Vale meetodiga õiged tulemused? Eesti keele morfoloogiline ühestamine statistika abil

Heiki-Jaan Kaalep, Tarmo Vaino

Kokkuvõte

Artikkel kirjeldab esimest katset kasutada eesti keele morfoloogilise mitmesuse vähendamiseks statistilist meetodit, varjatud Markovi mudelit (VMM), mille sobivus eesti keele jaoks on a priori kaheldav. Antakse lühiülevaade meetodist, kirjeldatakse treenimist ja testimist ning tulemusi.

1. Sissejuhatus

Morfoloogiline ühestamine seisneb morfoloogiliselt analüüsitud lause igale sõnale tema võimalike morfoloogiliste märgendite hulgast õige valimises. Näiteks morfoloogiliselt analüüsitud lausest:

```
Mees
   mees+0 //_S_ sg n, //
   mesi+s //_S_ sg in, //
peeti
   peet+0 //_S_ adt, sg p, //
   pida+ti //_V_ ti, //
kinni
   kinni+0 //_D_ //
```

saame peale ühestamist:

```
Mees
   mees+0 //_S_ sg n, //
peeti
   pida+ti //_V_ ti, //
kinni
   kinni+0 //_D_ //
```

Erinevad ühestamisalgoritmid lähtuvad teatavatest ühistest eeldustest [Merialdo 1994, lk 156]:

1. Igale sõnale sobib ainult teatav väike hulk morfoloogilisi märgendeid kõigi võimalike morfoloogiliste märgendite hulgast. Tüüpiliselt leitakse see morfoloogilise analüsaatori abil.
2. Kui sõnal lauses on mitu võimalikku morfoloogilist märgendit, siis teatavate lokaalset konteksti kasutavate reeglite põhjal on võimalik määratleda iga sõna jaoks korrektne märgend.

Oma töös kirjeldame algul ühte konkreetset ühestamisel kasutatavat mudelit – Varjatud Markovi Mudelit – mis tugineb tekstide põhjal tehtud statistikale ja peaaegu ei kasuta lingvistile intuitiivselt arusaadavaid reegleid sobivate märgendite valimisel. Hoolimata sellest,

et antud mudel kasutab traditsioonilisest keeleteadusest väga kaugeid meetodeid, annab ta paljude keelte peal üllatavalt häid tulemusi.

Me kirjeldame ka konkreetset katset eesti keele statistilisel ühestamisel. See hõlmab sobiva märgendite süsteemi valikut, ühestaja treenimist treeningkorpuse peal ja testimist testkorpuse peal.

Eesti keele tüpoloogilisest eripärast tulenevalt olid meil enne katset suured kahtlused, et Varjatud Markovi Mudel ei ole eesti keele peal hästi rakendatav, kuid katse tulemusena on need kahtlused peaaegu kadunud.

2. Ühestamisel kasutatud mudel

Ühestajates kasutatakse põhiliselt kaht tüüpi algoritme:
1. reeglitel põhinevaid [Brill 1992, Karlsson 1990],
2. stohhastilisi [Merialdo 1991, Cutting et al. 1992].

Mõnevõrra on kasutatud ka muid meetodeid, näiteks neurovõrke [Benello, Mackie ja Anderson 1989].

Reeglitel põhineva ühestaja korral formuleerib lingvist teatava hulga kontekstil põhinevaid reegleid (näiteks: verbile ei saa järgneda tagasõna). Ühestamiseks rakendatakse neid (sõltuvalt mudeli iseärasustest) suvalises või siis mingis kindlas järjekorras. See meetod eelistab mitte liiga suurt märgendite hulka. Suur märgendite hulk nõuab palju reegleid ja nende omavaheline kombinatoorika muudab reeglite koostamise keerukaks [Dermatas ja Kokkinakis 1995].

Statistilise ühestamismeetodi korral konstrueerikse esmalt ühestamisel kasutatavad tõenäosuste tabelid.
On meetodeid, mis tabelite konstrueerimiseks vajavad eelnevalt käsitsi ühestatud tekste. On ka meetodeid, mis konstrueerivad tabeleid ühestamata tekstidest lähtudes, iteratiivselt treeningtekste ühestades ja nende pealt tabeleid koostades. Meie kasutasime meetodit, mis vajab eelnevalt käsitsi ühestatud teksti -Varjatud Markovi Mudeli (VMM), ingl. k. Hidden Markov Model (HMM) nime all tuntud tõenäosuslikku mudelit, mille kohta vt. nt. [Weischedel et al 1993], [Armstrong et al 1996], [Dermatas ja Kokkinakis 1995]. Meie rakendasime VMMi tema puhtal klassikalisel kujul, mille puhul eeldatakse järgmist.

1. Lauset ei vaadelda kui sõnade järjestust, vaid kui mingite märgendite (M) järjestust. Märgendid võivad olla morfoloogilised, spetsiaalselt ühestamiseks kasutatavad, süntaktilised vm.
2. Kuna sõnal võib olla mitu M, siis konkreetsele lausele võib vastata mitu võimalikku Mide järjestust, aga ainult üks neist on õige.
3. Mõned järjestused on antud keeles tüüpilised, mõned mitte.
4. Võimalikest järjestustest tuleb valida kõige tüüpilisem, s.o. kõige tõenäolisem. See ongi antud lause puhul õige.
5. Iga konkreetse sõna puhul lauses tuleb tema võimalike Mide hulgast valida selline M, et lause Mide järjestus oleks kõige tõenäolisem.
6. Terve lause Mide järjestuse arvutamisel lähtutakse üksikute Mide tõenäosustest.
7. Üksiku sõna tõenäosus antud lauses ei sõltu kõigist teistest selle lause sõnadest, vaid ainult ühest (harvem kahest) eelnevast Mist. Seda nimetatakse Markovi sõltumatuse eelduseks.

Mitmed nendest eeldustest võivad tunduda mittepõhjendatud. Nt. eeldus nr. 2 ei kehti alati, st. lause (aga võibolla ka terve teksti konteksti) arvestades ei ole võimalik alati eelistada mingi sõna võimalike märgendite hulgast ühte. Näitena tegelikust elust võib tuua ühe ettekande pealkirja:
Jälgi koolimatemaatika arenguteel

Loomulik on kahtlus, et mudelit, mis lähtub eespool toodud eeldustest, ei saa eesti keele peal üldse kasutada. VMM arvestab ju ainult sõnade (õigemini märgendite) järjekorda ja äärmiselt piiratud konteksti.

3. Märgendid

Kuna VMM "näeb" ainult märgendeid ja nende tõenäosusi, siis märgendite süsteemi valik on peamine, mis eristab head VMM-ühestajat halvast. Samas ei ole olemas häid eeskirju, kuidas märgendite süsteemi teha; see on niivõrd keelespetsiifiline.

Märgendid, mida kasutab oma töös ühestaja (ÜM), ei pruugi olla samad mis morfoloogilise analüsaatori poolt sõnadele omistatavad märgendid (MM). ÜM on puhtalt ühestaja sisemine asi; nii talle sisendina antav tekst kui ka väljundina saadav tekst on ikka esialgsete, morfoloogiliste märgenditega. On võimalik, et eri MMidega sõnad esinevad ühestamise seisukohalt sarnastes kontekstides, nagu ka see, et sama MMiga sõnad erinevates kontekstides. Seega ühestamisel on sageli mõttekas mõned morfoloogilised klassid ühendada, mõned aga mitmeks lahutada. Nt. tavalised nimisõnad ja pärisnimed liita üheks klassiks, asesõnade puhul aga eristada isikulisi asesõnu teistest.

ÜM valimisel lähtutakse järgmistest nõudmistest.

1. ÜM peavad esindama süntaktiliselt selgelt eristuvaid klasse, s.t nad peavad oma kontekstis olema selgelt eristatavad. Oletame näiteks et meil on kaks ÜM, A ja B. A esineb tüüpiliselt kontekstis X, B aga kontekstis Y. Kui meil on nüüd tekstis sõna, mis oma kujult võib kuuluda nii tüüpi A kui B, siis konteksti X puhul saame ta ühestada tüüpi A, konteksti Y puhul aga tüüpi B. Kui kontekstid ei ole selgelt eristatavad (nt. A esineb vahetevahel ka kontekstis Y), siis konteksti põhjal märgendi valimine annab osal juhtudest vale tulemuse.
2. ÜM klassid peaksid olema küllalt suured, et statistika nende kontekstide suhtes oleks usaldusväärne. Samas on reaalsus, et tekstide maht, mida ühestamisalgoritmide koostamisel saab kasutada, on piiratud - raha, aja ja tegijate puudusega. Nii et meil on fikseeritud suurusega tekstimaterjal, milles esinevad sõnad on jaotatud mingi hulga ÜM klasside vahel. ÜM klasside suuruse nõue on samaväärne nõudega, et ÜM klasside arv oleks teatud piirides. Paljude keelte puhul peetakse õigustatuks ÜM arvu alla 100, ehkki näiteks rootsi keele puhul kasutatakse 180 märgendit [Källgren 1996].
3. ÜM tuleks valida nii, et ühestamisest oleks ikka tõesti kasu. Nt. tuleks panna nimetavas ja omastavas käändes sõnad eri klassidesse, sest nad on sageli vormilt homonüümsed ja neid saabki eristada ainult konteksti alusel.

Toodud kaalutlused on omavahel vastuolus.

Vastuolu esimese ja teise nõude vahel:
Esimene kaalutlus eeldab suurt ÜM hulka, teine nõuab aga väikest, st et harva esinevad morfoloogilised klassid tuleb kokku võtta suurematekslassideks või suuremate klassidega, mistõttu klasside süntaktiline iseloom "hägustub".

Vastuolu esimese ja kolmanda nõude vahel:
Lingvistiliselt pakub kindlasti huvi eristada nimetavat, omastavat ja osastavat käänet, kuid nende kontekst ei ole väga selgelt eristuv. Kui need klassid kokku võtta, saame lihtsamini ühestada, aga see oleks pigem "hägustamine".

Vastuolu teise ja kolmanda nõude vahel:
Kui ajada klassid väga suureks, nt. eristades ainult noomeneid ja verbe, siis saame küll usaldusväärse statistika ja head tulemused nt. verbide ja noomenite vahelise mitmesuse likvideerimisel, aga paljud lingvistiliselt olulised mitmesused jääksid lahenduseta.

Praegu kasutame 88 ÜM, mis on valitud järgmiselt. Eristatakse omadussõnu, põhiarvsõnu, järgarvsõnu, nimisõnu, pärisnimesid, isikulisi asesõnu, muid asesõnu, lühendeid, verbe, alistavaid ja rinnastavaid sidesõnu, hüüdsõnu, ees- ja tagasõnu, määrsõnu, punktuatsioonisümboleid ja tundmatuid sõnu.

Käändsõnade puhul eristatakse 5 käänet: nimetavat, omastavat , osastavat , lühikest sisseütlevat e. aditiivi ja "kõiki muid". Isikuliste asesõnade puhul eristatakse lisaks ka kolme isikut. Ei eristata ainsust ja mitmust.

Verbide puhul eristatakse kokku 13 ÜM-i: "ei", "ära", esimene pööre, teine pööre, kolmas pööre, kaudne kõneviis, "pole" ja "polnud", da-infinitiiv, 0-lõpuline vorm, tingiva kõneviisi vormid, käskiva kõneviisi vormid, ma-infinitiivi vormid, partitsiibid. Ei eristata ainsust ja mitmust ega aega.

4. Treenimine

Treeningkorpuseks oli G. Orwelli "1984" eestikeelne tõlge [Orwell 1990], v.a. Lisa, kokku 75 000 sõna. Kogu tekst oli morfoloogiliselt analüüsitud, seejärel ühestatud T. Puolakaise piirangut grammatikal põhineva ühestajaga [Puolakainen 1997] (mis jättis 16% sõnadest mitmeseks, kuid ühestas ligi 100%lise korrektsusega) ja lõpuks käsitsi kontrollitud ning ühestatud.

Treenimisel oli kaks faasi: esialgsete tabelite koostamine ja tabelite parandamine treeningkorpuse põhjal.
Treenimiseks kasutasime ISSCO ühestajat[Armstrong et al 1996], mis on vabalt kasutatav tarkvara ja mille saime ISSCO koduleheküljelt http://issco.unige.ch.

4.1. Esialgsete tabelite koostamine

Tabelite koostamiseks sai ühestaja 4 sisendit:
1. Morfoloogiliselt analüüsitud, kuid ühestamata tekst
2. Morfoloogiliselt analüüsitud ja ühestatud tekst
3. Ühestamisel kasutatavate märgendite (ÜM) loend
4. Teisendustabel morfoloogilistelt märgenditelt ÜM-idele

Mõlemal sisendtekstil olid märgitud lause algus ja lõpp ning kirjavahemärgid olid tõstetud sõnadest lahku.

Kõigepealt teisendati nii ühestatud kui ühestamata tekstis morfoloogilised märgendid ÜM-ideks. Morfoloogiliste märgendite hulka tähistame $M = \{m_1\ m_2 \ldots m_n\}$, kus m_i on üks ÜM. Seejärel vaadati mitteühese teksti kõiki sõnu ja koostati loend kõigist ÜM komplektidest, mis sõnadele olid omistatud. Neid ÜM komplekte nimetame mitmesusklassideks. Mitmesusklasside hulk $V = \{v_1\ v_2 \ldots v_q\}$ on loomulikult hulga M osahulkade hulk ($v_i \subseteq M$).

Ühestatud teksti põhjal arvutati:
1. Tõenäosuste vektor $E = \{e_1\ e_2 \ldots e_w\}$, kus e_i on tõenäosus, et m_i on lauses esimene morfoloogiline märgend.
2. Maatriks $P = \{p_{kl}\}$, kus p_{kl} on tõenäosus, et märgendile m_k eelneb märgend m_l.

Vaadeldes korraga nii ühestatud kui ühestamata teksti, koostati maatriks $X = \{x_{kl}\}$, kus x_{kl} on tõenäosus, et mitmesusklassi v_l kuuluvatest märgenditest tuleb valida märgend m_k

4.2. Tabelite parandamine

Järgmine samm oli treenimiseks kasutatud ühestamata teksti automaatne ühestamine, kasutades eelmisel sammul saadud tõenäosuste maatrikseid P, X ja vektorit E.

Kuidas see täpselt käib, on kirjeldatud lihtsustatud näite varal lisas 1.

Pärast treeningkorpuse automaatset ühestamist võrdlesime käsitsi ja automaatselt ühestatud tekste. Selgus, et 12,67 % sõnadest olid valede ÜMidega. Seejuures kõige sagedasemad vead olid nud- või tud-partitsiibi määramine verbi asemel omadussõnaks (24 % kõigist vigadest) ja "ei" määramine verbi asemel määrsõnaks (13%).

Nüüd algas tõenäosuste maatriksite kohandamine inimese poolt. Seda saaks teha nii, et kirjutada tõenäosusmaatriksitesse otse mingid muud numbrid sisse. ISSCO ühestaja puhul on siiski kaks mugavamat võimalust.

Esiteks on võimalik eraldi tabelisse kirjutades muuta sõnade kuuluvust ÜM klassi, ilma et peaks morfoloogilise analüsaatori väljundit teisendama. Nt. sõna "jooksul" esines käsitsi ühestatud tekstis 38 korda tagasõna funktsioonis ja 0 korda nimisõnana. Seega oleks põhjendatud tema määramine alati tagasõnaks; võimalik viga ka tundmatute tekstide ühestamisel on seejuures kaduvväike.

Muutsime järgmiste sõnade kuuluvust ÜM klassi:
jooksul, eest, ees, kohal, juures, abil, meelest, asemel, jaoks, vahele, peal, puhul, kõrval, hulgas, kallal, suhtes, vältel, kallale, ääres, saatel, kannul määrasime kuuluma ainult tagasõnaks, mitte kunagi nimisõnaks; *nagu, kui* määrasime kuuluma ainult alistavaks sidesõnaks, mitte kunagi määrsõnaks.

Teiseks on võimalik kaudselt muuta maatriksis P olevaid tõenäosusi. Ka selleks kasutatakse eraldi tabelit, kus on igal real on kaks ÜM-i ja + või – märgiga number, mis siis vastavalt suurendab või vähendab maatriksis P olevat tõenäosust, et selline ÜM-ide järgnevus on õige; tabelisse võib + või – märgiga numbri asemel ka otse tõenäosuse, s.t. arvu vahemikus 0 – 1.

Lisasime tabelisse read, mille mõte on järgmine:
Eessõnale ei saa järgneda verbi; verbile ei saa järgneda tagasõna; kui on valida nud-, tud-partitsiibi ja omadussõna vahel, eelista partitsiipi; kui "ei" puhul on valida määrsõna ja verbi

vahel, eelista kindlasti verbi; kui “ei” järel on verbivorm, eelista kindlasti seda mistahes muule tõlgendusele; omastavas käändes sõnale järgneb tagasõna, mitte määrsõna; osastavas ja omastavas käändes sõnale eelneb eessõna, mitte määrsõna.

Seejärel ühestasime uuesti treeningkorpust ning võrdlesime teda käsitsi ühestatud korpusega. Nüüd oli erinevusi 7,1%. Siinkohal katkestasime treenimise, et vaadata, kuidas tõlketeksti peal treenitud programm käitub originaalse ilukirjandusteksti peal.

5. Testimine

Testkorpuseks oli 2000 sõnaline väljavõte Vello Lattiku raamatust “Mihklipäeval. Mihklikuul” [Lattik 1983], mis oli kahe filoloogi poolt sõltumatult käsitsi ühestatud.

Testkorpuse ühestamise tulemused on toodud alljärgnevas kahes tabelis.

	Alguses	Pärast ühestamist
Keskmiselt tõlgendusi sõna kohta	1,72	1,02
Sõnu kokku	2005	2005
Morf. analüsaatori vigu	0,1%	0,1%
Mitteüheste sõnade protsent	42,49%	2,34%
Tõlgendusi kokku	3450	2052
Mitteüheseid sõnu	852	47

Tabelis toodud arvud käivad MM kohta, mitte ÜM kohta. See, et pärast ühestamist jääb osa sõnu mitmeseks, on seletatav sellega, et enne ühestamist võetakse mitu MM kokku üheks ÜMiks, kusjuures ühe sõna erinevad MMid teisenduvad tavaliselt siiski erinevateks ÜMideks. Kuid juhul, kui sõna erinevad MM teisenduvad üheks ÜMiks, MMide ühestamist ei toimugi. Nt. sõna *olema* on ainus verb, mille ainsuse ja mitmuse 3. pööre on homonüümsed – *on*. Kuna meie oma ÜMide valikul praegu ainsust ja mitmust ei erista, siis *on* jääb ühestamisest kõrvale.

Küsimus, kui korrektselt on arvuti ühestanud, ei ole nii lihtne kui esmapilgul paistab. Selgituseks on alljärgnev tabel, kus on kirjas, mitme % võrra eri tegijate poolt ühestatud tekstid üksteisest erinevad.

	arvuti – filoloog1	**arvuti – filoloog2**	**filoloog1 – filoloog2**
Ühestatud erinevalt	7,93%	8,68%	3,50%

Antud tabelis võib olla üllatuseks suur erinevus inimeste poolt ühestatud tekstide osas. Samas on see inimeste tehtud märgenduse mittekattuvus ligikaudu sama suur kui on kirjeldatud TREEBANK projekti puhul [Marcus et al 1990].

Võrdluse teiste keeltega, kus on kasutatud statistilisi meetodeid ühestamisel, näitab, et eesti keelele sobib antud meetod umbes sama hästi kui näiteks rootsi keelele, kus erinevus inimese ja arvuti poolt ühestamisel oli algul 7% [Källgren 1996].

Võrdlus hoopis teistel alustelt lähtuva ühestamismeetodiga – kitsenduste grammatikaga [Puolakainen 1997] näitab, et VMM esialgsed tulemused eesti keele peal ei jää peaaegu alla ühestajale, mis kasutab inimeste poolt formuleeritud reegleid.

6. Perspektiivid

Enne VMMi katsetamist eesti keele ühestamisel oli meil suuri kahtlusi, et eesti keele tüpoloogiast tingituna ei ole antud statistiline meetod eesti keele peal kasutatav. Katse lükkas meie kahtlused suures osas ümber ja annab aluse katsetada veelgi nii sama meetodiga kui teiste statistiliste meetoditega.

Me oletame, et VMMil põhinevat ühestajat saab eesti keele jaoks veelgi parandada, valides alljärgnevate võimaluste seast:

1. Valida võib-olla sobivam ÜMide süsteem
2. Täiustada tõenäosuslikke maatrikseid
3. ÜMi paaride asemel vaadelda kolmikuid, ehkki katsed paljude keeltega on näidanud, et see ei pruugi aidata
4. Arvestada lisaks ÜMide tõenäosustele mõnikord ka konkreetste sõnade tõenäosusi

Antud katse ei andnud veel tulemusi, mille alusel saaks otsustada, et just VMM on parim mudel eesti keele statistilisel ühestamisel. Seega võib olla perspektiivne ka muude statistiliste meetodite katsetamine.

Kirjandus

Lattik 1983 - Vello Lattik, “Mihklipäeval. Mihklikuul” Tallinn 1983, Eesti Raamat, lk. 4-10

Armstrong et al 1996 - Susan Armstrong, Gilbert Robert, Pierrette Bouillon, "Building a Language Model for POS Tagging" ftp://issco-ftp.unige.ch/pub/multext/tagger.doc.ps

Källgren 1996 – Gunnel Källgren, “Linguistic Indeterminacy as a Source of Errors in Tagging”, in COLING-96 proceedings, Copenhagen 1996, vol.2 pp. 676-680

Marcus et al 1990 – Marcus, M., Santorini, B, Magerman “First steps towards an annotated database of American English” In Readings for Tagging Linguistic Information in a text Corpus, ed. by Langendoen and Marcus, Tutorial for the 28th Annual Meeting of the ACL

Orwell 1990 – Orwell, G. “1984”, tlk. Elias Treeman. Loomingu Raamatukogu, Tallinn “Perioodika” 1990

Benello, J; Macki, A.; Anderson, J. A. (1989). “Syntactic category disambiguation with neural networks.” Computer Speech and Language, 3, 203-217.

Merialdo, B. (1994). “Tagging English text with a probabilistic model.” Computational Linguistics, 20(2), 155-171.

Merialdo, B. (1991). “Tagging text with a probabilistic model. “International Conference on Acoustics, Speech and Signal Processing, 809-812.

Dermatas, E.; Kokkinakis, G. (1995). “Automatic Stochastic Tagging of Natural Language Texts.” Computational Linguistics, 21(2), 137-163.

Brill, E. (1992). “A Simple rule-based part of speech tagger.” In Proceedings, Third Conference on Applied Natural Language Processing. Trento, Italy, 152-155.

Karlsson, F. (1990). “Constraint grammar as framework for parsing running text.” In Proceedings, Thirteenth International Conference on Computioanal Linguistics. Helsinki, Finland, 168-173.

Cutting, D.; Kupiec, J.; Pederson, J.; Sibun P. (1992). "A practical part-of-speech tagger." In Proceedings, Third Conference on Applied Natural Language Processing. Trento, Italy, 133-140.

Lisa 1. Lihtsustatud näide statistilise ühestamise kohta.

```
Mees
    mees+0 // NCSN //_S_ sg n, //
    mesi+s // NCS //_S_ sg in, //
peeti
    peet+0 // NCSA //_S_ (adt), //
    peet+0 // NCS1 //_S_ sg p, //
    pida+ti // VMP //_V_ ti, //
kinni
    kinni+0 // RR //_D_ //
```

Kaldjoonte vahel on algul ÜM ja selle järel MM
Antud näites piirdume ÜM hulgaga M = {NCSN, NCS, NCSA, NCS1, VMP, RR} ja mitmsesusklasside hulgaga V = { {NCSN, NCS}, {NCSA, NCS1,VMP}, {RR}}

Oletame, et treenimise käigus saadi järgmised tõenäosuste maatriksid:

Lauset alustavate märgendite tõenäosuste vektor E:

NCSN	NCS	NCSA	NCS1	VMP	RR
0,5	0,3	0,2	0,0	0,0	0,0

Maatriks X = $\{x_{kl}\}$, kus x_{kl} on tõenäosus, et mitmesusklassi v_l kuuluvatest märgenditest tuleb valida märgend m_k:

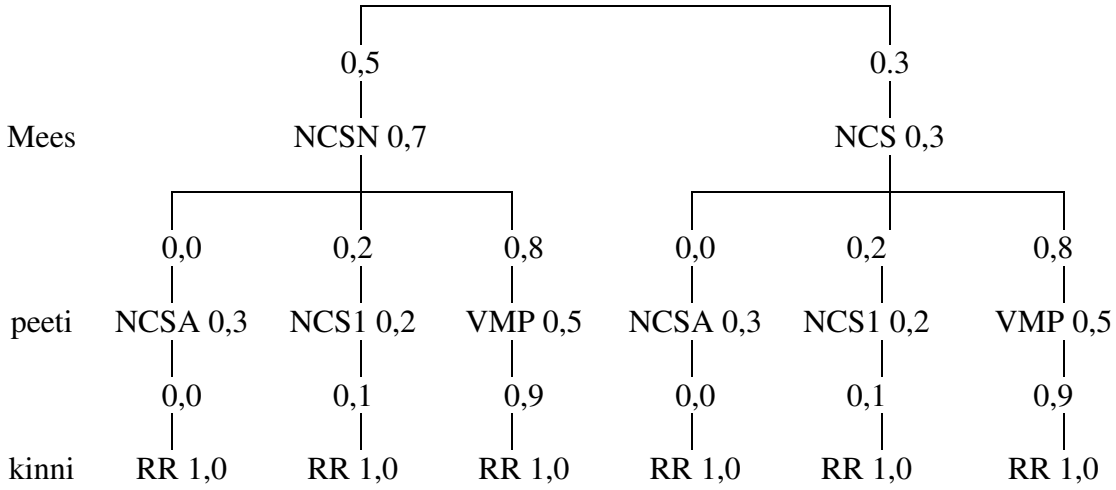
Ühestaja märgend	Mitmesusklass		
	{NCSN,NCS}	{NCSA,NCS1,VMP}	{RR}
NCSN	0,7	0,0	0,0
NCS	0,3	0,0	0,0
NCSA	0,0	0,3	0,0
NCS1	0,0	0,2	0,0
VMP	0,0	0,5	0,0
RR	0,0	0,0	1,0

Maatriks P = $\{p_{kl}\}$, kus p_{kl} on tõenäosus, et reas olevale märgendile m_k eelneb veerus olev märgend m_l.

	NCSN	NCS	NCSA	NCS1	VMP	RR
NCSN	0,0	0,0	0,0	0,0	0,0	0,0
NCS	0,0	0,0	0,0	0,0	0,0	0,0
NCSA	0,0	0,0	0,0	0,0	0,0	0,0
NCS1	0,2	0,1	0,1	0,0	0,0	0,0

VMP	0,8	0,1	0,1	0,0	0,0	0,0
RR	0,0	0,0	0,0	0,1	0,9	0,0

Alljärgnev graaf kirjeldab arvutuste käiku. Kaartele on märgitud ÜMide järgnevuste tõenäosused maatriksist P. ÜMide juurde on märgitud ÜMide valimise tõenäosused maatriksist X. Sõna *Mees* on lause esimene sõna ja seetõttu on tema ÜMide kohale märgitud tõenäosused vektorist E.



Alljärgnev tabel näitab, kuidas arvutatakse iga konkreetse ÜM järjendi tõenäosus, lähtudes eespool esitatud graafile kantud numbritest.

Ühestaja märgendite järjend	Selle järjendi tõenäosus
NCSN – NCS1 - RR	0,5*0,7*0,2*0,1*1,0=0,007
NCSN – VMP – RR	0,5*0,7*0,8*0,5*0,9*1,0=0,126
NCS – NCS1 – RR	0,3*0,3*0,2*0,2*0,1*1,0=0,00036
NCS – VMP – RR	0,3*0,3*0,8*0,5*0,9*1,0=0,0324

Maksimaalse tõenäosuse saab NCSN – VMP – RR, mis tagasi MMideks teisendatuna annab ühestatud tulemuseks

```
Mees
    mees+0 // NCSN //_S_ sg n, //
peeti
    pida+ti // VMP //_V_ ti, //
kinni
    kinni+0 // RR //_D_ //
```