EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 251–266

KORPUSTE TÜKELDAMINE: RAKENDUSI SILPIDE NING ALLKEELTEGA

Kairit Sirts, Leo Võhandu

Ülevaade. Keelekorpustes sisalduvat materjali on võimalik erineval moel tükeldada, andes sellega võimaluse uurida keele erinevaid tahkusid. Artiklis uurime kahte oma omadustelt väga erinevat tükeldust: teksti tükeldamist silpideks statistilise keelemudeli leidmise eesmärgil ning keele tükeldamist erinevateks allkeelteks eesmärgiga leida eesti keele põhisõnavara.

Silpidel baseeruv statistiline keelemudel hõlmab endas 500 kõige sagedamini esinenud silpi ning on kolmetasandiline, koosnedes silpide, silbipaaride ja silbikolmikute järgnevuse sagedustabelitest. Sagedustabel on oma olemuselt maatriks, mille ridadeks on kas silbid, silbipaarid või silbikolmikud ning veergudeks silbid. Ridade ja veergude ristumispunktides on arv, mis näitab, mitu korda vastav veeru silp esines tekstikorpuses vastava reaelemendi järel.

Eesti pseudokeele generaator on silpidel baseeruva statistilise keelemudeli rakendus. Eesti pseudokeele generaatorit kasutades on võimalik genereerida teksti, mis ei ole küll päris eesti keel, aga kahtlemata kõlab eesti keelena.

Silpide kategoriseerimise eesmärgiks on rühmitada silbid vastavalt nende võimalikele asukohtadele sõnas. Pakume välja algoritmi silpide automaatseks rühmitamiseks kasutades silpide sagedustabelit. Näitame eksperimentaalselt kümne silbi abil, kuidas silbid jagunevad algus-, lõpu- ja kesksilpideks.

Keelt võib tükeldada n-ö põhisõnavara sisaldavaks üldkeeleks ning erinevateks allkeelteks, mis sisaldavad vastavat oskussõnavara. Käesolevas artiklis arutleme, kas ja kuidas on käesoleval ajal defineeritud üldkeel. Ühtlasi pakume välja algoritmi sellise põhisõnavara üheseks määratlemiseks arvuti abil.

Võtmesõnad: arvutilingvistika, keelemudel, silbitamine, silbiseostus, graafesitus, silpide rühmitamine, üldkeel, allkeeled, eesti keel

1. Sissejuhatus

Selle artikli autorid on hariduselt ja mõtteviisilt informaatikud, mitte filoloogid, raallingvistid või keeleõpetajad. Meie igapäevase teadustöö põhieesmärgiks on otsida paljuparameetriliste objektsüsteemide peidetud struktuure ja korrapärasid. Loodud meetodid on semantikavabad ning võimaldavad tunnuste abil kirjeldatud objektsüsteemide olemust küllalt kiiresti avada.

Paari viimase aasta jooksul oleme katsetanud neid meetodeid ka keelekorpustel. Internetis paiknevaid Tartu Ülikooli töörühmade loodud materjale[1] kasutades oleme uurinud eesti keele silbistruktuuri (Võhandu, Sirts, Aab 2008, Sirts 2008) ja üritanud fikseerida esmaseid tähelepanekuid. Samas selgus, et mingit standardselt esinduslikku keelekorpust ei eksisteerigi. Kõrvaltvaataja pilgule avanes hoopis üpris selgelt vajadus formaalselt defineerida eesti keele allkeeled, sest kõikehõlmavate mudelite ehitamine on praegu ilmselt liiga raske.

Järgnevas kirjeldamegi kõigepealt tööd korpustega ja seejärel eelnevast tulenevalt mõningaid mõtteid eesti keele allkeelte defineerimisvõimalustest.

Arvutile võib keelt õpetada mitmel moel – programmeerides kogu grammatika reeglistiku ning andes ette terve sõnastiku või siis õpetades programmi olemasolevate keeleressursside abil. Nendeks ressurssideks on inimeste poolt realiseeritud keel näiteks kirjaliku keele tekstikogude ehk korpuste näol.

Korpuste baasil õppimine on induktiivne meetod. On olemas keeleressurss, mida on juba mingil moel kasutatud. Õppima asudes ei ole keelereeglistik teada ning see tuletatakse õpitud korpuste baasil vastavalt valitud õppemetoodikale. Tulemuseks võib olla reeglistik, mida me grammatikaõpikus ei kohta. Samuti võib keeleelementide valik, millel reeglistik baseerub, olla sootuks harjumatu.

Artiklis on keeleelementideks valitud silbid, millele on üles ehitatud eesti keele mudel. Mudeli silpidevahelised seostusreeglid esitatakse silpide järgnevuse statistiliste sageduste abil. Silpide järgnevuse sagedused on õpitud eesti kirjakeele korpuse abil.[2]

2. Korpuse töötlemine

2.1. Korpuse valik

Eesmärgiks oli saada kogum eesti keeles esinevaid silpe, mis oleks piisav, et ära katta suurem osa kogu keelest. Esmapilgul võib tunduda, et selle eesmärgi saavutamiseks on üsna ükskõik, milline korpus valida. Eeldusel, et korpus on piisavalt suur, võiks esinduslik silpide hulk igasuguse korpuse puhul n-ö pinnale ujuda, olgu siis tegemist ilukirjanduslike või ajakirjanduse tekstidega.

Olles läbi teinud silbitamise ning mudeli koostamise protsessi nii ilukirjanduse kui ka ajakirjanduse korpusega, võib öelda, et mõlemal valikul on nii oma eelised kui ka puudused. Ilukirjandustekstide korpuse puudusena võib välja tuua selle, et ilukirjanduse tekstide sõnavara on oluliselt laiem kui igapäevases (kõne)keeles kasutatav. Seda aspekti võib käsitleda loomulikult ka eelisena, kui ülesande püstitus nõnda sätestab. Antud juhul on eesmärgiks aga leida võimalikult kompaktne silpide hulk, mis võimalikult palju kataks n-ö tavakeelt. Sellise ülesande puhul on

[1] Vt http://www.cl.ut.ee/korpused/index.php?lang=et (28.12.2008).
[2] Vt http://www.cl.ut.ee/korpused/baaskorpus/ (28.12.2008).

ilukirjanduses kasutatava sõnavara, mida tavakeeles reeglina ei kasutata, olemasolu igal juhul võimalikuks puuduseks.

Ajakirjandustekstide eeliseks on see, et sõnavara hulk on väiksem ja lähedasem tava(kõne)keelele. Siiski on ajakirjandusetekstidel ka omad puudused. Seal leidub mitmeid sageli korduvaid sõnu, mis on omased just päevakajalistele ajaleheartiklitele, näiteks *Euroopa*, *sotsiaal-*, *aktuaal-* jne. Seetõttu omandavad sellistes sõnades sisalduvad silbid koos vastava järjestusega ebaproportsionaalselt kõrge sageduse. Peale selle leidub ajakirjanduslikes tekstides palju pärisnimesid ning võõrkeelseid sõnu. Kui võõrkeelsetest sõnadest tekkiv probleem on kergesti hoomatav, siis eestipäraste pärisnimede esinemine ei tundugi esialgu problemaatiline olevat. Kui hakata neid pärisnimesid aga lähemalt uurima, siis selgub, et need koosnevad tihti eesti keeles sagedamini kasutatavatest silpidest, neist moodustuv sõna aga ei olegi eesti keeles mujal kasutusel kui ainult selles konkreetses pärisnimes. Näitena võib tuua perekonnanime *Tammet*. Silbid *tam* ja *met* on mõlemad suhteliselt sagedalt esinevad, kuid järgnevus *tam-met* esineb ainult pärisnimes.

Võttes arvesse ülaltoodud kaalutlusi oleks kõige sobivam kasutada optimaalse silpide kogumi leidmiseks tasakaalustatud korpust, mis sisaldaks võrdsel määral nii ajakirjandusteksti kui ka ilukirjandust. Lisaks võib üritada teksti töödelda selliselt, et silpide kogumi leidmisel jäetakse vaatluse alt välja pärisnimed (tuvastatavad lause keskel esineva suure algustähe põhjal) ja võõrkeelsed sõnad (tuvastatavad võõrtähtede esinemise põhjal sõnas).

2.2. Korpuse silbitamine

Käesolevas artiklis kajastatud tulemuste aluseks on ajakirjandustekste sisaldav korpus.[3] Selleks, et korpuse tekstist saaksid silbid ning silpidest mudel, oli vaja korpust kõigepealt natuke töödelda. Kasutatud korpuse iga rida oli märgendatud allikaviitega, mis tuli enne silbitamist eemaldada. Ka olid täpitähed märgendatud vastavate koodidega, mis tulid asendada. Seejärel oli teksti silbitamiseks võimalik kasutada Eesti Keele Instituudi (EKI) loodud silbitamise tarkvara.[4] Järgmiseks ülesandeks oli kõikide korpuses esinenud silpide esinemise arvu kokkulugemine ning sageduse järgi järjestamine. Keelemudeli loomiseks kaasati viissada kõige sagedamini esinenud silpi, mille baasil loodi silpide järgnevuse sagedustabel.

Kirjeldatud protsessi käigus tekkis ka mitmeid probleeme. Põhilised probleemid, mis seoses korpusega esinesid, olid täpitähtede valed kodeeringud ning ohtrad õigekirjavead korpuse tekstis. Õigekirjavigade tõttu tekkis silpide vahele selliseid seoseid, mida eesti keeles tegelikult ei esine. Samas on selliste seoste kindlaks tegemine ning mudelist eemaldamine väga ajamahukas töö.

Et aimu anda, millisel kujul tekst korpuses on esitatud, toome siinkohal fragmendi (1).

(1) A JAE1990\ee0283 Järgnes ülekuulamine, mida viis läbi Tomingas.
AJAE1990\ee0283 Mis on teie nimi?
AJAE1990\ee0283 Mihhail Konstantinovitš Krupski (siin Tomingas eksib nimedega).
AJAE1990\ee0283 Kas teil on õde?

[3] Vt http://www.cl.ut.ee/korpused/baaskorpus/txt/1999aja_txt_elan.zip (28.12.2008).
[4] Vt http://www.eki.ee/tarkvara/silbitus/ (28.12.2008).

```
AJAE1990\ee0283   Jaa.
AJAE1990\ee0283   Mis nimi?
AJAE1990\ee0283   Nade&zcaron;da.
AJAE1990\ee0283   Kus teid vangi v&otilde;eti?
AJAE1990\ee0283   Gat&scaron;inas.
AJAE1990\ee0283   Kas olete olnud kohtu all?
AJAE1990\ee0283   Jaa.
```

Teksti mõistlikuks esitamiseks tuleb teha parasjagu eeltööd. Me kasutame J-keelseid[5] programmilõike, mille abil on teksti ettevalmistamine lihtne ja lühike:

```
tekst =: asenda_koodid tekst_vaikeseks kustuta_muster loe_fail 'c:\
j601\user\proov.txt'
silbid =: silbita tekst
```

Teksti puhastamisprogramm haarab rea lõpust ülakomade vahel oleval aadressil paikneva teksti, kustutab rea eesotsas oleva lisainformatsiooni, teeb siis kõik tähed väikeseks ja lõpuks asendab umlaudid normaaltähtedega.

Umlautide asendusprogramm on selline:

```
asenda_koodid =: 3 : 0
vana =: '&otilde;';'&auml;';'&ouml;';'&uuml;';'&scaron;';'&zcaron;'
uus =: 'õ';'ä';'ö';'ü';'š';'ž'
vanauus =: vana ,. Uus
y rplc vanauus
)
```

Silbitamise käigus tekkinud põhiline probleem oli see, et EKI silbitajaga ei ole võimalik korrektselt silbitada kõiki liitsõnu. Raskusi tekkis selliste liitsõnadega, mille puhul oleks tarvis eelnevalt teada, et tegemist on liitsõnaga ja mille puhul oleks tarvis silbitada iga liitsõnaosa eraldi. Näiteks võib tuua *võib-ol-la* vs. *või-bol-la*.

3. Statistiline keelemudel ja sellel baseeruvad rakendused

Eesti keele silbistruktuuri tundus olevat kõige lihtsam ja sobivam uurida silpide järgnevuse abil, modelleerides silpide seostusreeglid stohhastilise lõpliku automaadi abil, mis on esitatud maatrikskujul. Tegemist on ruutmaatriksiga, mille ridadeks ja veergudeks on teatud hulk välja valitud silpe ning rea ja veeru ristumiskoht näitab, mitu korda veerusilp järgnes analüüsitavas tekstis reasilbile.

(2) $T = ||tij||\ i = 0...n-1, j = 0...n-1$

Tõenäosus, et *j*-silp järgneb *i*-silbile on võrdne maatriksi *ij*-elemendi väärtuse ning *i*-rea summa jagatisega.

(3) $P(t_j|t_i) = t_{ij} / \sum_{j<n-1}(t_{ij})$

[5] Vt http://www.jsoftware.com/ (28.12.2008).

Edaspidi nimetame seda maatriksit sagedustabeliks. Sagedustabeli leidmiseks tuli kõigepealt välja valida, kui palju ja millised silbid analüüsi kaasata. Kokku esines silbitatud korpuses 7225 erinevat silpi, lisaks sõnade vahe ehk tühik, mis sai samuti defineeritud eraldi silbina. Analüüsi jaoks said valitud 501 kõige sagedamini esinenud silpi (500 silpi ning tühik), mis ühtekokku katsid 86 protsenti tekstist. Tühik eraldi silbina pakub huvi seetõttu, et meie huviks on modelleerida mitte ainult silpide järgnevust sõnas, vaid ka sõnade vahelised piirid. Ilma tühikuta oleks üpris keeruline aru saada, milliste silpide järel saabub sõna lõpp või milliste silpidega võiks sõna alustada. Viimase silbina pääses valitud silpide hulka *gib*, mis esines tekstis 266 korda. Muuhulgas jäeti analüüsist välja osa silpe, mis kuulusid 500 kõige sagedamini esinenud silbi hulka, kuid sisaldasid võõrtähti, ning mille kõrge esinemissagedus oli tingitud ajakirjandustekstide spetsiifikast. Välja jäeti silbid: *ca*, *co*, *fo*, *fi*, *fir*.

Silpe, mis oleks tulnud analüüsist välja arvata, on ka järelejäänud 500 silbi hulgas, aga nende tuvastamine ei ole enam nii lihtne. Selleks tuleks käsitsi läbi vaadata kõik silpide järgnevused, et kindlaks teha, millised neist esinevad ainult päris- või kohanimedes või on tekkinud liitsõnade mittekorrektsel silbitamisel.

Tabelis 1 toome esimesed viiskümmend kõige sagedamat silpi koos sagedustega.

Tabel 1. Sagedamini esinevad silbid

tühik	388338
le	21640
ta	21340
se	20458
ja	19081
ma	17316
te	16589
da	15611
li	15532
si	13784
a	13186
ga	12418
mi	12219
ti	11982
tu	11646
ne	11610
ni	11549

va	11311
o	11292
ri	10002
on	9973
e	9962
na	9420
ka	9094
gi	8527
ku	7758
la	7655
de	7336
me	7210
sa	6877
nud	6457
gu	6178
et	6012
ra	5988

di	5581
lu	5411
ko	5361
ju	5256
sel	5243
su	5147
du	5079
ei	5057
i	4832
ha	4575
ge	4461
ki	4446
kui	4176
vad	4107
he	3988
ü	3960

Välja valitud 500 silbi ja tühiku baasil koostasime erinevaid sagedustabeleid:

1. silpide järgnevuse sagedustabel;
2. silbipaaride järgnevuse sagedustabel;
3. silbikolmikute järgnevuse sagedustabel.

Silpide järgnevuse sagedustabel on seosmaatriks, mis näitab ära järgnevusseose ning selle tugevuse kahe silbi vahel.

Sagedustabeliks on nxn-maatriks, mille ridadeks ja veergudeks on silbid kindlaks määratud järjestuses. Antud juhul on silbid järjestatud esinemissageduse

järgi kahanevas järjekorras ning iga silp on kodeeritud oma järjekorranumbriga 0..n-1.

Olgu S = {s_0, s_1,..., s_{n-1}) analüüsitavate silpide hulk. Silpide s_i ja s_j vaheline järgnevusseos R(s_i, s_j) = 0, kui silp s_j ei järgnenud mitte kordagi analüüsitavas tekstis silbile s_i. Silpide s_i ja s_j vaheline järgnevusseos R(s_i, s_j) > 0, kui silp s_j järgnes vähemalt ühe korra analüüsitavas tekstis silbile s_i. Järgnevusseose R(s_i, s_j) väärtuseks on arv, mitu korda silp s_j järgnes tekstis silbile s_i.

Siinkohal esitame fragmendi silpide järgnevuse sagedustabelist (tabel 2), mis kajastab andmeid kümne enim esinenud silbi kohta.

Tabel 2. Silpide järgnevuse sagedustabel

	tühik	***le***	***ta***	***se***	***ja***	***ma***	***te***	***da***	***li***	***si***
tühik	0	730	4961	1830	10399	3271	4402	90	1124	800
le	13599	324	298	31	120	636	55	72	26	32
ta	6551	12	142	223	914	763	103	1382	130	55
se	12211	1096	121	19	13	157	188	1190	90	9
ja	14035	265	60	11	16	50	679	58	212	6
ma	8844	248	533	157	436	12	255	75	169	65
te	7431	1421	64	9	2	859	46	196	104	4
da	8990	214	262	17	354	309	48	456	21	302
li	4333	75	168	909	70	116	103	90	11	202
si	4479	187	247	215	37	83	40	94	60	20

Kuna ridade ja veergude summad on erinevad, siis ei ole veergude ega ridade väärtused otseselt võrreldavad. Selleks, et neid saaks omavahel võrrelda, tuleks read normeerida.

Silbipaaride järgnevuse sagedustabel on seosmaatriks, mis näitab ära järgnevusseose ning selle tugevuse kahe järjestikuse silbi ehk silbipaari ning üksiku silbi vahel.

Silbipaaride sagedustabeliks on mxn-maatriks, mille ridadeks on silbipaarid ja veergudeks silbid vastavas järjestuses. Nii silbid kui ka silbipaarid on järjestatud esinemissageduse järgi kahanevas järjekorras. Iga silp ja silbipaar on kodeeritud oma järjekorranumbriga vastavalt 0..n-1 ja 0..m-1.

Olgu SP = S x S = {s_is_j} silbipaaride hulk. Silbipaari s_is_j ja silbi s_k vaheline järgnevusseos R(s_is_j, s_k) = 0, kui silp s_k ei järgnenud mitte kordagi analüüsitavas tekstis silbipaarile s_is_j. Silbipaari s_is_j ja silbi s_k vaheline järgnevusseos R(s_is_j, s_k) > 0, kui silp s_k järgnes vähemalt ühe korra analüüsitavas tekstis silbipaarile s_is_j. Järgnevusseose R(s_is_j, s_k) väärtuseks on arv, mitu korda silp s_k järgnes silbipaarile s_is_j.

Uuritavas korpuses esines kokku 57092 erinevat silbipaari. Neist said analüüsi kaasatud 5000 kõige sagedamini esinevat silbipaari, mis koosnesid 500 enim esinenud silbist ja mis katsid ära 76 protsenti kogu analüüsitud tekstist.

Toome ära fragmendi silbipaaride sagedustabelist (tabel 3), mis kajastab kümne enam esinenud silbipaari ning silbi vahelisi seoseid.

Tabel 3. Silbipaaride järgnevuse sagedustabel

	tühik	***le***	***ta***	***se***	***ja***	***ma***	***te***	***da***	***li***	***si***
***ja* tühik**	0	24	148	67	146	131	197	12	51	46
***le* tühik**	0	16	158	64	392	120	116	0	58	28
***se* tühik**	0	32	138	36	395	88	96	1	38	44
tühik *ja*	9930	0	0	0	0	17	0	3	0	0
tühik *o*	29	1653	0	0	14	2482	18	114	1854	2
tühik *on*	9849	0	2	0	0	0	0	0	1	0
***ga* tühik**	0	28	112	49	182	124	165	3	33	28
***on* tühik**	0	11	126	55	20	73	146	0	29	18
***da* tühik**	0	12	114	57	277	91	134	1	31	7
***ma* tühik**	0	9	77	44	188	74	87	1	22	16

Silbikolmikute järgnevuse sagedustabel on seosmaatriks, mis näitab ära järgnevusseose ning selle tugevuse kolme järjestikuse silbi ehk silbikolmiku ning üksiku silbi vahel.

Silbikolmikute sagedustabeliks on mxn-maatriks, mille ridadeks on silbikolmikud ja veergudeks silbid vastavas järjestuses. Nii silbid kui ka silbikolmikud on järjestatud esinemissageduse järgi kahanevas järjekorras. Iga silp ja silbikolmik on kodeeritud oma järjekorranumbriga vastavalt 0..n-1 ja 0..m-1.

Olgu SK = $\{s_i s_j s_k\}$ silbikolmikute hulk. Silbikolmiku $s_i s_j s_k$ ja silbi s_l vaheline järgnevusseos $R(s_i s_j s_k, s_l) = 0$ siis, kui silp s_l ei järgnenud mitte kordagi analüüsitavas tekstis silbikolmikule $s_i s_j s_k$. Silbikolmiku $s_i s_j s_k$ ja silbi s_l vaheline järgnevusseos $R(s_i s_j s_k, s_l) > 0$ siis, kui silp s_l järgnes vähemalt ühe korra analüüsitavas tekstis silbikolmikule $s_i s_j s_k$. Järgnevusseose $R(s_i s_j s_{kj}, s_l)$ väärtuseks on arv, mitu korda silp s_l järgnes silbikolmikule $s_i s_j s_k$.

Uuritavas korpuses esines kokku 257240 erinevat silbikolmikut. Neist said analüüsi kaasatud 10000 kõige sagedamini esinevat silbikolmikut, mis koosnesid 500 enim esinenud silbist ning mis kokku katsid ära 51 protsenti kogu analüüsitud tekstist.

Toome ära fragmendi sagedustabelist, mis kajastab kümne enim esinenud silbikolmiku ja silbi vahelisi seoseid (tabel 4).

Tabel 4. Silbikolmikute järgnevuse sagedustabel

	tühik	***le***	***ta***	***se***	***ja***	***ma***	***te***	***da***	***li***	***si***
tühik *ja* tühik	0	17	83	57	20	103	166	6	36	39
tühik *on* tühik	0	11	126	55	20	73	146	0	29	18
tühik *et* tühik	0	8	169	39	1	76	153	0	8	11
tühik *ei* tühik	0	4	158	3	7	15	24	0	3	7
tühik *kui* tühik	0	4	115	26	2	67	57	0	6	9
tühik *ka* tühik	0	2	19	17	3	21	39	2	6	11
tühik *o ma*	1935	8	0	2	3	3	0	10	0	0
tühik *ta* tühik	0	1	22	21	7	11	18	0	24	5
tühik *ees ti*	1893	30	0	0	0	6	0	0	1	0
tühik *see* tühik	0	2	8	6	4	4	19	0	3	2

3.1. Eesti pseudokeele generaator

Eesti pseudokeele generaator on programm, mille abil saab genereerida eesti keelele sarnanevat keelt. Kuigi tegemist pole eesti keelega, on tema kõla vägagi sarnane eesti keelele. Eesti pseudokeele generaatori aluseks on silpidest koosnev statistiline keelemudel.

Pseudokeele teksti genereeritakse silp-silbi haaval ning iga järgmise silbi genereerimisel arvestatakse maksimaalselt kolme viimati genereeritud silbiga. Võimalusel kasutatakse järgmise silbi genereerimiseks silbikolmikute järgnevuse sagedustabelit. Kui see pole võimalik, siis üritatakse kasutada silbipaaride järgnevuse sagedustabelit ning kui ka see pole võimalik, siis kasutatakse silpide järgnevuse sagedustabelit.

Järgmine silp valitakse välja juhuslikult statistilise tõenäosuse alusel. Kuna ka sõnavahe on defineeritud silbina, millel on oma esinemise sagedus iga silbi ees ja järel, siis järgmise silbi juhuslikul valikul genereeritakse piisavalt ka sõnavahesid, mis tagab teksti liigendumise mõistliku pikkusega sõnadeks. Kuna muid teksti liigendamise märke (koma, punkt jms) mudelis ei ole, siis on genereeritud tekst lihtsalt sõnade jada ilma lauseteks liigendamiseta.

Toome näite pseudokeele generaatoriga genereeritud tekstist (2).

(2) ja tiiu vahel pakitunudki europarlamendi kevade asi medate inseni mitme positsioonist kui kasu mistada sellest arutamine on sul liiba pinnaga torma on hoopis venelased suureneb usa kinnisvarast seda istuda väga rusikapanu üle miljoni krooni esimest korda

3.2. Silpide kategoriseerimine

Genereerides teksti eesti pseudokeele generaatoriga võib juhtuda, et ühe sõna lõpusilbist alustatakse kohe järgmise sõnaga. Selleks, et taoliseid juhtumeid minimeerida, oleks vaja natuke heuristilist teadmist selle kohta, millised silbid millistes sõnaosades esineda võivad. Sellest eesmärgist lähtuvalt üritamegi jagada silbid rühmadesse ning defineerida, millal üks või teine silp sõnas ette tulla võib.

Katsetusteks valisime juhuslikult 10 silpi sagedaima 50 hulgast: *ti*, *va*, *ri*, *e*, *gi*, *la*, *di*, *sel*, *i*, *kui* järjekorraindeksitega vastavalt 13, 17, 19, 21, 24, 26, 34, 38, 42, 46 (vt tabel 5).

Tabel 5. Juhuslikult valitud silbid kategooriate eksperimentaalseks leidmiseks

	ti	***va***	***ri***	***e***	***gi***	***la***	***di***	***sel***	***i***	***kui***
ti	13	94	8	6	24	2	0	2	0	0
va	40	11	93	0	1	56	10	5	0	0
ri	285	38	3	35	23	4	5	7	0	2
e	31	11	708	0	7	554	1	0	0	0
gi	22	27	2	20	7	7	3	77	0	0
la	173	166	104	0	8	3	10	22	0	0
di	12	40	57	8	22	3	4	5	0	0
sel	0	0	0	0	170	1	0	0	0	0
i	1	27	2	0	8	0	2	0	0	0
kui	0	13	0	0	254	0	0	1	0	0

Defineerime silbi kvantitatiivse parameetri (kvantp) kui vastava silbi rea summa jagatise vastava silbi veerusummaga:

(3) $kvantp_k = \sum_{j=0...n-1} t_{kj} / \sum_{i=0...n-1} t_{ik}$

Defineerime silpide seostusmaatriksi K:

(4) $K = ||k_{ij}||, (k_{ij} = 1, t_{ij} > 0), (k_{ij} = 0, t_{ij} = 0)$

Defineerime silbi kvalitatiivse parameetri (kvalp) kui seostusmaatriksi K vastava silbi rea summa jagatise vastava silbi veerusummaga:

(5) $kvalp_k = \sum_{j=0...n-1} k_{kj} / \sum_{i=0...n-1} k_{ik}$

Sõnavahe ehk tühiku parameetri (tparam1) defineerime kui vastavate silpide väärtused tühiku reas jagatise vastavate silpide väärtustega tühiku veerus:

(6) $tparam_k = t_{ok} / t_{ko}$

Arvutame näiteandmetele kirjeldatud parameerite väärtused (tabel 6).

Tabel 6. Kategoriseerimise parameetrite väärtused

silp	kvanp	kvalp	tparam
ti	0.258232	0.875	0.0283775
va	0.505855	0.777778	2.33881
ri	0.411464	1.125	0.070059
e	19.0145	1.5	6.7093
gi	0.314885	0.8	0.00749951
la	0.771429	0.875	0.403023
di	4.31429	1.14286	0.119798
sel	1.43697	0.285714	1.77266
i	–	–	21.9409
kui	134	3	1.19471

Tabelis 7 defineerime järgmised silpide kategooriad.

Tabel 7. Silpide kategooriad

Lühend	Nimetus	Kirjeldus
VA	välistav algussilp	Esinevad ainult sõnade alguses, neile võib eelneda ainult sõnavahe.
TA	tugev algussilp	Sagedased sõnade alustajad, aga neile võib olla omistatud ka muid kategooriaid.
NA	nõrk algussilp	Võivad samuti esineda sõna alguses, aga neil on päris kindlasti veel omistatud ka muid kategooriaid. Sageli kuuluvad need silbid ka kesksilpide hulka.
KS	kesksilbid	Esinevad sõna keskel, reeglina on nad ka veel kas alustajad ja/või lõpetajad.
NL	nõrk lõpusilp	Lõpetajad, aga sageli võivad mängida ka kesksilbi rolli.
TL	tugev lõpusilp	Lõpetavad sageli sõnu. Peale neid võib tulla kas sõnavahe või veel üks tugev lõpusilp.
VL	välistav lõpusilp	Esinevad ainult sõna lõpus, neile võib järgneda ainult sõnavahe.

Selleks, et sõnu kvanp, kvalp ja tparam alusel kategooriatesse jagada, on vaja kindlaks määrata süsteemi parameetrite väärtused. Katseliselt defineerime süsteemi parameetrid järgmiselt (6).

(6) kvanp parameetrid:
a = 10
b = 1/a
c = 2,5
d = 1/i
kvalp parameetrid:
e = 1,5
f = 1/c
tparam parameetrid
g = 5
h = 1/e
i = 2
j = 1/g

Silpide kategooriad arvutatakse tabelis 8 esitatud reeglite abil.

Tabel 8. Kategooriate arvutamise reeglid

Kategooria	**Reeglid**
VA	kvanp = _
TA	(kvanp >= a JA kvalp > 1) VÕI (kvalp >= e JA kvanp > 1) VÕI tparam > g
NA	1 < kvanp < a VÕI 1 < kvalp < e VÕI h < tparam < j VÕI tparam > 1
KS	c < kvanp < d VÕI f < kvalp < e
NL	b < kvanp < 1 VÕI f < kvalp < 1 VÕI i < tparam < g
TL	(kvanp <= b JA kvalp < 1) VÕI (kvalp <= f JA kvanp < 1) VÕI tparam < h
VL	kvanp = 0

Katse tulemused on ära toodud tabelis 9.

Tabel 9. Silpide kategoriseerimise eksperimendi tulemused

Kategooria	**Silbid**
VA	*i*
TA	*e*, *i*, *kui*
NA	*va*, *ri*, *la*, *di*, *sel*
KS	*ti*, *va*, *ri*, *gi*, *la*, *di*, *sel*
NL	*va*, *la*
TL	*ti*, *ri*, *gi*, *di*, *sel*
VL	

Katse tulemused on mõnevõrra moonutatud, sest parameetrite kvanp, kvalp ja tparam arvutamisel on arvestatud ainult katsesilpe sisaldavat fragmenti sagedustabelist. Seetõttu langeb näiteks silp *ti* ainult kesksilbi ja tugeva lõpusilbi kategooriasse, samas kui keeleline vaist nõuaks selle silbi paigutamist ka algussilbi kategooriasse.

4. Eesti keele allkeelte formaalse defineerimise vajadusest

Eesti rahvuskeel tekkis 19. sajandi teisel poolel ning tagab suhtlemisvõimaluse kõigil elualadel (EE). Keelt võib tükeldada õige mitmeti. Erialasest kallutusest tingituna vaatleksime lähemalt seda jaotust, kus kirjakeel loetakse koosnevana üldkeelest ja oskuskeeltest (T. Erelt 1982: 17, Kull 2000: 143).

Eesti semiootilise mõtte suurmees Jakob Linzbach kirjutas 1916. aastal oma venekeelse raamatu "Filosoofilise keele printsiibid. Täpse keeleteaduse kogemus." 38. peatükile väga ilmeka pealkirja: "Keel ja teadus. Teaduse jagunemise paratamatus. Paljukeelsuse õigustus." Napilt seitsmel leheküljel annab J. Linzbach hiilgava ülevaate formaalselt kirjeldatud erikeelte tekke vajadusest keeruka maailma nähtuste ja protsesside lõpmatu hulga eri külgede täpsel kirjeldamisel ning võimalikult mitmekülgsel ja selgel esitamisel. J. Linzbach näitab seejuures, et on vaja tervet formaalsete reeglite kohaselt toimivate märgisüsteemide (keelte) kogumit. Seega pole J. Linzbachi arvates lootagi mingi universaalkeele teket, vaid igal juhul on tegu paljukeelsusega.

Eesti keele korpus koosneb faktiliselt mitmes allkeeles kirjutatud tekstidest. T. Hennoste ja K. Muischnek osundavad (2000), et baaskorpuse kategooriad on ajakirjandus, religioosne kirjandus, hobid ja harrastused, populaarkirjandus, esseed ja biograafiad, dokumendid, teadus, ilukirjandus, entsüklopeediad ja propaganda.

Kerge on märgata, et nende kategooriate tekstikäsitlus ja sõnavara on vägagi erinev. Kõigepealt on ilmne, et põhisõnavara süvaossa kuuluvad sõnad (tuumsõnad) on nii Wierzbicka kui R. Langackeri mõttes primitiivid (Luuk 2008). Iga allkeel on tekkinud nende primitiivide baasil loomuliku evolutsiooni tulemusel ja on paraku nii J. Linzbachi mõttes kui kaasaegse ontoloogilise süsteemikirjelduse aspektist tegelikult siiani täpselt fikseerimata.

Tsiteerime siinkohal T. Hennostet: "Eesti keele allkeelte teaduslik süstemaatiline määratlemine on olnud väga juhuslik (vt mõned varased katsed Rätsep 1976; Pajusalu 1992). Praktiliselt on kasutatud mõnda mõistet (*kirjakeel*, *ühiskeel*, *argikeel*, *kõnekeel*, *murre*), kusjuures need on üsna uduselt defineeritud ja praktilises kasutuses pigem intuitiivsed." (Hennoste 2000: 9) T. Hennoste (2002: 231) väidab, et tema (Hennoste 2000) ja K. Kerge (2000) allkeelte süsteemide skeemid on tugevalt ja põhimõtteliselt erinevad.

M. Erelt ja T. Hennoste avaldasid kogumikus "Tähendusepüüdja" paljuütleva pealkirjaga artikli "Vaja on veel üht eesti keele grammatikat" (M. Erelt, Hennoste 2002). Huvitavad on veel kogumikus "Tuumsõnade semantikast ja pragmaatikast" (R. Pajusalu jt 2004) avaldatud seisukohad, kus põhiliselt käsitletakse tuumsõna keskset, suhteid väljendavat osa, mida nimetatakse põhisõnavara operaatoriks.

Niipalju siis juhtfiloloogide vaadetest eesti keele allkeeltele ja keelele endale. Allkeelte kui piiritletud süsteemide formaalsete täiskirjelduste – ontoloogiate loomine on ilmselt tuleviku probleem ning nõuab filoloogide ja raallingvistide kõrval ka keeruliste infosüsteemide formaalkirjeldajate – ontoloogide otsest osavõttu.

Siinkohal piirduksime lihtsama, kuid siiski huvitava ülesandega. Kuidas määratleda eesti üldkeele põhisõnavara, millega saab kõike soovitavat selgelt ja täpselt üles kirjutada ning välja ütelda? Praegusel hetkel oleme veendunud, et põhisõnavara koostamisel tuleb kiire (võib-olla ligikaudse) lahenduse saamiseks kasutada juba

olemasolevaid avalikke, üldkättesaadavaid sõnastikke, mis kindlasti peavad olema käideldavad digikujul.

Kiire kõrvalpõige Keelevara koduleheküljele näitab, et praegu on üldnimekirjas kümme eesti keele sõnaraamatut.[6]

Kaks meile vajalikku põhisõnastikku, "Eesti kirjakeele seletussõnaraamat" (EKSS) ja "Võõrsõnastik" on Keelevara tasulises nn profipaketis andmebaasina käideldavad. Kuidas nende sõnastike abil üritada defineerida eesti üldkeele põhisõnavara, laskumata võõrsõnadesse ja nendega sageli seonduvatesse oskuskeelte sõnadesse? Usaldame EKI sõnastikumeistreid ja valime üldkeele sõnavarasse mitte EKSS-i märksõnad, vaid nende kirjeldamiseks kasutatud semantiliste kirjelduste sõnavara. Arvuti abil ei ole selle töö tegemine kuigi raske. Ilmselt on tekkiv sõnanimistu veel kõlbmatu, sest seal on vastavalt EKSS-i autorite subjektiivsusele sees ka võõrsõnu ja oskuskeelte sõnu.

Esimese lähendina võiksime defineerida eesti üldkeele põhisõnavara kui EKSS-i seletussõnad, millest elimineerime "Võõrsõnastiku" sõnad. Võõrsõnad on tavaliselt kas rahvusvahelised üldsõnad või väga sageli osutuvad mingi oskuskeele terminiteks. Oluline on seejuures veel asjaolu, et sageli on erinevates oskuskeeltes (metakeeltes) ühe ja sama sõna tähendus erineva semantikaga. Toome siinkohal triviaalse näite sõnaga *programm*. EE osundab *kava, eeskava, saatekava, tegevus-, toimimis- või juhtkava, õppekava, eeskiri, algoritm*. Kerge on endale ette kujutada, kuidas erinevatel elualadel on mängus selle sõna erinevad semantikad. (Muide, siit saaks omaette huvitava uuriva artikli EKSS-i toimetajate uskumustest selgitavate ja kõigile eelduslikult üldarusaadavate sõnade valiku osas.)

Umbes selline võiks siis olla üldkeele eestikeelne põhisõnavara. Nüüd saame püstitada uue probleemi. Milline peaks olema järgmine kõrgem keeleline tavatase, mis enamikku eestlasi rahuldaks ja annaks piisava stiililise mitmekesisuse esitusliku ja grammatilise lihtsuse juures? See ühiskeel võiks olla midagi soomlaste *selkokieli* (klaarkeel?!) ja inglaste *Plain English*-i mõtteviisi ja tasemega määratut. Inglise keele valdajatele võib samal teemal soovitada lugeda Arvi Parbo mõnusa eessõnaga varustatud ja otseselt eesti lugejale mõeldud Michael Haagenseni raamatut "Writing in Plain English" (2007).

Mida meil on eesti keele kohta taoliselt üldloetavalt vastu panna? Tingimisi ehk Martin Ehala ja Tiina Veismanni 2001. a ilmunud raamat "Noor keelekasutaja". Tõsisem koondlugu, mis oleks lihtne, põnev ja õhuke, on aga ikka kirjutamata.

Üldkeele sõnavaraline tase oleks määratud varem koostatud põhisõnavaraga, millele lisanduksid ühiskeelele omased ühesed võõrsõnad või laensõnad. Need looksid keskse tuuma ümber hägusa sõnapilve. Mis sellest kasu on?

Eesti lastele on see vajalik muu maailmaga lõimumiseks. Paar aastat tagasi tegi L. Võhandu arvutused, mis näitasid, et meie õpilane peab kogu kooliskäimise jooksul iga päev omandama keskmiselt 15 talle võõrast mõistet, võõrsõna ja võõrkeelset sõna. Kõik need sõnad vajavad memoreerimist, kordamist (efektiivne võõrkeeleõpe väidab, et uut sõna saab vabalt kasutada alles pärast 50-kordset kordamist). On päris ilmne, et selline omandamiskoormus on üpris suur. Võõramaalastele, kes eesti keelt õpivad, on see hägus sõnavaraline lisakiht vastupidiselt suhteliselt kergesti õpitav, sest mõisted on juba tuttavad. Nende õppekiirus kasvaks kindlasti märgatavalt.

[6] Vt http://www.keelevara.ee/teosed/ (28.12.2008).

Alles sellise filoloogide ja pedagoogide poolt hoolikalt läbi vaadatud ja heaks kiidetud üheselt määratud põhisõnavara abil saaks hakata oskuskeelte sõnavarasid korrektsemalt koostama, uurima ja ühestama. Mitmes oskuskeele komisjonis osalenuna võin[7] täie tõsidusega väita, et semantiline ühestamine pole sugugi triviaalne probleem.

5. Kokkuvõte

Kuigi silpidest koosneva statistiline keelemudeli loomine ning eesti keele tükeldamine allkeelteks tunduvad esmapilgul olevat täiesti erinevad uurimisvaldkonnad, siis on neil ka oluline ühisosa. Mõlema ülesande sisuks on keele tükeldamine mingil viisil: esimesel juhul tükeldamine silpideks eesmärgiga uurida silpide järgnevusi ning koostada nendel järgnevustel baseeruv keelemudel, teisel juhul tükeldamine allkeelteks eesmärgiga defineerida n-ö põhisõnavara ning erinevad oskussõnavarad.

Artiklis kirjeldasime silpidest koosnevad statistilise keelemudeli koostamist. Mudelisse sai kaasatud 500 sagedamini esinenud silpi, mis kattis 85% kogu korpusest. Varem oleme loonud sarnase mudeli ka 1000 silbi baasil, mis protsentuaalselt ei andnud olulist efekti. Mudeli hetkel oli silpide arvu piiravaks teguriks selle koostamiseks kasutatud arvutiprogrammi suur ressursitarve. Tulevikus on plaanis katsetada sellise mudeli loomist, mis sisaldaks peagu kõiki korpuses esinenud silpe, jättes välja ehk ainult need silbid, mis esinesid seal vaid ühe korra. Hinnanguliselt peaks selles mudelis silpide arv jääma 5000 ja 7000 vahele.

Kirjeldatud mudel on kolmetasandiline koosnedes silpide, silbipaaride ja silbikolmikute järgnevuse sagedustabelitest. On selge, et mida rohkem tasandeid mudelis on, seda täpsem ja adekvaatsem ta on. Artiklis kirjeldatakse eesti pseudokeele generaatorit, mis baseerub sellel kolmetasandilisel mudelil. Töö käigus sai pseudokeele generaatorit kasutatud muuhulgas ka selleks, et hinnata mudeli tasandite hulga piisavust. Ühetasandilist mudelit (silpide järgnevusi) kasutava pseudokeele generaatori väljund ei sarnanenud veel kuigivõrd eesti keelele, pigem oli tegemist üksteisele järgnevate silbijadadega, mis mõistlikke eestikeelseid sõnu ei moodustanud. Kahetasandilist mudelit kasutava pseudokeele generaatori väljund hakkas juba rohkem sarnanema eesti keelele, kuid ei tundunud siiski veel piisavalt hea. Antud hinnangud on loomulikult subjektiivsed, kuid kolmetasandilisele mudelile baseeruva generaatori väljund tundus piisavalt hea, et sellise tasandite arvuga piirduda. Edaspidi on kavas moodustada sarnane keelemudel ka kasutades (pseudo)morfeeme ning loodame, et just eesti pseudokeele generaatori väljund aitab hinnata, millisteks algosakesteks on taolise mudeli loomise puhul mõistlikum eesti keelt tükeldada.

Silpide kategooriatesse jagamisel oli hüpoteesiks, et eristuvad mingid konkreetsed silpide hulgad, mis esinevad sõnades ainult teatud positsioonil. Kuna selle ülesande lahendamisel aluseks olnud andmetabel (silpide järgnevuse sagedustabel) on oma mõõtmetelt liiga suur, et visuaalse vaatluse abil mingisuguseid järeldusi teha silpide grupeeruvuse kohta, siis tuletasime lihtsad valemid, mis, rakendatuna sagedustabelile, annavad hinnangu, millises positsioonis võib iga silp sõnades esineda. Selles artiklis on esitatud vaid väikese hulga silpide kategooriatesse jaotamine algoritmi näitlikustamiseks. Kõigi 500 silbi kategoriseerimise tulemused on

[7] Leo Võhandu (toimetaja märkus).

K. Sirtsu magistritöös (2008). Selgus, et vaid väike hulk eesti keeles esinevaid silpe on sellised, mis esinevad sõnades mingil kindlal positsioonil (kas ainult alguses või ainult lõpus). Enamik silpe on paraku universaalsed, mis võivad esineda erinevates sõnades erinevatel positsioonidel. Seega oleks tulevikus tarvis uurida muid algoritme, kuidas silpe väiksematesse ja hoomatavamatesse rühmadesse grupeerida ning muid tunnuseid, mille alusel seda teha.

Viimases alajaotuses arendatakse sissejuhatuses tehtud tähelepanekut, et eesti keele jaoks mingit standardselt esinduslikku keelekorpust ei eksisteerigi. Kõigepealt viidatakse peagu sajandivanustele J. Linzbachi mõtetele formaalsete reeglite kohaselt toimivate keelte terve kogumi kohta. Igal juhul on meil tegu sisulise paljukeelsusega. Teades, et võrguvarana on olemas mitmeid eesti keele sõnastikke, pakutakse välja üks suhteliselt lihtne tee eesti ühiskeele põhisõnavara eraldamiseks ja korrektseks korrastamiseks. Loodud baasile saab mitmeti ehitada konkreetsete allkeelte erisõnastikke.

Viidatud kirjandus

Ehala, Martin; Veismann, Tiina 2001. Noor keelekasutaja. Tallinn: Künnimees OÜ.

Erelt, Mati; Hennoste, Tiit 2002. Vaja on veel üht eesti keele grammatikat. – Renate Pajusalu, Tiit Hennoste (toim.). Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Tartu: TÜ Kirjastus, 126–131.

Erelt, Tiiu 1982. Eesti oskuskeel. Tallinn: Valgus.

Haagensen, Michael 2007. Writing in Plain English. Tallinn: Koolibri.

Hennoste, Tiit 2000. Allkeeled. – Hennoste, Tiit (toim.). Eesti keele allkeeled. Tartu Ülikooli eesti keele õppetooli toimetised 16. Tartu: TÜ Kirjastus, 9–56.

Hennoste, Tiit 2002. Keelekasutuse uurimine. – Emakeele Seltsi aastaraamat, 48 (2001), 217–262.

Hennoste, Tiit; Muischnek, Kadri 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Tiit Hennoste (toim.). Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: TÜ Kirjastus, 183–217.

Kerge, Krista 2000. Kirjakeel ja igapäevakeel. – Tiit Hennoste (toim.). Eesti keele allkeeled. Tartu Ülikooli eesti keele õppetooli toimetised 16. Tartu: TÜ Kirjastus, 75–110.

Kull, Rein 2000. Kirjakeel, oskuskeel, üldkeel. Tallinn: Eesti Keele Sihtasutus.

Linzbach, Jacob 1916. Printsipõ filosofskago jazõka. Opõt totšnago jazõkoznanija. Petrograd.

Luuk, Erkki 2008. Semantilised tasandid ja semantilised primitiivid. – Keel ja Kirjandus, 12, 949–967.

Pajusalu, Renate; Tragel, Ilona; Veismann, Ann; Vija, Maigi 2004. Tuumsõnade semantikat ja pragmaatikat. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 5. Tartu: Tartu Ülikooli kirjastus.

Sirts, Kairit 2008. Eesti keele silbisüsteemi uurimine J-keele vahenditega. Magistritöö. Tallinn: Tallinna Tehnikaülikool.

Võhandu, Leo; Sirts, Kairit; Aab, Eiki 2008. Eesti silbisüsteemi struktuurist. – Eesti Rakenduslingvistika Ühingu aastaraamat, 4, 263–269.

Kaudviited

Pajusalu, Karl 1992. Regional and Social Varieties of Estonian. – Ural-Altaische Jahrbücher. Ural-Altaic Yearbook, 64, 23–34.
Rätsep, Huno 1976. Lindu tuntakse laulust, inimest keelest. – Keel, mida me harime. Tallinn: Valgus, 116–120.

Võrgumaterjalid

Eesti keele korpused. http://www.cl.ut.ee/korpused/index.php?lang=et (28.12.2008).
Eesti Kirjakeele Korpus 1890-1990. http://www.cl.ut.ee/korpused/baaskorpus/ (28.12.2008).
Ajakirjandustekstid 1999. http://www.cl.ut.ee/korpused/baaskorpus/txt/1999aja_txt_elan.zip (28.12.2008).
EKI silbitamise tarkvara. http://www.eki.ee/tarkvara/silbitus/ (29.12.2008).
J programmeerimiskeel. http://www.jsoftware.com/ (28.12.2008).
Elektroonilised eesti keele sõnaraamatud. http://www.keelevara.ee/teosed/ (29.12.2008).

Kairit Sirts (Tallinna Tehnikaülikool). Uurimisteemaks on statistiline keelemudel ja selle rakendused.
kairit.sirts@hot.ee

Leo Võhandu (Tallinna Tehnikaülikool) uurimisvaldkonnad on andmeanalüüs, keerukate andmekogumite peidetud struktuuri avamine, graafiteooria.
leovoo@hot.ee

CUTTING THE TEXT CORPORA: APPLICATIONS WITH SYLLABLES AND SUB-LANGUAGES

Kairit Sirts, Leo Võhandu
Tallinn University of Technology

In this paper we study different aspects of language by using different cuts of language corpora. There are two particular cuts under observation, which are very different by their nature: mincing the text into syllables for developing a statistical language model and dividing the language into sub-languages for identifying the base vocabulary.

Our syllable based statistical language model includes the 500 most frequently observed syllables. It is a three-level model consisting of frequency tables for syllables, syllable pairs and syllable triplets. A frequency table is a matrix with syllables, syllable pairs or syllable triplets in rows and syllables in columns. The numbers in matrix cells show how many times the syllable in the column happened to follow the element in the row.

The Estonian pseudo language generator is an application of the syllable based statistical language model. Using the Estonian pseudo language generator it is possible to generate a text which is not fully Estonian, but definitely sounds like one.

The purpose of categorizing syllables is to assort the syllables according to their possible locations in a word. We propose an algorithm for automatic syllable grouping using the data in the syllable frequency table. We show experimentally how syllables are grouped into word-initial, word-internal and word-final syllables.

Language can be divided into general language using a base vocabulary and different sub-languages, which contain particular terminology. In this paper we discuss the definition of general language. We also propose an automatic algorithm for defining its base vocabulary.

Keywords: computational linguistics, syllabification, syllable association, graph representation, language model, syllable grouping, general language, sub-languages, Estonian