Püsiühendite leidmine suurtest tekstikorpustest

Heiki-Jaan Kaalep, Kadri Muischnek
Tartu Ülikool

Tööd on toetanud ETF oma grandiga 4352

Selles artiklis kirjeldatakse perifrastiliste verbide tuvastamist korpustes, kasutades keele- ja ülesandespetsiifilist tarkvara SENVA (Software for Extracting N-ary Verbal Associations), mis on keelest sõltumatu tarkvara SENTA (Software for Extracting N-ary Textual Associations) (Dias jt 2000) edasiarendus. Tulemuseks on 16 600 perifrastilisest verbist koosnev leksikon. Kirjeldatakse tarkvara, korpustest leitud kollokatsioonide hulgast "õigete" väljavalimise põhimõtteid ja antakse hinnang töö täpsusele ja saagile.

1. Sissejuhatus

Tekstis esinevate lausete edukaks automaatanalüüsiks ei piisa ainult morfoloogia- ja süntaksireeglite tundmisest ja kasutamisest. Hea tulemuse saamiseks peab tingimata arvestama ka selles keeles esinevate püsiühenditega. Selliste püsiühendite hulka kuuluvad ka perifrastilised verbid - ahel-, ühend- ja väljendverbid. Perifrastiliste verbide leksikon on vajalik igal keele automaatanalüüsi tasandil - nii morfoloogilisel, süntaktilisel kui ka semantilisel tasandil. Sellise leksikoni saab muidugi koostada olemasolevate sõnaraamatute või muude keeleressursside baasil. On siiski tuntud tõsiasi, et mitmesõnalised üksused on traditsioonilise lingvistika poolt sageli unarusse jäetud: leksikograafe huvitavad peamiselt üksiksõnad ja nende tähendused; grammatikuid on huvitanud keele üldisemad seaduspärasused.
Samas on loodud mitmesuguseid tarkvarasüsteeme, mis statistilisi meetodeid kasutades leiavad tekstikorpustest sõnad, õigemini sõnaühendid, mis esinevad üksteise naabruses sagedamini, kui võiks järeldada nende üksi esinemise sagedustest. Sellise tarkvara kasutamine võimaldab täiendada ja täiustada perifrastiliste verbide leksikoni ja vastata järgmistele küsimustele:
1. Kas tekstis on perifrastilisi verbe, mida meie sõnaraamatute põhjal koostatud andmebaasis ei ole?

2. Kas neid perifrastilisi verbe, mida sisaldab meie sõnaraamatute põhjal koostatud andmebaas, kasutatakse ka tänapäeva kirjalikes tekstides?

Oletada võib, et andmebaasi saab sellisel viisil paremaks muuta. Aga kui palju paremaks, seda on raske ennustada. Samuti on rakse enne töö algust teha oletusi selle mahu kohta.

Varasem katse 500 000-sõnalise ilukirjanduskorpusega (Kaalep, Muischnek 2002) näitas, et suhteliselt palju tekstides sageli kasutatavaid verbiühendeid puudus sõnaraamatute baasil koostatud andmebaasist. Nii otsustasimegi, et on otstarbekas kasutada andmebaasi täiendamiseks sama tarkvara, kuid palju suuremaid korpusi.

2. Andmebaas

Eesmärgiks oli koostada võimalikult ammendav andmebaas, mis sisaldaks kõiki sagedasemaid perifrastilisi verbe, mida tänapäeva tekstides kasutatakse. Sellise andmebaasi tegemiseks otsustasime kasutada olemasolevaid keeleressursse: nii inimkasutajale mõeldud sõnaraamatuid kui ka olemasolevaid keelekorpusi, mida suuremaid, seda parem.

Alustasime olemasolevatest inimkasutajale mõeldud sõnaraamatutest ja panime nende baasil kokku 10 800 verbiühendit sisaldava baasi, mis on väljas ka TÜ

Selleks kasutasime järgmisi sõnaraamatuid:

1. "Fraseoloogiasõnaraamat" (Õim 1993)
2. "Eesti kirjakeele seletussõnaraamat" (EKSS)
3. Filosofti tesaurus (_____)
4. Partikkelverbide loend "Das Estnische Partikelverb als Lehnübersetzung aus dem Deutschen" (Hasselblatt 1990)
5. "Eesti keele mõistelise sõnaraamatu" indeks (Saareste 1979)
6. "Sünonüümisõnastik" (Õim 1991)

Umbes 3000 verbiühendit selles andmebaasis on ühendverbid, ülejäänu moodustavad põhiliselt väljendverbid, aga on ka mõningaid ahelverbe. Ahelverbidest saab meie kasutatud meetodil korpustest leida ainult ma-infinitiivi ja verbi ühendeid, sest korpuse eeltöötluse käigus viiakse kõik tekstis esinevad verbid ma-infinitiivi kujule.

3. Korpus

Eeldame, et kui kasutada piisavalt suurt korpust, mis sisaldab tekste piisavalt erinevatest tekstiklassidest, siis selle baasil peaks saama koostada otsitava nähtuse enam-vähem ammendava käsitluse.

Kirjeldatava ülesande lahendamiseks kasutasime kolme erinevat korpust, mis on

1. 500 000-sõnalist osa 90ndate aastate ilukirjanduskorpusest. See korpus koosneb 2000-sõnalistest ilukirjandustekstide katketest aastatest 1992-1998.
2. 9,8 miljonit sõna ajalehetekste aastatest 1995-2001, peamiselt ajalehte "Postimees" 1998. aastast ja "Eesti Ekspress" 1998-2001, lisaks veel mõned numbrid "Maalehte", "Päevalehte" ja "Äripäeva". Siin ajalehekorpuses on ainult kvaliteetlehed, tabloide pole siia võetud.
3. 12,6 miljonit sõna Riigikogu toimetatud stenogramme aastatest 1995-2001. Ka see allkorpus esindab selgelt kirjalikku keelekasutust, mitte suulist kõnet. Mis sõnavarasse puutub, siis on nendes tekstides muidugi palju õiguskeele slängi, kuid ka üllatavalt palju mahlakaid ütlemisi, nagu *alt ära hüppama* või *sõnnikust saia tegema*.
On ilmne, et nende korpuste omavaheline tasakaal on ideaalist kaugel. Probleem on sama, mis praegu meie töörühmas koostatava suure eesti keele korpuse puhul: liiga vähe on ilukirjandust. Oleks muidugi olnud võimalik analüüsida SENTA abil ka ülejäänud olemasolevate allkorpuste ilukirjandust (olemas on ilukirjanduse allkorpused iga kümnendi kohta 1890-1990), kuid tahtsime jääda võimalikult tänapäevase keelekasutuse juurde ja lugesime ka 80ndate aastate ilukirjanduse "vananenuks".

4. SENVA - tarkvara perifrastiliste verbide leidmiseks

Perifrastilise verbi verbiosa võib tekstis olla mistahes vormis. Ülejäänud perifrastilise verbi komponendid on alati kindlas vormis. Sellise mitmesõnalise üksuse komponentide järjekord võib muutuda ja komponendid ei pruugi lauses paikneda vahetult üksteise järel. See kõik raskendab nende automaatset tekstis tuvastamist.
Seega tuleb lahendada järgmised ülesanded:
1. Vähendada töötlusesse minevate sõnaühendite hulka.
2. Leida sõnad, mis esinevad koos sagedamini kui võiks eeldada nende lihtsast esinemissagedusest (so kollokatsioonid).
3. Viimaste hulgast leida mitmesõnalisi verbe moodustavad kollokatsioonid.
Nende ülesannete lahendamiseks peame kombineerima lingvistilisi ning statistilisi meetodeid ja ka käsitsitööd. Järgnevalt kirjeldamegi automaatse analüüsi erinevaid etappe: korpuse ettevalmistamist, kollokatsioonide ekstraheerimist ja statistilist töötlust.

4.1. Korpuse ettevalmistamine

Kuna mitmesõnaliste verbide verbiosa pöördub lauses vabalt, aga tema määr- või nimisõnalised komponendid on alati kindlas vormis, siis tuleb korpuse ettevalmistamisel viia verbid ma-ifinitiivi kujule, kuid kõigi teiste sõnaliikide puhul säilitada nende lauses esinemise kuju. Selleks teostasime korpuse statistiliseks töötluseks ettevalmistamisel kõigepealt täieliku morfoloogilise analüüsi ja ühestamise. Morfoloogiliselt ühestatud korpuses on igale tekstisõnale lisatud tema morfoloogiline analüüs ja algvorm. Sellisest tekstist jäetakse nüüd alles kõigi verbide algvormid, kuid teiste sõnaliikide lauses kasutatud kujud. Sellisel töötlusel läheb paratamatult kaduma enamus ahelverbe.

4.2. Kollokatsioonide valik

Kollokatsioonide valik statistiliseks töötluseks on väga oluline, kuna täpse valikuga on võimalik tunduvalt vähendada programmi poolt väljastatavate kollokatsioonide hulka ja nii ka nende käsitsi töötlemiseks kuluvat aega. Töö sellel etapil on oluline mõju programmi töö kvaliteedile - täpsusele ja saagile. Teiselt poolt tuleb muidugi tunnistada, et niimoodi võime me välja praakida mõne ebatüüpilise ja seetõttu võib-olla just eriti huvitava sõnaühendi.

Lingvistiliselt töödeldud korpusest valisime kõik võimalikud kollokatsioonid järgmiste põhimõtete järgi.

1. Kollokatsiooni (sõnaühendi) enda pikkuse piirasime kaksikute ja kolmikuteni, sest sellised paistsid enamasti olevat eesti keele mitmesõnalised verbid.
2. Otsitava sõnaühendi komponentide vahel ei saa olla kirjavahemärke (punkti, koma, jutumärke jms) ega sulge.
3. Otsitava sõnaühendi komponentide vahel saab olla maksimaalselt teatud kindel arv sõnaühendisse mittekuuluvaid sõnu. Töötlesime korpust neli korda, võttes selleks distantsiks 0, 1, 2 ja 3 sõna.
4. Statistilisele töötlemisele lähevad ainult verbi (ma-infinitiivi) sisaldavad kollokatsioonid.
5. Ka verbe sisaldavate kollokatsioonide hulgast eemaldame need, mis sisaldavad teatud sõnu, mis ei saa olla mitmesõnalise verbi komponendiks, nimelt:
    - pärisnimesid,
    - asesõnu (mõne erandiga)
    - sidesõnu
    - abiverbe *olema* ja *ära*
    - teatud määrsõnu (praegu u 100 tk, nt *palju*, *taas*)

- teatud nimisõnu, õigemini tuli "väljaviskamise" nimekirja panna nimisõnade käändevormid. Need on kas nimisõnad, mis tüüpiliselt (ja eriti selles käändevormis) esinevad lauses vaba laiendina nt *öösel* või on liiga spetsiifilised, nt *advokaat* või *ministeerium*. Selles nimekirjas on praegu u 3000 sõnavormi.

Nn halbade määrsõnade ja nimisõna vormide nimekirja saime selliselt, et lasime programmil töötada ilma nende nimekirjadeta, saadud kollokatsioonide esikomponentidest (so mitte-verbilistest osadest) tegime sagedusloendi ja võrdlesime seda olemasoleva andmebaasi põhjal tehtud sarnase loendiga. Need sagedamad esikomponendid, mis polnud osalenud andmebaasis olevate mitmesõnaliste verbide moodustamisel, vaatasime käsitsi läbi ja märkisime ära need, mis meie arvates ei saa olla mitmesõnalise verbi komponendiks. Seega on need nimekirjad subjektiivsed ja pole välistatud, et me neid kasutades viskame ära nii mõnegi huvitava verbiühendi. Kuid kogu selle protsessi kõige töömahukam osa on saadud kollokatsioonide käsitsi läbivaatamine ja nii kasutasime kõiki võimalusi, et seda töömahtu natukenegi vähendada.

6. Lõpuks järjestatakse sõnad kollokatsioonide sees nii, et verb on selles viimasel kohal. Seda seepärast, et väljenditest *ei tahtnud pähe saada* ja *aga ikkagi saime pähe* saime eelmiste etappide tulemusel väljendid *pähe saama* ja *saama pähe*. Selleks, et need kokku võtta, tulebki sõnad kollokatsioonisiseselt järjestada kujule *pähe saama*. Alles siis saame ühesugused kollokatsioonid kokku lugeda. Seda läheb järgnevatel, statistilise analüüsi etappidel vaja.

4.3. Statistiline analüüs

Eelmisel etapil leitud kollokatsioonidel rakendame statistilist analüüsi. Selleks kasutame G. Diasi loodud keelest sõltumatut programmipaketti SENTA (Software for Extracting N-ary Textual Associations) (Dias jt., 2000), mida me eesti mitmesõnaliste verbide leidmiseks kohandasime. Alljärgnevalt kirjeldame SENTA tööpõhimõtet.

4.3.1 Ühise oodatavuse (ÜO) (Mutual Expectation) mõõt

Mitmesõnalised üksused on definitsiooni kohaselt sõnajadad, mis esinevad üksteise läheduses liiga sageli, et see saaks olla juhuslik. Sellest eeldusest lähtudes defineeritaksegi sõnajadasse kuuluvate sõnade kokkukuuluvuse määra kirjeldav matemaatiline mudel. Seda mudelit kasutatakse, et arvutada ühist oodatavust, mis omakorda tugineb normaliseeritud oodatavusel.

N sõna vahelist normaliseeritud oodatavust defineeritakse kui keskmist ootust, et teatud positsioonis esineb mingi kindel sõna, kui (*n-1*) positsioonis juba esinevad sõnad on teada. Nt. kolmiku *vahi alla võtma* [*vahi +1 alla +2 võtma*] keskmine ootus peab arvesse võtma , et *võtma* tuleb pärast *vahi alla*, aga ka seda, et *alla* esineb *vahi* ja *võtma* vahel ning et *vahi* esineb enne kui *alla võtma*. ÜO põhiidee on hinnata sõnaühendist ühe sõna väljajätmise hinda (kokkukuuluvuse mõttes). Mida tihedamalt on jada sõnad omavahel seotud, st mida vähem lubavad nad endi hulgast mõne eemaldamist, seda suurem on normaliseeritud oodatavus. Normaliseeritud oodatavus defineeritakse kui *n* liikmega sõnajada esinemise tõenäosus *p*, mis on jagatud kõigi selliste (*n-1*) liikmeliste sõnajadade tõenäosuste keskmisega, mis erinevad *n*-liikmelisest sõnajadast 1 sõna eemaldamise poolest.

$$NO = \frac{p(n - pikkusega_jada)}{\frac{1}{n}\sum p(n-1-pikkusega_jada)}$$

Seega, mida rohkem on tekstis selliseid (*n-1*) liikmelisi jadasid, mis esinevad kuskil mujal kui meid huvitava *n*-liikmelise jada koosseisus, seda suurem on nende tõenäosuste aritmeetiline keskmine ja seega seda väiksem on NO.
NO on seosetugevuse üldtuntud mõõdu, Dice'i koefitsiendi (Smadja 1993) üldistus n pikkusega sõnajadade jaoks; Dice'i koefitsient nimelt on võrdne NO-ga kahekomponendiliste sõnajadade (kus *x* ja *y* on sõnad) puhuks:

$$Dice(x, y) = \frac{p(x, y)}{\frac{1}{2}(p(x) + p(y))}$$

Daille (1995) on näidanud, et üheks tõhusaks kriteeriumiks mitmesõnaliste üksuste leidmisel on lihtne sagedus. Sellest eeldusest tulenevalt väidetakse, et kahest ühesuuruse NO-ga sõnajadast on see, kumb on sagedasem, ka tõenäolisem mitmesõnalise üksuse kandidaat:

$$\ddot{U}O = p(n - pikkusega_jada) \times NO(n - pikkusega_jada)$$

Arvutades ÜO-d N-sõnalise tekstikorpuse peal, kasutame ülaltooduga samaväärset valemit, mis tõenäosuste asemel sisaldab absoluutsagedusi *f*:

$$\ddot{U}O = \frac{f(n-pikkusega_jada)}{N} \times \frac{f(n-pikkusega_jada)}{\frac{1}{n}\sum f(n-1-pikkusega_jada)}$$

Kui oleme välja arvutanud ühe sõnajada ÜO ja temas sisalduva, ühe sõna võrra lühema sõnajada ÜO, siis kasutame GenLocalMaxs algoritmi otsustamaks, kumb neist on "see õige". See algoritm eeldab, et üks sõnajada on mitmesõnaline üksus või, antud juhul, fraasiverb, kui kokkukuuluvus seda moodustavate sõnade vahel pole väiksem tema alaosade kokkukuuluvusest ja kui see kokkukuuluvus ise on suurem pikema sõnajada osade kokkukuuluvusest, so kui see sõnajada ise ei ole mõne suurema püsiväljendi osa. Teiste sõnadega, üks sõnajada, ütleme *W*, on mitmesõnaline üksus või meie juhul fraasiverb, kui tema ühise oodatavuse väärtus, *ÜO(W)* on lokaalne maksimum. Olgu *n*-sõnalises jadas *W* sisalduvate *(n-1)*-sõnaliste jadade hulk Ω_{n-1} ja kogu *(n+1)*-sõnaliste jadade hulk, milles sisaldub *W*, Ω_{n+1}. Siis

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$

kui n=2 siis

kui *ÜO(W)* > *ÜO(y)*, siis *W* on mitmesõnaline üksus

kui N>2, siis

kui *ÜO(x)* ≤ ÜO*(W)* ja ÜO*(W)* > ÜO*(y)*, siis on *W* mitmesõnaline üksus

Seega, juhul kui:

sõnajada pikkus on 2, siis tema hinne peab olema suurem kui teda sisaldavatel pikematel sõnajadadel;

sõnajada pikkus on üle 2, siis saab tema hinnet lisaks võrrelda ka nende sõnajadade hinnetega, mida tema sisaldab; seejuures piisab, kui alamjadade hinded ei ületa tema oma.

Selgus, et meie ülesande jaoks oli GenLocalMaxs mõnikord liiga jäik. Kui suurendasime maksimaalset sõnade arvu, mis püsiühendi komponentide vahel võib olla, siis saime uusi häid püsiühendi kandidaate, aga mõned ka kaotasime. Seetõttu otsustasime töödelda tekstikorpust SENTAga 4 korda, iga kord erineva distantsiga (0

kuni 3), so kollokatsiooni moodustavate sõnade vahel võis olla 0 kuni 3 sõna. Seejärel liitsime saadud püsiühendi kandidaatide hulgad, et nad hiljem läbi vaadata. Sel moel suurenes tõenäosus, et korpusest on leitud head mitmesõnalised ühendid, aga samas suurenes ka väljundi maht.

5. "Õigete" sõnaühendite väljavalimine

Kõik sõnaühendid, mis läbisid GenLocalMaxsi filtri, lugesime me potentsiaalseteks mitmesõnalisteks verbideks, isegi need, mille sagedus oli kõigest 2 ja ÜO nii väike, et seda näidati nullina. See taktika õigustas ennast, sest nii mõnigi selline sõnaühend osutus tõesti ühend- või väljendverbiks, kuid teiselt poolt suurenes nii võimalike mitmesõnaliste verbide läbivaatamiseks kuluv aeg.

Sõnaühendid järjestati verbi järgi ja loendi vaatas läbi kaks inimest üksteisest sõltumatult, märkides ära mitmesõnalise verbina tundunud kollokatsioonid. Ootamatult selgus, et see, käsitsitöö etapp oli kogu töös üks raskemaid ja töömahukamaid. Raskusi tekitas nimelt piiri tõmbamine andmebaasi sobivate ja mittesobivate väljendite vahel. Illustreerime seda järgnevate näidete varal.

5.1. Näited

1. Verb *esitama*.

Saareste mõistelise sõnaraamatu indeksis on sõnaühendid *ettekäändeid esitama* ja *nõudeid esitama*. „Sünonüümisõnastikus“ on *väljakutset esitama* (*välja kutsuma* sünonüümina).

Oma korpustes ei kohanud me neist ühtegi, küll aga klišeetaolisi väljendeid *nooti esitama*, *territoriaalseid pretensioone esitama*, *süüdistust esitama*. Neil väljendeil on kahtlemata samasugune õigus olla väljendite sõnastikus kui eespool tooduil. Kui *väljakutset esitama* on loetud sobivaks sõnastikku võtmiseks, siis peaks sobima ka *palvet esitama* (= paluma), *pöördumist esitama* (= pöörduma), *taotlust esitama* (= taotlema), aga miks mitte ka *tagasiastumispalvet*, *seadusemuudatusi*, *pankrotiavaldust*, *nimekirja esitama*? On tõsi, et väga paljusid asju saab esitada, aga samuti on tõsi, et väga paljusid (nt. päikest, lauda) ei saa. Seega esitama esineb koos ainult teatud semantilisse klassi kuuluvate sõnadega. Samas, *esitama* sarnaneb oma abstraktsuses verbiga *võtma*: verbi ja nimisõna ühendile annab tähenduse peaaegu ainult nimisõna*, nt. viina võtma*, *naist võtma*, *kartuleid võtma*. Seega otsustasimegi lülitada andmebaasi suure hulga *esitama*-ga seotud väljendeid: esiteks sisaldub neis piiratud sõnavara, mis kuulub haldus- ja bürokraatiamaailma; teiseks tähistavad nad

sellesama haldus- ja bürokraatiamaailma konkreetseid toiminguid, sel moel sarnanedes terminitega.

Samas, kas sõnastikku ikka peaks kuuluma väljendid *tegevusaruannet, tuludeklaratsiooni, võlanõudeid, pankrotihagi esitama*? Üldkeele sõnastikku vist mitte, aga inimene, kes tahab haldus- ja bürokraatiamaailmas olla tõsiseltvõetav, peaks teadma, et korraldust antakse, ettepanekut esitatakse ja ülevaadet võib nii anda kui esitada.

2. Määrsõna *alla*.

Andmebaasis on 118 ühendverbi abimäärsõnaga *alla.* Nende hulgas on nii selliseid, kus verbi tähendus ühendis on hoopis teine kui verbi tähendus ilma määrsõnata, nt *alla ajama, alla kirjutama*, kuid enamuse moodustavad muidugi *alla* ja mitmesuguste liikumisverbide ühendid. Ühendverb *alla minema* on olnud nii EKSS-s, Hasselblatil kui ka Saareste mõistelise sõnaraamatu indeksis, *alla kukkuma* on "Sünonüümisõnastikus", Filosofti tesauruses ja Hasselblatil, mõlemad on leitud ka korpusest. Kuid selliseid verbi ja määrsõna ühendeid nagu *alla liikuma* ja *alla liuglema,* mida korpustest küll leiti, pole üheski sõnaraamatus. Vahe paistab siin olevat ainult selles, et *minema* ja *kukkuma* on sagedasemad verbid kui *liikuma* ja *liuglema*. Meie otsustasime kõik *alla* ja liikumisverbi ühendid andmebaasi lisada.

3. Nimisõna *aken*.

Üks kriteerium, mida kasutasime andmebaasi väljendite valimisel, oli väljendite „imelikkus“. Andmebaasi on võetud korpusest väljendid nagu

akna all istuma, akna all seisma, akna alla astuma, akna alla seisma, akna juurde astuma, aknaid pesema, aknaid sisse loopima, aknale ilmuma, aknast jälgima, aknast nägema, aknast paistma, aknast vaatama, aknast vahtima, aknast välja vaatama, aknast välja vahtima

Ükski neist pole varem esinenud mõnes andmebaasi aluseks olevas sõnastikus. Ühelt poolt on tõsi, et nende väljendite tähendus kujuneb lihtsalt osasõnade tähenduste summana. Samas, nii mõnigi väljend esindab selgelt omaette tegevust või selgelt piiritletud situatsiooni, nt. *aknaid sisse loopima* või *aknaid pesema*; mitme väljendi puhul on *aken* „imelikus“ käändes, nt. *aknale ilmuma* (miks mitte *aknasse ilmuma*?) või „imeliku“ eessõnaga, nt *akna alla astuma* (mis koha peale siis täpselt astuma?), mis võib tekitada segadust teise keele kõnelejale. Ühesõnaga, tegemist on mitmes mõttes ebatavaliste väljenditega ja seetõttu ongi nad andmebaasi võetud.

5.2. Printsiibid

Andmebaasi lisamiseks sobivate sõnaühendite valimisel lähtusime eelkõige põhimõttest, et tavalist, regulaarselt moodustatavat verbi ja muu(de) sõna(de) ühendit ei ole mõtet andmebaasi viia. Praeguses andmebaasis aga on sees mitmeid selliseid perifrastilisi verbe, mis on tulnud sinna juba sõnaraamatutest. Näiteks on „Sünonüümisõnastikus“ eraldi kirjetena sees ka sellised üksused nagu *ajakohaseks tegema*, sest tal on sünonüüm *ajakohastama*.

Rusikareegel oli, et pigem võtame andmebaasi natuke mitte-väljendeid kui jätame mõne vajaliku võtmata: välja visata on hiljem kergem kui uuesti kogu korpust läbi vaadata. Seetõttu võtsime mõned väljendid (nagu eeltoodud *akna*-ga väljendid) andmebaasi n-ö igaks juhuks, mitte niivõrd väljendite endi pärast kui just võtmise-jätmise problemaatilisuse tõttu.

Andmebaasi jaoks kõlbmatuteks tunnistati:

1. Enamus modaalverbi ja verbi ning samuti faasiverbi ja verbi ühendeid.
2. Ühendid, kus verbi kollokaat on vaba laiend, nt *aastal toimuma* või *äsja lõppema*. Sagedasemad sellised nimisõnavormid olid juba nn halbade kollokaatide nimekirjas.
3. Ühendid, kus kollokaati võib kasutada koos paljude samasse semantilisse klassi kuuluvate verbidega, nt *asju korraldama, asju organiseerima* jne.
4. Ühendid, kus samal verbil on palju samatüübilisi kollokaate.

Andmebaasi otsustati võtta kõik sellised sõnaühendid, kus verbi kasutatakse ülekantud tähenduses või kus kontekst muudab verbi tähendust. Tõsi küll, seda põhimõtet oli praktikas raske järgida. Kui lisada verbile mõni funktsioonisõna, partikkel, siis enamasti see muudab verbi tähendust. Sellistel juhtudel oli otsustamine kergem. Teisest küljest paisutas selline kõikvõimalike määrsõna ja verbi ühendite, millest osa on regulaarselt moodustatavad, andmebaasi mahtu (nt 118 väljendit sõnaga *alla*).

Täistähenduslikud sõnad aga lisavad alati tähendusele midagi. Mida üldisema tähendusega on verb (*saama, tegema* jpt), seda raskem on otsustada, kas sõnaühend tuleks andmebaasi võtta või mitte. Kas näiteks väljendites *palka saama* ja *AIDSi saama* on verbi või ka kogu sõnaühendi tähendus erinev? Neid kollokatsioone siiski andmebaasi ei võetud, küll on seal näiteks väljend *haiget saama* ja *nalja saama.*

Kui kollokatsiooni moodustasid harvaesinevad sõnad, siis oli sellisel sõnaühendil suurem võimalus andmebaasi pääseda, näiteks *õlgu kehitama*.

Kuna nii mõnigi kord oli keeruline otsustada, kas korpusest leitud sõnaühend kõlbaks väljendite leksikoni lisada või mitte, siis on meie andmebaas küllaltki ebaühtlane. Temas on väljendeid, mis sobiksid eri tüüpi leksikonidesse, aga kindlasti ka selliseid, mida ühtegi sõnastikku ei tohiks võtta.

6. Tulemused ja nende kontroll

Me töötlesime SENVAga kolme korpust: ilukirjandust, riigikogu stenogramme ja ajahehetekste. Järgnev tabel näitab igast korpusest leitud "heade" kollokatsioonide arvu, mis SENVA väljundist käsitsi välja sorteeriti.

	ilukirjandus	riigikogu	ajalehed
SENVA leitud mitmesõnalised verbid	3 000	5 800	8 500
nendest oli juba eelnevalt andmebaasis	1 900	2 600	4 200
polnud andmebaasis	1 100	3 200	4 300

Tabel 1. Mitmesõnalised verbid korpuses

Nagu näha, puudus andmebaasi aluseks olevatest sõnastikest oluline osa korpusest leitud sõnaühendeid. Just neid me otsisimegi ja see, et neid nii palju leidsime, õigustab kogu ettevõtmist. Kolme korpuse peale kokku tuvastati 9 900 erinevat mitmesõnalist verbi, millest 4 600 olid andmebaasis enne olemas ja 5 300 olid uued korpusest leitud kollokatsioonid.

Sõnaraamatute põhjal koostatud andmebaasis oli 11 300 kirjet, millest 6 700 ei õnnestunud korpustest leida. On tõenäoline, et paljud neist pakuvad küll ajaloolist huvi, kuid tuleks tänapäeva keele kajastamisele/analüüsimisele orienteeritud andmebaasist eemaldada.

Seda, et üllatavalt paljusid andmebaasis olevaid sõnaühendeid ei õnnestunud korpusest leida, võib (lisaks programmi puudustele) seletada kahe asjaoluga. Esiteks pööratakse sõnaraamatutes suurt tähelepanu idioomidele ja fraseologismidele, mis on (eriti mitte-ilukirjanduslikus) kirjutatud tekstis haruldased. Teiseks kalduvad andmebaasi aluseks olnud sõnaraamatud peegeldama ilukirjanduse keelt, ja eriti just ilukirjanduse keelt kuni 20. sajandi 80ndate aastateni - aga meie ilukirjanduse korpus oli teiste kasutatud korpustega võrreldes väike ja sisaldas 90ndate aastate ilukirjandust.

6.1 Täpsus

Mitmesõnaliste üksuste tuvastamine kümmet miljonit sõna sisaldavast korpusest osutus tunduvalt töömahukamaks ülesandeks kui me olime eeldanud oma eelnevast katsest poole miljoni sõnalise korpusega. Järgnev tabel näitab korpuste suurust, SENVA leitud kollokatsioonide koguhulka, "heade" kollokatsioonide hulka ja programmi töö täpsust.

	ilukirjandus	riigikogu	ajalehed
sõnu korpuses (miljonites)	0,5	12,6	9,8
SENVA tuvastatud kollokatsioonid	14 500	272 000	308 000
nendest mitmesõnalisi verbe	3 000	5 800	8 500
täpsus	21%	2%	3%

Näeme, et korpuse kahekümnekordne suurenemine tõi kaasa ka SENVA tuvastatud kollokatsioonide koguhulga suurenemise samas mahus, kuid "heade" kollokatsioonide hulk suurenes vaid 2-3 korda, see väljendub ka täpsuse olulises vähenemises.
Seda nähtust võib võrrelda erinevate üksiksõnade hulga kasvu vähenemisega korpuse suurenemisel.
Kuna meid huvitas maksimaalne saak, so tahtsime kätte saada võimalikult paljud korpuses esinevad mitmesõnalised verbid, siis ei üritanud me kollokatsioonide hulka piirata näiteks sagedusega, sest see oleks kaasa toonud haruldasemate mitmesõnaliste verbide väljajäämise.

6.2. Saak

Selleks, välja selgitada, kui palju korpuses tegelikult esinevatest mitmesõnalistest verbidest SENVA üles leiab, tegime järgmise katse. Püsiühendite andmebaasist valiti juhuslikult välja 500 ühend- ja väljendverbi. Nende esinemist korpuses kontrolliti käsitsi. Põhimõtteliselt peaks SENTA olema võimeline leidma korpusest neid kollokatsioone, mis esinevad seal vähemalt 2 korda, nii et kokkulugemisel arvestasimegi ainult selliseid.
Katse tulemused on esitatud järgnevas tabelis:

	ilukirjandus	riigikogu	ajalehed
mitmesõnalisi verbe	500	500	500
neist leidus korpuses	71	130	221
SENVA tuvastas	61	107	188
saak	86%	82%	85%

Tabel 3. SENVA saak erinevates korpustes

Sellest katsest võime järeldada, et 18-14% korpuses vähemalt kaks korda esinevatest mitmesõnalistest verbidest jääb SENVA poolt leidmata. Selle tavalisim põhjus peitub GenLocalMax algoritmis endas. Kui mitmesõnaline verb esineb küllalt sageli mingis kindlas kontekstis, siis see pikem kontekst domineerib lühema üle. Näiteks riigikogu stenogrammides esineb ühendverb *üles võtma* kontekstides *kutsuma üles võtma* ja *teemat üles võtma* nii sageli, et need kolmikud valib SENVA mitmesõnalise verbi kandidaatideks ja nii loobub kaksikust *üles võtma.*

Kõige sobivam viis sellise vea parandamiseks paistab olevat lingvistiliste kitsenduste lisamine programmile, mis eemaldaksid ebasobivaid kombinatsioone ja annaksid nii "headele" sõnaühenditele suurema võimaluse GenLocalMaxi poolt valitud saada.

7. Kokkuvõte

Olime seadnud endale eesmärgiks koostada mitmesõnaliste verbide võimalikult ammendav andmebaas, mis võiks olla aluseks erinevate leksikonide koostamisel, aga ka automaatsel süntaktilisel ja semantilisel analüüsil. Alustasime inimestele mõeldud sõnaraamatutest ja ühendasime seal leiduva info ühte andmebaasi. Kuid saadud baas sisaldas hulgaliselt vähekasutatavaid fraseologisme, samas puudusid sealt paljud tekstides laialt kasutatavad sõnaühendid. Probleemi lahendamiseks otsustasime otsida mitmesõnalisi verbe suurest, mitmekümne miljoni sõnalisest tekstikorpusest, kasutades selleks statistilist programmi SENVA, mis on eesti keele eripäradele ja antud ülesandele vastavaks kohandatud versioon programmist SENTA (Dias jt 2000). Programmi poolt leitud kollokatsioonid tuli käsitsi läbi vaadata, et otsustada, millised SENVA pakutud kandidaadid sobivad andmebaasi ja millised mitte. Selle töö tulemusena valmis mitmesõnaliste verbide andmebaas, milles on praegu 16 600 kirjet, neist 5 300 sellised, mida polnud andmebaasi aluseks olevates sõnaraamatutes, kuid mida SENVA leidis korpustest. Ehkki saadud andmebaas on mitmeti ebaühtlane, on ta siiski kasutatav kui toormaterjal erinevate teoreetiliste ja praktiliste küsimuste lahendamiseks, alates tõlkimisest ja lõpetades arvutilingvistikaga.

Viidatud kirjandus

Daille, B. Study and Implementation of Combined Technoques for Automatic Extraction of Terminology. Rmt: The Balancing Act: Combining Symbolic and

Statistical Approaches to Language. Cambridge, MA: London, England: MIT Press 1995, lk 49-66

Dias, G., Guilloré, S., Bassano, J. C., Lopes, J. G. P. Extraction Automatique d'unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire. Journal Traitement Automatique des Langues, 41/2 2000, lk 447-473

Kaaalep, H-J., Muischnek, K. Püsiühendite leidmine teksti abil. Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Tartu 2002, lk 172-184

Smadja, F. Retrieving collocations from text: XTRACT. Computational linguistics, 19/1 1993, lk 143-177