

TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

Arvutiteaduse instituut

Informaatika eriala

Aivi Kaljuvee

**Määruste ja täiendite eristamine
statistiliste meetoditega**

Magistritöö (20 AP)

Juhendaja: Kaili Müürisep

TARTU 2008

Sisukord

1. Sissejuhatus.....	4
2. Nimisõnalised määrused ja määruslikud täiendid	6
2.1 Määruse ja täiendi mõisted.....	6
2.2 Nimisõnalise määruse ja täiendi vormidest.....	6
2.3 Määruse ja täiendi kuuluvuse probleem inglise keeles.....	8
3. Inglise keelele loodud meetodid	9
3.1 Juhendamata õppel põhinevad meetodid.....	9
3.2 Juhendatud õppel põhinevad meetodid.....	9
3.2.1 Transformatsioonipõhine meetod.....	10
3.2.2 Taandamismudel.....	11
3.2.3 Otsustuspuu.....	12
3.2.4 Lähima naabri meetod.....	13
3.2.5 Tugivektormasinad.....	14
4. Valitud statistiliste meetodite rakendamine eesti keelele.....	16
4.1 Treeningandmete ettevalmistamine.....	16
4.2 Eksperimentide tulemused.....	20
4.2.1 Transformatsioonipõhine meetod.....	21
4.2.2 Taandamismudel.....	22
5. Järeldused ja plaanid edaspidiseks.....	26
6. Kokkuvõte.....	27
Abstract.....	28
Viited.....	29
Lisad.....	31
Lisa 1. CD: Andmete ettevalmistamiseks kasutatud skriptid. Treeningandmed.....	32
Lisa 2. Transformatsioonipõhise meetodiga leitud reeglid.....	34
Lisa 3. Sagedused taandamismudelis.....	38

1. Sissejuhatus

Eesti keele nimisõnalised määrused ja täiendid on sisu ja vormi poolest sarnased. Loomuliku keele töötuses süntaksi analüüsimisel on aga oluline neid lauseliikmeid eristada. Täiend annab lisainformatsiooni nimisõnale, määrus verbile – eksimine nimisõna täiendiks või määruseks liigitamisega toob kaasa teksti poolt edasi antava tähenduse valesimõistmise. Näiteks, kui anda lauses „*Ma nägin binokliga meest*“ sõnale *binokliga* vale analüüs, siis sellega on ka valesti määratud, kellele ese kuulus. Sõnale korrektse süntaktilise rolli leidmine vajab antud juhul lausest väljapoole jääva konteksti kasutamist. Probleem on siiski veidi lihtsam – tüüpiliselt on vajaminev info lauses eneses olemas. Näiteks lauses „*Ma nägin tumeda peaga meest*“ ei ole eelmise näitega sarnast ebaselgust, kuid see kahjuks ei tähenda, et automaatses süntaksianalüüsis oleks sõnale *peaga* põhja leidmine, mille laiendiks see on, lihtne ülesanne. Teades sõnade tähendusi, saab välistada selle nimisõna kokkukuulumise verbiga, kuid kuna süntaktilisel analüsaatoril puudub juurdepääs semantilisele infole, siis süntaktilise reegli koostamine, mis siin õige valiku teeks, ei ole alati võimalik. Just põhjusel, et nimisõnaliste määruste ja täiendite eristamine vajab reeglina semantilist infot, on see keeruliseks probleemiks.

Eesti keelele on loodud kitsenduste grammatikal põhinev süntaksianalüsaator [Müürisep, 2000]. Kui analüsaator ei suuda üheselt sõnale sisendis analüüsi anda, siis jäetakse väljundis sellele mitu märgendit, analüsaator on suutnud osad süntaktilised klassid, millesse sõna kuulumine on võimalik, konteksti põhjal välistada, kuid lõplikku valikut ei ole tehtud. Ülal kirjeldatud probleem on eesti keele analüsaatorile selliste mitmete analüüside suurimaks põhjustajaks [Müürisep, 2000].

Käesoleva töö eesmärk oli uurida probleemi lahendamise võimalikkust statistiliste meetodite abil. Varem statistilist lähenemist eesti keeles antud probleemi juures katsetatud ei ole. Kuna sarnane mitmesusprobleem eksisteerib ka inglise keele jaoks, on esimeseks sammuks olnud olemasolevate meetoditega tutvumine. Selles valdkonnas tehtud uurimustööst pakub ülevaate käesoleva töö teoreetiline osa. Hindamiseks, kas need algoritmid on sobivad eesti keelele, on tarvilik eksperimenteerida nende rakendamisega. Magistritöö raames on sellega algust tehtud – katsetatud on kahe meetodiga, mille kasutamine on võrreldes ülejäänutega vähem töömahukas ning mis vajavad keeletehnoloogilisi ressursse, mis on eesti keele jaoks juba olemas.

Magistritöö jätkub ülevaatega, kuidas on eesti keele grammatikas määratletud

määrused ja täiendid, ning põhjendatakse pikemalt, miks neid lauseliikmeid on automaatsel analüüsil raske eristada. Seejärel tuuakse paralleel sarnase inglise keele jaoks eksisteeriva probleemiga. Kolmandas peatükis tutvustatakse olemasolevaid algoritme, mis on loodud mitmesuste lahendamiseks inglise keelele, keskendudes juhendatud masinõppel põhinevatele meetoditele. Magistritöö raames tehtud praktilist tööd – valitud meetoditele treeningandmete ettevalmistamist ning meetodite rakendamist – kirjeldatakse neljandas peatükis.

2. Nimisõnalised määrused ja määruslikud täiendid

Käesolevas peatükis antakse lugejale ülevaade, kuidas on määratletud eesti keele grammatikas määrus ja määruslik täiend ning selgitatakse, miks neid lauseliikmeid on süntaktiliste reeglite abil raske eristada. Seejärel vaadeldakse täiendite ja määruste mitmesuse probleemile analoogset inglise keeles esinevat nähtust.

2.1 Määruse ja täiendi mõisted

Määrus ehk adverbiaal on eesti keele grammatikas verbi laiend, mis ei ole sihitis ega öeldistäide, määrus on samuti omadus- ja määrsõna laiend [Erelt jt., 1993, lk. 60]. Määrused on jagatud nende grammatiliselt olulistest semantilistest funktsioonidest lähtudes kolme rühma [Erelt jt., 1993][Erelt jt., 1997]:

- (i) tegevuse osalisi (*Tal on raamat, Jüri abiellus Tuulega, Isa lõhkus kirvega puid*),
- (ii) tegevuse asjaolusid (*Ta on tänna linnas, Ilm läheb külmaks, Rong sõitis kiiresti*),
- (iii) sündmuste põhjuslikke seoseid (*Käed on külmast kanged, Suure näljaga süüakse kõike*) väljendavateks määrusteks.

Täiendina ehk atribuudina mõistetakse nimisõna laiendiks olevat lauseliiget, mis väljendab nimisõnaga tähistatud olendi, eseme või nähtuse tunnust [Erelt jt., 1993, lk. 114]. Kui see laiend on nimisõnaga määruslikus seoses – st. lisab lausele informatsiooni, mida on võimalik edasi anda ka määrusega ning on ka määrusele omases käändes – siis on tegemist määrusliku täiendiga.

Määruslikku täiendit sisaldava nimisõna fraasi saab ümber sõnastades muuta määrusega lauseks. Näitena: „*Uks viib eluruumidesse*“, „*üks eluruumidesse*“. Vaatluse all olevad lauseliikmed kannavad näidetes sama tähendust ning ei erine ka vormilt. Just vormilise sarnasuse tõttu on neid automaatses süntaksianalüüsis keeruline eristada.

Eesti keele kitsenduste grammatikas [Müürisep, 2000] on täiendi ja määruse analüüs peamine mitmesuste põhjus. 4,5% kõigist sõnadest jäi mõlema märgendiga.

2.2 Nimisõnalise määruse ja täiendi vormidest

Nii nimisõnaline täiend kui ka määrus võivad esineda kõigis 14 käändes.

Nimetavas käändes nimisõna on väga harva täiendiks, kuid leidub siiski erandeid: aja- ja hulgasõnad ning deverbaalide laiendid (nt *kell neli, kümme krooni kilomeeter*,

dokumendid kaotanu, ülemus olemine, iga nädal maalkäimine), samuti on nimetavas käändes lisandid (*seltsimees Stalin, herra professor*) [Mihkla jt., 1974]. Sama võib öelda määruse kohta: määrus on harva nimetavas käändes, kuid hulga- ja ajamäärus ning võrdlust märkiv viisimäärus on erandid [Müürisep, 2000].

Omastavas käändes nimisõnaline täiend on väga sagedasti esinev ning selle tähendus on avar. Eristatakse subjektilist, objektilist ja adverbiaalset genitiivatribuuti [Erelt jt., 1993, lk. 121], kusjuures adverbiaalne atribuut võib väljendada nii valdajat, tegijat, reksiooni, kohta, aega, hulka kui ka otstarvet. Omastavas käändes määrus ei ole nii sagedane, väljendades peamiselt aega, hulka või laiendades *ne*-liitelisi omadussõnu (*pikkune, sarnane, alane, vastane*) ning nendest tuletatud *lt*-liitelisi määrsõnu [Müürisep, 2000].

Osastavas käändes nimisõnaline täiend märgib peamiselt põhisõna liiki ja laadi (*värvi, tõugu, tüüpi, sorti* jne). Samuti on osastav käände hulgafraaside laiendite tüüpkäanne (*kaks kinga, rühm õpilasi, trobikond pilte*) [Mihkla jt., 1974]. Käesolevas töös jäi viimatinimetatud täiendiliik vaatluse alt välja, kuna kasutatud korpuses märgendati neid kui hulgasõna laiendeid, mitte kui täiendeid. Määrus on osastavas käändes aja- ja hulga- ja viisimäärusena ning võrdlusena [Müürisep, 2000].

Kohakäänetes, rajavas, olevas, ilmaütlevas ja kaasaütlevas käändes nimisõna on tavaliselt määruslik täiend, mis võib paikneda nii põhisõna ees kui järel. Need käänded on ka määruste tüüpkääneteks. Sageli on määrusliku täiendi ja määruse eristamine võimalik ainult lauset mõistes või põhjalikku reksioonisõnastikku kasutades.

Näiteks lauses „... *kes pillub kalambuure viies keeles*“ on ka inimesel väga raske otsustada fraasi „*viies keeles*“ kuuluvuse üle. Lauses „*Varasem teade USA presidendiks kandideerinud Bob Dole'i kandideerimisest osutus ennatlikuks.*“ jäävad ainult lauseehitust vaadates täiendi ja määruse vahel mitmeks sõnavormid *presidendiks* ja *kandideerimisest*. Nende täpsem automaatne analüüs on võimalik ainult reksioonileksikoni põhjal (teade millestki ja kandideerima kellekski). Vahel on lisaks lause tähenduse mõistmisele lause analüüsiks vajalikud ka taustateadmised. Näiteks lausest „*Lapsed läksid koeraga onu vaatama.*“ ei selgu, kas koer on täiend (st kuulub onule) või määrus (läks koos lastega).

Eesti keele süntaktiliselt märgendatud korpust koostades märgendasid üht ja sama teksti paralleelselt kaks lingvistit ning nende märgenduste erinevusi võrreldi hiljem poolautomaatselt. Sel teel ühtlustati ka määruse ja täiendi mitmesed analüüsid korpusel [Uibo, Bick, 2005].

2.3 Määruse ja täiendi kuuluvuse probleem inglise keeles

Kui eesti keeles on see nimisõna, mille funktsionaalset rolli lauses – määruseks või täiendiks olemist – on raske üheselt määrata, siis inglise keeles on sarnase mitmesuse põhjustajaks prepositsioonifraas (PP). Lause analüüsil vaid süntaktilisi reegleid rakendades jääb ebaselgeks, kas fraas kuulub nimisõna või verbi juurde. Näited (PP alla joonitud): *I ate a pizza with friends, I ate a a pizza with anchovies*. Tuntud on see PP kuuluvuse probleemina (*ingl prepositional phrase attachment, PP attachment*).

Inglise keele prepositsioonifraasidele vastavad nii eesti keele obliikvakäändelised¹ nimisõnafraasid kui ka kaassõnafraasid (nii ees- kui tagasõnadega). Kuna eesti keele kaassõnafraase kasutatakse enamasti määrusena ning eesti keele süntaksianalüsaator ühestab neid heuristiliste reeglite abil [Müürisep, 2000], siis käesolevas töös eesti keele kaassõnafraasi kuuluvust ei uuritud.

Et korrektne analüüs PPga lausele sõltub paljus sõnade semantikast, siis on osutunud ammendavate grammatikareeglite loomine nähtuse kirjeldamiseks ülejõukäivaks ülesandeks. Statistiline lähenemine on probleemi lahendamisel olnud edukam. Järgnev peatükk annab ülevaate sel suunal tehtud uurimistööst.

¹ kohakäänded, saav, rajav, olev, ilma- ja kaasütlev kääne

3. Inglise keelele loodud meetodid

Inglise keele jaoks on probleemiga laialdaselt tegeletud. Arendatud on nii reeglipõhiseid kui statistilisi meetodeid. Paremaid tulemusi on andnud viimased.

Reeglipõhise lähenemise [Volk, 2001, lk. 31] võib jagada kaheks:

- (i) süntaktilised reeglid: peamiselt heuristilised reeglite kasutamine, mis näiteks eelistavad vähemate tippudega analüüsipuid või PP lisamist fraasi juurde, mis on sellele lauses vasakul konteksis lähim;
- (ii) semantilised reeglid: näiteks sõna poolt argumentidele esitatavate semantiliste piirangute kirjeldamine.

Statistilised meetodid jagunevad samuti kaheks: juhendatud ning juhendamata masinõppel (*ingl supervised, unsupervised learning*) põhinevad meetodid. Üldjuhul ei kasuta meetodites mitmesuste lahendamisel informatsiooni kogu lausest, vaid need töötavad reeglina neljaliikmeliste järjenditega, mis koosnevad:

- (i) prepositsioonist (tähisteks p),
- (ii) nimisõnast, mis on prepositsioonifraasile kuuluva noomenifraasi peasõna (n_2),
- (iii) verbifraasi peasõnast, mille laiendiks prepositsioonifraas olla võib (v),
- (iv) noomenifraasi peasõnast, mille laiendiks prepositsioonifraas olla võib (n_1).

Käesolevas töös nimetatakse edaspidi neid järjendeid ka nelikuteks.

3.1 Juhendamata õppel põhinevad meetodid

Hea ülevaate juhendamata õppel põhinevatest meetoditest annab [Volk, 2001, lk. 45]. Tulemused, milleni on jõutud juhendamata õppe kasutamisega inglise keelel, on aga jäänud alla juhendatud õppe meetoditega saavutatule. Kuna eesti keele jaoks on loodud süntaktiliselt märgendatud korpus, mille olemasolu on tingmuseks juhendatud masinõppe kasutamisele, siis käesolevas töös juhendamata õppel põhinevate meetoditega pikemalt ei tegeletud.

3.2 Juhendatud õppel põhinevad meetodid

Järgnevates alampeatükkides on toodud ülevaade olemasolevatest juhendatud õppel põhinevatest meetoditest PP mitmesuse probleemi lahendamiseks inglise keelele. Lisaks kasutatud masinõppevõtetega tutvumisele, on olnud eesmärgiks uurida, milline

informatsioon on eri autorite töödes valitud PP mitmesusejuhtude kirjeldamiseks. Paljud algoritmid piirduvad peatüki alguses defineeritud neljaliikmeliste järjendite kasutamisega. Hilisemates töödes on kaasatud semantiline info neliku sõnade kohta. Ka on uuritud, kas lausest peaks eraldama rohkem informatsiooni kui need neli sõna. Käesolevasse ülevaatesse on meetodid valitud selliselt, et kasutatud erinevad variandid oleksid kaetud.

Teadlikult on välja jäetud võimendusalgoritm (*ingl boosting*) [Abney jt., 1999], närvivõrkude kasutamise, maksimum entroopia mudeli (*ingl maximum entropy model*) ning mitmete pooltõenäosuslike meetodite kirjeldused.

Autorid on oma töödes üldjuhul kasutanud algoritmide treenimiseks ja testimiseks Penni Puudepanga I versiooni WSJ (*Wall Street Journal*) alamkorpust². Artiklites avaldatud tulemused on seega omavahel võrreldavad. Kui järgnevates alampeatükkides meetodite tutvustamisel ei ole treeningandmeid eraldi kirjeldatud, siis on kasutusel olnud nimetatud korpus.

3.2.1 Transformatsioonipõhine meetod

Transformatsioonipõhine õpe on ennekõike tuntud kui meetod sõnaliikide märgendamiseks või grammatika automaatseks genereerimiseks. Brill ja Resnik [Brill, Resnik, 1994] on aga kasutanud sama ideed ka PP mitmesuste lahendamisel. Meetod leiab treenimise tulemusena reeglid, mis otsustavad PP kuuluvuse. Reeglid on esitatud transformatsioonidena, mille üldkuju on järgmine:

loe PP seotuks noomeniga/verbiga, kui nelikus $v=s_1$ ja $n_1=s_2$ ja $p=s_3$ ja $n_2=s_4$,

kus s_i tähistab konkreetset sõnavormi ning tingimuses kontrollitakse kas ühte, kahte, kolme või kõiki nelja sõna mitmesuse nelikust. Reeglid eraldatakse ahne algoritmiga. Esmalt loetakse kõikides treeningkirjetes PP kas noomeni või verbi juurde kuuluvaks. Seejärel asutakse otsima vastust küsimusele, mida tuleks teha, et kirjeldatud teisenduste abil kuuluvused treeningnäidetele korrektseteks muuta. Hinnatakse, milline teisendus (kõikvõimalike lubatud teisenduste hulgast) treeningandmetele rakendatuna muudab kuuluvusi nii, et veaprotsent väheneb võimalikult palju. Kuuluvusi modifitseeritakse selle parima teisenduse järgi ning see saab ühtlasi treenimise tulemuseks oleva transformatsioonide hulga esimeseks liikmeks. Edasi liigutakse järgmisesse iteratsiooni, kus meetod otsib taas sel hetkel parimat transformatsiooni – siit tuleneb ka algoritmi liigitumine ahnete algoritmide alla. Tsükkel jätkub seni, kuni kõikide nelikute kuuluvused

2 Info korpuse kohta aadressil: <http://www.cis.upenn.edu/~treebank/>

ühtivad õigetega. Saadavate reeglite rakendamine uute mitmesuse juhtude klassifitseerimiseks toimub samas järjekorras, millega need eraldati.

Meetod töötab inglisekeelsetel tekstidel korrektsusega 80,8%. Tulemus paranes 81,8%ni, kui sõnad markeeriti lihtsal viisil Wordnetist pärineva semantilise infoga. Neliku noomenite ja verbidega seoti sõnade kõikvõimalike hüperonüümide (ülemmõistete) hulk, täpsemalt hüperonüümidele vastavad Wordneti sünohulgad, ning transformatsioonidesse lisati võimalus kontrollida sõna olemist hüponüümiks sünohulga sõnadele.

3.2.2 Taandamismudel

Taandamismudeli (ingl backed-off model) töö prepositsioonifraasi mitmesuste lahendamisel põhineb ideel, mille esmakordselt pakkus välja Katz [Katz, 1987]. Siis rakendati seda põhimõtet kõnetuvastuses n-gramm mudelite kohandamiseks hõredatele andmetele. Ülesandeks oli sõna esinemise tõenäosuse hindamine eelnenud n sõna põhjal. Kui korpus ei sisaldanud piisavalt näiteid, arvestati tõenäosuse määramisel (n-1) eelnenud sõna. Kui näiteid leidis jätkuvalt alla seatud läve, vähendati eelnenud sõnade järjendi pikkust veelgi. Selliselt rekursiivselt liikudes muudetakse esialgset ülesannet seni, kuni treeningandmetest saadav statistika on usaldusväärne. Collins ja Brooks [Collins, Brooks, 1995] võtsid taandamisvõtte kasutusele prepositsioonifraasi mitmesuste lahendamiseks. Neliku klassifitseerimiseks otsitakse treeningandmetest identseid nelikuid. Kui need leitakse, määrab kuuluvuse treeningnäidete enamuse kuuluvus. Vastasel juhul otsitakse näiteid, milles prepositsioon, verb või üks nimisõnadest erineb klassifitseeritava neliku vastavast sõnast. Kui sellist kolmikut ei leita, lubatakse erineda kahel sõnal, sealt edasi kolmel. Kui eelnenud rekursiooni etapid ei ole vastust andnud, loetakse neliku prepositsioonifraas noomeni juurde kuuluvaks, st kasutatakse vaikimisi väärtust.

Probleem, kuidas kombineerida ühel rekursiooni-etapil leitud näidete kuuluvus, lahendati kõikidele etappidele sarnaselt. Kuuluvuse määrab treeningandmete enamuse kuuluvus – kõiki treeningnäiteid arvestatakse võrdse kaaluga. Valemis (1) on näidatud arvutus etapil, mil sageduste leidmisel lubatakse erineda ühel sõnal, valemit (2) rakendatakse etapil, mil erinevad kaks sõna. Katsetused näitasid, et algoritmi töö tulemus on parim, kui nõutakse prepositsioonide kattumist treeningnäidetes ja klassifitseeritavas nelikus. Taandamismudeli täpsuseks saadi 84,5%. Eksperimentidest selgus lisaks, et sagedustele läve seadmine, st arvutuste rakendamine vaid siis, kui treeningnäidete hulk ületab valitud läve, vähendaks algoritmi töö täpsust.

$$\bar{p}(v, n1, p, n2) = \frac{\bar{f}(v, n1, p) + \bar{f}(v, p, n2) + \bar{f}(n1, p, n2)}{f(v, n1, p) + f(v, p, n2) + f(n1, p, n2)} \quad (1)$$

$$\bar{p}(v, n1, p, n2) = \frac{\bar{f}(n1, p) + \bar{f}(v, p) + \bar{f}(n2, p)}{f(n1, p) + f(v, p) + f(n2, p)} \quad (2)$$

f – nelikute arv treeningandmetes, milles funktsiooni argumentdega näidatud sõnad kattuvad klassifitseeritava neliku vastavate sõnadega
f̄ – loeb kokku samad nelikud, lisatingimuseks – nelikute kuulumine on nimisõna
p̄ – tõenäosus, et klassifitseeritava neliku kuulumiseks on nimisõna

3.2.3 Otsustuspuu

Stetina ja Nagao [Stetina, Nagao, 1997] lisavad treeningandmetele semantilise informatsiooni Wordnetist. Nende algoritm ei tööta mitte nelikutes olevate sõnavormidega vaid semantiliste klassidega, millesse sõnad Wordnetis on jagatud. Nelikus oleva nimisõna või verbi asetamine Wordneti hierarhiasse nõuab aga lisaks teadmist, millises tähenduses mitmetähenduslik sõna lauses esines. Mitmetähenduslikkuse probleemi lahendamiseks pakutakse välja juhendamata õppel põhinev ühestamisalgoritm, mis töötab samuti nelikutel, kaasamata laiemat konteksti lausetest. Meetod kasutab eeldust, et sarnases kontekstis esinedes on kaks erinevat sõna kasutatud tähendustes, mis on üksteisele semantiliselt lähimad. Semantilise kauguse kahe sõnatähenduse vahel leiab algoritm arvestades tähenduste omavahelist kaugust Wordneti hierarhias ning ka tähenduste kaugust hierarhia ülemmõistest. Ühe neliku sõnatähenduste leidmiseks valitakse ülejäänutest sellele sarnaseim (nelikute kauguse mõõduks on kombinatsioon neis sisalduvate sõnade kaugustest) ning tähendusteks määratakse üksteisele vastavate sõnade lähimad tähendused. Algoritmi töötab 72% korrektsusega, osutub, et edasiseks prepositsioonifraasi mitmesuse lahendamiseks, on see tulemus rahuldav.

Kui sõnade tähendusklassidesse kuulumine on määratud, moodustatakse prepositsioonifraasi mitmesuste lahendamiseks otsustuspuu (*ingl decision tree*). Puu struktuurina hoitud info kasutamine toimub järgmiselt [Mitchell, 1997]: (i) otsitavad väärtused paiknevad puu lehtedes; (ii) nendeni jõudmiseks läbitakse puu juurtipust alustades; (iii) iga sõlme juures kontrollitakse ühte tingimust ja vastavalt sellele, kuidas tingimus täidetud on, liigutakse sobivat kaart mööda järgmise sõlmeni. Otsitavaks väärtuseks nelikute puhul on prepositsioonifraasi kuulumine noomeni või verbi juurde. Tingimuseks, mida antud ülesande puhul puu läbimisel iga sõlme juures testitakse, on neliku noomeni või verb kuulumine teatavasse semantilisse klassi. Juurele lähemates sõlmedes on semantilised klassid üldisemad, muutudes lehtede suunas spetsiifilisemateks.

Iga prepositsiooni jaoks koostatakse oma otsustuspuu.

Puu konstrueerimiseks kasutatakse ID3 algoritmi [Quinlan, 1986]. Eesmärgiks on moodustada puu, mis klassifitseeriks treeningandmed korrektselt ja koosneks seejuures võimalikult väikesest arvust kaartest ja tippudest. Igale teele puu juurtipust leheni, st igale tingimuste järjendile, vastab treeringnäidete osahulk, mis neid tingimusi rahuldab. Selle osahulga klassifikatsiooni annab lehele märgitud väärtus. Seega iga sõlme juures jagatakse treeningnäited kuuluvateks sõlmest lähtuvatele alampuudele. Et piirata puu suurust, valitakse sõlmes kontrollitavaks tingimuseks see, mis jagab nelikud nii, et alampuudes domineeriks selgelt treeningnäidete prepositsioonifraasi kuulumine kas nimi- või tegusõnale. Puu loomist alustatakse juurtipust ja sõlmi lisatakse seni, kuni nelikute jagamisel on jõutud osahulgani, milles näidete prepositsioonifraasi kuuluvus on sama või peaaegu sama (erinevusi on vähem etteantud väärtusest).

Kirjeldatud meetod, mis sisaldab sõnadele semantilise informatsiooni lisamiseks ka sõnatähenduste ühestamist, saavutas Penni Puudepanga nelikutel treenides täpsuse 88,1%.

3.2.4 Lähima naabri meetod

Ülalkirjeldatud lähenemine lisas nelikutele semantilise info Wordnetist, selles jaotises tutvustatavad kaks meetodit – [Zavrel jt., 1997], [Zhao, Lin 2004] – kaasavad ülesande lahendamisse tähendusinfo ilma selle mahuka keeleressursi abita. Tähendusinfo pärineb korpustest, mis ei vaja eelnevat lingvistilist märgendamist. Kasutatakse sama eeldust, millel põhines eelnevalt kirjeldatud sõnatähenduste ühestamine – kui sõnad esinevad sarnases kontekstis, on nende tähendused lähedased. Vaadeldav kontekst on aga siin laiem. Meetodid defineerivad selle erinevalt. Zavrel jt. võtsid arvesse sõnast kaks positsiooni eespool ja kaks positsiooni tagapool esinenud sõnad. Zhao ja Lini algoritmis on kontekstiks sõnale vahetult eelnenud ja järgnenud sõnad. Kui eelneb või järgneb sõna, mis kuulub korpuse 20 sagedaseima sõna hulka, lisatakse konteksti ka selle kõrval olev sõna. Edasiseks kasutamiseks hoitakse mõlemas meetodis korpustest eraldatud info sõnade distributsiooni kohta vektoritena, mille iga dimensioon vastab ühele kontekstisõnale ja väärtus näitab selle sõna esinemise sagedust. Seega on prepositsioonifraasi mitmesuse lahendamise algoritmil kasutada nelikute sõnadele vastavad naturaalarvuliste väärtustega vektorid, mis iseloomustavad sõna distributsiooni.

Uue neliku klassifitseerimiseks eraldatakse treeningnäidetest sarnased nelikud ja leitakse kuuluvus nende kuuluvuse põhjal. Sellise põhimõttega masinõppe meetodit – ei

kasutada eelnevalt kogutud statistikat, vaid iga uue andmekirje jaoks töödeldakse seda alamhulka treeningandmetest, mis on kirjele lähedane – tuntakse kui lähima naabri meetodit (*ingl nearest neighbour*). Vaadeldavad kaks algoritmi arvutavad distributsioonivektorite põhjal nelikute vahelist kaugust erinevalt. Samuti erineb sarnaste kirjete arv, mis töötlemiseks valitakse. Ühine on aga see, et klassifitseeritavale nelikule vastavas alamhulgas ei panusta kirjed otsitavasse väärtusesse võrdselt, vaid erineva kaaluga, mille määrab taas, kui sarnane on kirje klassifitseeritavale nelikule.

Ajaliselt varasem Zavreli jt. meetod saavutab 84,4% täpsuse. Zhao ja Lini algoritmi tulemuseks on 86,5% täpsus. Viimatinimetatu on selle tulemusega parim teadaolev meetod, mis lisaks süntaktiliselt märgendatud korpusele ei vaja juurde keeleressurssi, mille ettevalmistamine on töömahukas – kasutatakse süntaksikorpust ja lingvistiliselt märgendamata keelekorpust.

3.2.5 Tugivektormasinad

Olteanu ja Moldovani [Olteanu, Moldovan, 2005] pakutud lähenemine erineb eelnevatest sellega, et mitmesuste lahendamiseks vaadeldakse laiemat lause konteksti. Eksperimendid on näidanud [Ratnaparkhi jt., 1994], et inimene määrab PP põhja märgatavalt edukamalt, kui nelikute asemel on teada terve lause. Nimetatud autorid näitavad, et enama info kaasamine lausest parandab tulemusi ka mitmesuste automaatsel lahendamisel. See info saadakse lause analüüsi puud kirjeldades. Arvesse võetakse näiteks verbi ja nimisõna fraasi NP (PP võimalik põhi) paiknemist üksteise suhtes lauses – treeningnäitesse lisatakse järjend puu tippude märgenditest, mis paiknevad NP ja verbi tippe ühendaval teel. Veel näiteid sellest, mida treeningkirjesse lisatakse: tipu märgend, mille tüüpiks on NP; lauses prepositsiooni ja NP peasõna vaheline kaugus; kas verb asub NP peasõnast ja PP nimisõnast lauses eespool jne.

Seega lisainfo saamiseks ei vajata uut keeleressurssi, kirjeldatakse lause süntaktilist struktuuri ja kasutatakse ikka ainult puudepanka. Autorid näitasid eksperimentidega aga ka seda, et puudepanka olemasolu ei ole tingimata vajalik – arvestatavad tulemused saadi ka siis, kui süntaktiline info pärineb parseri poolt genereeritavatest lausepuudest. Kasutatakse Charniaki parserit [Charniak, 2000], mis töötab inglise keelel vähem kui 40 sõnast koosnevatel lausetel korrektsusega 90,1%. Parseri väljundi kasutamiseks tuleb küll sellele käsitsi lisada info, milline on PP korrektne asukoht puu hierarhias, st parseri väljundis ei ole tarvis puu struktuuri muuta, lisatakse üksnes info, millise tipu juurde PP kuulub.

Süntaksilise info kasutamisele lisaks kaasatakse mitmesuste kirjeldamiseks informatsioon, mis on kättesaadav tohutust tekstikogust – Internetist. Tekstidest eraldatakse heuristiliste reeglitega näited PP-de kohta, mille paiknemine lausestruktuuris ei ole mitmene. Loendatakse mitu korda on PP seotud teatud verbiga ning mitmel korral teatud noomeniga. Samale PP-le, verbile ja noomenile vastavale treeningnäite kirjele saab üheks seda iseloomustavaks omaduseks nende sageduste suhe.

Treeningnäite kirjeldus, mis parima tulemuse andnud eksperimendis koosnes 17 komponendist, teisendatakse naturaali- ja reaalarvuliste väärtustega vektoriks. Valitud masinõppe meetod – *tugivektormasin* (ingl *support vector machine*) – opereerib neil vektoritel. Tugivektormasin [Cortes, Vapnik, 1995] vaatleb neid vektoreid, mille pikkuseks on n , kui punkte n -dimensionaalses ruumis ja otsib tasandit selles ruumis, mis treeningandmed parimal moel kahte jaotaks. Ühele poole tasandit jäävad andmepunktid, mis esindavad treeningnäiteid PP allumisest NP-le, teisele poole jäävad punktid, mille vektorid kirjeldavad PP kuulumist VP-le. Tasandile otsitakse sellist asendit, et kahe punktihulga lähimad punktid (Eukleidilise kauguse mõttes) – tugivektorid – asuksid tasandist võimalikult kaugel. Meetod eeldab, et andmepunktid on lineaarselt eraldatavad. Reeglina eelneb tasandi konstrueerimisele andmepunktide teisendamine enamadimensionaalsusega ruumi punktideks, mille järel võib ka lineaarse eraldatavuse tingimus täidetuks osutada. Võimaluseks on ka tasandi asetamine nii, et piiratud hulgal andmepunktidel on lubatud asuda vaele pool tasandit.

Kirjeldatud lähenemist treeniti ja testiti Penni Puudepanga II versioonil ning meetodi korrektsuseks saadi 93,6%. Osaliselt on tulemuse paranemine võrreldes eelnevate meetoditega seletatav asjaoluga, et treeningmaterjali maht oli siin ligi kaks korda suurem (Penni Puudepanga I versiooni vajaliku info eraldamiseks kasutada ei saanud). Samas oletavad autorid, et algoritmide tulemused neil kahel andmehulgal treenituna ei saa erineda üle 4%. Seega on siin alampeatükis kirjeldatud lähenemine seni teadaolevatest PP mitmesuste lahendamise meetoditest parim.

4. Valitud statistiliste meetodite rakendamine eesti keelele

Eesti keelele rakendamiseks valisin kaks meetodit: transformatsioonipõhise õppe ja taandamismudeli. Transformatsioonipõhise meetodi eeliseks on see, et treenimise tulemuseks on lingvistilised reeglid. Masinõppe algoritmid leiavad reeglina treenimise käigus tõenäosuseid, mida hiljem uute andmete töötlemisel kasutatakse, kuid need tõenäosused on inimesele raskesti tõlgendatavad. Teise valiku põhjuseks on see, et taandamismudel on inglise keelele välja töötatud meetoditest koos võimendus-algoritmiga [Abney jt., 1999] teadaolevalt parim, mis kasutab treeningandmetena vaid neljast sõnast koosnevat järjendit, lisamata treeningkirjetele muud informatsiooni. Alustasin määruste ja määruslike täiendite mitmesuse probleemi lahendamise võimalikkuse uurimist selliste meetodite kasutamisega, mille rakendamine on vähem töömahukas.

4.1 Treeningandmete ettevalmistamine

Inglise keelele loodud meetodites eeldatakse, et lauses prepositsioonifraasi mitmesuse lahendamisel piisab otsuse tegemiseks vaid nimi- ja tegusõna vaatlemisest, mille juurde fraas potentsiaalselt kuulub. Probleemi sarnaselt eesti keelele lihtsustades eeldame, et seda, kas nimisõna on lauses määrusliku täiendi või määruse rollis, saab hinnata, kui on teada nimi- ja tegusõna, mille laiendiks mitmene nimisõna olla saab. Algoritmidele sisendiks olevad treeningnäited mitmesustest, esitatakse siinses töös viieliikmeliste järjenditena, mis koosnevad:

- (i) mitmesest nimisõnast (tähisteks n_2)
- (ii) käändest, milles see kasutatud oli (k_2)
- (iii) verbist, mida mitmene nimisõna laiendab, kui see on lauses määruseks (v)
- (iv) nimisõnast, mida mitmene nimisõna laiendab, kui see on lauses täiendiks (n_1)
- (v) viimase käändest (k_1).

Sõnad on järjendites algvormis. Näidetesse pole lisatud informatsiooni selle kohta, kas nimisõna on lauses potentsiaalseks ees- või järeltäiendiks.

Treeningmaterjal eraldati eesti keele süntaktiliselt märgendatud korpusest³. Samad tekstid, ilma süntaktilise märgendusega, kuid koos morfoloogilise infoga, analüüsiti süntaksianalüsaatoriga. Kogutava treeningmaterjali andsid parseri väljundisse jäänud

3 Korpus on ligipääsetav aadressilt: <http://www.el.ut.ee/korpused/syntaksikorpus/>

mitmesused – väljund näitas mitmese sõna asukoha ning süntaksikorpuse selle sõna õige süntaktilise rolli lauses. Süntaksianalüsaatoris on lauseanalüüsi reeglid jaotatud heuristilisteks ja mitteheuristilisteks [Müürisep, 2000]. Erinevuseks on see, et esimesed väljendavad keelereegleid, mis ei ole üheselt tõesed – suure tõenäosusega reegel kehtib, kuid on ka erandeid. Treeningandmete mahu suurendamiseks kaasati treeningnäidetes sõnad, mille analüsaator jättis mitmeseks, kui töötas vaid tavaliste reeglitega. Testiti aga mitmesustel, mis pärinesid analüüsil kõiki reegleid kasutanud parseri väljundist.

Süntaksikorpuse ja analüsaatori märgendus on eesti keele kitsenduste grammatika märgendus, mis esitab sõnade madala pindmise funktsionaalse kirjelduse [Müürisep, 2000]. Seega on sõnadele teada süntaktiline roll lauses, nt alus, öeldis, sihitis, laienditele näidatakse ka põhja leidumise suund, kuid kasutada ei ole fraasistruktuuri puid. Mitmesuste leidmine automaatselt ei ole seetõttu elementaarne ülesanne.

Joonisel 1 on toodud lause eesti keele süntaktiliselt märgendatud korpusest. Lause iga sõna paikneb eraldi real, kus on kirjas nii sõnavorm, lemma ja morfoloogiline lõputunnus. Kaldkriipsude vahel paikneb morfoloogiline informatsioon (sõnaliik, kääne või pööre, arv jpm.). Süntaktiline märgend algab sümboliga @.

<s>
 Varasem varasem+0 // _A_ comp sg nom // @AN>
 teade teade+0 // _S_ com sg nom // @SUBJ
 USA USA+0 // _Y_ nominal ? // @NN>
 presidendiks president+ks // _S_ com sg tr // @ADVL
 kandideerinud kandideeri=nud+0 // _A_ pos // @VN>
 Bob Bob+0 // _S_ prop sg nom // @NN>
 Dole'i Dole+0 // _S_ prop sg gen // @NN>
 esinemisest esinemine+st // _S_ com sg el // @<NN
 osutus osutu+s // _V_ main indic impf ps3 sg ps af // @+FMV
 ennatlikuks ennatlik+ks // _A_ pos sg tr // @ADVL
 . . // _Z_ Fst //
 </s>

Süntaktilised märgendid:

@ADVL – määrus

@AN> – omadus- ja järgarvsõna eestäiendina

@+FMV – finiidne öeldis

@NN>, @<NN – nimi-, ase- ja põhiarvsõna ees- ja järeltäiendina

@SUBJ – alus

@VN> – partitsiip eestäiendina

Joonis 1. Näide süntaktiliselt märgendatud lausest.

Treeningmaterjali programseks kogumiseks tuli kasutada mitmeid lihtsustusi. Neist esimeseks oli otsus, et kirjavahemärgid ja sidesõnad näitavad ära vahemiku, kust tuleb otsida mitmese nimisõnaga kokkukuuluvaid sõnu. Kui need ei ole lauses osalausepiiri tähiseks, vaid rinnastavad muid lauseliikmeid, ei õnnestu leida treeningnäite järjendi kõiki komponente. Sel põhjusel on ~27%le treeningkirjetest leidmata verb ja ~13%le puudu nimisõna, millele mitmene sõna võis täiendiks olla. Edasi sai lahendamist vajavaks probleemiks asjaolu, et kirjavahemärkide või sidesõnadega piiratud lauselõigus võis esineda mitu tegusõna, kui käändelist verbivormi sisaldav lauselühend ei ole muust lausest komaga eraldatud. Kolmandaks raskuskohaks treeningnäidete koostamisel osutus tarvidus leida lausetest ühendverbid ning verbi ja noomeni püsiühendid. Püsiühenditega tuli arvestada, kuna mitmene sõna on semantiliselt seotud terve püsiühendi tähendusega, mitte tähendusega, mida omab verb üksinda .

Lauselühendeid puudutav lahendati järgmiselt:

- Kui lauses oli v-, tav-, nud- või tud-kesksõna⁴ kasutatud omadussõnalise täiendina, siis verbi mitmesuse kirjesse ei lisatud. Ei olnud võimalik otsustada, kas sõna laiendab kesksõna või öeldise rollis olevat verbi. Sel põhjusel puudub verb ~7% treeningnäidetele.

Näited:

Mitmene sõna (alla joonitud) kuulub kokku kesksõnaga:

Mõnes kiiruga laagriks kohendatud turbarabas võis tõesti põgenemisvõimalus avaneda.

Mitmene sõna kuulub kokku öeldisega.

Nad seisid täies aupaistes vastu selginevat taevast.

- Korrektses kasutuses on määruslikud des-, mata-, maks-, nuna- ja tuna-lühendid⁵ ülejäänud lausest eraldatud komaga siis, kui käändelise verbi laiendid asuvad selle järel. Kui need puuduvad või paiknevad lauses verbist eespool, siis koma ei kasutata. Koma puudumisel määrab mitmese sõna asukoht lauses, kas sellega kokkukuuluvat verbi on võimalik määrata. Kui sõna paikneb pöördelise ja käändelise verbi vahel, jääb treeningnäitele verb leidmata (~1% treeningandmetest), muudel juhtudel on mitmese sõna põhja asukoht ühene.

Näited:

Lühend lõpeb des-vormiga, järelikult on mitmene sõna öeldise laiendiks:

Mees kiskus koju jõudes sokid jalast.

Mitmene sõna on seotud käändelises vormis oleva verbiga, kuid treeningnäiteid eraldava programmi jaoks on mõlemad verbid võimalikud:

Ta korrrutas kolleegi lapsega rääkides pidevalt ...

- Muudel juhtudel, kui lauselõigus leidisid samaaegselt verbid pöördelises ja käändelises vormis, eelistati käändelises vormis olevat tegusõna ja treeningkirjesse lisati see.

Ühend- ja väljendverbide ning tugiverbiühendite eraldamine oli probleemiks ~42% korpusetekstide töötlemisel. Nimelt sellise mahuga korpuse alamosa ei olnud käesoleva töö tegemise hetkel märgendatud infoga verbiühendite leidumise ja nende asukoha kohta. Samuti ei olnud võimalik kasutada tarkvara, mis tekstile selle info lisaks. Seetõttu tuli

4 Parseri väljundis ei ole võimalik eristada v-, tav-lühendite kasutamist määrus- või täiendlühendina. Mitmesuste leidmisel loetakse v-, tav-vormid alati täiendlühendeid moodustavateks.

5 Parseri väljundis ei ole võimalik eristada mata-lühendi kasutamist määrus- või täiendlühendina. Mitmesuste leidmisel loetakse mata-vorm alati määruslühendit moodustavaks.

püsiühendite määramise ülesanne lahendada taas eesmärki oluliselt lihtsustades. Ühendite leidmisel toetutakse eesti keele verbikesksete püsiühendite andmebaasile⁶. Ühendverbide eraldamiseks otsitakse lauselõigust, kus mitmese sõna põhi arvatakse olevat, ühendverbi koosseisu potentsiaalset kuuluvat määrsõna. Iga määrsõna puhul kontrollitakse andmebaasist, kas see sõna moodustab verbiga ühendverbi. Sarnaselt toimitakse ka väljendverbide ja tugiverbiühendite otsimisel – kui lauselõigust leitakse nimisõna, mis andmebaasi järgi võib verbiga püsiühendi moodustada, lisatakse see väljendverb treeningkirjesse. Treeningkirjetesse sattusid seeläbi aga ka verbiga seotud kollokatsioonid, kuna käesoleva töö valmimise ajal oli ühendite andmebaasis need märgitud samuti väljendverbideks. Kirjeldatud robustset meetodit testiti verbiühenditega märgendatud korpuse alamosal. Kui lauses püsiühendit ei olnud, eksis meetodid kokku 2,0% juhtudest, pidades lauset ühendit omavaks. Kui lauses oli kasutatud ühendverbi, leiti see üles 87,3% juhtudest, kui tegemist oli verbi ja noomeni ühendiga, leiti ühend 54,7% juhtudest. Nõrgem tulemus verbi ja noomeni ühendite jaoks oli ootuspärane, kuna esitatud reegel määrab ühendi vaid siis, kui ühendi koosseisus olev nimisõna on lauses kasutatud käändes, millega see on andmebaasi kantud.

Viimaseks ettevalmistuseks enne algoritmide rakendamist sai treeningjärjenditesse kuuluvate pärisnimede asendamine märgendiga 'NIMI'.

Siinses alampeatükis kirjeldatud viisil koguti süntaksikorpusest 19968 treeningkirjet, kasutades mitmesuste leidmiseks mitteheuristiliste reeglitega töötanud süntaksianalüsaatori väljundit, kõikide reeglite kasutamine andis 14299 treeningnäidet. Esimeses andmehulgas oli 86,2% mitmesusenäidetest mitmene sõna korpuse järgi määrus, teises näitehulgas oli määruste osakaal 83,7%. Eraldatud treeningandmed ning tekstide töötlemiseks kasutatud programmid on toodud lisas 1.

4.2 Eksperimentide tulemused

Treeningandmete eraldamisele korpusest järgnes valitud kahe meetodi rakendamine. Töö tulemused on esitatud järgnevas kahes alampeatükis.

Testimisel järgisin k-jaotusega rist-valideerimise (*ingl k-fold cross-validation*) põhimõtet, milles k väärtuseks valisin 10. Jagasin väiksema treeningnäidete hulga (mitmesusenäited, mille andis kõiki reegleid kasutanud süntaksianalüsaatori väljund) kümneks osaks selliselt, et näiteid määruslike täiendite ja määruste kohta oleks osades

6 Andmebaas on ligipääsetav aadressilt: <http://www.el.ut.ee/ressursid/pysiyhendid/>

võrdselt. Igal testimiskorral kasutasin üht osa testimiseks ning ülejäänuid mitmesusenäiteid koos kirjetega suuremast andmehulgast treenimiseks (lisandusid mittekattuvad kirjed). Siin toodud tulemused on kümne testimiskorra keskmised.

4.2.1 Transformatsioonipõhine meetod

Eesti keelele rakendasin versiooni transformatsioonipõhise õppe meetodist, mis ei kasuta WordNeti abi. Algoritmi implementatsioon on saadaval Eric Brilli veebilehelt⁷. Tarkvara koodi ma ei muutnud, mistõttu testisin treeningandmetega, mille järjenditest olin eemaldanud ühe komponendi: käändeinfo nimisõnale, millele mitmene sõna on võimalikuks laiendiks.

Katsed näitasid, et erinev treeningandmete algväärtustamine, st. sisendis algoritmile kõikides treeningkirjetes mitmese sõna märkimine määruseks või täiendiks, mõjutas tulemusi minimaalselt, sest treenimise käigus eraldatud teisendused suures osas kattusid. Kui initsialiseerimisel märgiti kõik nimisõnad täienditeks, siis esimesena valitud ja rakendatud teisendused muutsid 93.2% kirjetele nimisõna määruseks. See on aga ligilähedaselt sama seisund, kust algoritm alustab tööd, kui algväärtustamine on vastupidine. Tabelis 1 on toodud näited eraldatud reeglitest. Tabeli esimeses veerus on antud reegel, teise veeru väärtus näitab kirjete arvu, millele mitmese sõna märgend muutus peale teisenduse rakendamist korrektseks (teisendust rakendati iteratsiooni etapil, mil see reegel osutus parimaks ja lisati treenimise tulemuseks olevasse järjestatud transformatsioonide hulka). Ülejäänud eraldatud reeglitega on võimalik tutvuda lisas 2.

n_2 on täiend, kui k_2 =omastav	79
n_2 on määrus, kui n_2 =kaks	15
n_2 on määrus, kui n_2 =paar	15
n_2 on täiend, kui n_1 =mees, k_2 =kaasaütlev	13
n_2 on määrus, kui n_1 =tema	8
n_2 on määrus, kui n_2 =minut	7
n_2 on määrus, kui n_2 =aeg	7
n_2 on täiend, kui n_1 =mälestus	7
n_2 on täiend, kui n_2 =puu, k_2 =seestütlev	7
n_2 on määrus, kui n_2 =pudel	6

Tabel 1: 10 parimat reeglit, kui algselt märgiti kõikides näidetes mitmene sõna määruseks. Väärtused teises tulbas näitavad mitme treeningkirje klassifikatsioon muudeti teisenduse rakendamisega korrektseks.

⁷ <http://research.microsoft.com/~brill/>

Järgnevalt on esitatud näited lausetest, milles esinevate mitmese lahendamisel olid leitud reeglid edukad ning samuti näited kordadest, kui sõna süntaktine klass määrati valesti. Näitelausega koos on toodud sellele vastav kirje testandmetes.

(i) Mitmene sõna (alla joonitud) märgiti korrektselt määruseks:

Eindra laseb oma käega esimese õlle vaadist välja.

n_1 =õlu, v =välja_laskma, n_2 =vaat, k_2 =seestütle

(ii) Mitmene sõna märgiti korrektselt täiendiks:

... saatis kevadel USA presidendi naisele Hillary Clintonile Heli Lätse laserplaadi ja sai hiljuti tänuga vastuse.

n_1 =vastus, v =saama, n_2 =tänu, k_2 =kaasaütlev

(iii) Mitmene sõna on lauses määrus, kuid märgiti täiendiks:

Seda, mida me võime endale lubada Nimetus baaris, ei saa me endale lubada varietees.

n_1 =NIMI, v =endale_lubama, n_2 =baar, k_2 =seesütlev

(iv) Mitmene sõna on lauses täiend, kuid märgiti määruseks:

Tõenäoliselt tekib aprillis autokaubanduses suur segadus hindadega.

n_1 =segadus, v =tekkima, n_2 =hind, k_2 =kaasaütlev

Algoritm töötas testkirjetel korrektsusega 85,7%. Kirjed, milles nimisõna oli korpuse põhjal määrus, klassifitseeriti korrektsusega 94,4%. Kui mitmene sõna oli täiend, oli tulemuseks 39,5%-line täpsus. Erinev algväärtustamine mõjutas tulemusi 0,1% võrra.

4.2.2 Taandamismudel

Taandamismudeli rakendamisel arvestasin sellega, et treeningnäidete järjendid koosnevad viiest elemendist. Seega tuli algoritmile lisada taandamisetapp, mis hindab nimisõna kuuluvust järjendi kõigi viie elemendi abil. Edasi otsisin sarnaselt Collinsi ja Brooksi artiklis [Collins, Brooks, 1995] kirjeldatule neid järjendi komponente, mida tuleb ülejäänud etappidel arvesse võtta. Kui treeningkorpuses ei leidu vaadeldavat viisikut, kasutab algoritm nelja või vähem elemendi koosesinemise sagedusi. Et hinnata milliste elementide kombinatsioonid viisikust on nimisõna kuuluvuse määramisel olulisemad, viisin läbi katse, milles mõõtsin kuuluvuse määramise täpsust erinevate tunnuste nelikute, kolmikute, paaride ja üksikute elementide kaupa (vt tabel 2). Täpsuse mõõtmisel kasutati ainult neid elementide kombinatsioone, millele leidis üheselt treeningkorpuses vastav kombinatsioon. Näiteks nelikule $\{n_1, v, n_2, k_2\}$ vastav tulemus 92,4% tähendab, et need

kirjed, millele treeningandmetes ei olnud identseid vasteid, kuid leidus esimese nimisõna käände poolest erinevaid näiteid, klassifitseeriti valemit (3) kasutades 92,4%lise korrektsusega.

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(n_1, v, n_2, k_2)}{f(n_1, v, n_2, k_2)} \quad (3)$$

f – järjendite arv treeningandmetes, milles funktsiooni argumentdega näidatud sõnad kattuvad klassifitseeritava järjendi vastavate sõnadega

f̄ – loeb kokku samad järjendid, aga mitmene sõna *n*₂ on lauses esinenud määrusena

p̄ – tõenäosus, et mitmene sõna *n*₂ on määrus

Seega näitavad tabeli väärtused, millise täpsusega töötab algoritm neil testkirjetel, mille kuuluvus määratakse taandamisetalpil, kus vastavad sagedused on kasutusel. Algoritmi valisin need kombinatsioonid, mis asuvad tabelis ridu eraldavast tumedast joonest ülalpool. Saadud algoritm on toodud joonisel 2. Sagedused, mida algoritm kasutab on toodud tutvumiseks lisa 3.

4 elementi		3 elementi		2 elementi		1 element	
<i>n</i> ₁ , <i>v</i> , <i>n</i> ₂ , <i>k</i> ₂	92,4%	<i>v</i> , <i>n</i> ₂ , <i>k</i> ₂	93,7%	<i>n</i> ₁ , <i>n</i> ₂	84,4%	<i>k</i> ₂	80,0%
<i>n</i> ₁ , <i>k</i> ₁ , <i>n</i> ₂ , <i>k</i> ₂	90,6%	<i>n</i> ₁ , <i>n</i> ₂ , <i>k</i> ₂	89,1%	<i>n</i> ₂ , <i>k</i> ₂	83,3%	<i>k</i> ₁	77,6%
<i>n</i> ₁ , <i>k</i> ₁ , <i>v</i> , <i>n</i> ₂	89,5%	<i>n</i> ₁ , <i>k</i> ₁ , <i>k</i> ₂	86,6%	<i>v</i> , <i>k</i> ₂	81,3%	<i>v</i>	70,6%
<i>k</i> ₁ , <i>v</i> , <i>n</i> ₂ , <i>k</i> ₂	89,4%	<i>n</i> ₁ , <i>k</i> ₁ , <i>n</i> ₂	86,5%	<i>k</i> ₁ , <i>k</i> ₂	81,3%	<i>n</i> ₂	70,0%
<i>n</i> ₁ , <i>k</i> ₁ , <i>v</i> , <i>k</i> ₂	86,3%	<i>k</i> ₁ , <i>n</i> ₂ , <i>k</i> ₂	85,6%	<i>n</i> ₁ , <i>k</i> ₂	78,1%	<i>n</i> ₁	68,4%
		<i>k</i> ₁ , <i>v</i> , <i>k</i> ₂	82,7%	<i>k</i> ₁ , <i>n</i> ₂	75,8%		
		<i>n</i> ₁ , <i>v</i> , <i>k</i> ₂	82,6%	<i>k</i> ₁ , <i>v</i>	73,9%		
		<i>k</i> ₁ , <i>v</i> , <i>n</i> ₂	79,4%	<i>n</i> ₁ , <i>k</i> ₁	73,6%		
		<i>n</i> ₁ , <i>v</i> , <i>n</i> ₂	79,0%	<i>n</i> ₁ , <i>v</i>	69,8%		
		<i>n</i> ₁ , <i>k</i> ₁ , <i>v</i>	76,0%	<i>v</i> , <i>n</i> ₂	69,3%		

Tabel 2: Mitmesuse järjendi komponentide panus algoritmi töö täpsusesse, pikem selgitus tekstis

if $f(n_1, k_1, v, n_2, k_2) > 0$

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(n_1, k_1, v, n_2, k_2)}{f(n_1, k_1, v, n_2, k_2)}$$

elseif $f(n_1, v, n_2, k_2) + f(n_1, k_1, n_2, k_2) + f(n_1, k_1, v, n_2) + f(k_1, v, n_2, k_2) > 0$

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(n_1, v, n_2, k_2) + \bar{f}(n_1, k_1, n_2, k_2) + \bar{f}(n_1, k_1, v, n_2) + \bar{f}(k_1, v, n_2, k_2)}{f(n_1, v, n_2, k_2) + f(n_1, k_1, n_2, k_2) + f(n_1, k_1, v, n_2) + f(k_1, v, n_2, k_2)}$$

elseif $f(v, n_2, k_2) + f(n_1, n_2, k_2) + f(n_1, k_1, k_2) + f(n_1, k_1, n_2) + f(k_1, n_2, k_2) > 0$

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(v, n_2, k_2) + \bar{f}(n_1, n_2, k_2) + \bar{f}(n_1, k_1, k_2) + \bar{f}(n_1, k_1, n_2) + \bar{f}(k_1, n_2, k_2)}{f(v, n_2, k_2) + f(n_1, n_2, k_2) + f(n_1, k_1, k_2) + f(n_1, k_1, n_2) + f(k_1, n_2, k_2)}$$

elseif $f(n_1, n_2) + f(n_2, k_2) + f(v, k_2) + f(k_1, k_2) > 0$

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(n_1, n_2) + \bar{f}(n_2, k_2) + \bar{f}(v, k_2) + \bar{f}(k_1, k_2)}{f(n_1, n_2) + f(n_2, k_2) + f(v, k_2) + f(k_1, k_2)}$$

else

$$\bar{p}(n_1, k_1, v, n_2, k_2) = \frac{\bar{f}(k_2)}{f(k_2)}$$

f – järjendite arvtreeningandmetes, milles funktsiooni argumentdega näidatud sõnad kattuvad klassifitseeritava järjendi vastavate sõnadega

\bar{f} – loeb kokku samad järjendid, aga mitmene sõna n_2 on lauses esinenud määrusena

\bar{p} – tõenäosus, et mitmene sõna n_2 on määrus

Joonis 2: Taandamismudeli algoritm

Taandamismudelis otsitakse kirje klassifitseerimiseks treeningnäidete hulgast võimalikult täpseid vasteid. Tabelis 2 toodud väärtused on tõenduseks sellise lähenemise õigsusest. Vaid teise veeru esimene lahter kolmikuga $\{v, n_2, k_2\}$ tekitab vastuolu. Väheema arvu kattuvuste lubamine on andnud parema tulemuse. Samuti on madal meetodi täpsus – 90,5% – neile järjenditele, millele leidub identseid vasteid treeningandmetes. Viimane väärtus annab aluse oletada, et korpuse märgenduses esineb vigu. Sellest tulenevalt katsetasin ka, mis muutub, kui loobuda täpsete vastete sageduste kasutamisest, st algoritm alustaks kirje klassifitseerimist kontrolliga, kas leidub treeningkirjeid, milles neli järjendi elementi on samad. Muutust tehes tõstsin ka kolmiku $\{v, n_2, k_2\}$ sageduste kasutamise algoritmi samasse osasse (eemaldasid nelikud, mis sisaldasid seda kolmikut).

Testimine andis esialgse algoritmi (joonisel 2) täpsuseks 86.5%. Kui testkirje mitmene sõna oli korpuse järgi lauses esinenud määrusena, eksis meetod 4,9%l juhtudest. Kui mitmene sõna oli kasutatud täiendina, oli veaprotsendiks 56,7%. Meetodi muudetud versiooni jaoks jäi kõiki kirjeid koos vaadates tulemus samaks – 86,4%, kuid veaprotsent tõusis täiendite jaoks 68,9%ni, langedes samas määruste jaoks 2,7%ni. Tabelites 3 ja 4 antud tulemused näitavad meetodi täpsust eri etappidel klassifitseeritud testkirjete jaoks. Viimast etappi, st. mitmesuse lahendamist treeningkirjete abil, milles on sama ainult mitmese sõna kääne, ei vajanud ükski testkirje.

<i>Kattuvused treening- ja testkirjes</i>	<i>Korrektus</i>	<i>Protsent kõigist testkirjetest</i>
5 komponenti	90.5%	4.0%
4 komponenti	90.1%	15.8%
3 komponenti	86,4%	50,2%
2 komponenti	84,1%	30,0%

Tabel 3: Testimistulemused, arvestades treening- ja testkirjetes kattuvate komponentide arvu

<i>Kattuvused treening- ja testkirjes</i>	<i>Korrektus</i>	<i>Protsent kõigist testkirjetest</i>
$\{v, n_2, k_2\}, \{n_1, k_1, n_2, k_2\}, \{n_1, k_1, v, n_2\}$	93,3%	27,2%
3 komponenti (va $\{v, n_2, k_2\}$)	86,2%	42,4%
2 komponenti	81,1%	30,4%

Tabel 4: Testimistulemused, taandamisetaapid muudetud

5. Järeldused ja plaanid edaspidiseks

Katsed kahe statistilise meetodi rakendamisel eesti keele nimisõnalise täiendi ja määruse eristamiseks andsid väga sarnase tulemuse: meetodi täpsus oli 85,5-85,7%. Määruse analüüsil eksiti harvem kui täiendi analüüsil, kuid see on selgitatav asjaoluga, et määrusi esineb korpuses rohkem. Kui transformatsioonipõhise analüüsi eeliseks on see, et süsteem genereerib loetavad reeglid, mida on võimalik käsitsi muuta (vt lisa 2), siis ka taandamismudeli genereeritud sageduste tabel (vt lisa 3) on tänuväärt materjal lingvistilisele uurimistööle.

Kindlasti oleks vaja katsetusi jätkata nii töös käsitletud transformatsioonipõhise kui ka taandamismudeliga treeningandmete ekstraktimise meetoodika parandamisega:

1. Kontrollida, kas treeningtulemused muutuvad, kui eristada treeningandmetes obliikvakäändelisi nimisõnu nimetavas, omastavas ja osastavas käändes olevaist. Kui obliikvakäändelised nimisõnad esinevad sagedamini määrusena, siis mitteobliikvakäändelised on enamasti täiendi rollis.
2. Kontrollida, kuidas mõjutab ees- ja järeltäiendi funktsiooni teadmine tulemust. Kahjuks ei ole võimalik pindmise märgenduse põhjal alati täpselt siduda täiendit tema põhjaga, selleks oleks vajalik teada nimisõnafraasi struktuuri. Praeguses käsitluses seoti täiend lähima võimaliku põhjaga.
3. Arvestada osalause piire täpsemini treeningandmetes, samuti leida lahendus partitsiipidega seotud määruste analüüsiks.
4. Vaadata üle verbikesksete püsiühendite liigid, mida töös arvestada. Praegu kasutatav andmebaas sisaldab ka kollokatsioone.
5. Kontrollida, kuidas mõjutab juhendamata õppimine süsteemi käitumist (õppida automaatselt märgendatud korpusest).
6. Ühildada süsteem eesti wordnetiga, et genereerida üldisemaid reegleid. Võibolla on vajalik lisaks nimede üldistamisele asendada ka põhiarvsõnad ühe märgendiga. Vajalik võiks olla ka wordnetist saadav informatsioon, kas nimisõnaga tähistatav on elus või eluta.
7. Katset võiks laiendada ka kaassõnafraasidele, kuigi on teada, et kaassõnafraas esineb täiendina veelgi harvem [Müürisep, 2000].
8. Kuigi kasutatud treeningkorpust on märgendanud kaks lingvistilt paralleelselt ning need lõpuks ühtlustatud, võib korpuses siiski esineda vigu. Oleks vajalik sagedasemate nelikute märgendus veel kord süstemaatiliselt üle kontrollida.

6. Kokkuvõte

Magistritöö tegeles alamprobleemiga automaatses süntaksianalüüsis – nimisõnaliste täiendite ja määruste eristamisega. Erilist tähelepanu vajab see ülesanne põhjusel, et üldjuhul sõltub korrektne analüüs sõnade reksioonist või semantikast. Kahe lauseliikme eristamine ei ole seetõttu süntaktiliste reeglite abil sageli võimalik. Analoogne probleem eksisteerib inglise keeles, kus on mitmesuste põhjustajaks prepositsioonifraas PP, mis võib olla samuti lauses nimisõna või verbi laiendiks. Inglise keelele on PP kuuluvuse määramisel parimaid tulemusi saavutatud statistilist lähenemist kasutades. Käesoleva töö teoreetilises osas on toodud ülevaate sel suunal tehtud uurimustööst. Magistritöö praktiline osa seisnes kahe algoritmi rakendamises eesti keelele: transformatsioonipõhine meetod ning taandamismudel.

Algoritmide kasutamisele eelnes treeningmaterjali eraldamine eesti keele süntaktiliselt märgendatud korpusest. Eeltöö jagunes järgmisteks etappideks:

- (i) süntaksianalüsaatori väljundist mitmese analüüsi saanud nimisõnade leidmine;
- (ii) osalausepiiride määramine, et teha kindlaks verb ja nimisõna, mille laiendiks mitmene sõna oma vormi poolest olla võis;
- (iii) hindamine, kas lauselühendite esinemine segab verbi leidmist;
- (iv) hindamine, kas verb on osa püsiühendist.

Meetoditega saavutatud tulemused oluliselt ei erinenud: transformatsioonipõhine meetod töötas testkirjetele 85.7%lise korrektsusega, taandamismudeli jaoks oli vastavaks väärtuseks 86.5%. Mõlemad algoritmid olid edukamad määruste leidmisel, vigade osakaal vastavalt 5,6% ja 4,9% (modifitseeritud algoritmiga 2,7%). Samas täiendite puhul olid veaprotsendid vastavalt 61,5% ja 56,7% (68,9%).

Using statistical methods to disambiguate between nominal attributes and adverbials

**Master thesis
Aivi Kaljuvee
Abstract**

The thesis targeted a problem in automatic syntax analysis – disambiguating between attributes and adverbials. The topic needs special attention as typically the correct analysis depends on government and words' semantics and syntactic rules are unable to describe these restrictions. There is an analogous problem in English that involves prepositional phrases (PP). For the same reasons the task of determining the correct attachment site – either noun or verb attachment – for a PP is a difficult task. The problem has been extensively explored for English and statistical methods have proven to be most successful. The thesis provides an overview of the research conducted. The practical task in this thesis was to experiment with applying two algorithms to Estonian, the chosen algorithms were the transformation-based learning method and back-off model.

The first step was to extract training data from the syntactically annotated corpus for Estonian. The preparation consisted of several subtasks:

- (i) using the output from syntactic parser to determine ambiguous nouns;
- (ii) determining the clause where the verb and noun are located to which the ambiguous noun possibly is attached to;
- (iii) eliminating the verb component from a training instance if the clause contains non-finite constructions;
- (iv) finding multi-word verbs.

The results obtained with the two methods do not differ considerably – accuracy for transformation based learning is 85.7%, the value for back-off model is 86.5%. The algorithms were both more successful with adverbials – error rate 5.6% for the transformation-based learning method and 4.9% (2.7% for a modified version of the algorithm) for the back-off model. For attributes the values were 61,5% and 56,7% (68,9%) respectively.

Viited

1. S. Abney, R.E. Schapire, Y. Singer. 1999. *Boosting Applied to Tagging and PP-attachment*. Kogumikus *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, lk 38–45
2. E. Brill, P. Resnik. 1994. *A rule-based approach to prepositional phrase attachment disambiguation*. Kogumikus *Proceedings of 15th annual conference on Computational Linguistics*, lk 1198–1204
3. E. Charniak. 2000. *A Maximum-Entropy-Inspired Parser*. Kogumikus *Proceedings of the North American Chapter of the Association for Computational Linguistics-2000*, lk 132–139
4. M. Collins, J. Brooks. 1995. *Prepositional phrase attachment through a backed-off model*. Kogumikus D. Yarowsky, K. Church (toimetajad), *Proceedings of the 3rd Workshop on Very Large Corpora*, lk 27-38, Cambridge, Massachusetts
5. C. Cortes, V. Vapnik. 1995. *Support-Vector Networks*. Kogumikus *Machine Learning* 20, 3, lk 273–297
6. M. Ereht, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare. 1993. *Eesti keele grammatika. II Süntaks*. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.
7. M. Ereht, T. Ereht, K. Ross. 1997. *Eesti keele käsiraamat*. Tallinn: Eesti Keele Sihtasutus.
8. S. Katz. 1987. *Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser*. Kogumikus *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35, 3
9. K. Mihkla, L. Rannut, E. Riikoja, A. Admann. 1974. *Eesti keele lauseõpetuse põhijooned I*. Tallinn „Valgus“.
10. T. M. Mitchell. 1997. *Machine Learning*, peatükk 6, lk 52-60, McGraw-Hill
11. K. Müürisep. 2000. *Eesti keele arvutigrammatika: süntaks*. *Dissertationes Mathematicae Universitatis Tartuensis* 22
12. M. Olteanu, D. Moldovan. 2005. *PP-attachment disambiguation using large context*. Kogumikus *Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, lk 273–280, Vancouver, Canada

13. J. R. Quinlan. 1986. *Induction of Decision Trees*. Kogumikus *Machine Learning* 1, 1, lk 81-106
14. A. Ratnaparkhi, J. Reynar, S. Roukos. 1994. *A Maximum Entropy Model for Prepositional Phrase Attachment*. Kogumikus *Proceedings of the Human Language Technology Workshop*, lk 250–255
15. J. Stetina, M. Nagao. 1997. *Corpus based PP attachment ambiguity resolution with a semantic dictionary*. Kogumikus J. Zhou, K. Church (toimetajad), *Proceedings of the 5th Workshop on Very Large Corpora*, lk 66-80, Peking, Hongkong
16. H. Uibo, E. Bick. 2005. *Treebank-based research and e-learning of Estonian syntax*. Kogumikus M. Langemets, P. Penjam (toimetajad), *Proceedings of Second Baltic Conference on Human Language Technologies*. lk. 195–200, Tallinn
17. M. Volk. 2001. *The automatic resolution of prepositional phrase attachment ambiguities in German*. Habilitationsschrift, Zürichi Ülikool
18. J. Zavrel, W. Daelemans, J. Veenstra. 1997. *Resolving PP attachment Ambiguities with Memory-Based Learning*. Kogumikus *Proceedings of the Conference on Computational Natural Language Learning*, lk 136-144, Madrid, Hispaania
19. S. Zhao, D. Lin. 2004. *A Nearest-Neighbor Method for Resolving PP-Attachment Ambiguity*. Kogumikus *Proceedings of the 1st International Joint Conference on Natural Language Processing*, lk 428–434 Sanya, Hiina

Lisad

Lisa 1. CD: Andmete ettevalmistamiseks kasutatud skriptid. Treeningandmed.

Lisa 2. Transformatsioonipõhise meetodiga leitud reeglid

Lisa 3. Sagedused taandamismudel

Lisa 1. CD: Andmete ettevalmistamiseks kasutatud skriptid. Treeningandmed.

Lisa 2. Transformatsioonipõhise meetodiga leitud reeglid

Toodud reeglid on treenimisel eraldatud vähemalt kahe andmekirje põhjal.

n2 on täiend, kui k2=omastav 79
n2 on määrus, kui n2=kaks 15
n2 on määrus, kui n2=paar 15
n2 on täiend, kui n1=mees, k2=kaasaütlev 13
n2 on määrus, kui n1=tema 8
n2 on määrus, kui n2=minut 7
n2 on määrus, kui n2=aeg 7
n2 on täiend, kui n1=mälestus 7
n2 on täiend, kui n2=puu, k2=seestütlev 7
n2 on määrus, kui n2=pudel 6
n2 on määrus, kui n1=oma k2=omastav 6
n2 on täiend, kui n2=milline 6
n2 on täiend, kui n1=lahkumine 6
n2 on määrus, kui n2=hulk 5
n2 on määrus, kui n2=neli 5
n2 on määrus, kui n1=minut 5
n2 on täiend, kui n2=see, k2=nimetav 5
n2 on täiend, kui n1=tüdruk, n2=juus 5
n2 on täiend, kui n1=liit, n2=NIMI 5
n2 on määrus, kui n1=see 4
n2 on määrus, kui n1=mina 4
n2 on määrus, kui n1=sina 4
n2 on määrus, kui v=istuma, k2=omastav 4
n2 on määrus, kui n1=kroon, n2=miljon 4
n2 on täiend, kui v=olema, n1=inimene 4
n2 on täiend, kui v=olema, n1=kulu
n2 on täiend, kui n1=erastamine, k2=kaasaütlev 4
n2 on täiend, kui n2=sekund, k2=seesütlev 4
n2 on täiend, kui n1=vang, n2=liige 4
n2 on täiend, kui v=olema, n2=NIMI, k2=sisseütlev 4
n2 on määrus, kui n2=õhtu 3
n2 on määrus, kui n1=tund 3
n2 on määrus, kui n1=leib 3
n2 on täiend, kui n2=looline 3
n2 on täiend, kui n2=erakondlane 3
n2 on täiend, kui n1=pääs 3

n2 on täiend, kui n1=jää=mine 3
n2 on täiend, kui n1=viide 3
n2 on täiend, kui n1=käimine 3
n2 on täiend, kui v=viskama, n1=mäng 3
n2 on täiend, kui n1=minister, k2=ilmaütlev 3
n2 on täiend, kui n1=leht, k2=kaasaütlev 3
n2 on täiend, kui n1=ohvitser, k2=kaasaütlev 3
n2 on täiend, kui n1=poiss, k2=alalütlev 3
n2 on täiend, kui n1=press, k2=seestütlev 3
n2 on täiend, kui n2=mitu, k2=osastav 3
n2 on täiend, kui n2=ala, k2=seestütlev 3
n2 on määrus, kui n2=pea 2
n2 on määrus, kui n2=koht 2
n2 on määrus, kui n2=pilk 2
n2 on määrus, kui n2=sada 2
n2 on määrus, kui n2=kimp 2
n2 on määrus, kui n2=pakk 2
n2 on määrus, kui n2=kolm_kümmend 2
n2 on määrus, kui v=loosi_tõmbama 2
n2 on määrus, kui v=veetma 2
n2 on määrus, kui n1=selline 2
n2 on määrus, kui n1=meeskond 2
n2 on määrus, kui n1=maa 2
n2 on määrus, kui n1=huvi 2
n2 on määrus, kui v=vaatama, k2=omastav 2
n2 on määrus, kui v=olema, n1=tege=mine 2
n2 on määrus, kui v=laekuma, n1=kuu 2
n2 on määrus, kui n1=kord, k2=omastav 2
n2 on täiend, kui n2=habe 2
n2 on täiend, kui n2=paneel 2
n2 on täiend, kui n2=kallak 2
n2 on täiend, kui n2=km 2
n2 on täiend, kui n2=kuppel 2
n2 on täiend, kui n2=täi 2
n2 on täiend, kui n2=dollar 2
n2 on täiend, kui n2=maalane 2
n2 on täiend, kui n2=keiser 2
n2 on täiend, kui n2=mugavus 2
n2 on täiend, kui n2=itaallane 2

n2 on täiend, kui n2=metall 2
n2 on täiend, kui n2=kont 2
n2 on täiend, kui n2=Ühenda=tud 2
n2 on täiend, kui n2=püksid 2
n2 on täiend, kui v=valu_tekitama 2
n2 on täiend, kui v=ilma_tegema 2
n2 on täiend, kui v=signaali_andma 2
n2 on täiend, kui v=kiitma 2
n2 on täiend, kui v=teistest_erinema 2
n2 on täiend, kui v=pätsima 2
n2 on täiend, kui v=üles_tegema 2
n2 on täiend, kui v=ära_tooma 2
n2 on täiend, kui n1=osalus 2
n2 on täiend, kui n1=tomat 2
n2 on täiend, kui n1=ümbrik 2
n2 on täiend, kui n1=astu=mine 2
n2 on täiend, kui n1=kuuldus 2
n2 on täiend, kui n1=suhtle=mine 2
n2 on täiend, kui n1=valdkond 2
n2 on täiend, kui n1=plakat 2
n2 on täiend, kui n1=paraad 2
n2 on täiend, kui n1=lõiga=tu 2
n2 on täiend, kui n1=vastus 2
n2 on täiend, kui n1=põgene=mine 2
n2 on täiend, kui n1=vestlus 2
n2 on täiend, kui v=tagasi_lükkama, k2=seestütlev 2
n2 on täiend, kui v=tuginema, k2=alalütlev 2
n2 on täiend, kui v=elama, k2=osastav 2
n2 on täiend, kui v=naasma, k2=adit 2
n2 on täiend, kui v=mõjutama, k2=kaasaütlev 2
n2 on täiend, kui v=voolama, k2=kaasaütlev 2
n2 on täiend, kui v=kõndima, k2=seesütlev 2
n2 on täiend, kui v=lükkama, k2=alalütlev 2
n2 on täiend, kui v=meenuma, k2=seestütlev 2
n2 on täiend, kui v=tähendama, k2=seestütlev 2
n2 on täiend, kui n1=kujutus, v=olema 2
n2 on täiend, kui n1=lisa, v=saama 2
n2 on täiend, kui n1=lehm, v=ära_tulema 2
n2 on täiend, kui n1=jooks, v=pakkuma 2

n2 on täiend, kui n1=teadmine, v=puuduma 2
n2 on täiend, kui n1=märk, v=olema 2
n2 on täiend, kui n1=küsimus, v=olema 2
n2 on täiend, kui n1= põgenik, v=korraldama 2
n2 on täiend, kui n1=hääl, k2=seestütlev 2
n2 on täiend, kui n1=majandus, k2=seestütlev 2
n2 on täiend, kui n1=kuju, k2=seesütlev 2
n2 on täiend, kui n1=minek, k2=alaleütlev 2
n2 on täiend, kui n1=minek, k2=sisseütlev 2
n2 on täiend, kui n1=võtja, k2=seestütlev 2
n2 on täiend, kui n1=lill, k2=kaasaütlev 2
n2 on täiend, kui n1=sina, k2=alalütlev 2
n2 on täiend, kui n1=kontroll, k2=alalütlev 2
n2 on täiend, kui n1=keegi, k2=seestütlev 2
n2 on täiend, kui n1=kõnelus k2=kaasaütlev 2
n2 on täiend, kui v=teatama, n2=rinne 2
n2 on täiend, kui v=olema, n2=staadion 2
n2 on täiend, kui v=panema, n2=ühiskond 2
n2 on täiend, kui v=mainima, n2=NIMI 2
n2 on täiend, kui n2=toime, k2=kaasaütlev 2
n2 on täiend, kui n2=kümme, k2=osastav 2
n2 on täiend, kui n2=ehitus, k2=kaasaütlev 2
n2 on täiend, kui n2=minister, k2=nimetav 2
n2 on täiend, kui n2=aken, k2=seestütlev 2
n2 on täiend, kui n2=nädal, k2=seesütlev 2
n2 on täiend, kui n2=tegija, k2=saav 2
n2 on täiend, kui n1=palju, n2=tema 2
n2 on täiend, kui n1=elu, n2=NIMI 2
n2 on täiend, kui n1=vend, n2=küla 2
n2 on täiend, kui n1=tund, n2=päev 2
n2 on täiend, kui n1=lisa, n2=NIMI 2
n2 on täiend, kui n1=olu, n2=seadus 2
n2 on täiend, kui v=olema, n2=elu, k2=kaasaütlev 2
n2 on täiend, kui n1=NIMI, n2=mees, k2=nimetav 2

Lisa 3. Sagedused taandamismudelis

Esitatud on nende järjendite sagedused, millele leidis treeningandemetes rohkem kui 5 näidet. Järjendi juures olevatest väärtustest esimene näitab kordade arvu, mil mitmene sõna oli lauses täiendiks, teine number kordade arvu, mil see oli määruseks.

n₁; v; n₂; k₂
olema;pluss;kaks;nom 0 17
olema;NIMI;sõna;ad 1 11
olema;mis;tema;ad 0 7
õitsema;august;juuni;el 0 7
paljundama;jaga=mine;puhmas;gen 7 0
õitsema;juuni;august;term 0 6
olema;iga;juht;ad 0 6

v; n₂; k₂	
olema;NIMI;in 7 31	õppima;kool;in 0 8
olema;kaks;nom 0 32	olema;aeg;nom 0 8
olema;tema;ad 2 24	olema;see;nom 6 2
olema;sõna;ad 2 17	paljundama;puhmas;gen 7 0
olema;NIMI;kom 8 10	avama;NIMI;#cap 1 6
olema;aasta;gen 11 4	üttelema;NIMI;in 1 6
meelde_tuletama;meel;adit 0 13	olema;amet;in 0 7
olema;tuba;in 0 12	üttelema;NIMI;#cap 7 0
istutama;aprill;in 0 11	olema;tema;gen 6 1
õitsema;juuni;el 0 11	olema;haigla;in 1 6
olema;käsi;in 1 10	olema;NIMI;ad 1 6
olema;NIMI;#cap 7 4	olema;kord;nom 0 7
olema;juht;ad 0 10	sõdima;NIMI;kom 0 6
olema;kodu;in 0 10	olema;kiri;in 0 6
olema;üks;nom 4 5	olema;kolm;gen 3 3
olema;mina;ad 0 9	olema;kümme;nom 1 5
olema;kolm;nom 0 9	olema;liit;in 0 6
saama;NIMI;el 0 9	olema;ilm;in 0 6
pähe_tulema;pea;adit 0 9	olema;see;#cap 6 0
olema;tema;el 3 6	koolis_käima;kool;in 0 6

silmas_pidama;silm;in 0 9	nägema;NIMI;in 0 6
olema;NIMI;el 3 6	unes_nägema;uni;in 0 6
istutama;mai;in 0 9	tegema;NIMI;in 1 5
olema;sina;ad 0 8	külla_tulema;küla;adit 0 6
olema;elu;in 0 8	õitsema;september;term 0 6
õitsema;august;term 0 8	viima;NIMI;ill 0 6
olema;cm;null 0 8	olema;kõla;in 1 5
kätte_võtma;käsi;adit 1 7	

n₁; n₂; k₂	
NIMI;NIMI;el 9 41	tüdruk;juus;kom 7 0
NIMI;sõna;ad 1 49	jaga=mine;puhmas;gen 7 0
NIMI;NIMI;in 6 32	NIMI;rand;in 0 7
NIMI;NIMI;all 5 27	kogu;NIMI;#cap 7 0
NIMI;NIMI;kom 6 21	NIMI;osa;in 0 7
NIMI;NIMI;ill 2 21	NIMI;kinnitus;ad 0 7
sama;aeg;ad 0 21	NIMI;riik;in 4 3
iga;juht;ad 1 17	kõik;kolm;nom 0 7
pluss;kaks;nom 0 17	juuni;august;term 0 7
NIMI;kool;in 0 13	NIMI;vald;in 2 5
NIMI;3;digit 0 12	meeter;sekund;in 5 2
kõrgus;cm;null 0 12	tund;neli;nom 0 6
NIMI;tema;all 0 12	arve;aasta;gen 5 1
mis;tema;ad 0 12	NIMI;kaks;nom 0 6
mis;tema;all 0 12	algus;aasta;gen 6 0
NIMI;2;digit 0 11	NIMI;hinnang;ad 0 6
mees;nägu;kom 11 0	aeg;sama;#cap 6 0
NIMI;NIMI;abl 3 8	oma;aeg;ad 0 6
üks;päev;ad 0 11	sugukond;cm;null 0 6
kroon;miljon;gen 1 10	klass;NIMI;#cap 3 3
see;juht;ad 0 10	NIMI;NIMI;nom 2 4
tema;sõna;ad 0 10	tund;paar;nom 0 6
see;hetk;ad 0 10	päev;hommik;ad 0 6
mis;see;el 0 10	september;kaugus;kom 0 6
NIMI;linn;in 1 9	kord;elu;in 3 3
mees;kittel;kom 9 0	kaugus;september;in 0 6
see;aeg;ad 0 9	NIMI;mees;all 1 5

NIMI;NIMI;#cap 7 2	leping;NIMI;kom 1 5
NIMI;NIMI;ad 0 9	august;kaugus;kom 0 6
meeter;kaugus;ad 0 8	aasta;viis;gen 5 1
aasta;NIMI;in 1 7	see;nädal;ad 0 6
aasta;kaks;nom 0 8	nädal;kaks;nom 0 6
sama;hetk;ad 0 8	aeg;tükk;nom 0 6
august;juuni;el 0 8	oma;koht;all 3 3
vihka=mine;kaks;gen 1 7	kuu;viis;gen 3 3
iga;päev;nom 0 7	aasta;viis;nom 0 6
NIMI;voodi;el 0 7	mina;NIMI;kom 0 6
sõrm;mitu;nom 0 7	taim;cm;null 0 6
aasta;algus;in 0 7	aasta;kolm;nom 0 6
algus;sajand;gen 7 0	see;mina;all 0 6

n₁;k₁;k₂		
NIMI;gen;in 16 157	teine;el;el 0 12	NIMI;gen;term 1 6
NIMI;gen;ad 6 134	leping;gen;kom 0 12	oma;null;el 0 7
NIMI;gen;el 21 76	tema;gen;adit 0 12	NIMI;part;in 0 7
NIMI;gen;all 8 71	mina;gen;el 0 12	inimene;gen;el 1 6
NIMI;nom;all 9 60	aeg;gen;gen 2 10	mees;nom;el 0 7
NIMI;gen;kom 9 60	päev;gen;ad 1 11	osa;part;el 0 7
kroon;part;digit 7 53	miski;part;in 0 12	maa;gen;ad 0 7
NIMI;nom;kom 5 50	tuba;adit;el 0 12	tee;gen;in 0 7
aasta;part;nom 1 53	dollar;part;digit 1 11	jaga=mine;gen;gen 7 0
tema;nom;kom 1 50	päev;part;nom 1 10	kõik;el;el 0 7
see;ad;ad 4 47	oma;null;kom 1 10	sama;null;ad 0 7
NIMI;nom;digit 2 46	NIMI;nom;abl 2 9	mees;nom;gen 6 1
NIMI;nom;el 8 33	oma;gen;gen 3 8	mees;all;kom 6 1
NIMI;gen;digit 2 36	meeter;gen;ad 0 11	ise;null;el 0 7
NIMI;gen;adit 1 34	riik;gen;in 2 9	naine;gen;in 0 7
NIMI;nom;in 5 29	tüdruk;nom;kom 9 2	leping;gen;ad 0 7
oma;gen;el 1 31	juuni;el;term 0 11	laps;all;el 0 7
NIMI;gen;ill 1 31	aasta;part;in 0 11	tema;null;all 0 7
oma;gen;kom 4 28	mis;null;ad 0 11	päev;ad;in 1 6
NIMI;nom;ad 0 30	tema;gen;kom 0 11	mõte;nom;el 0 6
tema;gen;in 1 28	miski;part;el 0 11	käsi;adit;el 0 6
tund;part;nom 0 28	kell;nom;digit 0 11	sugukond;adit;null 0 6

tema;gen;el 2 26	too;ad;ad 0 11	sõna;part;nom 0 6
mis;nom;ad 0 28	mõni;gen;kom 0 10	NIMI;all;tr 0 6
NIMI;gen;tr 0 28	kroon;part;nom 0 10	miski;part;adit 0 6
mees;nom;kom 25 2	tema;gen;ad 1 9	mees;part;el 1 5
oma;gen;in 1 26	oma;refl;el 0 10	pilk;gen;all 0 6
iga;ad;ad 0 27	NIMI;ill;all 1 9	mööde;kom;digit 0 6
üks;ad;ad 0 27	kuu;part;nom 0 10	poiss;nom;kom 3 3
aasta;gen;gen 8 19	kroon;gen;gen 2 8	seadus;gen;ad 1 5
NIMI;gen;nom 2 24	see;part;all 0 10	asi;nom;in 1 5
see;gen;kom 2 24	minut;part;nom 0 10	tema>null;ad 0 6
iga;nom;nom 0 26	kes;nom;all 0 10	õhtu;ad;in 1 5
NIMI;nom;nom 11 14	see;nom;all 0 10	NIMI;el;kom 0 6
mis;nom;all 0 25	mõni;in;in 0 9	hind;nom;digit 0 6
NIMI;gen;abl 4 20	NIMI;all;el 0 9	mina;part;kom 0 6
sama;ad;ad 0 24	mees;nom;in 3 6	ministeerium;gen;in 2
oma;gen;all 1 22	miski;part;kom 0 9	4
kroon;gen;digit 11 12	NIMI;ad;el 0 9	maa;ad;term 0 6
lisa;tr;all 12 10	teine>null;in 0 9	riik;gen;all 1 5
see;nom;nom 0 21	kroon;gen;in 0 9	taim;nom>null 0 6
see;in;in 4 17	tema;#cap;ad 0 9	minut;gen;gen 2 4
NIMI;in;in 1 20	naine;nom;kom 5 4	vihka=mine;gen;tr 0 6
aasta;gen;in 1 19	see>null;el 1 8	pea;gen;el 0 6
mis;nom;el 0 20	mis;nom;nom 0 9	mingi;gen;kom 0 6
kord;part;nom 0 20	kõik;nom;in 0 9	NIMI;el;all 0 6
mina;nom;kom 1 19	mees;part;kom 7 2	aken;gen;adit 0 6
see;el;el 1 18	aeg;ad;#cap 9 0	inimene;part;nom 0 6
pluss;nom;nom 0 17	oma;gen;adit 1 8	september;term;el 0 6
aru;part;el 0 17	NIMI;nom;ill 0 8	rubla;part;nom 0 6
mis;nom;in 0 17	algus;in;gen 8 0	inimene;gen;in 0 6
NIMI;gen;gen 10 7	vihka=mine;gen;gen 1 7	jutt;gen;el 0 6
NIMI;all;adit 2 15	aasta;part;digit 1 7	kord;part;kom 1 5
NIMI;nom;adit 1 16	august;term;el 0 8	oma>null;all 3 3
aasta;gen;el 1 15	mees;part;nom 0 8	partei;gen;kom 0 6
oma>null;in 2 14	kell;nom;nom 0 8	maa;gen;in 1 5
kord;part;in 2 14	NIMI;part;kom 0 8	see;nom;ad 0 6
NIMI;nom;gen 8 8	aeg;part;in 0 8	pea;gen;kom 0 6
teine;in;in 2 14	NIMI;ad;in 2 6	mina;gen;kom 0 6
meeter;gen;gen 8 8	maa;gen;el 1 7	mees;gen;adit 0 6
see;#cap;ad 0 15	see;all;all 0 8	laps;gen;all 0 6

NIMI;ill;el 0 15	mina;gen;in 1 7	iga>null;ad 1 5
tema;gen;all 0 15	käsi;gen;all 0 8	sama;in;in 0 6
NIMI;kom;in 0 14	üks;el;el 0 8	mõni;nom;nom 0 6
mingi;ad;ad 0 14	sõrm;part;nom 0 8	kahju;nom;el 0 6
mis;nom;kom 0 14	NIMI;in;kom 1 7	oma;gen;ill 0 6
tema;part;kom 0 14	tema;nom;in 0 8	taim;nom;digit 0 6
protsent;part;digit 1 13	raamat;gen;el 0 8	nägu;gen;kom 1 5
NIMI;in;el 1 13	meeter;gen;digit 5 3	hää;nom;el 5 1
NIMI;el;ill 1 13	nädal;part;nom 0 8	see;part;ad 0 6
aeg;part;nom 0 14	liblikas;part;digit 0 8	mina;gen;all 0 6
üks;in;in 1 12	meeter;part;digit 1 7	kes>null;in 0 6
kord;nom;in 4 9	kroon;term;digit 0 8	tonn;part;digit 1 5
oma;gen;ad 0 13	aasta;ad;digit 4 4	sina;nom;kom 0 6
kaugus;kom>null 13 0	partei;gen;el 1 7	õu;el;adit 0 6
mis>null;in 0 13	kogu;gen;ad 1 7	august;in;kom 0 6
see>null;ad 1 12	nädal;gen;ad 1 7	linn;gen;kom 0 6
see>null;in 0 13	üks>null;in 0 8	maja;gen;ad 0 6
kaugus;kom;in 0 13	kuu;gen;gen 4 4	leht;gen;el 0 6
mees;gen;el 1 12	koht;all;in 0 8	inimene;nom;in 3 3
tund;part;in 3 10	oma;refl;kom 0 7	september;in;kom 0 6
NIMI;in;ad 0 13	iga;in;in 0 7	kodu;adit;el 1 5
kroon;part;in 1 11	see;gen;el 2 5	kroon;kom;digit 0 6
aasta;gen;ad 1 11	oma;gen;nom 0 7	ise;ad;ad 0 6
tund;gen;gen 1 11	samm;part;nom 0 7	plaan;el;digit 0 6
kõrgus;nom>null 0 12	iga>null;nom 0 7	mina>null;in 0 6
kõik;nom;nom 2 10	tema>null;in 1 6	jutt;nom;el 5 1
meeter;part;nom 0 12	voodi;el;all 0 7	tema;gen;nom 1 5
kõrgus;nom;digit 0 12	NIMI;el;adit 1 6	mis;gen;ad 0 6

n₁; k₁; n₂	
NIMI;nom;NIMI 24 79	NIMI;gen;aasta 1 6
NIMI;gen;NIMI 13 39	tüdruk;nom;juus 7 0
NIMI;gen;sõna 1 49	august;term;juuni 0 7
mis;nom;tema 0 24	NIMI;nom;2 0 7
NIMI;el;NIMI 4 15	hää;nom;ekraan 4 3
NIMI;ill;NIMI 1 17	NIMI;gen;valitsus 1 6

pluss;nom;kaks 0 17	aasta;part;viis 0 7
NIMI;nom;tema 0 15	mis;nom;NIMI 0 7
NIMI;gen;kool 1 14	NIMI;gen;rand 0 7
iga;ad;juht 0 15	sõrm;part;mitu 0 7
NIMI;gen;riik 6 7	NIMI;gen;tänav 1 6
mis;nom;see 0 13	see;nom;mina 0 7
kõrgus;nom;cm 0 12	jaga=mine;gen;puhmas 7 0
vihka=mine;gen;kaks 1 11	see;ad;aeg 0 7
NIMI;gen;vald 2 10	NIMI;gen;kinnitus 0 7
sama;ad;aeg 0 12	kaugus;kom;september 0 6
NIMI;gen;linn 0 11	leping;gen;sõlmi=mine 0 6
tema;gen;nägu 0 11	see;#cap;nädal 0 6
see;ad;hetk 0 10	aeg;ad;sama 6 0
üks;ad;päev 0 10	nädal;part;kaks 0 6
partei;gen;liige 0 10	aasta;ad;NIMI 1 5
mis;nom;sina 0 10	aeg;part;tükk 0 6
meeter;gen;kaugus 0 9	tund;part;paar 0 6
NIMI;gen;osa 1 8	päev;gen;hommik 0 6
aasta;gen;algus 0 9	september;in;kaugus 0 6
NIMI;nom;3 0 9	aasta;part;kolm 0 6
mõte;nom;pea 0 9	mees;gen;nägu 3 3
kroon;gen;miljon 1 8	august;in;kaugus 0 6
hobune;gen;selg 0 8	leping;gen;NIMI 0 6
tema;#cap;sõna 0 8	taim;nom;cm 0 6
riie;nom;selg 0 8	NIMI;gen;hinnang 0 6
sama;ad;hetk 0 8	mees;nom;nägu 6 0
aasta;part;kaks 0 8	NIMI;gen;aeg 0 6
NIMI;gen;turg 2 6	sugukond;adit;cm 0 6
NIMI;gen;nägu 1 7	NIMI;gen;mees 0 6
juuni;el;august 0 7	NIMI;gen;saar 0 6
NIMI;nom;mina 0 7	NIMI;nom;mees 3 3
NIMI;in;NIMI 0 7	tund;part;neli 0 6

n₂; k₂			
NIMI;in 42 296	selg;adit 0 16	kinnitus;ad 0 9	hüpe;in 0 7
NIMI;#cap 98 82	aken;el 0 15	mäng;in 0 9	300;digit 1 6
NIMI;el 22 144	kogu;in 0 15	kõik;all 3 6	trepp;el 0 7

kaks;nom 1 134	mina;kom 1 14	august;term 0 9	kr;null 0 7
NIMI;kom 34 97	kool;adit 0 15	kolm_kümmend;nom	külg;all 0 7
NIMI;ill 12 87	pihk;adit 0 15	0 9	kell;nom 0 7
NIMI;all 13 83	uks;el 0 15	laps;all 1 8	1996;digit 1 6
tema;all 1 83	külaline;all 0	3000;digit 0 9	aasta;el 1 6
sõna;ad 2 81	15	huvi;in 0 9	mälu;in 0 7
aasta;gen 65 17	maa;abl 2 13	kittel;kom 9 0	turg;all 2 5
kolm;nom 2 79	hulk;nom 0 15	komme;ad 0 9	kiirus;kom 0 7
NIMI;gen 63 13	juuni;el 0 15	ema;kom 1 8	valitsus;all 1 6
aeg;ad 0 69	päev;in 3 12	saamine;tr 2 7	sama;#cap 7 0
kodu;adit 0 58	kr/kg;null 0 15	jalg;el 0 9	kiri;kom 3 4
kord;nom 2 55	asi;el 0 15	60;digit 0 9	leping;gen 1 6
tuba;adit 1 51	keel;in 0 15	torn;el 0 9	kaks_kümmend;nom
päev;ad 0 52	aprill;in 0 15	peo;in 0 9	0 7
paar;nom 0 51	saal;adit 0 15	sõda;in 2 7	nädal;nom 0 7
2;digit 4 46	15;digit 4 11	seis;in 0 9	16;digit 1 6
cm;null 0 49	maa;el 2 13	kael;in 0 9	laud;el 2 5
tema;el 10 39	seadus;kom 3 12	teadvus;in 0 9	kontor;el 0 7
mina;all 0 47	laud;ad 3 12	ise;refl 0 9	poliitika;in 1 6
käsi;in 5 40	hinnang;ad 0 15	voodi;in 0 9	vald;adit 0 6
aasta;ad 2 43	kivi;kom 0 14	keskus;el 0 9	uks;all 0 6
tema;gen 40 4	kasum;el 0 14	lõpp;ad 0 9	valdkond;in 0 6
kaks;gen 12 32	köök;adit 0 14	kool;el 1 8	talu;in 0 6
k;null 0 43	andmed;ad 1 13	ekraan;el 4 5	anne;tr 0 6
tema;ad 3 40	nurk;in 2 12	lõpp;tr 0 9	päev;adit 0 6
NIMI;nom 4 38	suund;in 1 13	kümme;gen 3 6	vahetus;ad 0 6
aeg;nom 0 42	miljon;gen 3 11	küla;el 1 8	vari;adit 0 6
juht;ad 1 40	25;digit 2 12	kohus;ill 0 9	meel;kom 0 6
3;digit 0 41	nurk;adit 0 14	teine;adit 0 9	rida;adit 0 6
neli;nom 2 38	naine;kom 2 12	osa;gen 1 8	8;digit 0 6
6;digit 0 39	amet;in 0 14	teine;kom 1 8	jalg;adit 0 6
NIMI;ad 1 37	õhk;adit 0 14	raamat;el 1 8	x;null 0 6
päev;nom 2 36	koht;ad 1 13	2000;digit 1 8	iroonia;kom 0 6
kool;in 0 38	selg;el 0 13	pool;all 0 9	11;digit 0 6
NIMI;adit 4 34	mees;el 4 9	mees;ad 0 8	kohtumine;ad 1 5
voodi;el 0 37	see;all 1 12	sahtel;ill 0 8	pesa;in 2 4
tema;kom 3 34	elu;adit 0 13	tunne;kom 0 8	vastavus;adit 2 4
lõpp;term 0 36	aasta;nom 3 10	laps;kom 1 7	komisjon;in 0 6
viis;nom 1 35	nädal;gen 12 1	pudel;el 1 7	2.00;digit 0 6

m;null 5 31	õu;el 0 13	1500;digit 2 6	töö;kom 2 4
üks;nom 13 22	büroo;el 0 13	vars;kom 2 6	puhk;ad 0 6
see;el 10 25	külg;el 0 13	nina;kom 5 3	väide;ad 0 6
pea;adit 2 32	tee;ad 2 11	minut;gen 0 8	liige;all 1 5
käsi;kom 1 33	rand;adit 2 11	18;digit 2 6	sein;all 0 6
elu;in 5 29	puu;el 12 1	liit;kom 1 7	pool_teist;nom 0 6
1;digit 1 32	30;digit 1 12	vesi;kom 2 6	vangla;el 0 6
käsi;adit 1 32	sina;ad 0 13	sõber;kom 0 8	jooks;in 0 6
hetk;ad 0 33	sõrm;kom 1 12	jaam;in 1 7	pilt;ad 0 6
maja;in 0 32	koht;el 1 11	järg;ad 1 7	kohvik;ill 0 6
tuba;in 2 30	kael;adit 0 12	9;digit 1 7	liige;el 5 1
kodu;in 0 32	teine;all 0 12	pudel;gen 1 7	mai;el 0 6
mina;ad 0 31	ala;ad 3 9	44;digit 0 8	nädal;in 5 1
linn;in 2 29	nädal;ad 0 12	juuni;gen 8 0	jõud;el 0 6
10;digit 1 29	tund;nom 0 12	80;digit 0 8	sada;nom 0 6
mõte;in 3 27	vari;in 0 12	pistik;kom 0 8	vari;el 0 6
koht;in 0 28	maa;kom 1 11	kogu;adit 2 6	22;digit 0 6
nägu;kom 19 9	kaugus;ad 0 12	kg;null 0 8	tütar;kom 0 6
maa;ad 2 26	tuhat;nom 0 12	mees;nom 2 6	mõte;kom 0 6
algus;in 1 26	seeme;kom 0 12	sõber;all 0 8	piits;kom 2 4
mai;in 0 26	aed;in 1 11	tükk;nom 0 8	kelder;in 1 5
töö;all 3 23	silm;kom 1 11	kolm;all 1 7	köök;in 0 6
käsi;el 0 26	juus;kom 10 2	kaks;ad 1 7	tervis;tr 0 6
meel;adit 0 26	aeg;el 1 11	abi;adit 0 8	lagendik;all 0 6
50;digit 3 23	pea;in 1 11	töö;el 3 5	koridor;in 0 6
silm;in 0 26	see;gen 7 5	inimene;el 1 7	september;term 0 6
linn;adit 1 25	paik;in 0 12	töö;in 1 7	mitu;gen 2 4
kuus;nom 1 25	ots;el 1 11	kuu;gen 5 3	34;digit 0 6
100;digit 1 25	uni;in 2 10	osakond;adit 0 8	tükk;tr 0 6
ilm;in 4 21	sina;kom 0 12	hetk;gen 3 5	üks;el 4 2
pea;el 3 22	asutus;in 0 12	paik;el 0 8	kong;in 0 6
see;#cap 21 4	algus;el 1 11	aeg;tr 0 8	13;digit 1 5
suu;el 0 25	mina;gen 11 1	mood;ad 0 8	sein;ad 3 3
osa;in 2 23	jalg;kom 2 9	lugu;el 3 5	rind;all 0 6
õhtu;ad 1 24	hind;kom 1 10	lugu;in 0 8	kaks;part 0 6
selg;in 0 24	asi;kom 1 10	taimla;in 0 8	sahtel;el 0 6
aeg;gen 8 16	ölg;all 0 11	lõpp;el 0 8	osa;nom 0 6
4;digit 0 24	jõud;kom 0 11	lõpp;in 0 8	
5;digit 1 23	kord;adit 0 11	plaan;el 1 7	

päev;gen 12 11	ruum;el 0 11	talu;adit 0 8	pilk;kom 0 6
kümme;nom 1 22	ruum;in 3 8	kord;ad 0 8	24;digit 0 6
mees;all 2 21	koht;abl 0 11	vesi;adit 0 8	juuli;term 0 6
linn;el 1 22	haigla;in 1 10	küla;in 1 7	pool;ad 0 6
august;in 0 23	laen;tr 1 10	kukal;ill 0 8	piirkond;in 0 6
viis;gen 12 11	tuba;el 1 10	laud;in 2 6	naine;gen 3 3
7;digit 0 23	ema;all 0 11	mitu;part 6 2	jalg;in 2 4
mina;el 6 16	õigus;kom 5 6	raamat;in 1 7	tankla;in 0 6
NIMI;abl 4 18	õu;adit 0 11	kõri;el 1 7	töö;ad 0 6
koht;all 3 19	tulevik;in 0 11	põrand;ad 2 6	vesi;el 0 6
sajand;gen 22 0	tasku;in 3 8	kaaslane;kom 1 6	leht;kom 2 4
alus;ad 3 18	seadus;in 1 10	puhmas;gen 7 0	lähedus;in 0 6
saal;in 2 19	konverents;ad 0 11	määr;ad 0 7	õhtu;term 0 6
mees;kom 1 20	11	rida;gen 3 4	esik;in 0 6
tänav;ad 2 19	valgus;in 1 10	aeg;part 0 7	kõla;in 1 5
päev;el 3 18	mitu;nom 0 11	park;in 0 7	kord;part 0 6
asi;in 2 19	valimine;ad 0 10	kate;tr 0 7	km;null 2 4
hommik;ad 3 18	naine;all 0 10	pood;in 0 7	põlv;all 0 6
rand;in 1 20	nägu;ad 2 8	nägu;in 0 7	35;digit 0 6
see;kom 1 20	17;digit 0 10	suvi;ad 1 6	rida;nom 0 6
pool;nom 1 19	vald;in 3 7	peenar;all 1 6	aktsia;kom 1 5
mets;in 2 18	kaks;all 0 10	liit;in 0 7	keha;kom 2 4
paar;gen 2 18	tasku;ill 0 10	rusikas;kom 0 7	tund;gen 1 5
silm;adit 0 20	partei;gen 7 3	rida;in 3 4	400;digit 0 6
200;digit 1 19	häda;in 0 10	kantselei;ill 0 7	linn;all 1 5
20;digit 3 17	12;digit 0 10	7	elu;kom 2 4
laud;all 0 19	maja;adit 0 10	põhjus;ad 0 7	aadress;ad 1 5
küsimus;in 2 17	tase;ad 1 9	isa;kom 1 6	jutt;kom 0 6
riik;in 5 14	see;in 0 10	kogu;el 0 7	saar;ad 0 6
hääli;kom 2 17	klaas;el 6 4	paar;ad 1 6	keel;adit 0 6
seitse;nom 1 18	neli;gen 2 8	viis;abl 0 7	ekraan;ad 2 4
aasta;in 3 16	nimi;kom 6 4	büroo;ill 0 7	tasu;abes 2 4
see;nom 16 2	küsimus;all 3 7	süda;ill 0 7	nägu;all 0 6
september;in 0 18	19;digit 0 10	raha;kom 0 7	leht;all 1 5
18	hool;kom 0 10	riik;adit 1 6	trepp;ad 2 4
pool;gen 5 13	kuu;in 1 9	tee;abl 0 7	1.49;digit 0 6
elu;el 4 14	arve;ill 1 9	tee;all 1 6	poeg;all 0 6
kord;in 2 16	kabinet;adit 0 10	vorm;in 2 5	kallas;all 0 6
põrand;all 0 18	10	kirjandus;in 0 7	üks;gen 6 0

pea;kom 2 16	keskus;in 3 7	paber;all 0 7	11-18;digit 0 6
kolm;gen 8 10	ha;null 2 8	asi;nom 0 7	kaheksa;nom 0 6
hetk;nom 0 17	süda;in 0 10	nurk;el 0 7	mälu;el 1 5
nägu;adit 1 16	ministeerium;in 3 7	tänav;all 0 7	kord;el 2 4
maa;all 3 14	voodi;ill 0 10	keha;in 0 7	osakond;el 4 2
kaugus;kom 0 17	toime;adit 0 10	1994;digit 0 7	26;digit 0 6
ots;in 1 16	tund;in 1 9	kõne;in 1 6	leht;el 0 6
ulatus;in 0 17	leht;in 1 9	põld;ad 0 7	piir;ad 2 4
40;digit 0 17	sügis;ad 0 10	surm;term 0 7	rada;ad 0 6
sina;all 0 17	inimene;all 1 9	6,5;digit 1 6	hoone;in 2 4
tasku;el 0 17	ruum;adit 0 9	raamat;kom 1 6	auk;el 0 6
küla;adit 2 14	meel;in 0 9	silm;el 0 7	kauss;adit 0 6
kiri;in 1 15	kaks_teist;nom 0 9	sekund;in 5 2	tulemus;es 0 6
turg;ad 2 14	veerand;nom 0 9	inimene;kom 1 6	isik;all 2 4
aed;adit 0 16	nägu;el 0 9	teine;null 5 2	tema;term 0 6
sõna;kom 0 16	mets;adit 0 9	kallas;ad 2 5	inimene;nom 1 5
	asi;ill 1 8	pink;all 0 7	liige;kom 0 6
	liigutus;kom 0 9		kassa;ill 0 6

n₁ ; n₂		
NIMI;NIMI 45 169	päev;päev 0 9	hääI;ekraan 4 3
NIMI;sõna 1 50	NIMI;silm 0 9	NIMI;valitsus 1 6
mis;tema 0 32	aasta;kümme 3 6	NIMI;leht 2 5
sama;aeg 0 21	maa;NIMI 1 8	riik;kolm 2 5
iga;juht 1 17	mõte;pea 0 9	aeg;tükk 0 7
NIMI;linn 1 17	hobune;selg 0 9	sõrm;mitu 0 7
NIMI;kool 1 16	vägi;NIMI 2 6	mees;NIMI 1 6
pluss;kaks 0 17	NIMI;kord 0 8	NIMI;rand 0 7
mees;nägu 12 5	päev;NIMI 0 8	jaga=mine;puhmas 7 0
see;aeg 0 16	NIMI;mets 0 8	NIMI;elu 1 6
NIMI;tema 0 16	august;juuni 0 8	kroon;NIMI 2 5
aasta;NIMI 2 14	mina;NIMI 0 8	NIMI;meel 0 6
NIMI;riik 6 9	sama;hetk 0 8	taim;cm 0 6
mis;see 0 14	NIMI;turg 2 6	NIMI;saar 0 6
aasta;viis 5 9	leping;NIMI 2 6	päev;paar 0 6

see;hetk 0 13	riie;selg 0 8	kaugus;september 0 6
NIMI;aeg 0 13	meeter;kaks 0 8	august;kaugus 0 6
kroon;miljon 1 12	kogu;NIMI 8 0	aasta;kolm 0 6
NIMI;mees 3 9	NIMI;päev 1 7	NIMI;maa 1 5
NIMI;aasta 1 11	aasta;paar 0 8	linn;NIMI 3 3
NIMI;osa 1 11	meeter;sekund 5 3	kord;NIMI 0 6
NIMI;3 0 12	see;NIMI 1 7	leping;sõlmi=mine 0 6
mis;sina 0 12	tund;pool 1 7	kuu;viis 3 3
vihka=mine;kaks 1 11	NIMI;koht 1 6	käsi;käsi 0 6
NIMI;vald 2 10	NIMI;voodi 0 7	algus;aasta 6 0
kõrgus;cm 0 12	aeg;hulk 0 7	liit;NIMI 6 0
riik;NIMI 8 3	tund;neli 0 7	oma;naine 2 4
tema;nägu 0 11	NIMI;kinnitus 0 7	päev;hommik 0 6
NIMI;nägu 2 9	NIMI;tuba 0 7	klass;NIMI 3 3
tund;paar 0 11	oma;koht 3 4	NIMI;küla 0 6
NIMI;2 0 11	teine;ots 1 6	NIMI;hinnang 0 6
üks;päev 0 11	tüdruk;juus 7 0	arve;aasta 5 1
see;tema 1 9	ametnik;büroo 0 7	see;kolm 1 5
mis;NIMI 0 10	kõik;kolm 0 7	tema;käsi 0 6
NIMI;pea 1 9	see;nädal 0 7	aeg;sama 6 0
see;juht 0 10	algus;sajand 7 0	kes;NIMI 0 6
partei;liige 0 10	NIMI;kaks 1 6	minut;viis 1 5
aasta;kaks 1 9	iga;päev 0 7	ise;NIMI 0 6
see;mina 0 10	sõda;NIMI 3 4	sugukond;cm 0 6
aeg;NIMI 1 9	kord;paar 0 7	kord;elu 3 3
tema;sõna 0 10	juuni;august 0 7	ise;tema 1 5
aasta;algus 0 9	see;kord 0 7	NIMI;nurk 0 6
NIMI;mina 0 9	oma;NIMI 2 5	september;kaugus 0 6
tema;NIMI 2 7	NIMI;käsi 0 7	teine;kaks 3 3
meeter;kaugus 0 9	see;aasta 0 7	minister;NIMI 1 5
mees;kittel 9 0	nädal;kaks 0 7	sõna;paar 0 6
oma;aeg 0 9	NIMI;tänav 1 6	NIMI;liit 0 6

v;k₂		
olema;in 39 357	juhtuma;kom 1 9	looma;in 0 7
olema;ad 14 221	viibima;in 0 10	jääma;ad 0 7
olema;kom 54 137	müüma;digit 1 9	paljundama;gen 7 0

olema;nom 15 167	õppima;in 2 8	rääkima;in 1 6
olema;el 42 113	saatma;adit 0 10	tegelema;gen 4 3
olema;gen 70 47	tulema;kom 2 8	jõudma;ill 0 7
olema;digit 10 77	ütleva;el 1 9	teatama;digit 0 7
olema;all 18 44	alustama;kom 2 8	ütleva;#cap 7 0
saama;el 6 47	tooma;el 2 8	sõitma;ill 2 5
olema;tr 1 50	otsima;el 0 10	täitma;kom 0 6
rääkima;el 2 40	jääma;kom 2 8	puudutama;kom 1 5
nägema;in 3 28	rääkima;all 1 9	viskama;gen 2 4
toimuma;in 1 28	veetma;in 1 9	pakkuma;all 0 6
andma;all 1 26	pähe_tulema;adit 0 9	tõstma;el 1 5
võtma;el 1 25	istuma;gen 0 9	kõnelema;in 0 6
ütleva;all 2 24	kostma;el 0 9	maksma;kom 0 6
tegema;in 1 23	andma;in 3 6	kutsuma;adit 0 6
rääkima;kom 0 23	mõtleva;kom 0 9	lähenema;all 0 6
aru_saama;el 0 23	lahkuma;el 0 9	andma;adit 0 6
maksma;digit 1 21	võitma;in 0 9	kostma;ad 0 6
saama;in 3 19	jätkuma;in 2 7	viima;term 0 6
saama;tr 2 20	pääsema;adit 1 8	toetama;el 1 5
saama;digit 1 21	teatama;el 0 9	andma;ad 2 4
tegema;kom 3 19	siduma;kom 0 9	toetama;digit 0 6
panema;all 2 20	vestleva;kom 0 9	hukkuma;digit 0 6
käima;in 0 21	ütleva;ad 1 8	tekkima;ad 0 6
olema;#cap 17 4	seisma;ad 0 9	ootama;in 1 5
kirjutama;el 4 17	kuuluma;all 1 8	jõudma;all 0 6
õitsema;term 0 21	käima;ad 0 8	hoidma;ad 0 6
õitsema;el 0 21	lugema;el 1 7	andma;tr 0 6
istutama;in 0 20	sõitma;kom 1 7	näitama;all 0 6
leidma;el 1 19	andma;el 2 6	võtma;ill 0 6
saama;ad 0 19	pidama;ad 2 6	ootama;adit 0 6
minema;adit 0 19	avama;#cap 2 6	vaatama;adit 0 6
ütleva;in 3 16	juhtuma;in 1 7	avama;ad 1 5
tulema;el 1 17	alistama;digit 0 8	kandma;digit 0 6
tooma;adit 0 18	saama;adit 0 8	käes_hoidma;in 1 5
elama;in 2 16	maksma;all 2 6	elama;el 2 4
teadma;el 1 17	kätte_võtma;adit 1 7	üle_andma;gen 6 0
istuma;in 1 16	asetama;all 0 8	tabama;el 1 5
viima;ill 0 17	teatama;ad 2 6	maksma;null 0 6
olema;adit 6 11	mõtleva;all 0 8	sõitma;all 1 5

olema;ill 6 10	võtma;kom 3 5	läbi_käima;el 0 6
töötama;in 1 15	vaatama;in 2 6	tõusma;ad 0 6
saama;kom 1 15	tulema;ill 0 8	jõudma;digit 0 6
minema;all 0 15	pakkuma;in 1 7	ilmuma;gen 4 2
kasvama;digit 1 14	jätma;tr 0 8	nõustuma;kom 0 6
minema;ill 0 15	algama;el 0 8	mängima;kom 0 6
tegema;el 3 12	langema;all 0 8	kutsuma;all 0 6
jõudma;term 0 14	saatma;all 1 7	viskama;kom 2 4
jääma;all 0 14	ilmuma;in 1 7	kuuluma;digit 1 5
olema;null 4 10	olema;term 0 8	seisma;nom 1 5
kuulma;el 2 12	jääma;adit 0 8	alandama;digit 0 6
istuma;ad 0 13	jääma;tr 0 8	vabastama;el 0 6
saama;gen 4 9	kehtestama;digit 0 8	tulema;adit 0 6
tegelema;kom 1 12	ulatuma;term 0 8	väheneb;digit 0 6
meelde_tuletama;adit 0 13	teadma;in 2 6	viima;el 0 6
kasvama;ad 0 13	laskma;adit 0 8	harjuma;kom 0 6
jätma;adit 0 13	tegema;ad 0 8	kuulma;in 0 6
sõitma;adit 0 13	sõdima;kom 0 8	koolis_käima;in 0 6
viima;adit 0 12	kinnitama;all 2 5	laskma;kom 1 5
kõnelema;el 0 12	maksma;in 0 7	vaatama;gen 1 5
kuuluma;adit 0 12	võtma;gen 2 5	jõudma;el 2 4
hoidma;in 0 12	alustama;ad 1 6	võtma;all 1 5
minema;in 1 11	tulema;in 1 6	olema;part 2 4
üttelema;kom 2 10	katma;kom 0 7	saama;nom 0 6
saama;abl 1 11	nägema;gen 4 3	külla_tulema;adit 0 6
tegema;all 1 11	tulema;tr 0 7	küsima;el 0 6
seisma;in 0 12	seisma;kom 4 3	jääma;ill 0 6
jõudma;adit 0 12	tekitama;in 1 6	esitama;ad 0 6
ostma;el 1 10	vaatama;all 0 7	nõudma;abl 1 5
pidama;tr 1 10	kerkima;el 0 7	kohtuma;kom 0 6
sööma;el 0 11	teatama;all 1 6	tähendama;in 0 6
jätma;all 0 11	muutuma;in 1 6	lõpetama;in 2 4
tundma;in 0 11	kasutama;ad 0 7	puuduma;in 0 6
lõppema;kom 0 11	nõudma;ad 1 6	tungima;adit 0 6
saama;all 2 9	kasutama;in 1 6	nägema;el 0 6
jääma;in 2 9	pidama;in 1 6	vahtima;all 0 6
nägema;ad 1 10	püsima;ad 0 7	unes_nägema;in 0 6
hakkama;kom 1 10	märku_andma;kom 3 4	heitma;el 0 6
	püsima;in 1 6	lamama;in 0 6

leidma;in 1 10	laekuma;digit 1 6	juhtima;el 1 5
minema;kom 1 10	viibima;ad 0 7	viima;all 0 6
meeldima;all 0 11	kirjutama;all 0 7	panema;adit 0 6
loobuma;el 0 11	vaatama;ad 1 6	hoidma;kom 0 6
avama;digit 0 10	valitsema;in 0 7	võrdlema;kom 0 6
võtma;tr 0 10	astuma;adit 0 7	tulema;all 0 6
silmas_pidama;in 0 10	sattuma;adit 0 7	toimuma;kom 1 5
kandma;el 0 10	viitama;all 0 7	tegema;tr 0 6
pakkuma;digit 1 9	viskama;all 0 7	töötama;gen 2 4
kasutama;tr 0 10	viima;kom 1 6	alustama;digit 3 3
vastu_võtma;in 0 10	algama;in 0 7	selga_panema;adit 1 5
olema;abes 2 8	nägema;part 1 6	
tooma;all 0 10		

k₁;k₂			
gen;in 70 738	ad;gen 53 8	kom;ad 3 26	abl;kom 2 11
gen;el 91 701	el;kom 13 48	kom>null 14 15	adit;all 1 12
gen;kom 81 533	all;adit 4 56	part;es 1 27	ad;tr 0 13
gen;ad 35 565	all;in 11 49	el;ad 4 24	el;es 0 13
nom;in 96 400	ad;el 7 52	ad;nom 1 27	tr;tr 0 13
gen;all 30 457	part;part 14 44	tr;in 3 24	tr;ill 0 13
nom;kom 168 303	gen;term 1 57	el;nom 1 26	ill;adit 0 12
part;nom 3 385	in;el 6 49	all;ad 3 22	ill;all 1 11
part;in 35 352	all;all 4 51	part;term 1 24	ad;ill 1 11
gen;gen 188 188	all;kom 8 47	tr;el 7 18	all;#cap 10 2
nom;el 136 239	kom;el 11 43	abl;el 0 25	part;abes 2 9
gen;tr 9 346	in;nom 4 49	tr;gen 19 5	abl;ill 0 11
gen;adit 8 341	kom;gen 26 27	tr;adit 0 24	#cap;in 0 11
part;kom 64 280	gen;es 1 51	part;abl 1 23	term;gen 4 7
nom;all 41 277	adit;kom 4 48	el;abl 3 20	adit;nom 0 10
part;el 51 263	el;term 1 50	nom;abes 11 11	kom;abl 0 10
ad;ad 13 238	all;tr 0 50	adit;in 3 19	adit;ill 3 7
nom;gen 185 60	kom;adit 3 46	ill;kom 2 20	part;#cap 8 2
part;digit 18	tr;all 17 31	abl;abl 1 20	ad;term 0 10
215	in;digit 6 41	tr;kom 7 14	term;in 4 6
nom;nom 44 164	el;gen 31 16	gen>null 6 15	es;el 2 7
nom;ad 24 182	term;el 4 43	abl;adit 0 21	el;#cap 7 2

nom;digit 14 183	el;in 9 38	all;gen 11 10	all;abl 0 9
gen;ill 7 190	ad;digit 6 40	kom;ill 1 19	ad>null 3 6
in;in 20 158	kom;digit 4 42	in;adit 0 20	ill;in 2 7
part;adit 3 163	kom;all 7 38	part>null 4 16	adit;gen 7 2
nom;adit 11 148	null;el 4 41	null;nom 0 20	el>null 2 7
gen;digit 32 121	nom>null 1 43	nom;es 3 17	refl;kom 0 8
gen;nom 14 132	nom;tr 5 38	term;digit 4 14	adit>null 1 7
in;kom 30 107	nom;abl 4 39	adit;tr 0 18	ill;tr 0 8
part;all 16 110	null;kom 4 38	abl;digit 0 18	kom;#cap 7 1
ad;in 17 101	in;ad 2 39	all;es 0 17	ad;abl 0 7
null;in 3 110	gen;#cap 29 10	kom;nom 1 15	null;#cap 2 5
part;gen 74 36	el;digit 0 39	in>null 9 7	null;digit 0 7
kom;in 6 103	in;all 3 36	all;ill 1 15	es;in 0 7
el;el 15 88	all;digit 4 35	in;#cap 11 5	in;term 0 7
part;ad 15 85	tr;digit 1 37	kom;tr 0 16	in;ill 0 7
gen;abl 9 86	part;tr 4 34	null;ill 0 16	tr;nom 0 7
all;el 9 86	nom;#cap 31 7	gen;abes 3 12	abl;in 1 5
el;adit 4 83	#cap;ad 0 36	null;adit 2 13	ad;es 0 6
nom;ill 15 66	null;all 5 30	adit;adit 2 12	null;gen 3 3
abl;all 1 73	el;ill 2 32	refl;el 0 14	ill;ill 1 5
adit;el 4 70	nom;term 3 31	kom;kom 4 10	#cap;el 0 6
el;all 0 67	ad;all 0 34	in;tr 0 14	abl;tr 0 6
null;ad 3 64	ill;el 1 31	el;tr 0 14	es;adit 0 6
ad;kom 18 48	ad;adit 0 32	tr;ad 0 13	
in;gen 50 14	ad;#cap 29 0	all;nom 0 13	
part;ill 6 57			

k₂
in 280 2604
el 349 2064
kom 407 1798
ad 103 1494
all 126 1351
digit 90 1279
nom 76 1185
gen 659 539
adit 38 1052

```
tr 18 570  
ill 38 501  
abl 20 256  
null 48 203  
#cap 144 102  
term 6 218  
es 5 176  
part 14 75  
abes 21 64  
refl 0 9
```