EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 4, 263–269

EESTI SILBISÜSTEEMI STRUKTUURIST

Leo Võhandu, Kairit Sirts, Eik Aab

Ülevaade. Artiklis uurime eesti keele silpide sagedusmaatriksit, mis on loodud eesti kirjakeele korpuse baasil, kasutades ilukirjandustekste aastatest 1988–1998. Vaatluse all on 1000 kõige sagedamini esinevat silpi. Eksperimentaalsete uurimismeetoditena kasutasime Hamiltoni tee ning sotsiaalse võrgustiku klastrite leidmist eesmärgiga uurida silbisüsteemi sisemist varjatud struktuuri. Hamiltoni tee leidmiseks vajaliku programmi loomiseks kasutasime programmeerimiskeelt J-6.01. Programm ühendab kõik silbid järjestikusse maksimaalse ühenduste summaga jadasse ilma silpe kordamata. Teise meetodina leidsime sotsiaalse võrgustiku klastrid 100 ja 1000 kõige sagedamini esineva silbi jaoks. Selles lühikeses artiklis esitame graafiliselt ainult 100 silbi võrgustiku ilma täpsemate selgitusteta selle leidmise matemaatilise meetodi või esituse interpretatsiooni kohta.

Võtmesõnad: arvutilingvistika, silbitamine, silbiseostus, graafesitus, Hamiltoni tee, silpide sotsiaalne võrgustik, eesti keel

1. Sissejuhatus

Raallingvistika areng näitab üha veenvamalt, et mida rohkem me suudame tekstikäsitlust automatiseerida, seda kasulikumad ja põnevamad on tulemused. Seejuures on oluline, et me peame üritama võimalikult vähe kasutada üldlingvistilisi eelteadmisi. Miks? Nõnda toimides me jälgime teaduse üldist arengujoont. Formaalsete meetodite õitseaeg sai alguse reaalse maailma suhteliselt lihtsalt kirjeldatavate või modelleeritavate objektsüsteemide uurimisest (nt mehaanika, optika). Tekkinud probleemide lahendamiseks loodi tasapisi rida meetodeid, mis töötasid nn polünomiaalses ajas. Samade meetodite ülekanne hägusat või diskreetset modelleerimist nõudvatele objektsüsteemidele ei õnnestunud aga kuigi hästi. Peamiseks põhjuseks oli see, et hägusad süsteemid on põhiliselt iseorganiseeruvad, mitte kindlate reeglite järgi konstrueeritavad. Piltlikult öeldes, Loojal ei olnud olemas formaalset süsteemi (matemaatikat), mille abil ta oleks saanud deterministlikku maailma ehitada.

Kellelgi meist ei ole valemit otsa ees ega südamesse peidetud, mille kohaselt me peaksime käituma.

Iseorganiseeruvuse idee võimaldas käsitleda protsesse ja nähtusi, milles nn avatud süsteemi sisemise organisatsiooni keerukus kasvab, ilma et seda välismõjud otseselt juhiksid või haldaksid. Termini *iseorganiseeruvus* võttis ilmsi kasutusele 1947. a psühhiaater ja insener W. Ross Ashby. Laiemalt tuli see sõna käibele 1948. a Norbert Wieneri küberneetikaraamatu kaudu (eesti keeles Wiener 1961). Asjast huvitusid ka filosoofid ja nii avaldas Noam Chomsky ning Hilary Putnami õpetaja Nelson Goodman 1951. a raamatu "The Structure of Appearance" ("Nähtumuse struktuur"). Matemaatika ju sellega tegelebki, et leida nähtumust kirjeldavaid või peidetud struktuure. Et selline lähenemine viljakas on, näitab kasvõi akadeemik Jaan Einasto grupi töö tumeda aine uurimisel astronoomias (Näha pole ju midagi!).

Nendest ideedest lähtudes üritamegi tungida eestikeelse teksti kui nähtuse peidetud struktuuri, seejuures võimalikult iseorganiseeruvuse ideed kasutades. Loomulikult jääb see ekskursioon pinnapealseks, aga ehk leidub järeltegijaid ja ümbermõtlejaid.

2. Andmestik

Võtsime Tartu Ülikoolis loodud kirjakeele korpuse 1988–1998 ilukirjandusosa (vt ka Hennoste jt 2001).[1] Kasutades Eesti Keele Instituudi silbitajat[2] ja J-keelt[3] koostasime kõigi silbipaaride jaoks järgnevuste sagedustabeli. Enne silbitamist teisendasime teksti 8-bitilisse UTF-formaati.[4] Täpsemaks analüüsimiseks võtsime 1000 sagedasemat silpi pluss tühiku iseseisva üksusena. Saadud sagedustabeli veerus on reas esitatud silbile järgnenud silp. Töötlustarkvara loomiseks kasutasime J-keele versiooni 6.01. Et kogu sagedustabeli esitamiseks selles lühikäsitluses ei ole ruumi, siis esitame tervest 1001x1001 sagedustabelist vaid selle vasakpoolse ülemise 10x10 nurga (kõige tihedama osa tabelist).

Tabel 1. Silpide sagedustabel

	le	*ta*	*ma*	*ja*	*se*	*da*	*o*	*li*	*ga*	*te*
le	535	545	1024	98	24	125	96	72	707	191
ta	37	146	1438	582	62	1599	0	76	1750	129
ma	719	1179	37	604	115	145	0	214	608	207
ja	343	174	119	23	17	71	0	176	262	330
se	1419	179	254	23	15	2713	18	113	1392	167
da	598	346	407	93	23	402	0	22	93	93
o	2709	3	4261	21	10	41	0	8661	16	13
li	147	266	190	58	858	76	33	56	106	138
ga	92	274	181	38	26	86	0	17	37	20
te	1011	112	2628	1	8	1272	28	77	779	52

Tabelit 1 vaadeldes on näha, et esineb rida loomulikke järgnevuspaare (paarile järgnev arv näitab paari esinemissagedust):

[1] Vt http://www.cl.ut.ee/korpused/baaskorpus/ (10.10.2007).
[2] Vt http://www.eki.ee/tarkvara/silbitus/ (10.10.2007).
[3] J-keel = Jsoftware, vt http://www.jsoftware.com/stable.htm (10.10.2007).
[4] Vt http://www.chmaas.handshake.de/delphi/freeware/xvi32/xvi32.htm (10.10.2007).

o-li 8661, *se-da* 2713, *o-le* 2709, *te-ma* 2628, *ta-ga* 1750, *ta-da* 1599, *ta-ma* 1438, *se-le* 1419 jne.

Mida taoliste tabelitega teha? (Eriti, kui need tabelid nii suured on!)

Andmeanalüüsi üldteoorias on selliste maatriksite korrastamiseks mitmeid meetodeid. Enamasti baseeruvad need meetodid mingite seoste või mõõtude maksimeerimisel.

3. Silpide järjestamisest. Hamiltoni tee

Kõige esimene meetod on mõtteliselt lihtne ja üritab silbid niimoodi järjestada, et nad kõik oleksid üksteisele järgnevad ja iga silbipaari koosesinemiste arvude summa üle kogu uue järjestuse oleks maksimaalne. Seda kõiki silpe ühendavat optimaalset sidusat teed nimetatakse *Hamiltoni teeks*. Graafiteoorias ja diskreetse optimeerimise valdkonnas on Hamiltoni tee leidmine hästituntud ülesanne ja kuulub nn NP-keerukate probleemide hulka. Mida see tähendab? N objektist koosneval süsteemil võib kõiki objekte paigutada N! erinevasse järjestusse. Kui väikese objektide arvu korral on nende järjestuste leidmine lihtne töö (nt kolme objekti a, b, c korral on järjestusi 6: abc, acb, bac, bca, cab, cba), siis juba N=10 korral on järjestusi 10!=3628800. Kui oletada, et iga järjestuse ja selle naaberseoste sageduste summa leidmiseks kulub ainult 1 sekund, siis kümne objekti kõigi Hamiltoni teede leidmiseks ja hindamiseks kuluks ei rohkem ega vähem kui 42 ööpäeva!

Selgub, et näiteks N=20 puhul tuleks kõikide variantide läbimise korral tööd teha $7{,}7.10^{10}$ aastat. Tööaeg on nii suur, et tulemus ei huvitakski enam kedagi. (Vahemärkusena lisame, et J. Einasto andmetel kõrvetab Päike meid kõiki juba umbes miljardi aasta pärast olematusse ja Maa ise hävib *ca* 5 miljardi aasta pärast.) Nüüd peaks selge olema, et 1000 objekti korral pole mingit lootustki täpse lahendi leidmiseks. Arvutiteaduses on rida ligikaudseid meetodeid, mis päris korralikke lähendeid annavad.

Me kasutame ühte J-keelset programmi, mis võimaldab Hamiltoni teed leida ka üpris suurte maatriksite korral. Selle keele plussiks on ülilühikesed programmid (mis ei tarvitse küll kergesti arusaadavad olla). Kui meie silpide sagedustabeli nimeks on *mm*, siis vastav Hamiltoni tee programm, mis stardib i-ndast silbist, läbib kordusteta kõik ülejäänud silbid ja trükib selle jada ka välja, ei ole kuigi pikk:

```
imax=. i.>./
ntg=. dyad define
imax 0(}:x)}({:x){y
)
hp=.]`($:@(],ntg&mm))@.((#mm)&~:@#)
```

Näiteks hp 6 annab tulemuseks silbijada

o li se da le ma ta ga ja te

Toome näitena veel mõned silbijadad:

Ja te ma ta ga se da o li
Te ma ja le o li se da ta ga

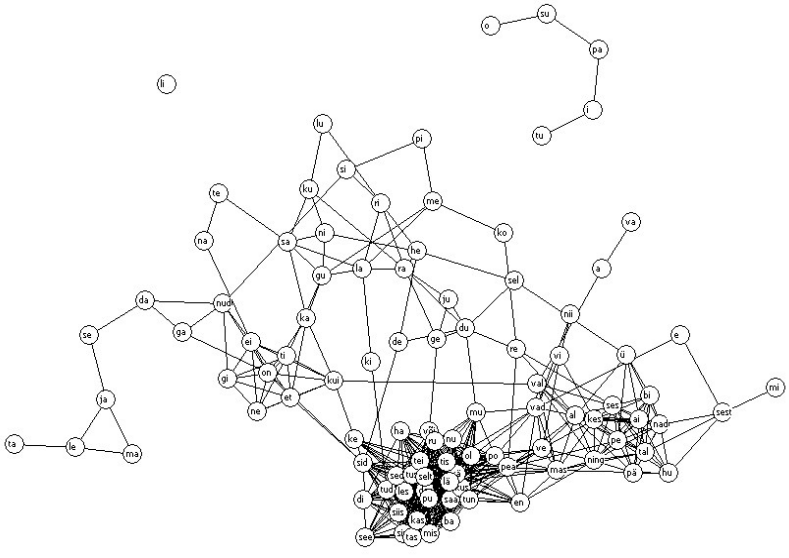
Näeme, et silpide seostus on vägagi suupärane. Huvitav on neid "jorusid" võrrelda üle 40 aasta tagasi Leo Võhandu informatsiooniteooria seminaris Tartu Ülikoolis Priit Järve (1965: 97–103) poolt saadud tulemustega eesti keele fraaside entroopiamudelite kohta. Kahe- ja kolmetäheliste ühendite puhul saadud fraasid olid järgmised:

> *Siserdamega mibul ta lid neiijud ka onndagine*
> *Tas peanud lestemate kõnniigla kaltsed aseg*

Ilmne on, et kahetäheliste silpide ühenduvus on parem 3-grammidest saadud tulemusest. Oleks põnev, kui mõni üliõpilane otsiks P. Järve artikli välja ja üritaks silpidega teha sama testi piisavalt mahuka tekstikorpuse peal.

4. Silbid süsteemina

Eelmises jaotuses vaadeldud Hamiltoni tee kõrval on veel terve rida meetodeid, mis silbisüsteemi sisemist peidetud struktuuri avada aitaksid. Me vaatleme siin vaid ühte Eik Aabi poolt loodud silpide sotsiaalvõrgulise sõprusmudeli tulemuspilti. Ideeks on valida antud silbile lähim naaber mitte paaritise koosesinemise sageduse, vaid 100- või 1000-silbilise koosesinemise alusel. Nii saadud sarnasusmõõt näitab, kuivõrd sarnaselt kaks vaadeldavat silpi käituvad ülejäänud sagedaste silpide suhtes. Teisisõnu, me otsime silpe, mille käitumismuster silpide kogumis on "globaalselt" analoogiline. Iga silp otsib teiste silpide hulgast endale "iidolit". Iga silbi jaoks leitakse efektiivsus e=R/E, kus R on sõprade arv ja E on nendele kokku kuluv energia (kauguste summa). Iidoliks valitakse sõprade hulgast need, kellel on suurem efektiivsus (ehk tihedus tema ümber on suurem). Algoritmi täielik kirjeldamine ei oleks siinkohal sobilik. Küll aga esitame 100 sagedasema silbi graafilise pildi.



Joonis 1. Silpide sotsiaalne võrgustik

5. Mis on sellest kõigest kasu?

Silbistruktuuri uurimine on alles algstaadiumis, seepärast on praegu raske kõiki võimalikke rakenduskohti välja tuua. Nimetame vaid mõned:

- Markovi peitmudelite õppematerjali minimeerimine;
- kõrge katteväärtusega lihttekstide moodustamine. Kasulik oleks see näiteks meie keele õpetamisel võõramaalastele;
- keel kui "Small World" mudelite klassi kuuluv süsteem;
- vanade tekstide silbistruktuuri uurimisel saab infot nii fonotaktika kui silbitaktika muutumise trendidest;
- teades sagedusseostuste süsteemi nihkeid, saab ennustada edasisi muutusi (nn iteratiivne ennustamine) jne jne.

6. Mida edasi teha?

Juba mainitud teemade uurimise kõrval tasub ilmselt tõsist tähelepanu osutada ka morfeemidele. Soomes loodi huvitav süsteem Morfessor morfeemide automaatseks eraldamiseks, et vältida kõnetuvastuses sõnastikuväliste sõnade suurt hulka (Siivola jt 2003). Soome keele jaoks töötas see süsteem efektiivselt. Türgi ja suahiili keele puhul ei olnud tulemused nii head.

Eks me teksti saa tükeldada just nende kolme struktuurielemendi – sõnade, silpide ja morfeemide abil. Milline nendest meie käsitluse jaoks sobivaim on, seda näitab tulevik.

Kirjandus

Hennoste, Tiit; Kaalep, Heiki-Jaan; Muischnek, Kadri; Paldre, Leho; Vaino, Tarmo 2001. The Tartu University Corpus of Estonian Literary Language. – Tõnu Seilenthal, Anu Nurk, Triinu Palo (eds). Congressus Nonus Fenno-Ugristarum, Pars IV, Dissertationes sectionum: Linguistica. I. Tartu, 337–344.

Järve, Priit 1965. Eesti keele kirjapildi entroopiast ja liiasusest. – Keel ja struktuur 1. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu, 97–103.

Siivola, Vesa; Hirsimäki, Teemu; Creutz, Mathias; Kurimo, Mikko 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. – Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003). Geneva, Switzerland, 2293–2296.

Wiener, Norbert 1961. Küberneetika. Tallinn: Valgus.

Korpused ja tarkvara

Eesti kirjakeele korpus 1890–1990. http://www.cl.ut.ee/korpused/baaskorpus/ (10.10.2007).

Eesti Keele Instituudi tarkvara. Silbitus. http://www.eki.ee/tarkvara/silbitus/ (10.10.2007).

J-keel = Jsoftware, http://www.jsoftware.com/stable.htm (10.10.2007).

Freeware Hex Editor XV132. Version 2.51. http://www.chmaas.handshake.de/delphi/freeware/xvi32/xvi32.htm (10.10.2007).

Leo Võhandu on lõpetanud Tartu Ülikooli matemaatikaosakonna, praegu Tallinna Tehnikaülikooli emeriitprofessor. Uurimisvaldkonnad: andmeanalüüs, keerukate andmekogumite peidetud struktuuri avamine, graafiteooria.
leovoo@hot.ee

Kairit Sirts (Tallinna Tehnikaülikool) on informaatika eriala magistrant. Uurib eestikeelse silbisüsteemi peidetud formaalset struktuuri.
karambula@hot.ee

Eik Aab (Tallinna Tehnikaülikool) on informaatika eriala doktorant, teemaks iseorganiseeruvate sotsiaalsüsteemide uurimismeetodid.
eik.aab@innofusion.ee

A PRELIMINARY STRUCTURAL VIEW OF THE ESTONIAN SYLLABLE SYSTEM

Leo Võhandu, Kairit Sirts, Eik Aab
Tallinn University of Technology

Using the Corpus of Estonian Literary Language and specifically the selection of fiction texts from years 1988–1998 we have studied the frequency table of Estonian syllables (1000 most frequent syllables). As an experimental study we have used Hamilton Path (HP) and Social Network Clusters to study the inner structure of syllable system. For HP we have created and represented a program in J-6.01 which connects all syllables into a sequential path with the maximal connections sum and without a syllable repetition. As another method we have created SNC networks for 100 and 1000 most frequent syllables. In this short article we have presented graphically only the 100 syllable network without explaining exactly either the mathematical method or networks interpretation.

Keywords: computational linguistics, syllabification, syllable association, graph representation, Hamilton path, syllable social network, Estonian