustada, kas ja mida muuta. Tabel 3 esitab väljavõtte 9. klassi sõnavara iseloomustavast tabelist.

Tabel 3. Väljavõte 9. kl. sõnavarast

Sõna	Uus	Pikk	Abstraktne	Harv	Õpikuis
ahhetama	+	-	-	+	1
ahistaja	+	-	-	+	1
ahistav	+	-	-	+	2
ahmima	+	-	-	-	1
ahv	-	-	-	+	13
ahvikari	+	-	-	+	1
aiandus	+	-	-	+	2
aiapidamine	+	+	-	+	1
aids	+	-	+	+	1
aim	+	-	+	-	3
aina	-	-	-	-	11

Plussmärk veerus „Uus“ näitab, et sõna pole varasemates klassides esinenud. Plussmärk veerus „Abstraktne“ näitab, et sõna abstraktsus on 3-pallilisel skaalal maksimaalne e. 3. Plussmärk veerus „Harv“ näitab, et sõna ei kuulu sagedussõnastiku 7400 sagedasema sõna hulka.

Viimane veerg näitab, mitu korda sõna kõigis õpikutes (1-9. klass) kokku kasutati.

Tabelis 3 toodud sõna „ahhetama“ esineb 9. klassis esmakordselt õpikuis. See pole pikk sõna, abstraktsust pole võimalik määrata, sõna esineb keeles harva ja õpikuis ka vaid üks kord. Kui autor soovib vähendada selle õppetüki keerukust, kus see sõna esineb, võib ta selle välja jätta.

Teine sõna „ahistaja“ on samasuguste karakteristikutega, kuid eeltoodud soovitustele lisaks võiks oletada, et see on keeles kasvava tähtsusega — sel juhul võiks sõna võimaluse korral hoopis rohkem korrata.

Sõna „ahmima“ on vaid ühest aspektist keerukas, kuid ta esineb õpikuis ka vaid ühe korra ja siinkohal oleks võimalus vähendada üks kord esinevate sõnade arvu õpikus. Või tuleks teda õpikus hoopis rohkem kasutada, sest ta on üldkeeles küllalt sage?

Kõige rohkem keerukuse näitajaid on selles tabelis sõnadel „aiapidamine“ ja „aids“. „Aiapidamine“ esineb õpikuis küll vaid ühel korral, kuid selle väljajätmise soovituse vastu räägib fakt, et see on liitsõna ja oma osade kaudu mõistetav. Sõna „aids“ on mitmeti keerukas, kuid järsku väärib see teema rohkem tähelepanu?

Sõna „aina“ pole ühegi näitaja järgi keerukas ja esineb õpikuis 11 korda. Nii võib see kindlasti jääda.

Peale sõnade, mida autorid ise kasutavad, võiksid huvi pakkuda ka sõnad, mida nad ei kasuta, aga mis üldkeeles on küllalt sagedased. Nii andsimegi neile loendid: sagedased sõnad, mida õpikutes polnud (135 sõna sagedussõnastiku sageduselt esimese 2000 hulgast ja 650 sõna teise 2000 sõna hulgast). Need sõnad kuulusid peamiselt ajakirjandusse, nt: „esimees, linnavalitsus, partei, kohaselt, toetus, juhatus, otsekui, ametnik, senine, nentima“. Kuid oli ka selliseid sõnu nagu „saatus, kaotus, kütus, äsja, juut, tõsiasi, tartlane“. Huvitav oli asjaolu, et õpikutes esines sõna „fakt“, kuid ei esinenud sõna „tõsiasi“.

Autorid said ka loendid pikavõitu lausetest, sest lause liigne pikkus tekitab tekstis asjatut keerulisust. Tabelis 4 on toodud sõnade arv lauses, mida lugesime pikaks.

Tabel 4. Pikaks loetud laused klassiti

Klass	4	5	6	7	8	9
Lause sõnades	10	11	12	13	14	15

Omaette küsimus oli: kas õpikute tekstides kujutatakse mehi ja naisi võrdse sagedusega või eelistatakse üht sugupoolt teisele? Seejuures tekib kohe küsimus: kuidas mehi-naisi tähistavaid sõnu automaatselt leida? Tegelikult on tegemist just sellise ülesandega, mida PAELa abil on lihtne lahendada. Nimelt peame me algul koostama loendi mehi ja naisi tähistavatest sõnadest (s.h. pärisnimedest). Kui see on tehtud, siis on ülesanne analoogiline teksti sõnadele abstraktsuste määramisega: teksti alusel tehtud tabelisse, kus esimese veerus on tekstis esinenud sõnad ja ülejäänud veergudes nende sõnade mitmesugused tunnused, lisandub üks veerg. Seal on kirjas, kas sõna tähistab meessoost või naissoost olendit.

Mehi-naisi tähistavate pärisnimede loendi tegime käsitsi. Üldnimede loendi tegemisel pidime aga arvestama, et kuna eesti keeles võib tuletuse ja liitsõnamoodustuse abil uusi sõnu juurde

moodustada, siis käsitsi me ammendavat loendit luua ei suuda. Proovisime seda teha poolkäsitsi järgmise algoritmi järgi:

1. Võta mingi mehi-naisi tähistavate lihtsõnade loend. Meil olid kasutada mõned loendid, mis olid tehtud käsitsi algklassi lugemike alusel.
2. Analüüsi (automaatselt) tekstikorpuses olevaid sõnu.
2.1. Kui liitsõna viimane komponent tähistab meest või naist, siis tähistab ka sõna ise sama sooga olendit.
2.2. Kui liitsõna esimene komponent on „nais-“ või „mees-“, siis see määrab sõna poolt tähistatava olendi soo (nt. „naisdiktor“, „meesjuuksur“).
2.3. Kui sõna lõpus on naissugu tähistav liide („anna-“, „tar-“), siis see määrab sõna poolt tähistatava olendi soo.
3. Kontrolli saadud loend käsitsi üle.

Selgus, et nii saadud loend on siiski natuke problemaatiline. Esiteks on terve hulk liitsõnu, mille poolt tähistatavatel ei ole soorollidega midagi pistmist, nt. „külmapoiss“, „lumelell“, „lumememm“, „autoisa“. Teiseks on hulk lihtsõnu, mida kasutatakse valdavalt teatud sooliste inimeste tähistamiseks, nt „traktorist“ on mees, „blondiin“ on naine, aga kas „kiilakas“ on kindlasti mees ja „feminist“ kindlasti naine? Oletasime, et meie tekstide puhul – põhikooli õpikud – pole selliste probleemsete sõnade hulk siiski suur ja me ei pea oma loendit nende osas täiustama. Kolmandaks, sõnad võivad liitsõnaosadena oma soolisust muuta, nt. „meeskond“ näib tähistavat meeste kollektiivi vastandina „naiskonnale“, aga „võttemeeskond“ jälle mitte.

Ülaltoodud mehi-naisi tähistavate sõnade määramisega kaasnevast ebatäpsusest hoolimata tegime nende kohta statistikat. Tulemused on tabelis 5.

Tabel 5. Mehi-naisi tähistavate erinevate sõnade arv õpikutes kokku

	M	N
Pärisnimed	177	158
Üldnimed	85	62

Nagu näeme, on mehi tähistavaid sõnu kasutusel rohkem kui naisi tähistavaid. Lisaks tuleb märkida, et:

1. Ka mehi tähistavate sõnade kasutuskordi on õpikutes rohkem
2. 4-6. kl räägitakse inimestest rohkem kui 7-9. kl
3. 4-6. kl räägitakse rohkem meestest, 7-9. kl naistest

6. Kokkuvõte

1. Õpikud väärivad ja vajavad analüüsimist
2. Selleks on loodud tarkvara-pakett PAEL, mis on ühelt poolt paindlik ja kohandatav vastavalt ülesandele, teiselt poolt arvestab eesti keele spetsiifikat
3. PAELa on praktiliselt kasutatud eesti keele õpikute käsikirjade analüüsil (PHARE)
4. Käsikirjad olid pigem ebaühtlased ja keerulised kui lihtsad
5. Analüüsi tulemused on edastatud autoritele

PAELa head omadused on:

1. Kohandatav paljude ülesannete jaoks
2. Arvestab eesti keele spetsiifikat
3. Oskaja käes tõhus töövahend

Tema halvaks omaduseks on, et ta nõuab spetsiifilisi oskusi – Unixi kasutamist ja oskust shelli skriptidest aru saada, et neid vajadusel modifitseerida. Perspektiivis on kavas kohandada PAEL tavakasutaja-sõbralikumaks.

7. Kirjandus

Baayen, R. H. 2001 Word Frequency Distributions. (Text, Speech and Language Technology, vol. 18, series editors Nancy Ide and Jean Véronis), Dordrecht, Kluwer Academic Publishers.

Juurak, R. 2001 Kommentaar OECD raportile. - Haridus 1, lk 3.

Kaalep, H-J., Muischnek, K. 2002. Eesti kirjakeele sagedussõnastik. Tartu: TÜ kirjastus

Kaalep, H-J., Vaino, T. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. Kogumikus " Arvutuslingvistikalt inimesele" Tartu, lk 87 – 99

Kaalep, H-J., Vaino, T. 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. - Keel ja Kirjandus 1, lk 30-38.

Mikk, J. 2000. Textbook: Research and Writing. Frankfurt am Main, 2000, p. 190-195

Mikk J., Mikk E., Tirmaste J. 1991. Computerised readability analysis of textbook of English. - Problems of textbook effectivity. Toim J. Mikk Tartu: University of Tartu, lk 112 – 121.