EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 4, 53–65

KÕNETEHNOLOOGIA VAJAB ŽANRILIST LÄHENEMIST

Krista Kerge, Hille Pajupuu, Kairi Tamuri, Heidi Meier

Ülevaade. Oleme varasemas uurimuses näidanud, et ettelugemisel seostuvad pausid ja hingamine kui etteloetud teksti loomulikkuse tunnused tugevasti funktsionaalstiiliga (ajakirjandustekst, ilukirjandustekst). Tekstiuurimuse andmestik aitab muuta loomulikuks sünteeskõne, kuid ka tuvastada tekstiliiki ehk žanrit, eristada funktsionaalstiile ja autorite sõnastuslaadi ning muuta otstarbekamaks sõnastiku ja grammatika kasutuse teksti automaatses analüüsis. Artiklis osutatakse hingamise ja pausidega seostuvale ja teistele eesti teksti formaliseeritavate parameetrite uurimustele, mida keeletehnoloogia arenduses seni arvestatud ei ole. Varasema kontrollimiseks võrreldakse ilukirjandusteksti kahe žanri ettelugemise pause ja hingamist ajakirjandusuudise seniste andmetega. Tulemused kinnitavad, et pauside ja sissehingamise kestus ning nende seotus lause süntaktilise liigendusega erineb mitte ainult funktsionaalstiilide, vaid ka žanrite kaupa. Süntaktilist liigendust kannavad seejuures peamiselt kirjavahemärgid. Ilukirjanduse puhul mängib pauside tekkes kaasa teksti lugejapoolne interpretatsioon. Tekst–kõne-sünteesis, mis on mõeldud tekstide ettelugemiseks, tuleks pauside ja hingamise genereerimisel lähtuda pikema ajakirjandusliku uudise kui muudeltki parameetritelt neutraalse tekstiliigi andmetest.*

Võtmesõnad: pausid, hingamine, funktsionaalstiil, žanr, uudised, ilukirjanduskeel, kõnesüntees, eesti keel

* See artikkel on valminud tänu Eesti Teadusfondi grandile nr 6742 ja riiklikule programmile "Eesti keele keeletehnoloogiline tugi".

Tekstiliikide parameetrid ja kõnetehnoloogia

Juba sünteeskõne intonatsiooni ja pauside modelleerimisega seoses (vt Mihkla jt 2005) tekkis huvi, kas ja kuidas sõltuvad etteloetud teksti paralingvistilised muutujad, näiteks hingamine ja pauseerimine keelekasutuse valdkonnast, millega seostuvad funktsionaalstiilid ehk allkeeled, ja žanrist, mis on kindla registri tekstiliigiline väljund. Et selline sõltuvus on olemas, sai kinnitust etteloetud teksti pauside ja hingamise detailsel analüüsil (vt Pajupuu, Kerge 2006).

Keeleruumi formaalselt mõõdetava leksikogrammatilise ja paralingvistilise varieeruvuse ideoloogiat ning seniste tekstiuurimuste andmestikku ei ole eesti kõnetehnoloogias kuigivõrd rakendatud, see võiks aga kasulikuks osutuda (nt Barnbrook 2007). Esimene kohalik katse nimetatud suunas oligi eelviidatud hingamise ja pauside määramine ajakirjandus- ja ilukirjandustekstis, kus vaatluse all olid pikemad raadiouudised ning etteloetud kriminaalromaan (Mihkla jt 2005, Pajupuu, Kerge 2006). Seda uurimust on ilukirjanduse poolel laiendatud armastusromaaniga (Tamuri 2007) ning võetud tulemused hiljem kokku (Kerge jt, ilmumas). Viimast uurimust tutvustame allpool üksikasjalikumalt.

Ettevalmistatud suulisele žanrile, nimelt loengule keskendub eesti suulise keele uurimist alustav kirjutis (Hennoste 1994); ettevalmistatud spontaanse kõne ajaorganisatsiooniline analüüs on toonud loengužanri taas päevakorrale (Meister, Lippus, ilmumas). Seda kõike on aga ilmselgelt vähe, kui võtta arvesse andmeid, mida kirjalike või kirjalikult ette valmistatud žanrite kohta on teada – just sellistele žanritele võiks kõnetehnoloogia arendus toetuda, seades sihiks keele, sealhulgas kirjakeele jätkusuutlikkuse, mille võti on Mart Rannuti (2003) järgi kõnetehnoloogia, k.a asjalik suuline arvutisuhtlus.

Tekstiuurimise andmestikku peaks peale kõnesünteesi vajama ka teksti automaatne analüüs jm keeletehnoloogilise rakenduse valdkonnad. Eesti teksti uurimusi, mis võiks seejuures pakkuda vähemalt tuge või ideid, on enamgi (vt ülevaadet allpool). Žanri tuvastamine aitaks kohati piirata süntaktilisel automaatanalüüsil või tõlkimisel kasutatavat leksikoni ja selle üksuste tõlgendusvõimalusi, seades ühtlasi tõenäosusjärjekorda grammatiliste vormide-mallide kasutuse ja tõlgenduse kindla allkeele ja sellele omase leksika kontekstis. Žanriomaseid lingvistilisi tunnuseid ei tule tekstiliigi tuvastuse juures arvesse võtta süsteemselt, piisaks kõige tugevamini korreleeruvate tunnuste koosesinemuse modelleerimisest. Tekst–kõne-sünteesis võib omakorda toetuda keskmiste näitajatega ehk n-ö neutraalselt mõjuvatele žanritele (mis on odavam lahendus ja ilmselt vastuvõetav näiteks teksti sisu kiirest hõlmamisest huvitatud nägemispuudega inimestele). Süntesaatori kasutusala laienedes on võimalik rakendada žanri kiirtuvastust teksti põhjal ning programmeerida mitu lugemis- või asjaliku kõne mudelit, mis võtavad arvesse etteloetava või loodava teksti liiki seal, kus spontaansele kõnele toetumine mõjuks argikeelsena. See aga on ideaal n-ö pikemas plaanis, kui peaks tähtsustuma suuline suhtlus arvuti ja tema kasutaja vahel. Rõhutatagu siinkohal, et spontaanse kõne analüüsiga ei ürita siinkirjutajad tegelda ega toetu niisiis ka sellekohastele uurimustele, mida Eestis on juba rohkesti.[1]

[1] Vt nt: Publikatsioonid. Tartu Ülikooli suulise kõne uurimisrühm. http://www.cl.ut.ee/yllitised/ (3.01.2008).

Eesti tekstiliikide uuritud formaliseeritavad parameetrid

Nüüdiseesti allkeelte – nende nimi on kõnetehnoloogia ja korpuslingvistika ingliskeelses kirjanduses sageli *text-type* või *functional style* – süntaks muutub seniste uurimuste järgi ajas (vt Kerge 2002b: 44 jm, uustrükk Kerge 2003a [4]) ning erineb tugevasti ka sünkroonplaanis (vt Kerge 2002a). Sajandivahetuse allkeelte parameetrite keskmisi esindab tekstilause pikkuse ja sisemise keerukuse, samuti teksti abstraktsust tõstvate *mine*-tarindite osatähtsuse poolest ajakirjandustekst. Näiteks kõigub eesti tekstilause pikkus u 9 ja 25 sõne vahel, ajakirjandusteksti lause on aga u 14 sõnet pikk. Kirjavahemärke ja sidendeid, mis viitavad lause tinglikule liigendatusastmele, on allkeeliti u 1–4, ajakirjanduses u 1,5. Regulaarset *mine*-nominalisatsiooni esineb teksti sõnede hulgas u 1–3,5%, ajakirjanduskeeles 1,8% (vt Kerge 2003a: 40–45).[2] Analoogilisi uurimusi on erakondade valimisplatvormide kohta (Rääk 2002), seaduste kohta (Rätsep 2003) ning autoristiili kohta žanriti (Kerge 2003b, Põld 2007).

Lisaks on hulk muid lingvistilisi tunnuseid, mis eristavad tekstiliike ning ühtlasi iseloomustavad teksti laadi[3] (s.o jutustavat, kirjeldavat, seletavat ja argumenteerivat teksti). Heidi Meieri (2003a) uurimuse järgi (vt ka Meier 2002, 2003b) on igale tekstiliigile omased tugevalt korreleeruvate tunnuste kimbud. Kuna tunnused on mehhaaniliselt tuvastatavad, teeb see võimalikuks tekstiliigi automaatse määramise.

H. Meier on allkeelte kaupa uurinud näiteks isikuliste asesõnade, lühendite, *mine*-tuletiste, 2–4- ja enam kui 14-täheliste sõnade esinemissagedust; sõnaalgulise *g*, *b*, *d* ning võõrtähtede järgi hõlpsasti tuvastatavate võõrsõnade sagedust; rinnastavate ja alistavate sidesõnade esinemust ja sagedust; tekstiliikide sagedasimaid sõnavorme jpm, andes üsna hea ülevaate eesti tekstiruumi allkeelte parameetritest. Kõik parameetrid on ülevaatlikult esitatud ka tabelitena. Žanritest on uuritud esseed. (Täpsemalt Meier 2003a.)

Kogu see andmestik võiks kasulikuks osutuda ka eesti kõnetehnoloogia rakendustes või pakkuda selles valdkonnas just žanrituvastuse ideid. Sama eesmärki täidab nt Ülle Viksi ja Indrek Heina (2001) seadustekstianalüüs. Tuvastuskriteeriumide edasisel valikul pakuvad tuge ja eeskuju ka Douglas Biberi jt korpuseuuringud, mille põhjal tekstid oma retoorilise tüübi järgi moodustavad eelmainitud tekstilaadiga sarnaseid dimensioone (vt nt Biber 1995, Biber, Conrad 2001 vm).

Eestis on žanriti uuritud ka sõnaliigisuhteid. Seaduse andmeid kajastavad Ü. Viks ja I. Hein (2001). Teksti loomulikkusuuringute raames on sõnaliigisuhteid uurinud Krista Kerge, Hille Pajupuu ja Rene Altrov (2007), kes vaatlevad Frances Heyligheni ja Jean-Marc Dewaele (2002) meetodil, kui suur osatähtsus on nimisõnadel ning nimisõnaga seostuvana omadus- ja kaassõnadel *versus* muudel sõnaliikidel. Sõnaliigisuhete järgi on arvutatud tekstiliigi formaalsusindeksid, mille järgi iga tekst paigutub formaalsuse–kontekstuaalsuse kontiinuumis: nimisõnade ja nendega seotud sõnaliikide suur osatähtsus viitab teksti suuremale formaalsusele ja ühetimõistetavusele; koos muude sõnaliikide osatähtsusega kasvab kontekstuaalsus ja mitmetimõistetavus (vt Heylighen, Dewaele 2002). Uurimusega on eesti žanri-

[2] *mine*-tuletiste osatähtsuselt on õigustekst esindatud äärmusliku näitajaga 6,8%, mida sama uurimuse järgi u 30% ulatuses võiks vältida kui asjatut nominaalstiili.

[3] Ingliskeelses kirjanduses viidatakse teksti laadile terminiga *text-type*, saksakeelses *Textsort*, kuid termin *teksti laad* näib eesti keeles kõige paremini viitavat žanriülestele kvalitatiivsetele erijoontele. *Tekstitüüp* kui üldsõna on olnud vabamas kasutuses juhul, kui ei ole põhjust kasutada täpse sisuga termineid (nt dialoog on kord suuline, kord kirjalik, kord intervjuužanr, kord kauplusedialoog, kord vaba vestlus). Samamoodi kasutatakse sõna *tekstitüüp* siinses artiklis.

test siiani hõlmatud kirjalik essee ning asjalik suuline monoloog ja dialoog (Kerge jt 2007), õigusaktid (Kerge 2006); hulk kirjavahetust: elektroonilised ametikirjad töötajate vahel ja asutusest välja (Rais 2007), ametikirjad asutuse sees, asutuste vahel ja asutuselt kodanikele, samuti poiste ja tüdrukute meilid (Kerge 2006); lisaks suuline hoidedialoog (Kerge 2006) ja parlamendikõned (Pajupuu 2007). Eksamiesseede ning tigukirja- ja meilivahetuse taustal on nüüdseks uuritud ka isikustiili, täpsemalt Jaan Kaplinski esseistikat ja kirjavahetust (Põld 2007). Žanriti erineb selgelt nii teksti tihedus (nimisõnade osatähtsus) kui ka kontekstuaalsus. Sellekohasedki andmed võiksid leida laiemat rakendust.

Loomulik kõneliigendus ja žanriuuringute tähtsustumine

Kõnesünteesi kontekstis on Eestis viimastel aastatel tõsiselt tegeldud lauseintonatsiooni loomulikustamisega, uurides selle reeglistamisvõimalusi (Mihkla jt 2003), ning jõutud statistiliste hinnangute eelistamiseni (Mihkla jt 2004). Ka pause ja pausieelseid pikendusi kui kõneliigenduselemente on üritatud mitmel meetodil modelleerida (Mihkla 2005, Mihkla, Kuusik 2005).

Loomulikus kõnes on samas iseenesestmõistetav ja enamasti ka kuulda hingamine. Täiesti ilmselt ei saa hingamispause näiteks sünteeskõnesse lisada juhuslikult, arvestada tuleb, et hingamispausid on seotud füsioloogiliste teguritega: ühest sissehingamisest teiseni peab kuluma niipalju aega, kuni sissehingatud õhuga rääkida saab. Seega ei tohiks sünteeskõne sissehingamissagedus erineda oluliselt inimese omast: kõne kiirust arvestades tuleks määrata hingamispausidele sobiv vahemaa, leida neile õige koht ning õige kestus. Taustal peetakse silmas, et tegemist on tekste ette lugeva ehk *Text-to-Speech* tüüpi (TTS) süntesaatoriga.

Mõte süntesaator hingama panna tuli kümmekond aastat tagasi Douglas H. Whalenilt (1994), kes leidis, et hingamise lisamine teeks inimestele teksti jälgimise kergemaks: ilma hingamiseta sünteeskõne tüütab kuulajat, mõjub üksluiselt ning ebaloomulikult. Ka Nick Campbell on kõnesüntesaatorite puudusena täheldanud võimetust anda edasi selliseid kõnes esinevaid helisid nagu naer või hingamine. N. Campbelli arvates tõstaks nende lisamine sünteeskõne vastuvõetavust oluliselt (Campbell 1998). Ometigi suur osa tänapäeva süntesaatoreid ei naera ega hinga.

Hingamist kõnes on uurinud François Grosjean ja Maryann Collins (1979) ja toonud välja kolm olulist aspekti:

1) kõneleja ühendab oma hingamismustri lause planeerimisega, st kohandab füsioloogilised vajadused lingvistiliste vajadustega, et tagada kõne soravus;
2) kõneleja kontrollib oma hingamist, st ta jaotab kõne regulaarsetesse hingamis-rühmadesse;
3) kõneleja hingamine on sõltuv süntaksist, st kõneleja hingab ainult siis, kui lausungi moodustajastruktuur seda lubab.

F. Grosjeani ja M. Collinsi (1979) järgi pole pausid määratud otseselt vajadusega hingata. Küll aga sõltub hingamine pausidest. Sageli teevad kõnelejad pikemaid pause süntaktiliselt olulistes kohtades, nt fraasipiiril, kus esinevad tavaliselt just

hingamispausid. Kuigi tegemist on suhteliselt ammuse uurimusega, pole rakenduslingvistikas (sh kõnesünteesis) seda teadmist ära kasutatud.

Kõnesünteesiga seoses on pauside olemust uurinud Chiu-Yu Tseng (2002) ja leidnud, et kõnes organiseeruvad pausid hierarhiliselt: kõige pikemad pausid markeerivad prosoodilise rühma piire.[4] Prosoodiline rühm on võrdne vähemalt ühe hingamisrühmaga, kuid võib neid sisaldada rohkemgi. Hingamisrühma sees on omakorda lühemaid hingamiseta pause (Tseng, Chou 1999). Neist kahest uurimusest saab küll andmeid kirjavahemärkide kohale langenud pauside kestustest, ent ei saa teada, kuhu lugedes pause tegelikult tehakse. C.-Y. Tseng (2002) küll ütleb, et pausid võivad, kuid ei pruugi kokku langeda kirjavahemärkidega, ent ei käsitle seda teemat sügavamalt: me ei saa teada, kui palju pause langeb kokku kirjavahemärkidega, või millised neist on hingamispausid, millised mitte. Hingamispause ei uuri C.-Y. Tseng üldse lähemalt, näiteks ei saa me teada, kui suure osa hingamispausist moodustab sissehingamine, kui suure osa vaikus.

Pidades silmas mitut eesmärki, nt kõne loomulikkuse mõõtmist ja hindamist ning kõnesünteesi loomulikkuse tõstmist, püüdsime nendele küsimustele vastused leida, defineerides hingamis- ja muud pausid ning arvestades uurimise juures žanrilist eripära. Uurimuste rida puudutab kahe allkeele, ajakirjandus- ja ilukirjanduskeele etteloetud teksti.

Pauside pikkust, sagedust ja asukohta eri tüüpi tekstides (spontaanne dialoog, professionaalne ja mitteprofessionaalne etteloetud uudis) on rakenduslikul eesmärgil uurinud ka Sofia Gustafson-Čapková ja Beáta Megyesi (2001). Uurimusest järeldub, et pauside pikkus ja sagedus sõltub tekstitüübist: viimastest moodustub kontiinuum, mille ühte suunda jääb dialoog (pausid pikemad, sagedamini esinevad) ja teise professionaalne uudiste ettelugemine (pausid lühemad, harvem esinevad).

Valdavalt tehakse pause olulistes kohtades, nt süntaktilistel piiridel ja semantiliselt tähtsate sõnade juures (viimased olid sõnumile aluseks). Monoloogides varieerivad kõnelejad tavaliselt pauside pikkust ja asukohta vastavalt informatsiooni struktuurile (Gustafson-Čapková, Megyesi 2001).

Gunnar Fant kolleegidega (2002, 2003) on uurinud pauside kestust ilukirjanduse ja uudiste lugemisel. Tulemused näitavad, et kõik uudiste pausid on ilukirjanduse omadest lühemad.

Kuigi kumbki vahetult nimetatud uurimustest ei tegele hingamispausidega, annavad need uurimused selge vihje vajadusele läheneda kõne uurimisele žanrist lähtuvalt.

Seni avaldatud selleteemalised eesti tulemused (Pajupuu, Kerge 2006) osutavad, et nii pausid kui ka hingamine on seotud teksti kompositsiooniga, s.o teksti osade ja neis sisalduvate lõikude ja lausete piiriga, ning tekstilause süntaktilise liigendusega, mida eesti keeles enamasti markeerivad kirjavahemärgid. Alustavalt võeti vaatluse alla kaks etteloetud teksti liiki, pikemad ajakirjanduslikud uudised ning võrdluseks ilukirjandusest kriminaalromaan. Tulemused suunasid väitma, et sünteeskõne pause ja hingamist tuleks modelleerida ajakirjandusliku uudise andmetele toetudes (Pajupuu, Kerge samas). Et just ilukirjandusteksti andmed osutusid ootamatuks, suurendati uurimise järgmises etapis ilukirjandusliku materjali mahtu ja lisati armastusromaan (Tamuri 2007). Selle uurimuse tulemusi järgnevas ka tutvustatakse.

[4] Prosoodiline rühm on kõneüksus, mis koosneb vähemalt ühest hingamisrühmast (sissehingamisest sissehingamiseni). Kirjalikus tekstis vastab sellele lõik (graafiliselt iseloomustab seda taane). Sama fenomen ilmneb ka kirjalikku teksti lugedes.

Pausid ja hingamine romaani allžanrites

Allpool on täpsustatult uuritud pauside paiknemist ja kestust etteloetud ilukirjandustekstis, täpsemini kahes romaani allžanris (armastusromaani ja kriminaalromaani katkendid, vastavalt 19 lauset, 413 sõnet ja 39 lauset, 647 sõnet). Ilukirjandustekstid salvestati vaikses ruumis Edirol R-09-ga loetuna professionaalsete näitlejate poolt (1 mees ja 1 naine, mõlemad lugesid samu tekste). Ajakirjandusliku uudise andmed pärinevad varasemast (pikemate uudiste terviklikud katkendid, 45 lauset, 666 sõnet), tekstid on Eesti Raadio professionaalsete uudistelugejate stuudiosalvestised, reaalselt loetud uudised (1 naine ja 1 mees, mehe ja naise tekstid olid erinevad). Ideaalis võinuks materjal hõlmata rohkem žanreid ja rohkem ettelugejaid, ent pidades silmas oma eemärki – mõõta kõne loomulikkust ja pakkuda lähtematerjali sünteeskõne pauside genereerimiseks – pidasime piisavaks tüüpilistele žanritele keskendumist. Lähtusime seisukohast, et iga emakeelekõneleja loetud teksti võib pidada loomulikuks, selgitamaks välja pauseerimise üldisi tendentse.

Kõnes mõõdetud pausid rühmitasime kaheks: 1) hingamispausid; 2) hingamiseta pausid. Hingamispaus koosneb sissehingamisest, millele eelneb ja/või järgneb vaikus. Hingamiseta pausiks lugesime vähemalt 30 ms kestva vaikuse. Praat 4.5 (Boersma, Weenink 2006) kasutades mõõtsime mõlemas tekstiliigis pausi kestuse, hingamispausis lisaks ka sissehingamise kestuse. Hingamispauside keskmised kestused tekstiliigiti on esitatud tabelis 1. Otsisime ka pauside ja süntaksi seost, arvestades seda, et süntesaator tunneb ära ennekõike graafiliselt markeeritud lauseliigenduse (kirjavahemärgid, lõigupiir, kindel kirjavahemärgita sidend, nimi jne) ja et praktikas on näiteks kirjavahemärgiõpetust omakorda seotud intonatsiooni ja pauseerimisega.

Tabel 1. Hingamispausi ja sissehingamise keskmine kestus tekstilugejatel tekstiliigiti[5]

Meeshääl			Naishääl		
Paus t/sd (ms)[6]	**Sisse-hingamine** t/sd (ms)	**Sissehingamise osatähtsus pausi kestuses**	**Paus** t/sd (ms)	**Sisse-hingamine** t/sd (ms)	**Sissehingamise osatähtsus pausi kestuses**
AJAKIRJANDUSTEKST: UUDIS					
617/262	276/112	45%	597/256	269/99	45%
ILUKIRJANDUSTEKST: ARMASTUSROMAAN					
590/234	380/145	64%	606/240	374/174	62%
ILUKIRJANDUSTEKST: KRIMINAALROMAAN					
675/345	449/259	66%	780/482	362/222	46%

Ajakirjandustekstis ei erine hingamispauside keskmised kestused dimensioonil meeshääl–naishääl. Hingamise osatähtsus hingamispausis on 45%. Sellest sarnasusest johtuvalt käsitleme edasises ajakirjanduspauside analüüsis mees- ja naishäält koos.

Ka etteloetud armastusromaanis ei erine hingamispauside keskmised kestused dimensioonil meeshääl–naishääl: hingamispauside kestus on sarnane uudiste omadele, oluliselt erineb aga hingamise osatähtsus pausi kogukestusest, mis on üle 60%. Kui uudistes püütakse sisse hingata võimalikult kiiresti ja isegi kuuldamatult, siis ilukirjandust lugedes sellist vajadust pole, vastupidi, sissehingamine on sageli

[5] Välja on jäetud prosoodilise rühma ehk lõigupiiri pausid kui eriliselt pikad. Nende keskmine kestus on antud tabelis 3.

[6] *t* on pausi kogukestus või sissehingamise kestus, *sd* on standardhälve.

kuuldav ja pikk. Seega võime oletada, et hingamine on ilukirjandust lugedes meeleolu loomise teenistuses.

Kriminaalromaani hingamispausid on armastusromaani pausidest märkimisväärselt pikemad, suurem on ka standardhälve. Võimalik, et erineva pikkusega hingamispausidega lisatakse loetavasse põnevust. Selles žanris tuleb esile ka mees- ja naislugeja erinevus: mees hingab sisse naisest oluliselt pikemalt (sissehingamise osatähtsus vastavalt 66% ja 46%). Et armastusromaani lugedes sellist erinevust välja ei tulnud, siis ilmselt pole põhjus lugejate füsioloogilises erinevuses (nt naise väiksemas kopsumahus), vaid pigem näitlejate erinevates põnevuse tekitamise võtetes: meesnäitleja hingamispausi täidab suuresti sissehingamine, naisnäitleja oma aga vaikus. Ilukirjandužanriti esile tulnud lugejaerinevuste tõttu ja põhjusel, et nii mees kui ka naine loevad samu tekste, käsitleme mees ja naishäält edaspidi lahus.

Pauside ja süntaksi seose analüüsi tulemused on esitatud tabelis 2.

Tabel 2. Pauside vastavus kirjavahemärkidele

Kirjavahemärk	**AJAKIRJANDUSTEKST: UUDIS** (meeshääl + naishääl)		
	Kirjavahemärkide arv tekstis	Pauside arv kirjavahemärgi kohal / neist hingamispause	
punkt lõikude vahel (kirjapildis taane)	14	14/14	
punkt lausete vahel	31	31/31	
koma	37	27/13	
mõttekriips	8	8/6	
koolon	6	6/5	
semikoolon	1	1/1	
paus kirjavahemärgita (ekstrapaus)		19/5	
Kirjavahemärk	**ILUKIRJANDUSTEKST: ARMASTUSROMAAN**		
	Kirjavahemärkide arv tekstis	Pauside arv kirjavahemärgi kohal / neist hingamispause	
		meeshääl	naishääl
punkt lõikude vahel (kirjapildis taane)	12	12/6	12/11
punkt lausete vahel	7	7/4	7/3
koma	45	16/8	20/12
mõttekriips	3	3/2	3/3
koolon	2	2/1	2/2
semikoolon	6	5/4	6/4
paus kirjavahemärgita (ekstrapaus)		49/22	9/0
Kirjavahemärk	**ILUKIRJANDUSTEKST: KRIMINAALROMAAN**		
	Kirjavahemärkide arv tekstis	Pauside arv kirjavahemärgi kohal / neist hingamispause	
		meeshääl	naishääl
punkt lõikude vahel (kirjapildis taane)	5	5/0	5/2
punkt lausete vahel	31	31/17	31/21
koma	58	12/7	31/11
mõttekriips	9	7/5	9/4
koolon	1	0	1/1
semikoolon	0	0	0
paus kirjavahemärgita (ekstrapaus)		41/24	23/7

Kirjavahemärkide seos pausidega on ilmne (vt tabel 2), üldjuhul tehakse nii ajakirjandusteksti kui ka ilukirjandusteksti lugedes kirjavahemärgi kohale paus. Kirjavahemärkidest tehakse kõige harvem paus koma kohal. Tekstiliikide erinevus on siiski märkimisväärne: ajakirjandustekstis tehakse paus 90% kõigist kirjavahemärkidest, seejuures 80% neist pausidest kasutatakse sissehingamiseks. Punkti ja lõigupiiri kohal tehakse ajakirjandustekstis alati paus ja hingatakse alati ka sisse, kuid ilukirjandustekstis ei ole ükski kirjavahemärk "kohustuslik" pausi ja sissehingamise koht. Isegi lõigupiir (mis ajaliselt on väga pikk) võidakse ületada sissehingamata. Ilukirjandustekstis langeb paus kirjavahemärgiga kokku umbes pooltel juhtudel, erinevused ilmnevad nii žanriti kui ka ettelugejati.

Armastusromaani puhul teeb meesnäitleja pausi kõigist kirjavahemärkidest 60% kohal ning kasutab neist 55,5% sissehingamiseks. Naisnäitleja teeb pausi kõigist kirjavahemärkidest 67% kohal ja kasutab neist 70% sissehingamiseks. Kriminaalromaani puhul teeb meesnäitleja pausi kõigist kirjavahemärkidest 53% kohal ning kasutab neist 53% sissehingamiseks. Naisnäitleja teeb pausi kõigist kirjavahemärkidest 74% kohal ja kasutab neist 51% sissehingamiseks.

Ilukirjandust lugedes tehakse erinevalt ajakirjandustekstist palju nn ekstrapause (pausid mujal kui kirjavahemärkide kohal). Nende hulk ei sõltu niivõrd žanrist, kuivõrd teksti ettelugejast. Uudiste lugemisel moodustavad ekstrapausid 18% kõigist pausidest, meesnäitlejal armastusromaani lugedes 50,5% ja kriminaalromaani lugedes 40% kõigist pausidest, naisnäitleja teeb ekstrapause vähem: armastusromaanis 15% ja kriminaalromaanis 22% kõigist pausidest. Ajakirjandustekstis on ka ekstrapausid suhteliselt reeglipärased: neid tehakse sidesõnade *ja/ning*, nimede ja numbrite ees. Ilukirjandustekstis on ekstrapauside tekkekohad seevastu peamiselt sisulised: ilma teksti sisu mõistmata on nad reeglistamatud.

Tekstide lugemisel võib peamiseks žanrierinevuseks pidada niisiis seda, et uudiselugejad järgivad pause tehes küllaltki täpselt teksti vormiseiku (kirjavahemärke, numbritega kirjutatud arvsõnu, suurtähelisi nimesid), ilukirjandustekstide lugejad aga lähtuvad pigem teksti sisust kui vormist.

Ajakirjandusteksti võib seega tõenäoliselt võtta kui etteloetava teksti etaloni, millele toetada teistegi pragmaatilise orientatsiooniga ehk tarbetekstide esitus – ajakirjandus esindab eesti tekstiruumi keskmisi näitajaid (vt Kerge 2002a, 2003a) ja mõjub kõige neutraalsemalt, millise tahes žanri esitusega oleks tegemist.

Ajakirjandusteksti liigendust, punktuatsiooni, lugeja hingamist ja ekstrapause arvesse võttes moodustub kuus kestuselt erinevat pausirühma, mis – nagu Tseng'i ja tema kolleegi uurimusteski (Tseng 1999, Tseng, Chou 2002) – on hierarhilist laadi (vt tabel 3). Nende pausirühmade kestuste ligikaudseid väärtusi oleme soovitanud aluseks võtta neutraalsust taotleva sünteeskõne genereerimisel (vt Kerge, Pajupuu 2006).

Tabel 3. Pausirühmade keskmised kestused ajakirjandustekstis

Pausirühmad	**Pausi keskmine kestus** (ms)	**Hingamise keskmine kestus** (ms)
prosoodilise rühma paus (lõigupiir)	988	330
punktipaus	613	273
kooloni-, semikooloni-, mõttekriipsupaus	486	227
komapaus sissehingamisega	398	222
komapaus sissehingamiseta	170	–
ekstrapaus (*ja*/*ning*, nime, numbri ees)	69	–

Kokkuvõte

Uurimus näitab, et hingamispausid on vähemalt ettelugemise juures tugevasti süntaktiliselt tingitud siis, kui tegemist ei ole markantselt ekspressiivse tekstiga. Žanrilised erinevused on seniuuritus selgelt väiksemad kui erinevused allkeelte vahel: ilukirjandusproosa ettelugemises on oluline teksti sisuline interpretatsioon, mille nimel teinekord ohverdatakse isegi hingamine.

Spetsiaalsed ilukirjandusproosa lugemise reeglid tuleks seega kujundada veelgi kitsamalt aluselt kui žanr (siin romaan), see aga nõuaks eeltööna mahukat etteloetud kirjanduse korpust. Kuni sellist pole ja asjaomaseid reegleid kujundada ei saa, võib süntesaator kunstproosat ette lugeda samal moel nagu ajakirjandustekste, s.o ilma eriliste emotsioonideta. Seega, kuni emotsionaalse kõne korpus areneb (vt Altrov 2007), tuleks tekst–kõne süntesaatori pause modelleerides toetuda graafiliselt tuvastatavatele süntaktilise liigenduse elementidele. Nende tugevat seost pauseerimisega näitab süntaktiliselt mitmeti keskmiseks osutunud tekstiliik, pikema neutraalse ajakirjandusliku uudise tekst.

Sellistel ja muudel artikli algupooles viidatud andmetel võib väita, et milline tahes rakenduslik keelekäsitelu nõuab lähenemist vähemalt allkeelte ja ideaalis žanrite kaupa. Lisaks sünteeskõnele võiks ka muudes keeletehnoloogilistes rakendustes aktiivsemalt kasutada tekstiuurimuse ülal viidatud andmeid ja pidada silmas, et keskmist eesti keelt ei ole olemas: on tema olukohased variandid, mis avalduvad keelekasutuse valdkonnale omastes žanrites, ja see nõuab žanrilist lähenemist keelele ka arvutuslingvistikas, nagu seda on võimaluste piires arvestatud kirjakeele korpuste kujundamisel.

Kirjandus

Altrov, Rene 2007. Emotsionaalse kõne korpuse loomine eesti keele tekst–kõne sünteesi jaoks. Tekstimaterjali evalvatsioon *viha* näitel. Magistritöö. Tartu Ülikooli filoloogiateaduskond. Eesti ja soome-ugri keeleteaduse osakond. http://dspace.utlib.ee/dspace/bitstream/10062/2739/1/altrov_rene.pdf (3.01.2008).

Barnbrook, Geoff 2007. ... Uncovering the SECRET life of language: an introduction to text exploration tools. – Plenary presentation. The 3rd Baltic Conference on Human Language Technologies. Kaunas, Lithuania, October 4-5, 2007.

Biber, Douglas 1995. Dimensions of Register Variation. A Cross-Linguistic Comparison. Cambridge: Cambridge University Press.

Biber, Douglas; Conrad, Susan 2001. Register variation: A corpus approach. – Deborah Schiffrin, Deborah Tannen, Heidi Hamilton (eds). The Handbook of Discourse Analysis. Oxford: Blackwell, 175–96.

Boersma, Paul; Weenink, David 2006. Praat: doing phonetics by computer (Version 4.5.08). Computer program. http://www.praat.org/ (3.01.2007).

Campbell, Nick 1998. Where is the Information in Speech? (And to What Extent can it be Modelled in Synthesis?). – Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis. Australia: Jenolan Caves, 17–20.

Fant, Gunnar; Kruckenberg, Ann; Gustafson, Kjell; Liljencrants, Johan 2002. A new approach to intonation analysis and synthesis of Swedish. – Proceedings. Speech Prosody 2002, Aix en Provence, France, 11-13 April 2002, 283–286. http://aune.lpl.univ-aix.fr/sp2002/papers.htm (3.01.2008).

Fant, Gunnar; Kruckenberg, Anita; Ferreira, Joana B. 2003. Individual variations in pausing. A study of read speech. – PHONUM, Reports in Phonetics 9. Umeå University, 193–196. http://www.ling.umu.se/fonetik2003/ (3.01.2008).

Grosjean, François; Collins, Maryann 1979. Breathing, pausing and reading. – Phonetica 36, 98–114.

Gustafson-Čapková, Sofia; Megyesi, Beata 2001. A Comparative study of pauses in dialogues and read speech. – Paul Dalsgaard, Børge Lindberg, Henrik Benner, Zheng-hua Tan (eds). EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, September 3–7. Aalborg, Denmark, 931–934.

Hennoste, Tiit 1994. Prospektiivsed minimaalhesitatsioonid eesti keele suulises tekstis. – K. Pajusalu, V. Ylivakkuri (toim.). Lähivertailuja 7. Turun Yliopiston suomalaisen ja yleisen kielitieden laitoksen julkaisuja 44. Turku, 33–51.

Heylighen, Frances; Dewaele, Jean-Marc 2002. Variation in the contextuality of language. An empirical measure. – Foundations of Science 7, 293–340.

Kerge, Krista 2002a. Kirjakeele kasutusvaldkondade süntaktiline keerukus. – Reet Kasik (toim.). Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu Ülikooli eesti keele õppetooli toimetised 23. Tartu: Tartu Ülikool, 29–46.

Kerge Krista 2002b. Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX sajandil. TPÜ eesti filoloogia osakonna veebitoimetised. Lingvistika 1. Tallinn: TPÜ Kirjastus. http://www.tlu.ee/fil/veebitoimetised/pdf/lingvistika1.pdf (3.01.2007). Uustrükk Kerge 2003a [4].

Kerge, Krista 2003a. Keele variatiivsus ja *mine*-tuletus allkeelte süntaktilise keerukuse tegurina. Tallinna Pedagoogikaülikooli humanitaarteaduste dissertatsioonid 10. Tallinn: TPÜ Kirjastus.

Kerge Krista 2003b. Autori stiil ja allkeele tekst. – Reet Kasik (toim.). Tekstid ja taustad II. Tekstianalüüsi vaatepunkte. Tartu Ülikooli eesti keele õppetooli toimetised 26. Tartu: Tartu Ülikool, 59–89.

Kerge, Krista 2006. Kontekstuaalsus – on selles mingit seaduspära? – Ettekanne ja teesid. Tekstipäev 2006, 20. detsember, Tartu Ülikool.

Kerge, Krista; Pajupuu, Hille; Altrov, Rene 2007. Tekst, kontekstuaalsus ja kultuur. – Keel ja Kirjandus 8, 624–637.
Kerge, Krista; Pajupuu, Hille; Tamuri, Kairi. Ilmumas 2008. Where should TTS-synthesizer pause and breath? – The 3rd Baltic Conference on Human Language Technologies, 4-5 October. Kaunas: Vytautas Magnus University and Institute of Lithuanian Language.
Meier, Heidi 2002. Olulisi aspekte tekstitüübivõrdluses. – Reet Kasik (toim.). Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu Ülikooli eesti keele õppetooli toimetised 23. Tartu: Tartu Ülikooli Kirjastus, 101–114.
Meier, Heidi. 2003a. Essee asend allkeelte tekstitüübivõrdluses. Magistritöö. Käsikiri Tallinna Ülikooli eesti filoloogia osakonnas. Tallinn: Tallinna Pedagoogikaülikool.
Meier, Heidi 2003b. Essee allkeelte võrdluses. – Reet Kasik (toim.). Tekstid ja taustad II. Tekstianalüüsi vaatepunkte. Tartu: Tartu Ülikooli Kirjastus, 116–135.
Meister, Einar; Lippus, Pärtel. Ilmumas 2008. On temporal organization of spontaneous Estonian: preliminary analyses results of lecture speech. – The 3rd Baltic Conference on Human Language Technologies, 4-5 October. Kaunas: Vytautas Magnus University and Institute of Lithuanian Language.
Mihkla, Meelis 2005. Modelling pauses and boundary lengthenings in synthetic speech. – The 2nd Baltic Conference on Human Language Technologies, April 4–5, 2005. Proceedings. Tallinn, 305–310.
Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista 2003. Modelling and perception of the Estonian general questions with the *kas*-particle. – Proceedings of 15th ICPhS. Barcelona, 539–542.
Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista; Kuusik, Jüri 2004. Prosody modelling for Estonian text-to-speech synthesis. – The 1st Baltic Conference on Human Language Technologies. The Baltic Perspective, Riga, Latvia, April 21–22 2004. Riga, 127–131.
Mihkla, Meelis; Kerge, Krista; Pajupuu, Hille 2005. Statistical modelling of intonation and breaks for Estonian text-to-speech synthesizer. – Robert Vich (ed.). Electronic Speech Signal Processing. Proceedings of the 16th Conference Joined with the 15th Czech-German Workshop "Speech Processing". Prague, Sept. 26-28, 2005. Studientexte zur Sprachkommunikation 36. Dresden: TUDpress, 91–98.
Mihkla, Meelis; Kuusik, Jüri 2005. Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis. – Linguistica Uralica 2, 91–97.
Pajupuu, Hille 2007. Sõnaliik ja kontekstuaalsuse variatsioonid. – Ettekanne ja teesid. VI rakenduslingvistika kevadkonverents "Keel ja leksikon", 26.-27. aprill 2007, Tallinn. http://www.rakenduslingvistika.ee/word_doc/070405Teesid-2007-Keeljaleksikon.doc (3.01.2008).
Pajupuu, Hille; Kerge, Krista 2006. Hingav süntesaator ja pausid tekstiliigiti. – Keel ja Kirjandus 3, 202–210.
Põld, Karin 2007. Kaplinski tekstide kontekstuaalsus ja süntaks. Seminaritöö. Käsikiri Tallinna Ülikooli eesti filoloogia osakonnas. Tallinn: Tallinna Ülikool.
Rais, Kairi 2007. E-suhtlus ja selle õpetamine koolis (äriühingu näitel). Bakalaureusetöö. Käsikiri Tallinna Ülikooli eesti filoloogia osakonnas. Tallinn: Tallinna Ülikool.
Rannut, Mart 2003. Eesti keele jätkusuutlikkusest. – Helle Metslang (koost.). Eesti kirjakeele kasutusvaldkondade seisundi uuringud. Tallinna Pedagoogikaülikooli eesti filoloogia osakonna toimetised 4. Tallinn: TPÜ Kirjastus, 214–235.
Rätsep, Siret 2002. Haldusõiguse keel seaduskeele üldises raamis. Bakalaureusetöö. Käsikiri Tallinna Ülikooli eesti filoloogia osakonnas. Tallinn: Tallinna Pedagoogikaülikool.
Rääk, Kristjan 2002. Poliitiliste programmide keelest. – Reet Kasik (toim.). Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu Ülikooli eesti keele õppetooli toimetised 23. Tartu: Tartu Ülikooli Kirjastus, 115–127.

Tamuri, Kairi 2007. Pausid etteloetud ilukirjandustekstis. Magistritöö. Käsikiri Tallinna Ülikooli eesti filoloogia osakonnas. Tallinn: Tallinna Ülikool.
Tseng, Chiu-Yu 2002. The prosodic status of breaks in running speech: examination and evaluation. – Proceedings. Speech Prosody 2002, Aix en Provence, France, 11-13 April 2002, 667–670. http://aune.lpl.univ-aix.fr/sp2002/papers.htm (3.01.2008).
Tseng, Chiu-Yu; Chou, Fu-Chiang 1999. A prosodic labeling system for Mandarin speech database. – Proceedings of ICPh99. San Francisco, California, 2379–2382.
Viks, Ülle; Hein, Indrek 2001. Sõnavormide kasutus teadustekstides. – EKI. Publikatsioonid. http://www.eki.ee/teemad/ (12.12.2007).
Whalen, Douglas H. 1994. Computerized Speech Pauses for a Breath. – The Futurist, May/June, 28 (3), 7.

Krista Kerge (Tallinna Ülikool) uurimisvaldkonnad on keele variatiivsus, tekstianalüüs, rakenduslingvistika (L1 ja L2 omandamine, õigus- ja haldussuhtlus, kõne paralingvistiliste komponentide ja süntaksi seosed).
krista.kerge@tlu.ee

Hille Pajupuu (Eesti Keele Instituut) uurimisvaldkondadeks on kõneakustika, kultuuridevaheline kommunikatsioon, keeletestimine.
hille.pajupuu@eki.ee

Kairi Tamuri (Eesti Keele Instituut) uurimisvaldkonnaks on kõneakustika: tekstiliigid, emotsionaalne kõne, pausid ja hingamine.
kairi.tamuri@eki.ee

Heidi Meieri (vabakutseline uurija) uurimisvaldkond on tekstianalüüs.
heidi.meier@gmail.com

SPEECH TECHNOLOGY NEEDS A GENRE-BASED APPROACH

Krista Kerge, Hille Pajupuu, Kairi Tamuri, Heidi Meier
Tallinn University, Institute of the Estonian Language

In reading out a text, pauses and breathing as two of the naturalness parameters of the read-out text are closely related to its functional style (journalism, fiction). The results of text study may, apart from contributing to the naturalness of synthetic speech, help recognize the genre of the text, differentiate between functional styles as well as authors, and economize on the use of word lists and grammar in text automatic analysis. The article refers to some studies that relate to breathing and pauses as well as to some other formalizable parameters of Estonian texts, yet have not been considered in language technologies. To verify the above thesis the pauses and breathing in two genres of read-out texts of fiction are compared to the available data of a news text. The results prove that the duration of pauses and inhalations as well as their relation to the syntactic structure of the sentence does differ not only according to functional style but also according to genre. The syntactic structure is mainly indicated by punctuation marks. In fiction pausing partly depends on reader's interpretation. In text-to-speech synthesis for the effect of reading aloud breaks and breathing should be generated from the data of extended news texts as a parametrically neutral genre.

Keywords: pauses, breathing, functional style, genre, news, fiction, speech synthesis, Estonian