EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 157–172

EESTI KEELE PÜSIÜHENDID ARVUTILINGVISTIKAS: MIKS JA KUIDAS

Heiki-Jaan Kaalep, Kadri Muischnek

Ülevaade. Artikkel räägib püsiühendite automaattöötlusest arvutilingvistikas. Püsiühendi all mõeldakse siin kahe või enama sõna(vormi) ühendit, mida mingi tähenduse väljendamiseks on tavaks koos kasutada; selle definitsiooni alla mahuvad nii idiomaatilised kui ka kollokatiivsed ühendid. Arvutilingvistikas on püsiühendid probleemiks, sest nad komplitseerivad teksti alt-üles analüüsimudelit, mille järgi lause struktuuri ja tähenduse ehituskiviks on üksiksõna. Artikkel annab ülevaate püsiühendite automaattöötluse kolmest etapist – püsiühendite tuvastamisest, nende leksikoni koostamisest ja püsiühendite märgendamisest tekstis. Nende ülesannete lahendamiseks on arvutilingvistikas välja töötatud tüüpilised meetodid, kuid need meetodid on eesti keele kui vaba sõnajärjega morfoloogiliselt keeruka keele analüüsil rakendatavad ainult teatud reservatsioonide ja modifikatsioonidega. Artiklis analüüsitaksegi eesti keele "erivajadusi" selles vallas.*

Võtmesõnad: arvutilingvistika, püsiühendid, püsiühendite tuvastamine, püsiühendite leksikon, püsiühendite märgendamine, eesti keel

1. Sissejuhatus

1.1. Mis on püsiühend?

Termini *püsiühend* inglise vaste *multiword expression* võttis arvutilingvistikas kasutusele Ivan A. Sag koos kaasautoritega 2002. aastal avaldatud paljuütleva pealkirjaga artiklis "Multiword expressions: A pain in the neck for the NLP" (Sag jt 2002).[1] Püsiühendi all mõeldakse kahe või enama sõna(vormi) ühendit, mida mingi tähenduse väljendamiseks on tavaks koos kasutada. Need on keelendid, mida inimese mälus arvatavasti, aga hea arvutilingvistilise tarkvara leksikonis kindlasti talletatakse tervikutena. Samas on *püsiühend* omamoodi katustermin, mille alla

* Artikli valmimist on toetanud sihtfinantseeritav teadusteema SF0180078s08. Autorid tänavad anonüümseid retsensente asjatundlike märkuste ja kommentaaride eest.

[1] Eesti traditsioonis on püsiühendi kohta kasutatud ka terminit *fraseem* (vt nt EKK 2007: 679).

koondatud sõnaühendite grupid erinevad üksteisest nii oma püsivuse astme, tähenduse moodustumise viisi kui ka süntaktilise struktuuri poolest.

Miks on püsiühendid arvutilingvistikas omaette probleemiks?

Võrreldes lauseid (1) ja (2) näeme, et nende sõnavormiline koostis on täpselt sama. Kui läheme ainult selle teadmisega edasi süntaktilisse analüüsi, siis saavad nii sõnavorm *aru* kui ka sõnavorm *piima* süntaktilise objekti analüüsi. Kuid need laused on erineva predikaadi, erineva argumendistruktuuri ja erineva tegevusobjektiga: ühe lause süntaktiliseks keskmeks on lihtverb *saama*, kuid teisel hoopis püsiühend: väljendverb *aru saama*.

(1) Peeter ei saanud ülesandest aru.
(2) Peeter ei saanud poest piima.

Seega: kui morfoloogia-tasandil ehk võibki käsitleda iga tühikutevahelist stringi omaette analüüsiüksusena, mis saab oma sõnaliigi ja grammatiliste kategooriate analüüsi, siis edasi, süntaktilise ja semantilise analüüsi jaoks, on oluline mitmesõnalise leksikaalse üksuse või mitmesõnalise minimaalse semantilise üksuse tunnistamine ja äratundmine.

Kui sagedased on püsiühendid tekstides? Tabelis 1 on toodud andmed lihtverbide ja verbikesksete püsiühendite esinemise kohta u 314 000 sõnast koosnevas tekstikorpuses.

Tabel 1. Verbikesksete püsiühendite hulk tekstikorpuses

Tekstiklass	Sõnesid	Verbikeskseid püsiühendeid	Üksi esinevaid põhiverbe
Ilu	104000	4000	16800
Aja	111000	2600	14500
Hor	98000	2000	12600
Kokku	314300	8600	42900

Tekstiklassi märkivad lühendid tabelis 1: ilu – ilukirjandustekst, aja – ajakirjandustekst, hor – populaarteadusliku ajakirja Horisont tekstid. Põhiverbide all on siin mõeldud mitte-abiverbe (s.t nt verbi liitajavormist *oli teinud* läks arvesse ainult *teinud*) ja mitte-modaalverbe (s.t ühendist *sai teha* läks arvesse ainult *teha*). Andmetest järeldub, et ilukirjanduse, ajakirjanduse ja populaarteaduse peale kokku on keskmiselt enam-vähem iga viies põhiverb mingi verbikeskse püsiühendi osa; ilukirjandustekstis on seda iga neljas põhiverb.

1.2. Püsiühendite liigitusest

Kuna erinevat liiki püsiühendid käituvad tekstis erinevalt ja järelikult tuleb neile ka automaattöötlusel erinevalt läheneda, siis vaadeldakse selles osas veidi lähemalt püsiühendite liigitust ja selle liigituse aluseid.

Nagu juba eespool kirjas, on *püsiühend* omamoodi katustermin, mille alla koondatud sõnaühendite grupid erinevad üksteisest oma püsivuse astme, süntaktilise struktuuri ja tähenduse moodustumise viisi poolest.

Püsivuse astme all mõistetakse järgmisi tunnuseid. Kas püsiühendi komponendid esinevad tekstis alati samas järjekorras, kas komponendid on alati kõrvuti või võib nende vahel olla püsiühendisse mittekuuluvaid sõnu? Nii on adverbifraas

läbi ja lõhki mõnes mõttes tühikuid sisaldav muutumatu sõna: selle püsiühendi komponentide vorm tekstis ei muutu ja komponentide järjestus on samuti alati sama, kuju **lõhki ja läbi* ei esine.

Püsivuse astmega on lähedalt seotud ka see, kas püsiühend on ainukordne sõnaühend või moodustatakse ta ühe sõna kombineerumisel mingi sõnade loendi või semantilisse välja kuuluvate sõnadega. Nii on väljendverb *lööb lokku* ainukordne ühend, kuid ühendid *lööb/laseb/viskab hundiratast* on moodustatud käändsõna kombineerumisel kindlasse sõnaloendisse kuuluvate verbidega ning ühendid *ajab marru/raevu/vihale* verbi kombineerumisel ühte semantilisse välja kuuluvate sõnavormidega.

Oma süntaktiliselt struktuurilt võivad püsiühendid olla nii noomenifraasid – *Egiptuse nuhtlus*, *löök allapoole vööd*; adverbifraasid – *läbi ja lõhki*, *maani täis*; adpositsioonifraasid – (kellegi) *käe läbi*, *metsa poole*, kui ka verbi ja tema seotud laiendi püsivad ühendid – *jalga laskma*, *läbi saama*, *kõnet pidama*. Verbist ja tema laiendist koosnevate püsiühendite hulka saab omakorda jagada laiendi sõnaliigilise või fraasiliigilise kuuluvuse järgi (adverb või afiksaaladverb *vs*. noomen või noomenifraas *vs*. adpositsioonifraas) või süntaktilise (formaalselt objekti positsioonis *vs*. muu seotud laiend) kuuluvuse järgi.

Järgnevalt vaatleme püsiühendi tähenduse moodustumise viise ja püsiühendite liigitumist selle alusel.

Kui püsiühendi tähendus ei ole teda moodustavate sõnade tähenduste summa, on tegu idioomiga (nt *laskis jalga*, *käis alla*, *lai leht*). Idiomaatilisi ühendeid saab edasi liigitada, üks võimalik liigitus on näiteks läbipaistev–läbipaistmatu idiomaatiline ühend. Kui sõnad esinevad ühendis oma tavatähenduses, on tegu kollokatsiooniga (nt *kissitab silmi*, *kange kohv*). Probleemiks on siin muidugi tavatähenduse piiritlemine, eriti väga polüseemsete või veidi laialivalguva ja ebamäärase tähendusega sõnade puhul, nagu seda on näiteks verbid *tegema* ja *ajama* või substantiiv *asi*. Kas liigitada väljend *ajas asju* kollokatiivseks või idiomaatiliseks ühendiks?

Lõpuks sellest, kuidas need püsiühendite liigitamise alused omavahel kombineeruvad.

Vaatleme lihtsuse mõttes ainult verbikeskseid püsiühendeid. Tähenduse moodustumine jagab nad idiomaatilisteks ja kollokatiivseteks ühenditeks. Idiomaatilised verbikesksed püsiühendid (eesti traditsioonis *väljendverbid*) jagunevad läbipaistmatuteks (*peab lugu*, *saab vatti*, *tuleb toime*) ja läbipaistvateks (*laseb* mingi ettevõtmise *põhja*, *valib* mingi *tee*, *võtab sõna*) idioomideks vastavalt sellele, kas keelekasutaja on võimeline mingit ühendit mõistma ilma seda eelnevalt vormitähenduse paarina omandamata või mitte. Läbipaistvus on skalaarne tunnus (vt nt Moon 1998: 23) ja keelekasutajati erinev.

Püsivuse astmelt on läbipaistmatud idioomid üldjuhul ainukordsed ühendid. Läbipaistvate idioomide hulgas esineb ka neid, mis moodustatakse ühe sõna kombineerimisel mingi sõnaloendi või semantilise väljaga, ja kollokatsioonide hulgas on nii ainukordseid ühendeid kui ka ühe sõna kombinatsioone sõnaloendi või semantilise väljaga. Kõigi eesti keele verbikesksete püsiühendite komponentide järjestus muutub sõltuvalt lausetüübist ning komponentide vahel võib olla püsiühendisse mittekuuluvaid sõnu.

Laiendi sõnaliik jagab verbikesksed püsiühendid verbi ja (afiksaal)adverbi ühenditeks (*ühendverbid*) ning verbi ja noomeni(fraasi) või adpositsioonifraasi ühenditeks. Kuigi ühendverbe käsitletakse tavaliselt homogeense hulgana, on

nendegi seas nii idiomaatilisi kui ka mitte-idiomaatilisi ühendeid (nt *sai kaotusest üle* vs. *hüppas kraavist üle*). Huno Rätsep (1978) jagab verbi ja adverbi ühendid ainukordseteks (*Toots **kirjutas** naabri pealt **maha***) ja korrapärasteks (*alla/üles/sisse/välja* jne *tulema/minema/jooksma* jne) ühendverbideks. Esimesed moodustavad süntaktilise terviku, millest sõltuvad seotud laiendid (s.t on tervikuna predikaadiks), ent korrapäraste ühendverbide adverbilised komponendid ei kuulu H. Rätsepa järgi lihtlause verbaalsesse tsentrumisse, vaid on verbi seotud laiendid (Rätsep 1978: 28–29).

Verbi ja käändsõna püsivate ühendite hulgas on samuti nii idiomaatilisi (nt needsamad *peab lugu*, *võtab sõna*) kui ka kollokatiivseid (*vastab küsimusele*, *kehitab õlgu*) ühendeid. Omaette rühmana eristuvad siin tugiverbiühendid, s.t verbi püsivad ühendid tegevust väljendava noomeniga, kus ühendi põhitähenduse annab noomen, verbi osaks on vaid verbile omaste grammatiliste tähenduste väljendamine ja noomeni sidumine muude osalistega selles lauses, nt *teeb tööd*, *peavad sõda*. Kui järgida aritmeetika-metafoori – fraaside või lausungite tähendus on seotud sõnade tähendustega nagu summa on seotud liidetavatega –, siis tugiverbiühendite puhul on tugiverbi enda tähendus null.

Püsiühendite piiritlemisest ja liigitamisest on palju kirjutatud, kuid vähemalt Rosamund Mooni (1998: 2) ega Wolfgang Fleischeri (1982: 8) väitel pole selles kirjanduses üldaktsepteeritud ega üldkasutatavat terminoloogiat, eriti terminid *kollokatsioon* ja *idioom* on erinevates käsitlustes erineva mahuga. Üldiselt paistab korduvat väide, et püsiühendid moodustavad sellise jada, mille ühes otsas on täiesti fikseerunud ühendid, millele saab tähenduse omistada ainult tervikuna, ja teises otsas kollokatiivsed ühendid, mille komponendid on oma põhitähendustes ja ühendi kui terviku tähendus moodustub tema osade tähenduste summast, kuid neid on mingi tähenduse väljendamiseks tavaks koos kasutada (nt Benson jt 1986: 252–254, Moon 1998: 19).

Praktikas tähendab see seda, et püsiühendite hulga alamhulkadeks jagamisel jääb piirialadele ikkagi teatud hulk vaieldavalt klassifitseeritud väljendeid.

1.3. Millest räägib see artikkel?

See artikkel räägib püsiühendite töötlemisest arvutilingvistikas, kusjuures keskendutakse selle töö kolmele etapile: püsiühendite tuvastamisele (artikli 2. osas), nende leksikoni koostamisele ja püsiühendite märgendamisele tekstis. Kuna viimased kaks etappi on omavahel tihedalt seotud, käsitletakse neid koos artikli 3. osas.

Paar terminoloogilist märkust. Edaspidi kasutame väljendit *püsiühendite tuvastamine* tähistamaks sõnaühendite loendi moodustamist tekstikorpuse põhjal. Inglise keeles on käibel terminid *collocation / multi-word expression / multi-word unit extraction*. Kasutame väljendit *püsiühendite märgendamine* tähistamaks püsiühendite eksplitsiitset tähistamist tekstis.

Esmapilgul jääb mulje, et püsiühendite töötlemise kaks etappi – nende tuvastamine ja püsiühendite märgendamine tekstis – moodustavad omamoodi ringtsükli: tekstist tuvastatud püsiühendid tuleb (kas samas või mõnes teises) tekstis uuesti märgendada. Milleks need kaks etappi, kas tuvastamise käigus ei saaks tuvastatavaid püsiühendeid kohe ka märgendada? Vastus on eitav, sest nagu järgnevas 2. osas täpsemalt kirjeldatud, põhineb püsiühendite tuvastamine sagedusel ja statistikal,

mis võimaldavad küll öelda, et selles tekstikorpuses esinevad näiteks sõnad *järgi* ja *vaatama* nii sageli üksteise naabruses, et tõenäoliselt on tegemist püsiühendiga. Kuid püsiühendite tekstist tuvastamise statistilised meetodid, millest osas 2.2 täpsemalt juttu tuleb, on võimetud otsustama, kas muutumatu sõna *järgi* ja verbi *vaatama* vormid moodustavad püsiühendi igas üksikus lauses, näiteks lausetes (3) ja (4).

(3) **Vaatan** kohe märkmikust **järgi**.
(4) Statistika **järgi vaatab** ETV saateid üle 60% eestimaalastest.

Otsustada, kas püsiühendi potentsiaalsed komponendid nendes lausetes kokku kuuluvad ja püsiühendi moodustavad või mitte, saab ka statistiliste meetodite abil, mis aga erinevad meetoditest, mida kasutatakse püsiühendite tuvastamisel.

Artikkel üritab läheneda teemale võimalikult laialt ja analüüsida igat tüüpi püsiühendite arvutitöötlusega seotud probleeme. Ent kuna autorid ise on põhjalikumalt tegelenud just verbikesksete püsiühenditega, siis on nendega seotud problemaatikat käsitletud suurema detailsusega.

2. Püsiühendite tuvastamine

Lühidalt, tekstikorpuse põhjal püsiühendite loendi moodustamiseks tuleb lahendada järgmised ülesanded. Kõigepealt tuleb tekstikorpusest leida n-ö ühendikandidaadid, s.o sõnapaarid, ka -kolmikud või isegi -nelikud, mille komponendid paiknevad tekstis üksteise (kindlalt defineeritud) naabruses või on omavahel seotud mingil muul moel, näiteks süntaktiliselt. Nende ühendikandidaatide hulgast leitakse tõenäolised püsiühendid kas lihtsa sageduse abil (esinevad sageli koos, järelikult kuuluvad kokku) või mõnda statistilist meetodit kasutades. Saadud tõenäoliste püsiühendite loend vajab pea alati inimese poolt ülevaatamist, enne kui selle põhjal saab koostada püsiühendite leksikoni või andmebaasi. Viimase loomisel liigitatakse püsiühendid alamhulkadesse ning lisatakse sõnaühendite loendile mitmesugust grammatilist ja/või kontekstuaalset infot, mis peaks hõlbustama nende püsiühendite märgendamist tekstis.

2.1. Ühendikandidaatide moodustamine

Püsiühendite tekstist tuvastamist alustatakse n-ö ühendikandidaatide moodustamisega. Olgu meil lause (5):

(5) Karjane kaotas lambad silmist, kuid leidis nad õhtu hakul jälle metsast üles.

Kui me eelnevalt oleme teksti jaganud lauseteks, kuid mingil muul moel pole teksti töödelnud ning me ei sea ühendikandidaatide vahel esinevate sõnade arvule mingit ülempiiri, siis saame sellest lausest väga palju ühendikandidaate. Näiteks ühendikandidaadid, mille üks komponent on verbivorm *kaotas*, on järgmised: *karjane kaotas*, *kaotas lambad*, *kaotas silmist*, *kaotas kuid*, *kaotas leidis*, *kaotas nad*, *kaotas õhtu*, *kaotas hakul*, *kaotas jälle*, *kaotas metsast*, *kaotas üles*, *lambad kaotas*, *silmist kaotas*, *kuid kaotas*, *leidis kaotas*, *nad kaotas*, *õhtu kaotas*, *hakul kaotas*,

jälle kaotas, *metsast kaotas*, *üles kaotas*. On selge, et sellisel ühendikandidaatide moodustamise viisil on palju puudusi.

Esiteks, valesid paare, s.t müra on liiga palju, sest paaride moodustamise kontekst on liiga pikk. Teiseks, moodustades ainult sõnavormide paare, hajub koosesinemise sagedus sama ühendi eri muutevormide vahel. Kolmandaks, sellisel viisil moodustatud kandidaatpaaride komponentide järjekord järgib jäigalt sõnavormide järjekorda lauses. Näiteks lausetes (5), (6) ja (7) esineb ühendverb *üles leidma* kokku kolmel korral, kuid kirjeldatud viisil kandidaatpaare moodustades saaksime kolm erinevat ühendit: *leidis üles*, *leidnud üles* ja *üles leidnud*.

(6) Karjane ei leidnud lambaid üles.
(7) Kui karjane oli lambad üles leidnud, läks ta nendega koju.

Loetletud probleemid ei ole üllatavad, kui me eeldame, et ühes lauses olevad sõnad on omavahel seotud, kuid sõnade lauses käitumise kohta ei ole meil mingeid teadmisi. Tegelikult saame oma keelealaseid teadmisi ja hüpoteese kasutades kandidaate siiski täpsemalt valida.

Liiga pika konteksti vältimise lihtsaim viis on eeldada, et tihedamalt seotud sõnad on ka lauses üksteisele lähemal, ja määrata kindlaks kandidaatühendi komponentide maksimaalne kaugus üksteisest. Praktikas lubatakse tavaliselt komponentide vahele maksimaalselt neli sõna. Kuid näites (5) on ühendverbi komponentide *leidis* ja *üles* vahel viis sõna. Parem viis oleks piirata kandidaatpaaride moodustamist osalausepiiridega, kuid osalausete piiride määramine pole triviaalne ülesanne – see eeldab teksti täielikku morfoloogilist ühestamist ja vähemalt osalist süntaktilist analüüsi.

Selleks, et ühendid *leidis üles* ja *leidnud üles* suudetaks lugeda sama ühendi muutevormideks, on tekstikorpus, millest püsiühendeid otsitakse, vaja morfoloogiliselt ühestada, s.t lisada igale tekstisõnale selles kontekstis ainuõige info tema lemma ja grammatiliste kategooriate kohta. Samas, ekslik oleks arvata, et morfoloogiliselt ühestatud korpust kasutades saame tekstisõnad kõrvale jätta ning tegeleda ainult lemmade koosesinemistega. Ühendverbide kui muutumatu sõna ja tekstis muutuva verbi ühendite tuvastamiseks võib tõesti kõik tekstisõnad asendada lemmadega, s.t *leidis* → *leidma*, *leidnud* → *leidma* ja *üles* → *üles*. Kuid verbi ja noomeni kindla muutevormi püsivate ühendite, näiteks väljendverbide puhul on asi teisiti. Näiteks näites (5) esineva väljendverbi *silmist kaotama* leidmiseks tuleb tekstis esinev verbivorm asendada lemmaga, kuid kui käändevorm *silmist* asendatakse tema lemmaga *silm*, saame ühesugused sõnapaarid *silm kaotama* lausetest (5), (8) ja (9), s.t eemaldatakse statistika poolt kasutatav info.

(8) Ta kaotas oma alluvate silmis igasuguse usalduse.
(9) Ta kaotas enne surma silma.

Veelgi keerulisemaks teeb lemma *vs*. tekstisõna valiku asjaolu, et paljude verbi ja noomeni püsivate ühendite nominaalne komponent muutub tekstis vastavalt objekti käändevahelduse reeglitele (nt *pidas kõne* vs. *ei pidanud kõnet*, *sirutas abistava käe* vs. *ei sirutanud abistavat kätt*), vt täpsemalt (Muischnek 2006).

Verbiühendite tuvastamisel tuleb arvestada ka eesti keelele omast vaba sõnajärge. Lausetest (6) ja (7) saame muude hulgas sõnapaarid *üles leidma* ja *leidma üles*, mis tuleb enne statistilisse töötlusse suunamist kas samale kujule viia või siis kasutada meetodit, mis ei arvesta paariliste järjekorda.

Eelpool öeldust saab järeldada, et korraga püüda tekstikorpusest kätte saada kõiki seal esinevaid püsiühendite liike on küllalt keeruline, sest nad käituvad tekstis niivõrd erinevalt. Näiteks vastupidiselt verbiühenditele nimisõnafraaside nagu *kange kohv* komponentide järjekord tekstis ei muutu, ka ei saa nende vahel olla muid sõnu; ühend käändub tekstis, kuid ühendi komponendid on alati samas käändes ja arvus (v.a neli viimast käänet). Ja vastupidi, ühendi *hullu lehma tõbi* esikomponendid tekstis ei muutu. Nii et lihtsam on läheneda püsiühendite eri liikidele n-ö individuaalselt. Üks häid tulemusi andev meetod on süntaktiliselt analüüsitud korpuse kasutamine, piisab ka osalisest süntaktilisest analüüsist. Süntaktilise analüüsi põhjal saab tuvastada näiteks verbi ja tema objekti, verbi ja tema muude seotud laiendite paare, noomenifraasilisi püsiühendeid jms.

2.2. "Tõeliste" püsiühendite väljasõelumine

Kui tekstikorpuse põhjal on püsiühendikandidaadid moodustatud, järgneb n-ö müra väljafiltreerimine ja püsiühendikandidaatide järjestamine. Mida keerulisemat meetodit kasutades (osalausepiiride arvestamine, morfoloogiline, süntaktiline analüüs) on moodustatud püsiühendite kandidaadid, seda vähem on vaja hiljem jõupingutusi teha müra väljafiltreerimiseks. Mis on müra? Näiteks moodustades analüüsimata tekstist sõnapaare maksimaalse distantsiga neli sõna ja järjestades need sõnapaarid sageduse alusel, on sagedusloendi tipus sellised sõnavormipaarid nagu *see on*, *ta on*, *ta oli*, *ja on*, *ja et*, s.t sagedusloendi tipu paarid koosnevad väga sagedaste sõnavormide kombinatsioonidest. Aitab siin nn stopp-sõnade loend, s.t loend sõnadest või sõnavormidest, mida sisaldavad paarid ei ole kunagi otsitavad püsiühendid. Tavalised stopp-sõnad on enamik asesõnu, enamik sidesõnu, verbi *olema* vormid, sellised adverbid nagu *ikka*, *enam*, *ainult* jne.

Kui otsitakse teatud tüüpi püsiühendeid, näiteks verbiühendeid, saab kasutada nn morfoloogilist filtrit, s.t statistilisse töötlusse suunatakse ainult need paarid, mille üks komponent on verb.

Nüüd järgneb statistilise töötluse etapp. Lihtsaim statistiline meetod on lihtne sagedusloend, seda saab üpris edukalt kasutada siis, kui vähemalt üks otsitava püsiühendi komponent ei kuulu väga sagedaste sõnade hulka. Näiteks annab lihtsa sageduse kasutamine paremaid tulemusi väljendverbide kui ühendverbide puhul. Põhjuseks on asjaolu, et püsiühendeid kalduvad moodustama just sellised sagedased verbid nagu *tegema*, *saama*, *pidama* jt. Ka ühendverbi koosseisus esinev partikkel on väljaspool ühendverbi tekstis sagedasem kui väljendverbi komponendiks olev käändsõnavorm. Seega võivad partikkel ja verb küllaltki sageli esineda samas osalauses ilma tegelikult kokku kuulumata (vt näide 4), käändsõnavormi ja verbi puhul esineb sellist "müra" tunduvalt vähem.

Sõnadevahelise seose tugevuse arvutamise aluseks keerulisemate statistiliste meetodite puhul on järgmine mõttekäik. Teades sõnaühendit moodustavate üksiksõnade esinemissagedusi ühes tekstikorpuses, saame arvutada, kui sageli satuksid need kaks sõna üksteise naabrusse (nt samasse osalausesse) eeldusel, et sõnad esinevad tekstis juhuslikult. Seda teoreetilist/hüpoteetilist koosesinemise sagedust nimetatakse *oodatavaks sageduseks*. Tekstikorpusest leiame *tegeliku* (*empiirilise*) *sageduse*, mis näitab, kui sageli need sõnad tegelikult üksteise naabruses esine-

vad. On mitmesuguseid meetodeid, mis võimaldavad hinnata oodatava ja tegeliku sageduse erinevuste olulisust.

Arvutilingvistikas on vastavaid valemeid kirjeldatud üle 80 (nt Pecina 2005) ja nende hulgast just käsiloleva ülesande lahendamiseks sobivaima leidmine pole lihtne. Põhjaliku ülevaate nendest seosetugevuse arvutusviisidest annab Stefan Evert (2004) oma doktoritöös, kus analüüsitakse rohkem kui 30 statistikut. Korpuslingvistika käsiraamatu kollokatsioone käsitlevas artiklis ütleb S. Evert (2008), et on pea võimatu soovitada üht, igasuguste andmete jaoks alati sobivat statistikut ning soovitab valida mitu, kuna need pakuvad koosesinemise andmestikule erinevaid vaatenurki.

Eesti keele jaoks on rakendatud seosetugevuse mõõdikut nimega ühine oodatavus tuvastamaks verbikeskseid püsiühendeid tekstikorpuses (Kaalep, Muischnek 2003).

Iga selliselt saadud püsiühendikandidaatide loend vajab inimese poolt ülevaatamist. Mida rohkem tööd on tehtud korpuse eeltöötlemisel (osalausepiirid, morfoloogiline ja süntaktiline analüüs), stopp-sõnade loendi koostamisel ja just selle materjali jaoks sobiva statistiku(te) valimisel, seda "puhtam" on tulemuseks saadud sõnaühendite loend.

3. Püsiühendite automaatne märgendamine tekstis: programm ja andmebaas

Artikli see osa räägib püsiühendite andmebaasipõhisest märgendamisest. Andmebaasi ülesehitus ja seal esitatava info hulk sõltuvad märgendamisprogrammi algoritmist ja vastupidi: tarkvara loomisel tuleb otsustada, millist infot püsiühendite käitumisomaduste kohta peab sisaldama andmebaas ja mida saab hallata programmiga. Edasi analüüsitaksegi osas 3.1 kõigepealt neid nähtusi, millega püsiühendite märgendamise tarkvara peab toime tulema. Siis, osas 3.2 arutletakse, kuidas seda infot otstarbekalt jagada programmi ja andmebaasi vahel.

3.1. Püsiühendid tekstis

Mingi keelenähtuse automaatse märgendaja loomise eeltööks on märgendatava nähtuse käitumisomaduste uurimine. Püsiühendi leksikaalse andmebaasi põhise märgendamise korral on oluline teada, kas püsiühendid esinevad tekstis täpselt sellistena, nagu nad andmebaasi aluseks olevas väljendiloendis kirjas on, või, olenevalt püsiühendi tüübist, muutuvad suurema või väiksema vabadusega.

Inglise keele keskses arvutilingvistikas peetakse reegliks sellist olukorda, et püsiühendid käituvad nagu "tühikutega sõnad" (nt ee *läbi ja lõhki*), mis esinevad alati samal kujul, ja alles hiljaaegu on avastatud, et ka näiteks idiomaatilised püsiühendid käituvad tekstis palju mitmekesisemalt kui seni arvatud (vt nt Riehemann 2001 3. ptk). Siinjuures kehtib üldine seaduspärasus, et mida läbipaistmatum on sõnaühendi tähendus, seda vähem ta (eesti keele puhul siiski ainult tema käändsõnaline komponent) tekstis varieeruda saab.

Pikemalt saab verbikesksete püsiühendite varieerumisest tekstis lugeda artiklist (Muischnek 2006). Siinkohal peatume lühidalt püsiühendite märgendamise seisukohalt olulistel asjaoludel.

Eesti keele puhul on ilmselt olulisim muutemorfoloogia – kas otsitav keelend koosneb muutumatutest, käänd- või pöördsõnadest ja kuidas käänd- ja pöördsõnad selle ühendi koosseisus muutekategooriatega kombineeruvad. Eesti keele verbikeskse püsiühendi süntaktiliseks tuumaks olev verb kombineerub üldjuhul vabalt kõigi verbi jaoks relevantsete morfoloogiliste kategooriatega, tema käitumist piiravad pigem sellised tegurid, mis piiravad üldse verbide kombineerumist grammatiliste kategooriatega (näiteks üldiselt ei saa impersonaali moodustada verbist, mille tegevussubjektiks ei ole inimene, nt *sajatakse*), aga mitte verbi püsiühendisse kuulumisest tingitud piirangud. Siiski on verbikesksete püsiühendite hulgas ka selliseid, peamiselt pragmaatilise funktsiooniga ühendeid, mis esinevad ainult imperatiivis (nt *võta näpust*, *võta või jäta*, *võta üht ja viska teist*), mida tuleb automaattöötlusel kohelda muutumatute stringidena.

Verbikeskse püsiühendi käändsõnaline komponent on enamasti n-ö kivistunud mingisse kindlasse käände- ja arvuvormi. Nii näiteks esineb ühendi *joonde ajama* käändsõnaline komponent ainult ainsuse lühikeses sisseütlevas; väljend pole võimalik kujul **joonesse ajama* või **joontesse ajama* või **joonele ajama*. Ent teatud tingimustel võib verbikeskse püsiühendi käändsõnaline komponent siiski käändes ja/või arvus varieeruda. Järgnevalt vaatlemegi neid varieerumisvõimalusi lähemalt.

3.1.1. Varieerumine käändes

Suur osa verbikesksete püsiühendite käändsõnalistest komponentidest on vormiliselt verbi objektiks (nt *saab aru*, *lööb lokku*, *teeb otsuse*, *paneb punkti*). Eesti keele objekti iseloomustab teadagi käändevaheldus vastavalt objekti käändevahelduse reeglitele. Üldine reegel on see, et mida idiomaatilisem, kivinenum on väljend, seda kindlamini on tema käändsõnaline komponent kivistunud objekti markeerimata käändesse – partitiivi. Läbipaistmatutes idioomides totaalobjekti ei esine, läbipaistvate idiomaatiliste ühendite puhul on aga umbes neljandik objektidest tekstikorpuses totaalsed, s.t genitiivis või nominatiivis (nt *rääkis augu pähe*, *vahtis silmad peast*, *andis rohelise tee*, *tegi puhta töö*).

Tugiverbiühendites esitab objektnoomen subjekti poolt sooritatavat tegevust (nt *tegi tööd*, *ajas juttu*, *tegi otsuse*, *sai alguse* jne). Objekti käändevaheldus ei sõltu siin mitte niivõrd verbi, kuivõrd tegevust väljendava noomeni tähendusest – kas kirjeldatav tegevus on teeline, s.t kas tähenduse oluliseks tunnuseks on tegevuse tulemuslikkus (nt otsustamine, algamine), või ateeline, s.t tegevuse tulemuslikkus ei ole tähenduse oluliseks osaks (nt töötamine, vestlemine). Kui objektiga väljendatav tegevus on ateeline, on objekt alati partsiaalne ja sarnaneb selles mõttes ainesõnaga (10). Kui objektiga väljendatav tegevus on teeline, otsustab tema käändevahelduse lause perfektiivne/imperfektiivne aspekt (11-12).

(10) Pühapäeval **pidasid** nad suvilas **pidu**.
(11) President **pidas** piduliku **kõne**.
(12) President **pidas** parajasti **kõnet**, kui ..

3.1.2. Varieerumine arvus

Nii nagu varieerumine käändes, sõltub ka püsiühendi nominaalse komponendi arvuvahelduse võimalus püsiühendi tüübist. Läbipaistmatu idiomaatilise ühendi nominaalne komponent arvus ei muutu. Tugiverbiühendite ja kollokatiivsete ühendite objektnoomeni arvuvahelduse võimalus sõltub objektnoomenist: on kollokatiivse ühendi objektiks ainesõna (nt *nõudis õigust*, *soovis õnne*), arvuvaheldust ei toimu. Tugiverbiühendid käituvad vormiliselt sama üldreegli järgi, ent ainesõna sarnaselt mitmuses mitte esinev objektnoomen väljendab ateelist tegevust, mille puhul on rõhk tegevusel endal, mitte selle tulemuslikkusel (nt *tegi tööd*, *avaldas mõju*, *andis abi*).

Kõige heterogeensemad on selles suhtes jällegi läbipaistvad idiomaatilised ühendid, mille hulgas leidub selliseid sõnapaare, mille nominaalne komponent võib olla nii ainsuses kui ka mitmuses (nt *kortsutab kulmu~kulme*, *teeb silma~silmi*, *heidab varju~varje*, *toob ohvri~ohvreid*). Predikaadina toimiva ühendi nominaalne komponent võib n-ö ühilduda arvus subjektiga (*tema murrab pead* vs. *nemad murravad päid*). Siiski esineb läbipaistvate idiomaatiliste ühendite varieerumist arvus vähem kui varieerumist käändes.

3.1.3. Püsiühendi komponentide paigutus

Lisaks muutemorfoloogiale on oluline ka püsiühendi komponentide võimalik sõnajärg – näiteks käändsõnafraasilistel püsiühenditel on see püsiv: sellises järjekorras nagu sõnad leksikoni on kantud, esinevad nad ka tekstis. Ent verbiühendi komponentide omavaheline järjestus sõltub lausetüübist ning verbiühendi puhul tuleb arvestada ka sellega, et püsiühendi komponentide vahel võib olla mitu püsiühendisse mittekuuluvat sõna (vt näiteid 5, 6 ja 7); verbi ja noomeni või (afiksaal) adverbi püsiva ühendi komponendid võivad asuda lausa teine teises (osa)lause otsas (13).

(13) **Saa** nüüd oma kaotusest ometi ükskord **üle**!

Kas püsiühendid saavad ületada osalausepiire? Vastus on jah, kuid harva. Nagu näha näitelausest (14), võivad osalausepiire ületada isegi läbipaistmatud väljendverbid ja muidugi ka tugiverbiühendid (15, 17) ning kollokatiivsed ühendid (16). Näidetes (14) ja (15) jätkub püsiühendit sisaldav pealause pärast kõrvallauset ja püsiühendi komponendid asuvad siiski samas osalauses (kuigi seda, et pärast kõrvallauset jätkub sama osalause, on automaatselt raske kindlaks teha), kuid tugiverbiühendite (17) ja kollokatiivsete ühendite (16) nominaalseid komponente saab aga laiendada püsiühendi verbilist komponenti sisaldava relatiivlausega, sellisel juhul asuvad püsiühendi komponendid tõesti eri osalausetes.

(14) Pealegi **lasid** mõlemad taksojuhid, kes minu autot blokeerisid, **jalga**, ja ..
(15) Samal hetkel **tundsid** mõlemad, nii Luik kui Sergejev, **kergendust** ..
(16) Naine nendib, et sai haiglast **abi**, mida **vajas**.
(17) See teema läbis presidendi **kõnet**, mille ta **pidas** ..

3.2. Kuidas varieerumisega toime tulla: mida esitada andmebaasis ja mis jätta programmi hooleks?

Kui märgendatava nähtuse keelelised omadused on kaardistatud, tuleb vastu võtta põhimõttelised otsused selle kohta, kuidas jagada märgendatava keelendi variatiivsusega toimetulekuks vajalikud ülesanded optimaalseimal viisil andmebaasi ja märgendusprogrammi vahel. Samuti tuleb leida optimaalne tekstiüksus, mille piires püsiühendi komponente otsida. Artikli selles osas analüüsitaksegi esiteks osalausepiiridega arvestamise vajadust ning teiseks sõnajärje-, käände- ning arvuvahelduse haldamist püsiühendite märgendamise tarkvara poolt.

Püsiühendite alaliigid erinevad üksteisest selle poolest, kas nende märgendamisel on vajalik osalausepiiride eelnev märgendamine või mitte. Käändsõnafraasiline püsiühend saab nagunii koosneda ainult üksteisele vahetult järgnevatest komponentidest, nii et tuvastamata osalausepiirid nende märgendamisel segadust ei tekita.

Nagu eelmises osas näidatud, võivad verbikesksed püsiühendid ületada osalausepiire, kuigi harva. Kuid lubades püsiühendite automaatsel märgendajal otsida potentsiaalse ühendi komponente üle osalausepiiride, põhjustame palju müra. Nii koosneb lause (18) kahest rinnastatud osalausest, milles mõlemas on predikaadiks väljendverb, vastavalt *nõu pidama* ja *aru saama*. Kuid verbid *pidama* ja *saama* ning käändsõnavormid *nõu* ja *aru* võivad kombineeruda ka verbikeskseteks püsiühenditeks *aru pidama* ja *nõu saama*. Seega, kui lubame tarkvaral otsida ühendi komponente erinevatest osalausetest, siis peame rohkem jõupingutusi tegema selle nimel, et tarkvara suudaks eristada tõelisi püsiühendeid püsiühendite potentsiaalsete komponentide juhuslikest koosesinemistest. Kui aga piirame püsiühendi võimaliku esinemispiirkonna osalausega, siis lähevad kaotsi lausetes (14–17) esinenud ühendid.

(18) Valitsus **pidas nõu** ja **sai** siis **aru**, et ..

Nagu kirjeldatud osas 3.1, käituvad püsiühendite erinevad liigid erinevalt selle suhtes, kas ja kuidas nad on võimelised tekstides morfoloogiliselt muutuma. Selle muutumisega toimetulekuks on mitu võimalust.

Esiteks võib kõik võimalikud muutevormid esitada andmebaasis ja otsida neid puhtast tekstist kui erinevaid püsiühendeid. See paisutab andmebaasi mahtu ja muudab tema struktuuri keerulisemaks: info, mis kuulub iga püsiühendi juurde, peab olema kajastatud ka iga esinemisvormi juures. Väljendi otsimine tekstist on seejuures aga lihtne.

Teine võimalus on kasutada sisendina morfoloogiliselt ühestatud teksti ja andmebaasis hoida iga väljendi juures info tema komponentide algvormi ja temaga tekstis kombineeruda võivate morfoloogiliste kategooriate kohta. See eeldab, et sisendteksti töötlemise programmid, näiteks morfoloogiline ühestaja, teevad oma töös väga vähe vigu.

Kui andmebaas on suur ja heterogeenne, nagu on näiteks eesti keele verbikesksete püsiühendite andmebaas,[2] mis sisaldab ühendverbe, idiomaatilisi väljendverbe, tugiverbiühendeid ja kollokatiivseid verbi ja käändsõna ühendeid, siis tuleb sealsed kirjed varustada mingi infoga, mis ütleks märgendusprogrammile näiteks ühendi

[2] http://www.cl.ut.ee/ressursid/pysiyhendid/ (12.02.2009).

üle saama kohta: see on ühendverb, genereeri/otsi kõiki verbi vorme, kuid mitteverbilist komponenti ära muuda; aga ühendi *otsust tegema* kohta: see on tugiverbiühend, mille nominaalne komponent muutub arvus ja objektikäänetes, genereeri/otsi kõiki verbi vorme ja käändsõnalise komponendi vorme objektikäänetes nii ainsuses kui mitmuses.

Nicole Grégoire (2007) on lahendanud selle probleemi hollandi keele püsiühendite leksikoni koostades nii, et on jaganud kõik püsiühendid nn ekvivalentsiklassidesse (ingl *Equivalence Class*), mille kõik liikmed käituvad tekstides täpselt ühte moodi. Nii saab püsiühendite andmebaasis korraga määrata terve klassi jaoks, millised vormid genereerida või milliseid vorme otsida.

On selge, et mida täpsematesse klassidesse on leksikon jaotatud, seda täpsem on tulemus. Nagu osas 3.1 kirjeldatud, on eesti keele verbikesksete püsiühendite hulgas homogeensed klassid ühendverbid ja läbipaistmatud idiomaatilised ühendid. Seevastu näiteks läbipaistvad idiomaatilised ühendid on heterogeenne klass ja vajab täpsemat liigitamist vastavalt ühendi käändsõnalise komponendi muutumisvõimele.

Kui püsiühendite märgendamise programm on leidnud, et samas osalauses esinevad koos andmebaasis oleva püsiühendi komponendid nõutavates vormides, siis tuleb veel otsustada, kas nad kuuluvad tõesti kokku või esinevad samas osalauses juhuslikult. Näiteks on eesti keele ühendverbide afiksaaladverbiline komponent sageli kasutatav ka adpositsioonina. Kui püsiühendeid üritatakse märgendada morfoloogiliselt ühestamata tekstis, puudub seal info tekstisõna sõnaliigilise kuuluvuse kohta. Nii on näites (19) olemas ühendverbi *üle kuulama* komponendid, mis seda ühendverbi ometi ei moodusta.

(19) Poole kõrvaga **kuulas** ta teise kurtmist oma raske elu **üle**.

Verbikesksesse püsiühendisse kuuluda võiv käändsõnavorm esineb üldiselt harva püsiühendit moodustava verbiga samas osalauses püsiühendit moodustamata, aga vahel siiski. Nii ei ole lauses (20) tegelikult väljendit *kätt paluma*, kuigi selle komponendid seal mõlemad esinevad.

(20) Vaadake **palun** minu **kätt**!

Näites (19) esitatud probleemi lahendamiseks piisab morfoloogilisest ühestamisest, mis annab teada, et sõnavorm *üle* on kaassõna. Kõige kindlam rohi lause (20) tüüpi vigade vastu on süntaksianalüüs, piisab ka osalisest ja pindmisest analüüsist, mis ütleb, et sõnavorm *kätt* on verbi *vaatama*, mitte verbi *paluma* objekt.

Raske on ka automaatselt eristada sama sõnaühendi idiomaatilist ja sõnasõnalist kasutust. Üldiselt on väidetud, et kui mingid sõnavormid koos esinedes moodustavad idioomi, siis sõnasõnalises tähenduses neid sõnavorme samas süntaktilises suhtes ei kasutata (vt näiteid 21 ja 22), välja arvatud juhul kui sõnaühend ühe leksikaalse üksusena on polüseemne (näited 23 *vs.* 24).

(21) **Tegi näo**, et ei saa midagi aru.
(22) Lilla värv **teeb** talvevalguses **näo** kahvatuks.
(23) Ta **neelas** tableti kähku **alla**.
(24) Vaikides **neelas** ta solvangu **alla**.

Paul Cook jt (2007) aga väidavad, et inglise keeles esinevad *ca* 40% fraasidest, millel on idiomaatiline tähendus, tekstis sõnasõnalises tähenduses. Järelikult lisandub idiomaatiliste ühendite puhul püsiühendite tekstis märgendamisel veel üks alamülesanne: otsustada, kas sõnaühendit on kasutatud idiomaatilises või sõnasõnalises tähenduses.

Eristamaks sama väljendi idiomaatilist ja mitte-idiomaatilist kasutust pakuvad mainitud autorid välja süntaktilisel jäikusel (ingl *syntactic fixedness*) põhineva meetodi. Selle all mõtlevad autorid seda, et idiomaatiline ühend esineb tavaliselt tekstis vähestes n-ö kanoonilistes vormides ja et sõnasõnalise tähendusega ühend, vastupidi, varieerub rohkem. Süntaktiliseks jäikuseks võiks pidada näiteks nimisõna atribuudiga laiendamise võimatust või piiratust. Nii saab idiomaatilisse ühendisse *üle piiri minema* (tähenduses 'liialdama, mõõdu- või sündsusetunnet minetama') kuuluvat nimisõna *piir* laiendada ainult atribuudiga *igasugune* (25). Sama ühendi sõnasõnalise kasutuse puhul on atribuudi valik aga piiramatu (nt 26).

(25) Lapsed **läksid** oma ülemeelikusega **üle igasuguse piiri**.
(26) Jalakäijad võivad **üle Eesti-Läti piiri minna** enda valitud kohas.

Graham Katz ja Eugenie Giesbrecht (2006) püüavad automaatselt eristada sama sõnaühendi idiomaatilisi ja mitte-idiomaatilisi tähendusi konteksti põhjal. Aluseks on eeldus, et mitte-idiomaatilises tähenduses kasutatud sõnaühendi naabruses peaksid sageli esinema samad sõnad, mis esinevad selle sõnaühendi komponentide naabruses siis, kui need komponendid esinevad tekstis iseseisvalt. Eeldus on tuletatud sõnatähenduste ühestamisel kasutatavast seaduspärast, et mitmetähendusliku sõna tähenduse konkreetses kontekstis saab järeldada temaga koos esinevatest sõnadest (vt nt Schütze 1998).

Kirjeldatud meetod lähtub püsiühendiga samas lauses või osalauses esinevatest sõnavormidest või lemmadest. Ilmselt oleks veelgi otstarbekam kasutada mingit infot lauses või osalauses esinevate sõnade grammatiliste kategooriate või süntaktiliste funktsioonide kohta. Loob ju lauses (27) esinev sõnasõnalise tähendusega ühend *maha saama* argumendistruktuuri, milles tegevusobjekti on väljendatud süntaktilise objektiga (*seda plekki*), ent lauses (28) esinev idiomaatiline *maha saama* argumendistruktuuri, milles tegevusobjekti on väljendatud hoopis komitatiivse adverbiaaliga (*elu pikima kõnega*).

(27) Ega seda plekki enam riidelt **maha ei saa**.
(28) Fidel **sai** möödunud suvel **maha** elu pikima kõnega.

4. Kokkuvõtteks

Püsiühendite automaattöötlusel tuleb lahendada esiteks küsimus, mis on püsiühend, ja teiseks tuleb need püsiühendid tekstis ära tunda. Arvuti- või korpuslingvistikas tähendab see nende märgendamist. Seejuures küsimus püsiühendite olemusest ei ole pelgalt teoreetiline, defineerimise probleem, vaid arvutilingvistikas tähendab see ka definitsiooni rakendamist: püsiühendite loendi koostamist.

Püsiühendite loendi koostamisel kasutatakse statistilisi meetodeid, mis võimaldavad tekstikorpusest leida sõnu, mis esinevad koos küllalt sageli, et võiks kahtlustada nende poolt püsiühendi moodustamist. Seejuures tuleb arvestada, et

sõltuvalt püsiühendi tüübist valitaks sobiv statistiline meetod ja et "sõnade koosesinemine" nõuab keeletegelikkusega arvestamist: kas "sõna" all mõista sõnavormi või lemmat, kas "koosesinemine" tähendab sõnadevahelise kauguse, süntaktilise seose ja/või sõnade järjekorra arvestamist või ignoreerimist lauses.

Püsiühendite äratundmine ehk märgendamine tekstis tähendab seda, et iga lause puhul kontrollitakse, kas seal esineb mingis loendis, nt püsiühendite andmebaasis olevaid väljendeid. Püsiühendite märgendaja peab toime tulema esiteks sellega, et püsiühendid ei esine tekstis alati täpselt samas vormis, mis loendis, ning teiseks sellega, et püsiühendi komponentide olemasolu lauses ei tähenda automaatselt, et nad seal ka püsiühendi moodustavad.

Eesti keele verbikesksete püsiühendite puhul verbi vormi valik ei ole kuidagi kitsendatud asjaoluga, et verb kuulub püsiühendi koosseisu. Verbikeskse püsiühendi käändsõnaline komponent võib varieeruda seda enam, mida läbipaistvamalt on ühendi kui terviku tähendus moodustatud tema komponentide tähenduste põhjal.

Püsiühendisse kuuluvate üksiksõnade sagedus tekstis on erinev. Sõnad, mis esinevad tekstis sageli, esinevad sageli ka samas lauses koos ilma püsiühendit moodustamata. Valdavalt puudutab see muutumatuid sõnu (nt *üle*), mis võivad olla kasutusel nii kaas- kui määrsõnana.

Et eristada sõnaühendi idiomaatilist kasutust mitte-idiomaatilisest, võib arvesse võtta sõnaühendi vormi jäikust – mida rohkem väljendi vorm tekstikorpuses varieerub, seda tõenäolisem on, et antud korpuses kasutavad autorid seda väljendit mitte-idiomaatilisena.

Viidatud kirjandus

Benson, Morton; Benson, Evellyn; Ilson, Robert (Eds.) 1986. BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam: John Benjamins.

Cook, Paul; Fazly, Asfaneh; Stevenson, Suzanne 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. – Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. June 28, 2007. Prague, 41–48.

Evert, Stefan 2004. The Statistics of Word Cooccurences: Word Pairs and Collocations. PhD dissertation. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf (19.08.2008).

Evert, Stefan 2008. Corpora and collocations. – Anke Lüdeling, Merja Kytö (Eds.). Corpus Linguistics. An International Handbook, Vol. 1. Handbücher zur Sprach- und Kommunikationswissenschaft 29.1. Berlin: Mouton de Gruyter, 1212–1248. [Extended manuscript http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf (19.08.2008).]

Fleischer, Wolfgang 1982. Phraseologie der deutschen Gegenwartssprache. Lepzig: WEB Bibliographisches Institut.

Grégoire, Nicole 2007. Design and implementation of a lexicon of Dutch multiword expressions. – Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. June 28, 2007. Prague, 17–24.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2003. Püsiühendite leidmine suurtest tekstikorpustest. – Margit Langemets, Heete Sahkai, Maria-Maren Sepper (toim.). Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Tallinn: Eesti Keele Sihtasutus, 101–118.

Katz, Graham; Giesbrecht, Eugenie 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. – Multiword Expressions: Identifying and Exploiting Underlying Properties. Proceedings of the Workshop. ACL/COLING-06. July 23, 2006. Sydney, 12–19.

Moon, Rosamund 1998. Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford: Clarendon Press.

Muischnek, Kadri 2006. Eesti keele verbikesksed püsiühendid tekstikorpuses. – Emakeele Seltsi aastaraamat, 51 (2005), 80–106.

Pecina, Pavel 2005. An extensive empirical study of collocation extraction methods. – 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). Proceedings of the Student Research Workshop, June 2005. Ann Arbor, Michigan, 13–18. http://ufal.mff.cuni.cz/~pecina/publications/ (21.08.2008).

Riehemann, Susanne 2001. A Constructional Approach to Idioms and Word Formation. PhD dissertation. Stanford University. http://doors.stanford.edu/~sr/sr-diss.pdf (18.09.2008).

Rätsep, Huno 1978. Eesti keele lihtlause tüübid. Tallinn: Valgus.

Sag, Ivan A.; Baldwin, Timothy; Francis, Bond; Copstake, Ann; Flickinger, Dan 2002. Multiword expressions: A pain in the neck for NLP. – Alexander Gelbukh (Ed.). Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002). Mexico City, Mexico, 1–15. http://lingo.stanford.edu/pubs/WP-2001-03.pdf (21.08.2008).

Schütze, Hinrich 1998. Automatic word sense discrimination. – Computational Linguistics, 24 (1), 97–124.

Heiki-Jaan Kaalep (Tartu Ülikool). Peamised uurimisvaldkonnad on korpuslingvistika, arvutimorfoloogia, elektroonilised sõnastikud, püsiühendid arvutilingvistikas.
heiki-jaan.kaalep@ut.ee

Kadri Muischnek (Tartu Ülikool). Peamised uurimisvaldkonnad on korpuslingvistika ja eesti keele korpuste koostamine; püsiühendid lingvistikas ja arvutilingvistikas; eesti keele süntaktiline struktuur ja selle formaliseerimine.
kadri.muischnek@ut.ee

ESTONIAN MULTIWORD EXPRESSIONS IN COMPUTATIONAL LINGUISTICS

Heiki-Jaan Kaalep, Kadri Muischnek
University of Tartu

Multiword expressions are known to pose problems for natural languge analysis. By *multiword expressions* we mean combinations of two or more word(form)s that are habitually used together to express a certain meaning; the term covers both idiomatic and collocational word combinations. This article concentrates on three main tasks in multiword expression processing: extraction, lexicon compilation and annotation. The standard methods for solving these tasks are analysed from the viewpoint of automatic analysis of Estonian, a language with a rich and complicated morphological structure and a free word (or constituent) order.

Keywords: computational linguistics, multiword expressions, multiword expression extraction, lexicon of multi-word expressions, multi-word expression annotation, Estonian