Püsiühendite leidmine teksti abil

Heiki-Jaan Kaalep
Kadri Muischnek

tööd on osaliselt finantseerinud ETF (grant nr 4352)

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. Lisaks on nende tundmine vajalik ka leksikograafias, keeleõppes jm. Kjellmer (1991) on näiteks väitnud, et meie mentaalne leksikon ei koosne mitte ainult üksikutest sõnadest vaid ka pikematest üksustest.

Momendil ei teata eesti keele püsiühenditest kuigi palju. Õigemini - teatakse küll, kuid see teave on suunatud peamiselt inimesele, keda huvitavad keelekasutuse nüansid ja väljendusrikkus. Nii sisaldavad sellealased publikatsioonid (nt EKSS, Hasselblatt 1990, Õim 1993, Õim 1998) küll suurel hulgal fraseologisme ja idioome, kuid pole keele automaattöötluses kuigi lihtsalt kasutatavad. Esiteks seetõttu, et nende viimine automaatselt töödeldavale kujule pole lihtne, ja teiseks seepärast, et paljusid nendes sõnaraamatutes toodud väljendeid kasutatakse tegelikes tekstides küllaltki harva. Näiteks ei leidu tänapäeva eesti kirjakeele korpuse 90ndate aastate allkorpuses (u 1 miljon sõna) kordagi selliseid väljendeid nagu *peenike peos* või *astla vastu üles lööma.*

Nii et meie ees on kaks küsimust:
1) Kas tekstides leidub veel (ja kui palju) püsiühendeid, mida sõnaraamatutes (ka fraseoloogiale orienteeritutes) ei esitata?
2) Kui laialt kasutatavad on erinevad püsiühendid?
Alustasime vastuse otsimisest esimesele küsimusele, sest alles siis, kui meil on olemas loend võimalikest püsiühenditest, saame vastata küsimusele, kui sageli neist igaühte kasutatakse. Seejuures otsustasime töövahendina kasutada statistilist meetodit kasutavat arvutiprogrammi, mis tekstikorpusele rakendatuna lihtsustab lingvisti tööd.

Kuna kirjeldatavas eksperimendis kasutatakse püsiühendite leidmiseks tekstist statistilisi meetodeid, on püsiühend siin võrdsustatud kollokatsiooniga. Kollokatsioon on sõnaühend, mis on defineeritud selle järgi, et teda moodustavad sõnad esinevad tekstides koos sagedamini, kui võiks eeldada nende eraldi esinemise sagedustest. Kollokatsioonid võivad olla väga erinevad nii neid moodustavate sõnade arvu poolest kui ka nende sõnade süntaktiliste funktsioonide ja omavaheliste seoste poolest. Nendeks võivad olla nii idioomid (nt *hambasse puhuma*), mida sõnaraamatud esitavad põhjalikult, kuid mida tekstides

harva esineb; ühend- ja väljendverbid, mida samuti sõnaraamatutes tüüpiliselt esitatakse (*üle saama, õppust võtma*); mitmesugused nimisõnafraasid (nt *rohelised mehikesed*). Lisaks eelpoolnimetatutele on kollokatsioonid näiteks veel kindla verbi ja nimisõna seosed (nt puid lõhutakse, mitte ei tehta katki), mis võõrkeeleõppijatele suurt peavalu valmistavad. Kollokatsioone moodustavad sõnad ei pruugi paikneda lauses vahetult üksteise järel. See kõik teeb nende automaatse tuvastamise tekstis keeruliseks.

Sõnaühendite või kollokatsioonide leidmiseks tekstis kasutatakse sõnade omavahelise seotuse mõõte. Enamus sõnade omavahelise seotuse leidmise meetodeid põhineb ideel lükata ümber hüpotees, et sõnad A ja B on üksteisest sõltumatud. Statistilisi meetodeid kasutataksegi, et mõõta kõrvalekallet sellest hüpoteesist. Mida suurem on kõrvalekalle, seda tugevam on seos sõnade A ja B vahel. Seega mõõdab sõnade omavahelist seotust valem, mis võrdleb sõnade A ja B tegelikke sagedusi mingis etteantud suurusega naabruses ja nende eeldatavaid sagedusi üksteise naabruses, mida saab tuletada nende sõnade sagedusest kogu tekstis. Väga hea lühikese kokkuvõtte enamkasutatavatest leksikaalsete seoste (ja seega ka kollokatsioonide) leidmise meetoditest esitab (Evert 2001). Evert ja Krenn (Evert ja Krenn 2001) on võrrelnud erinevaid meetodeid ja leidnud, et ükski nende poolt vaadeldud meetod polnud oluliselt parem kui teised. Nad väidavad ka, et madala sagedusega sõnapaaride leidmiseks polegi sobivat meetodit. Nagu siin artiklis edaspidi näidatakse, on tekstis kord-paar esinevaid sõnaühendeid võimalik tuvastada küll, aga selle hinnaks on suur käsitsitöö maht kõigi võimalike kandidaatide käsitsi kontrollimisel.

Mitmesõnaliste üksuste tuvastamine elektroonilisest tekstikorpusest ei ole nii lihtne kui esmapilgul paistab: programm võib "leida" tekstist väljendeid, mis koosnevad sõnadest, mis küll võivad selles tekstis sageli koos esineda, aga mingit mõttelist tervikut ei moodusta. Samas võib programm jätta tuvastamata selliseid sõnaühendeid, mida teksti käsitsi läbi vaatav lingvist peaks omavahel kokku kuuluvateks. Selleks, et statistikal põhinev programm võimalikult häid tulemusi annaks, peab ta arvestama ka analüüsitavate tekstide ja otsitavate väljendite lingvistiliste omadustega.

Selles artiklis kirjeldataksegi katset kombineerida lingvistilisi ja statistilisi meetodeid ühend- ja väljendverbide tuvastamiseks eestikeelses tekstikorpuses. Lühiduse mõttes ja ka analoogia põhjal ingliskeelse väljendiga *phrasal verb* on neid ühend- ja väljendverbe koos nimetatud siin fraasiverbideks. Miks ei ole kasutatud väljendit *perifrastiline verb*? Mõiste *perifrastiline verb* hõlmab ka modaalverbi ja infiniidi ühendeid (EKG II: 19), mille tekstis tuvastamine ei olnud kirjeldatava töö eesmärgiks. Programmi töö kontrollimiseks võrreldi korpusest leitud fraasiverbe püsiühendite andmebaasiga (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html), see võimaldas anda hinnanguid nii programmi tööle kui ka püsiühendite andmebaasile.

2. Tarkvara

Eesti fraasiverbide otsimiseks kohandasime G. Diasi loodud tarkvarapaketti SENTA (Software for Extracting N-ary Textual Associations – *n*-kohaliste sõnaühendite ekstraheerimise tarkvara) (Dias et al.

2000). SENTA kasutab keerulist matemaatilist valemit ja lokaalse maksimumi leidmise algoritmi, et hinnata tekstis esinevate sõnade kokkukuuluvust. Allpool kirjeldame neid lühidalt; põhjalikuma ülevaate annab (Dias et al. 2000).

2.1. 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.

2.1.1. Normaliseeritud oodatavus NO (*Normalised Expectation)*

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*. Olukorda kirjeldab tabel 1, kus iga rida tähistab üht võimalikku ootust.

Tabel 1: Näide ootustest, mida tuleb arvestada normaliseeritud oodatavuse hindamisel

Oodatav sõna	**Teades lünklikku kolmikut**
vahi	[_____ *+1 alla +2 võtma*]
alla	[*vahi +1* _____ *+2 võtma*]
võtma	[*vahi +1 alla +2* _____]

Normaliseeritud oodatavuse põhiideeks on hinnata ühe sõna jadast väljajätmise maksumust (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, 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.

2.1.2. Ühine oodatavus (ÜO) (*Mutual Expectation)*

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:

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

2.2. GenLocalMaxs algoritm

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.

3. Tekstikorpus

Kirjeldatavas katses kasutati üht osa tänapäeva eesti kirjakeele korpusest (http://www.cl.ut.ee/ee/corpusb/), nimelt 500 000-sõnalist väljavõtet 90ndate aastate ilukirjanduse allkorpusest. 90ndate aastate ilukirjanduskorpuses on kokku 611 000 sõna ja ta sisaldab ilukirjandustekste aastatest 1991-1998. Kuna eesti algupärase proosa kogutoodang on nii väike, on

korpusse valitud üks katke igast eesti keeles ilmunud proosaraamatust, lisaks ilukirjandust ajakirjast 'Looming'. Korpusse viidud tekstikatkete maht on ca 2000 sõna (või ka vähem, kui nt novell juhtus lühem olema), valikud arvestavad lõigupiire ja pole seepärast täpselt 2000 sõna pikkused. Korpuse koostamispõhimõtete kohta vt lähemalt nt (Hennoste, Muischnek 2000).

Tahtsime oma eksperimendis kasutada võimalikult tänapäevast teksti, valida oli ilukirjanduse ja ajakirjanduse korpuste vahel. Ilukirjandust eelistasime ajakirjandusele seepärast, et vastandina ajalehekeelele või suulisele keelekasutusele on see alati olnud traditsioonilise leksikograafia ja leksikoloogia põhiline "inspiratsiooniallikas". Eeldasime, et kasutades sama tüüpi teksti kui sõnaraamatute koostajad, õnnestub paremini võrrelda eelnevalt sõnaraamatute põhjal koostatud andmebaasi ja SENTA tulemusi.

4. Püsiühendite andmebaas

Püsiühendite andmebaas (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html) koosneb momendil väljendite andmebaasist ja fraasiverbide andmebaasist. Fraasiverbide andmebaasis, millega kirjeldatav eksperiment tehti, oli 2001. a. lõpu seisuga u 12 200 kirjet.

Andmebaas on koostatud järgmiste inimkasutajale mõeldud sõnaraamatute baasil:

1. "Fraseoloogiasõnaraamat" (Õim 1993)
2. "Eesti kirjakeele seletussõnaraamat" (EKSS)
3. Filosofti tesaurus (http://www.filosoft.ee)
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)

Fraasiverbid on meie andmebaasis jagatud ühend- ja väljendverbideks.
Ühendverb koosneb traditsioonilise grammatika järgi verbist ja selle tähendust muutvast abimäärsõnast. Käesoleva andmebaasi koostamisel on ühendverbideks nimetatud kõiki selliseid määrsõna ja verbi ühendeid, mida on sõnastikes esitatud omaette (ala)märksõnadena. Ka SENTA leitud määrsõna ja verbi ühenditest on andmebaasi võetud mitte ainult need sõnaühendid, kus (abi)määrsõna muudab verbi tähendust, vaid igasugused ühendid, mis tunduvad moodustavat omaette mõiste või mida nt inglise keelde ei saa tõlkida sõna-sõnalt (nt *istuli kukkuma*, *juurde tellima*). Seega võib öelda, et ühendverbide hulka on siinses andmebaasis tinglikult loetud igasugused määrsõna ja verbi ühendid v.a. verbi *olema* ja määrsõna ühendid. Verbi *olema* ühendid on fraasiverbide andmebaasist välja jäetud.

Ühendverbide andmebaasis leidub ka selliseid muutumatu sõna ja verbi ühendeid, kus muutumatu sõna ei ole tõenäoliselt mitte määrsõnaks, vaid kaassõnaks; selline ühend nõuab kindlas vormis käändsõna. Nt sellistes sõnaühendites, nagu *üle naerma* või *küüsi jätma* funktsioneerib muutumatu sõna alati

kaassõnana, mitte määrsõnana. Muidugi on selline tõlgendus ainult oletuslik, see, kuidas sellised sõnaühendid esinevad tegelikus keelekasutuses, vajab korpusel testimist.

Eesti keele grammatika (EKG II) nimetab väljendverbideks neid perifrastilisi verbe, mille sisuliseks tuumaks on noomen. Kõik väljendverbid on ainukordsed ühendid, mis moodustavad idiomaatilise tähendusterviku (EKG II: 20). Andmebaasi koostamisel on ka väljendverbi mõistet käsitletud võimalikult laialt: väljendverbidena on andmebaasi sisse võetud kõik nimisõna(de) ja verbi ühendid, mis on sõnastikes esitatud omaette (ala)märksõnadena. Samuti nagu ühendverbide puhul, on ka SENTA poolt tekstist tuvastatud nimisõna(de) ja verbi ühendite puhul andmebaasi võetud kõik ühendid, mis tundusid moodustavat omaette mõiste (nt *aega kulutama*, *imet tegema*, *bussi ootama*). Väljendverbide alla on andmebaasis tinglikult viidud ka kahe verbi ühendid (finiitverb+infiniit), nt *ajama panema, nahutada saama*. Kuid põhjustel, mida edaspidi täpsemalt selgitatakse, ei ole võimalik finiitverbi ja infiniidi ühendeid korpusest SENTA abil leida.

Enne katset võis oletada, et selline statistikal põhinev programm nagu SENTA ei anna eestikeelsele tekstile rakendatuna kuigi häid tulemusi, sest kokkukuuluvad sõnad võivad tekstis asuda üksteisest kaugel ja alati mitte samas järjekorras (*tahtis aru saada, sai aru*). Lisaks sellele ei esine kokkukuuluvad sõnad alati samal kujul (vrd nt *sai aru, ei saanud aru; rohelised mehikesed, rohelistele mehikestele)*. Võis eeldada, et programm leiab tekstist palju mõttetuid väljendeid, so sõnu, mis esinevad küll sageli koos samas osalauses, kuid ei moodusta ühte mõttelist tervikut. Statistikal põhinevad kollokatsioonide leidmise programmid on koostatud leidmaks korduvaid ja tõenäolisi seoseid sõnavormide vahel, nad ei arvesta erinevate keelte spetsiifikat.

Püsiühendite leidmisel eestikeelses tekstis tuleb kindlasti arvestada lemmatiseerimise probleemiga. Kas tekstis olevad sõnavormid on otstarbekam viia lemma kujule või jätta nad sellisteks, nagu nad tekstis esinevad? See sõltub sellest, milliseid püsiühendeid soovitakse leida. Näiteks nimisõnafraase tuvastades tuleb arvestada nii ühilduvate kui genitiivsete ja ka järeltäiendite ga. Kirjeldatava ülesande - ühend- ja väljendverbide leidmise - puhul tuli teksti ettevalmistamisel viia tekstis esinevad verbivormid lemma kujule ja ülejäänud tekstisõned jätta tekstis kasutatud kujule. Siit leiame vastuse küsimusele, miks ei ole SENTAga tuvastatud fraasiverbide seas finiitse verbivormi ja da-infinitiivi ühendeid (nt *tunda saama*), aga on ma-infinitiivi ja finiitse verbivormi ühendeid (nt *magama heitma*).

4. Eksperiment

Eksperimendis kasutati eelpoolkirjeldatud 500 000 -sõnalist tekstikorpust ja fraasiverbide andmebaasi.

Ühend- ja väljendverbide korpusest leidmiseks töödeldi teksti järgnevalt:
1. Tekstid analüüsiti morfoloogiliselt ja ühestati. Iga sõnavormi juures oli seejärel tema lemma ja info tema sõnaliigi, arvu, käände või pöörde jms kohta.

2. Verbidel säilitati lemma ja eemaldati tekstis esinenud sõnavorm ja muu morfoloogiline info, teistel sõnaliikidel hoiti alles sõnavorm ja eemaldati morfoloogiaanalüsaatori poolt lisatud info.
3. Leiti kõik võimalikud kollokatsioonid.
4. Eemaldati käesoleva ülesande seisukohalt mitteolulised kollokatsioonid, so kõik kollokatsioonid, mis ei sisalda verbi; asesõna sisaldavad kollokatsioonid v.a. mõned erandid (muidu oleks kõige sagedasem kollokatsioon *tema olema*), samuti eemaldati kirjavahemärke sisaldavad kollokatsioonid (SENTA jaoks on kirjavahemärk samasugune sümbol nagu täht, number vms), mõningaid adverbe sisaldavad kollokatsioonid. Nende eemaldatavate adverbide nimekiri kujunes töö käigus ja sinna kuuluvad näiteks sõnad *alati, muidugi, ikka* - so sellised adverbid, mis ei saa kuuluda ühendverbi koosseisu, kuid on tekstis küllalt sagedased.
5. Arvutati ÜO ja GenLocalMaxs; nende põhjal tehti leitud kollokatsioonide hulgast lõplik valik, mis läks programmi väljundisse.

Kirjeldatavas eksperimendis töödeldi tekstikorpust SENTAga 4 korda, iga kord erineva distantsiga (0 kuni 3), so kollokatsiooni moodustavate sõnade vahel võis olla 0 kuni 3 sõna. Selliselt saadud ühend- ja väljendverbi kandidaatide nimekirja võrreldi püsiühendite andmebaasiga, need väljendid, mida andmebaasis ei olnud, vaadati käsitsi üle ja otsustati igaühe puhul eraldi, kas tegemist on fraasiverbiga või mitte. See osa tööst - SENTA väljundist "mõistlike" verbiühendite väljavalimine - nõudis kõige rohkem inimtööd, sest programmi töö täpsus on vaid 19%.

5. Tulemused

SENTA leidis 500 000-sõnalisest tekstikorpusest 13 100 fraasiverbikandidaati. Neist 2500, so 19% olid nö mõistlikud kandidaadid, sellised, mida võiks meie kriteeriumite järgi nimetada ühend- või väljendverbideks. Nendest 2500-st 1630 olid püsiühendite andmebaasis juba olemas ja 870 oli selliseid, mis püsiühendite andmebaasis puudusid. Tabelis 2 on esitatud mõned neist fraasiverbidest, mis olid püsiühendite andmebaasis ja/või mida SENTA leidis tekstikorpusest. Nagu näeme, leidub sõnastikes ühelt poolt väljendeid, mida korpuses ei esine, teiselt poolt on aga korpuses küllalt igapäevaseid väljendeid, mida sõnastikud ei esita.

Tabel 2. Ühend- ja väljendverbe fraasiverbide andmebaasis ja SENTA väljundis

Püsiühend	**Andmebaas**	**SENTA väljundis**
abiellu astuma	+	-
abiellu heitma	+	-
abielu rikkuma	+	-
abielu sõlmima	+	-
abielu lahutama	-	+
andeks andma	+	+
andeks paluma	+	-
andeks saama	-	+

allkirja andma	-	+
hulluks minema	+	+
hulluks ajama	-	+
külla minema	+	-
külla tulema	+	+
külla kutsuma	-	+

Näeme, et SENTA leitud 2500-st ühend- ja väljendverbist olid andmebaasis olemas ainult 2/3. Pealegi on 500 000-sõnaline tekstikorpus selliste ülesannete lahendamiseks väikesevõitu, suuremate tekstihulkade kasutamisel võib loota veel olulisemat lisa olemasolevale andmebaasile.

6. SENTA töö kontroll

Töötades SENTAga nagu musta kastiga, kuhu antakse tekstikorpus sisse ja saadakse väljundiks potentsiaalsete püsiühendite loend, ei saa me teada, kuivõrd on tulemused usaldatavad, so kui palju tekstis tegelikult esinevaid püsiühendeid SENTA leiab ja kui palju SENTA leitud kollokatsioonidest tegelikult tekstis olemas on. Selle väljaselgitamiseks tehti järgmine katse. Püsiühendite andmebaasist valiti juhuslikult välja 500 ühend- ja väljendverbi. Nende esinemist korpuses kontrolliti käsitsi. Selgus, et nendest 500-st püsiühendist esines tekstikorpuses 131. Põhimõtteliselt peaks SENTA olema võimeline leidma korpusest neid kollokatsioone, mis esinevad seal vähemalt 2 korda. Selliseid ühend- ja väljendverbe oli 500-st 71.

Tekstikorpust SENTAga töödeldes tehti 4 katset, iga kord lubati erinevat distantsi (0 kuni 3 sõna) kollokatsiooni moodustavate sõnade vahel, lisaks veel kombineeritud distants so vahemik 0 kuni 3 sõna. Kontrolliti, mitu fraasi 71 hulgast SENTA leidis. Nende katsete tulemused on esitatud tabelis 2.

Tabel 3. SENTA leitud ühend- ja väljendverbide hulga sõltuvus lubatud distantsist püsiühendit moodustavate sõnade vahel.

Distants	0	1	2	3	kombineeritud
Püsiühendeid	45	46	50	52	57

Tabelist 3 näeme, et mida pikem on lubatud distants, seda rohkem õigeid väljendeid suudab SENTA korpusest leida. Kuid väärib märkimist, et distantsi pikendamisel ei leia SENTA enam mõningaid fraase, mida ta leidis lühema distantsi puhul. Kui suurendada distantsi kahelt sõnalt kolmele, toimub järsk muutus: SENTA ei leia enam mõningaid püsiühendeid, mis on korpuses sagedased. Nende 19 püsiühendi hulgas, mida SENTA distantsi suurenedes enam üles ei leidnud, esinesid 12 korpuses kaks korda, kuid viis püsiühendit olid üsna sagedased (vt tabel 4). "Koosesinemisi" antud tabelis tähistab juhtumeid, kus mõlemad sõnad esinevad samas lauses arvestamata seda, kas nad moodustavad ühend- või väljendverbi või mitte.

Tabel 4: Korpuses sageli esinevad püsiühendid, mida SENTA 3-sõnalise distantsi puhul enam ei tuvastanud

sõnaühend	koosesinemisi	püsiühendeid
ette näitama	10	9
hakkama saama	95	58
suitsu tegema	11	9
ära kasutama	21	19
ära maksma	12	9

Distantsidega 0, 1 ja 2 ei leidnud SENTA tabelis 4 toodud fraasidest ainult väljendit *ära maksma* (kuigi distantsi 3 puhul leidis ta rohkem väikese sagedusega püsiühendeid kui distantside 0, 1 ja 2 puhul).

SENTA andis välja ka selliseid püsiühendeid, mida sõnaraamatute põhjal koostatud andmebaasis ei olnud, aga mis sisaldasid endas andmebaasis leiduvaid fraase. Näiteks leidus andmebaasis väljend *ära maksma*, mida SENTA tekstikorpusest ei leidnud. Küll aga leidis ta väljendid *arve ära maksma* ja *võlga ära maksma*. See näib viitavat asjaolule, et ühendverbi *ära maksma* kasutataksegi põhiliselt nendes kontekstides. Kas ongi siis tegemist ühendverbiga *ära maksma* või hoopis kahe väljendverbiga?

Kontrollides käsitsi seda, kuidas suudab SENTA leida andmebaasist juhuslikult valitud viitsadat väljendit, selgus, et SENTA leidis tekstikorpusest ka mõned fraasiverbid, mida seal tegelikult ei olnud. Kuidas see võimalik on? Nimelt võivad ühend- või väljendverbi osad esineda samas osalauses ka ilma mõistelist tervikut moodustamata, so ilma kokku kuulumata. Nii leidis SENTA näiteks ühendverbi *tagasi tegema* lausest *Tagasi jõudes teeme sotid selgeks*, mis loeti veaks SENTA tulemuse hindamisel. Nagu arvata võib, kasvab selliste vigade hulk distantsi pikenedes.

Milliseid järeldusi saab teha nende 500 väljendi käsitsi kontrollimisest? Oletagem, et need 131 väljendit 500-st, mis korpuses esinesid, moodustavad juhusliku valiku kõigist korpuses esinevatest fraasidest. SENTA peaks suutma leida need väljendid, mis esinevad korpuses vähemalt 2 korda, antud juhul siis 71 131-st. Kasutades kombineeritud distantsi, võime me eeldada, et SENTA leiab korpuses olevatest väljenditest 57/71 = 80% neid, mis esinevad seal vähemalt kaks korda ja peaaegu 99% neist, mis esinevad kolm ja rohkem korda. Lisaks veel 8/60=12% neist, mis esinevad korpuses ühe korra. Seega on saak (*recall*) väga hea, aga täpsus kehv - mäletatavasti ainult 19%.

Kokkuvõte

Tekstikorpusest fraasiverbide leidmine pole kerge ülesanne. Siin näidati, kuidas lahendada seda ülesannet kombineerides lingvistilisi ja statistilisi meetodeid, kusjuures väljund vajab käsitsi toimetamist.

Artiklis kirjeldati statistilisel tõenäosusel põhinevat sõnade omavahelise seotuse mõõtude leidmise programmi SENTA (Software for Extracting N-ary Textual Associations) ja selle eestikeelsele tekstile rakendamise katset. Paremate tulemuste saamiseks tuli teksti eelnevalt töödelda: verbid viia tekstis lemma kujule, muudel sõnaliikidel säilitada tekstis kasutatud sõnavorm. Eestikeelseks tekstiks valiti 500 000-sõnaline osa tänapäeva eesti keele korpuse 90ndate aastate ilukirjanduse allkorpusest.

SENTA väljundit võrreldi mitmesuguste sõnaraamatute baasil koostatud eesti keele fraasiverbide andmebaasiga. SENTA töö kontrollimiseks teostati eksperiment, kus fraasiverbide andmebaasist juhuslikult valitud 500 ühendi esinemist kasutatud tekstikorpuses kontrolliti käsitsi. Kuigi kasutatud meetodi täpsus oli madal (19%), kompenseeris selle suur saak (*recall*) - SENTA leidis 99% ühenditest, mis esinesid korpuses kolm ja rohkem korda, 80% neist, mis esinesid kaks korda ning 12% ühenditest, mis esinesid üks kord.

Kuigi võis arvata, et eesti keele vaba sõnajärje ja keeruka morfoloogia tõttu ei tööta selline statistikal põhinev programm nagu SENTA eesti keelele rakendatuna eriti hästi, võib tulemusi vaadates järeldada, et see arvamus oli ekslik. See tähendab, et SENTA on leksikograafile hea abivahend: vaadata käsitsi läbi 13 100 fraasiverbi kandidaati on hoopis lihtsam kui otsida fraasiverbe 500 000-sõnalisest korpusest.

Viited

Daille B. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*. The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Ed. by J. Klavans and P. Resnik. Cambridge, MA; London, England: MIT Press 1995, pp 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, Vol 41:2, Christian Jacquemin (ed.). Paris, France, 2000, pp 447-473.

EKG II - Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika II Süntaks. Lisa: Kiri. ETA KKI Tallinn 1993

EKSS - *Eesti kirjakeele seletussõnaraamat (A-Žüriivaba).* ETA KKI, Tallinn, 1988-2000

Evert, S. On lexical association measures. http://www.collocations.de/EK/am-html/index.html 2001

Evert, S., Krenn, B. Methods for the Qualitative Evaluation of Lexical Association Measures. Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of ACL. CNRS, Toulouse, France, 2001, pp 188-196

Filosoft - *Tesaurus.* http://ee.www.ee/Tesa

Hasselblatt, C. Das Estnische Partikelverb als lehnübersetzung aus dem Deutschen. Wiesbaden 1990.

Hennoste, T., Muischnek K. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toimetaja Tiit Hennoste. Tartu 2000, lk 183-218.

Kjellmer, G. A mint of phrases. English Corpus Linguistics. Edited by Karin Aijmer and Bengt Altenberg. Longman 1991 pp 111-127

Saareste, A. Eesti keele mõistelise sõnaraamatu indeks. Finsk-ugriska institutionen, Uppsala 1979.

Õim, A. Fraseoloogiasõnaraamat. ETA KKI, Tallinn 1993

Õim, A. Sünonüümisõnastik. Tallinn, 1991

Õim, A. Väljendiraamat. Eesti Keele Sihtasutus. Tallinn, 1998