TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKA TEADUSKOND
Arvutiteaduse instituut
Infotehnoloogia õppekava

Andreas Orlo

Jalgpalli Frameneti tõlkimine eesti keelde ja jalgpalli tekstikorpuse koostamine

Bakalaureusetöö (6 EAP)

Juhendaja: Neeme Kahusk

Tartu 2014

Jalgpalli Frameneti tõlkimine eesti keelde ja jalgpalli tekstikorpuse koostamine

Lühikokkuvõte:

Kicktionary on mitmekeelne jalgpalli-alane leksikaal-semantiline ressurss, mis põhineb FrameNeti freimisemantikal. Kicktionary pakub freimi semantilist informatsiooni, mis võimaldab kasutajal mõista jalgpalli kirjeldamiseks kasutatud sõnade tähendusi. Töö tulemuseks on freimileksikon, mis sisaldab eesti keelde tõlgitud Kicktonary freime. Töö teiseks tulemuseks on jalgpalli tekstikorpus, mis on ka morfoloogiliselt ja süntaktiliselt ühestatud. Käesolev töö on eelduseks, et ühestatud jalgpallitekste saaks koostatud freimidega automaatselt märgendada.

Võtmesõnad:

Freim, korpus, ühestamine

Translating football Framenet into Estonian and making of football corpus

Abstract:

Kicktionary is a multlilingual lexical-semantic resource that is based on FrameNet's framesemantics. Kicktionary provides frame semantic information that allows users to understand the meanings of words as they are used to describe soccer. The outcome of this work is a frame lexicon that consists of Kicktionary frames, which are translated into Estonian. Another product of this work is a corpus of football texts, which is also morphologically and syntactically disambiguated. This work is a prerequisite for automatic annotation of the disambiguated text with the translated frames.

Keywords:

Frame, corpus, disambiguation

Sisukord

1 Sissejuhatus

Jalgpallis on palju termineid, mis on kasutusel ka paljudes teistes valdkondades. Et tekstis vahet teha, kas meil on tegemist jalgpalliterminitega või et leida teksis olevat informatsiooni, näiteks vastuseid küsimustele, oleks vaja jalgpallialast leksikaalset ressurssi. Üheks selliseks leksikaalseks ressursiks on Kicktionary[1], mis on freimisemantikal põhinev jalgpallisõnastik. Kicktionary on saadaval kolmes keeles: inglise, saksa ja prantsuse keeles. Seega on selle bakalaureusetöö üks eesmärke kohandada Kicktionary eesti keelde. Samuti puuduvad meil tekstikorpused, mis sisaldavad ainult jalgpalli-alaseid tekste. Bakalaureusetöö teiseks eesmärgiks on koguda kokku selliseid tekste erinevatest allikatest ning siduda nad terviklikeks korpusteks. Käesolev töö on eelduseks, et jalgpallitekste saaks automaatselt märgendada.

Töö koosneb neljast peatükist. Esimeses peatükis antakse töös kasutatud mõistete definitsioonid. Teises peatükis tutvustatakse FrameNet[2] projekti. Kolmandas peatükis antakse ülevaade Kicktionaryst, kirjeldatakse Kicktionary tõlkimist eesti keelde ning freimide koostamist. Neljandas peatükis selgitatakse jalgpalli tekstikorpuse koostamist.

Lisana on esitatud freimileksikon, kuhu kuuluvad kõik eesti keelde tõlgitud Kicktionary freimid. Teiseks on lisatud töö käigus koostatud tekstikorpus tavatekstina, mis koosneb ühestamata jalgpallitekstidest. Kolmanda lisana on esitatud morfoloogiliselt ja süntaktiliselt ühestatud korpus, kuhu kuuluvad morfoloogiliselt ja süntaktiliselt ühestatud jalgpallitekstid. Neljanda lisana on toodud nimekiri kõikidest freimi elementidest ja nende eestikeelsetest vastetest.

Selles töös on kasutatud kahte viitamisviisi: tsitaadid ja allmärkused. Allmärkustega on viidatud veebilehtedele ja töös kasutatud vahendite veebilehtedele ning muudel juhtudel on kasutatud tsitaatviiteid.

[1] http://kicktionary.de/index.html (viimati vaadatud 09.05.2014)
[2] https://framenet.icsi.berkeley.edu/fndrupal/home (viimati vaadatud 09.05.2014)

2 Mõisted

Selles peatükis on seletatud töös esinevate mõistete tähendused.

Freim (*frame*) – andmekeskne teadmuse esitus, mis seostab objekti mingi tunnusomaduste kogumiga. Freimi aluseks on hüptees (oletus): inimene säilitab oma mälus üldistatud teadmisi - suurt hulka nn. stereotüüpe, mida ta tunneb oma isiklikust kogemusest või mõnest muust allikast.[1] Freimid ehk raamid on kontekstimudelid, milles vastavalt olukorrale vahetuvad osalised ja muutujad, kuid püsib situatsiooni taust. Freim on loend tingimustest, mis peavad olema täidetud, et seda situatsiooni võiks nimetada teatud sõnaga.[2]

Freimi elemendid - Igal freimil on tuum- ja lisaelemendid, millest võib mõelda kui semantilistest rollidest [3]. Freimi elemendid, mis on vajalikud freimi tähenduse jaoks, on tuumelemendid, lisaelemendid on üldiselt väljendid aja, koha ja viisi kohta [4].

Leksikaalne üksused - on sõnad, mis on seotud kindla tähendusega. Kui sõnal on mitu tähendust, siis tüüpiliselt on mitu leksikaalset üksust seotud erinevate freimidega.[3] Leksikaalsed üksused on sõnad, mis kutsuvad freimi esile. Näiteks leksikaalsed üksused, mis kutsuvad esile freimi *Hit*, on verbid lööma ja tabama. Lisaks freimile on iga leksikaalne üksus seotud kindlate freimi elementidega [3].

Keelekorpus - on kirjaliku või suulise kõne kogum. Keeleteaduses on sõna *korpus* alla enne arvutite kasutuselevõttu tavaliselt mõeldud keeleainese kogumikku, mida kasutatakse uurimistöös materjalina vastandina autori enda intuitsioonil põhinevatele üldistustele. Arvutiajastul on korpustena hakatud mõistma peamiselt polüfunktsionaalseid elektroonilisel kujul olevaid tekstikogusid, millesse kuuluvad tekstid on validut eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest. Lühidalt: korpus on loomuliku keele tekstide kogu, mis on koostatud iseloomustamaks keele hetkeseisu või muutumist.[5]

3 FrameNet

FrameNet on freimisemantikal baseeruv projekt, mille käivitas Charles Fillmore California Ülikoolist Berkeleys 1997. a [5]. Charles Fillmore sündis 1929. aastal. Ta on töötanud 10 aastat Ohio Ülikoolis ja ühe aasta Stanfordi ülikoolis. California Berkeley ülikooli lingvistika osakonda asus ta tööle aastal 1971, kus ta on praegu emeriitprofessor. Charles Fillmore'i peamised uurimisvaldkonnad on süntaks ja leksikaalne semantika, rajanud on ta käändegrammatika ja freimisemantika ning on üks konstruktsioonigrammatika olulisi esindajaid. Siiani on ta tegev FrameNeti projektis.[2]

FrameNet on projekt, mis koostab freimisemantika teoorial põhinevat leksikaalset ressurssi. FrameNet ehitab leksikaalset ingliskeelset andmebaasi, mis on nii inim- kui ka masinloetav [6]. See andmebaas sisaldab umbes 1200 semantilist freimi, 13000 leksikaalset üksust ja üle 190000 näitelause [3]. FrameNet töötab korraga nii sõnaraamatu kui ka tesaurusena. Sõnaraamatu erijooned on nt definitsioonid, süntaktiline informatsioon freimielementide kohta ja näitelaused. Sarnaselt tesaurusele on sõnad seotud freimidega, milles nad osalevad, freimid on omakorda seotud sõnadega ja teiste freimidega.[2] FrameNeti projekt on töös rahvusvahelises arvutiteaduse instituudis (ICSI[3]) Berkeley linnas, Californias.[3]

FreimNeti kasutavad loomuliku keele töötlejad (nt sõnatähenduste ühestamisel, masintõlkes), leksikograafid, keeleõpetajad ja –õppijad. See projekt olnud väga mõjukas – sellele on pühendatud terve 16s International Journal of Lexicography. Ning FrameNeti tüüpi projektidega ja sellega seotud projektidega tegeletakse aktiivselt ka teistes keeltes (kaasa arvatud eesti keeles).[2]

[3] http://www.icsi.berkeley.edu/icsi/ (viimati vaadatud 09.05.2014)

4 Kicktionary

Kictionary on mitmekeelne elektrooniline leksikaal-semantiline ressurss. See sisaldab umbes 2000 jalgpalliterminit inglise, saksa ja prantsuse keeles, mis on struktureeritud stseenide ja freimide hierarhiasse.[7] Kicktionary koostati aastatel 2005 ja 2006, kui Kicktionary autor Thomas Schmidt külastas FrameNeti projekti rahvusvahelises arvutiteaduse instituudis [8].

Kicktionary rajamine tugines järgmistele teoreetilistele punktidele [8]:

- Charles Fillmore'i koostatud freimisemantika teooriale.
- FrameNeti projekti metoodikale, et ehitada suurt freimisemantikal põhinevat semantilist andmebaasi.
- Seelbachi ja Gaston Grossi tööle jalgpallikeele leksikograafiast.

Kicktionary peamine eesmärk oli (ja on) uurida, kuidas keelelised teooriad leksikaal-semantikast, keelekorpuste meetoditest, hüperteksti ja hüpermeedia tehnoloogiatest ning arvuti keelekasutuse tehnikatest on aidanud koostada leksikaalseid ressursse, mis on paremad kui traditsioonilised paberkujul sõnastikud.[8]

Kui tavaline sõnastik annab meile definitsioonid, häädlused ja sõnaliigid, siis semantiliselt märgendatud sõnastik nagu Kicktionary annab meile vajaliku konteksti, et näidata sõna tähendust nii, nagu see kehtib just jalgpalli puhul. Kicktionary pakub freimi semantilist informatsiooni, mis võimaldab kasutajal mõista sõnade nüansse ja tähendusi, mida kasutatakse jalgpalli kirjeldamiseks.[9]

Kicktionary sisaldab ligikaudu 1900 leksikaalset üksust, nendeks on nimisõnad, verbid, omadussõnad ja idiomaatilised väljendid. Iga leksikaalse üksuse jaoks on üks kuni kümme märgendatud näitelauset, mis on võetud korpusest, kuhu kuuluvad jalgpallimängu raportid. Märgendid identifitseerivad leksikaalset üksust ning samuti ka selle argumente, milleks võivad olla abitegusõnad või eessõnad.[8]

Leksikaalsed üksused on nende semantika ja argumentide struktuuri analüüsi põhjal grupeeritud ligikaudu sajasse freimi selliselt, et leksikaalsed üksused samas freimis jagavad olulisi semantilisi ja süntaktilisi tunnuseid. Freimid omakorda on jaotatud 16sse stseeni, kus iga stseen tähistab jalgpallimängu prototüüpilist sündmust nagu näiteks värav või üks ühele olukord.[8]

4.1 Kicktionary tõlkimine

Kicktionary tõlkimisel on võetud kõik Kicktionarys olevad freimid ja tõlgitud nad inglise keelest eesti keelde. Tõlkimisel on kasutatud Inglise-Eesti sõnaraamatuid, mis asuvad veebilehtedel aare.edu.ee ja translate.google.ee. Tõlgitud on leksikaalsed üksused ja freimi elmendid. Näiteks üks freim nimega Match kajastub Kicktionary lehel[4] järgnevalt:

Joonis 1. Näide freimist Kicktionarys.

Veebilehel on kõik freimid kirjeldatud HTML kujul. Seal on kirjas kõik vajalik info, et saaks koostada eestikeelseid freime. Lehelt saab välja lugeda freimi nime, leksikaalsed üksused ja freimi elemendid. Antud näite puhul on freimi nimi *Match*, selle all on kirjas leksikaalsed üksused (*lexical units*), mis on nii saksa, inglise kui ka prantsuse keeles. Leksikaalsetest ükustest allpool näeme freimi elemente (*frame elements*).

[4] http://kicktionary.de/FRAMES/Frame_Match.html (viimati vaadatud 14.05.2014)

4.2 Freimide koostamine

Freimide koostamisel kasutatakse andmete märgistuskeelt XML. Kõik freimid on koondatud kokku ühte freimileksikoni. Freimileksikon algab märgendiga <framelexicon> ning lõppeb märgendiga </framelexicon>, kõik freimid jäävad nende märgendite vahele. Ühe XML kujul freimi struktuur näeb välja järgmine:

```
<frame name="home game">
    <LexicalUnits>
        <lu lemma="kodumäng" />
        <lu lemma="kodus" />
        <lu lemma="võõrustaja" />
    </LexicalUnits>
    <Elements>
        <element name="võõrustaja" optional="true" />
        <element name="külaline" optional="true" />
        <element name="väljak" optional="true" />
        <element name="asukoht" optional="true" />
        <element name="aeg" optional="true" />
        <element name="mäng" optional="true" />
    </Elements>
</frame>
```

Joonis 2. Näide XML kujul freimist.

Freim algab märgendiga <frame> ning on määratud atribuudiga *name*, mille väärtus on freimi nimi inglise keeles. Märgendite <LexicalUnits> ja </LexicalUnits> vahele jäävad leksikaalsed üksused (nimisõnad, verbid, omadussõnad). Märgendite <Elements> ja </Elements> vahele jäävad freimi elemendid. Elemendid on määratud atribuutidega *name*, mis on elemendi nimi, ja *optional*, mis näitab kas freimi element on kohustuslik (*false*) või mitte (*true*). Freim lõppeb märgendiga </frame>. Igal freimil on vastavalt leksikaalsetele

üksustele kindlad freimi elemendid. Freimi elemendid võivad üle mitme freimi korduda, kuid mõni element võib olla ka spetsifiiline ainult ühele kindlale freimile. Nimekiri kõikidest Kicktionary freimi elementidest ja nendele vastavatest eestikeelsetest tõlgetest on toodud ära lisas 4.

Kokku on selliselt koostatud freime freimileksikonis 103 tükki. Kicktionarys on neid küll 104, aga freimi *Mark* veebileht Kicktionarys ei avanenud. Leksikaalseid üksuseid on kõikide freimide peale kokku 380. Freimid asuvad failis frames.xml (vt lisa 1). Järgnevalt on toodud nimekiri kõikidest tõlgitud freimidest:

finish, ball and goal, feign, follow up, goal kickoff, hit, intervene, miss goal, save, shoot at, shot, shot supports, bad pass, being free, connect, control, flick on, intercept, pass, pass back, pass combination, supply pass, award goal, celebrate goal, concede goal, goal, convert chance, multiple goals, overcome goalkeeper, own goal, prepare goal, score goal, beat, challenge, deny, lose ball, one on one, take on, trick, advantage, concede compensation, dissent, foul, give card, offside, receive card, referee decision, sanction, set piece", simulation, win compensation, chance, create chance, miss chance, ball bounce, ball escape, ball land, ball move, goalkeeper advance, player move, player move with ball, bring off, bring on, substitute, move, confusion, defence shot, mistake, lead, match quality, possession, score, spectator activity, tactics, trail, away game, defeat, draw, elimination, home game, match, match temporal subdivision, progression, result, start end match, victory, competition, competition stage, deploy, start, suspension, ball, body parts, coach, equipment, player, referee, spectators, team, bench, field, goal target, stadium.

5 Jalgpalli tekstikorpus

Et koostatud freimide abil saaks tekste märgenda, siis on vaja eestikeelset jalgpalli tekstikorpust. Kuna hetkel selline korpus puudub, siis töö üheks tulemuseks ongi üks selline korpus, mis sisaldab ainult eestikeelseid jalgpallitekste. Korpuse koostamine koosnes kahest osast: tekstide kogumisest ja nende ühestamisest.

5.1 Tekstide kogumine

Jalgpallitekstide kogumisel on kasutatud viite uudisteportaali: Õhtuleht[5], Soccernet[6], ERR[7], Postimees[8] ja Delfi[9]. Lisaks on kasutatud ka olemasolevast morfoloogiliselt analüüsitud ja ühestatud korpusest pärit tekste[10]. Uudised on valitud ajalises järjestuses (alustades koostamise kuupäeval värskeimatest ja liikudes vanemate poole). Tekstid on valitud freimide seisukohalt, st valitud on uudised, mis kirjeldavad rohkem jalgpallimatši ning sisaldavad freimides kasutuselolevaid leksikaalseid üksuseid.

Tekstid on jaotatud tekstifailidesse ja failid on koondatud ühte korpusesse (vt lisa 2). Failid korpuses on koostatud iga portaali tekstidest eraldi, see tähendab, et ei leidu koos ühes failis tekste, mis on võetud nii Õhtulehest kui ka Soccernetist. Uudiste eraldajateks tekstifailides on üks tühi rida. Failide nimedes on kasutatud vastavate uudisteportaalide nimesid, ehk teisisõnu kõik Soccernetist võetud jalgpallitekstid on failis nimega soccernet.txt, ERR-st võetud tekstid on failis err.txt, Õhtulehes võetud tekstid on failis nimega ohtuleht.txt, Postimehest postimees.txt ja Delfist delfi.txt.

Korpus tavatekstina koosneb nendest viiest tekstifailist: ohtuleht.txt, soccernet.txt, err.txt, postimees.txt ja delfi.txt. Korpuses on viie faili peale kokku 102 uudisteksti, korpus koosneb 1095 lausest, 15699 sõnast ja 94022 tähemärgist. Lausete, sõnade ja tähemärkide loendamisel on kasutatud veebis olevat tööriista TextMechanic[11].

[5] http://www.ohtuleht.ee/
[6] http://www.soccernet.ee/
[7] http://www.err.ee/
[8] http://www.postimees.ee/
[9] http://www.delfi.ee/
[10] http://www.cl.ut.ee/korpused/morfkorpus/myh01/
[11] http://textmechanic.com/Count-Text.html

Õhtulehe puhul on tekstid võetud spordirubriigi kategooriast jalgpall[12]. Õhtulehe tekstifail ohtuleht.txt sisaldab jalgpallitekste 21-st uudisest, failis on 187 lauset, 3283 sõna ja 19554 tähemärki.

Soccernetis on jalgpalliuudised jaotatud kahte kategooriasse „Eesti ja eestlased võõrsil“ ning „Rahvusvaheline“. Uudiseid on valitud mõlemast kategooriast. Tekstifail soccernet.txt koosneb 20-st jalgpalliuudisest, selles on 178 lauset, 2717 sõna ja 16235 tähemärki.

Eesti Rahvusringhäälingust on uudistekste võetud ERR-i spordiportaalist[13] ning kategooriast jalgpall. Tekstifail err.txt koosneb 20 uudistekstist ning sisaldab 226 lauset, 2678 sõna ja 16311 tähemärki.

Postimehest on tekstid võetud rubriigist sport ja kategooriast jalgpall[14]. Tekstifail postimees.txt sisaldab 21 uudist, selles on 266 lauset, 3898 sõna ja 23391 tähemärki.

Delfi puhul on uudiseid võetud samuti spordirubriigi kategooriast jalgpall[15]. Failis delfi.txt on tekstid 20 jalgpalliuudise kohta, kokku koosneb fail 238 lausest, 3123 sõnast ja 18531 tähemärgist.

5.2 Tekstide ühestamine

Korpusest on kasu ainult siis, kui saame sealt suhteliselt lihtsalt kätte meile vajaliku info. Aga selleks, et seda vajalikku infot kätte saada, peab sageli alustama info lisamisest korpusesse. Seega: kui soovitakse, et korpus ei jääks ainult elektrooniliste tekstide arhiiviks, tuleb tekstidele lisada info nende ülesehituse kohta (peatükid, pealkirjad, lõigud, laused jne.), samuti andmed morfoloogilise ja süntaktilise analüüsi tulemuste kohta jne.[5]

Et tekste oleks võimalik hiljem ka freimidega märgendada, on nad vaja kõigepealt ühestada. Morfoloogilise ühestamise käigus lisatakse igale sõnale kõik tema morfoloogilised teisendused ja seejärel valitakse välja antud konteksti sobiv sõnavorm. Süntaktilise analüüsi käigus lisatakse sõnavormile kõikvõimalikud süntaktilised märgendid ja pärast seda eemaldatatakse süntaktilise ühestamise käigus sõnavormilt konteksti põhjal kõik lubamatud süntaktilised märgendid.

[12] http://www.ohtuleht.ee/sport/jalgpall
[13] http://sport.err.ee/
[14] http://sport.postimees.ee/rubriik/154
[15] http://sport.delfi.ee/news/jalgpall/

Tekstide ühestamisel kasutasin eesti keele süntaksianalüsaatorit[16], mis on kirjutatud Tartu Ülikoolis Tiina Puolakaineni ja Kaili Müürisepa poolt. Programm võimaldab sisestatud tekste morgoloogiliselt ja süntaktiliselt analüüsida. Süntaksianalüsaator võtab sisendiks tekstifaili ja väljastab morfoloogiliselt ja süntaktiliselt ühestatud faili, mille laiendiks on .snx. Kõik morfoloogiliselt ja süntaktiliselt ühestatud failid on salvestatud ühte korpusesse (vt lisa 3) ning salvestatud samasuguse nimega, mis oli tema sisendiks olnud tekstifaililgi. Lisaks on korpuses fail nimega morf.kym, mis on saadud juba varem morfoloogiliselt analüüsitud ja ühestatud korpusest. Sellesse faili on kopeeritud kõik jalgpalli puudutavad märgendatud laused.

Eesti keele süntaksianalüsaatoris jagatakse teksti töötlemine järgmisteks etappideks [10]:

- Eeltöötlus, mille käigus tuntakse ära lausete lõpud, tehakse kirjavahemärkide analüüs ja teisendatakse tekst morfoloogiaanalüsaatori jaoks sobivale kujule.
- Morfoloogiline analüüs - leitakse sõnavormi tüvi ning lõpud ja neile vastav sõnaliik, kääne või pööre. Kui sõnavorm on mitmeti tõlgendatav, antakse selle kõik tõlgendused.
- Morfoloogiline ühestamine - konteksti info põhjal leitakse sõnavormi paljude tõlgenduste seast korrektne tõlgendus.
- Osalause piiride määramine - konteksti info, kirjavahemärkide ja morfoloogilise info põhjal leitakse liitlausetes osalausete piirid. See etapp toimub paralleelselt morfoloogilise ühestamisega, sest mõlemad on teineteisest väga sõltuvad.
- Süntaktiliste märgendite lisamine - morfoloogilise info ja konteksti põhjal lisatakse sõnavormile kõik võimalikud süntaktilised märgendid. Osaliselt võidakse märgendeid lisada ka juba morfoloogilise analüüsi käigus leksikonist või enne morfoloogilist ühestamist.
- Süntaktiline ühestamine - konteksti põhjal eemaldatakse sõnavormilt kõik lubamatud süntaktilised märgendid.

Näiteks lause *Eelmisel nädalal lõi Taani esiliigas tähtsa värava Kaimar Saag* on pärast ühestamist järgmine:

[16] http://www.cs.ut.ee/~kaili/parser/

```
Eelmisel    eelmine+l //_A_ pos sg ad  #cap // **CLB  @AN>
nädalal    nädal+l //_S_ com sg ad  //  @ADVL
lõi    loo+i //_V_ main indic impf ps3 sg ps af  #FinV #NGP-P //  @+FMV
löö+i //_V_ main indic impf ps3 sg ps af  #FinV #NGP-P //  @+FMV
Taani    Taani+0 //_S_ prop sg gen #cap //  @OBJ @NN>
esiliigas    esi_liiga+s //_S_ com sg in  //  @ADVL
tähtsa    tähtis+0 //_A_ pos sg gen  //  @AN>
värava    värav+0 //_S_ com sg gen  //  @OBJ @NN>
Kaimar    Kaimar+0 //_S_ prop sg nom  #cap #? //  @SUBJ @NN>
Saag    saag+0 //_S_ com sg nom  #cap //  @SUBJ
```

Joonis 3. Näide morfoloogiliselt ja süntaktiliselt ühestatud lausest.

Igal real on sõnavorm ja selle grammatiline kirjeldus: kaldkriipsude vahel on morfoloogilised märgendid ja rea lõpus süntaktilised [11].

Väljundi üldkuju on järgmine:

```
sõnavorm    tüvi+lõpp // morfoloogiline info // süntaktilised märgendid
```

- <sõnavorm> on sõna sellisena, nagu ta algselt esines.
- <tüvi> on lemma e. algvormi tüvi: käändsõnadel ainsuse nimetav (kui seda ei ole olemas, siis mitmuse nimetav), pöördsõnadel ma-infinitiivi tüvi ilma (ma-lõputa).
- <lõpp> on sõna lõpp, kusjuures mitmuse tunnus on temaga liitunud (nagu seda on käsitletud ka Ülle Viksi "Väikeses vormisõnastikus"); partikkel GI/KI, kui ta esineb, on lihtsalt lõppu "kleepunud"; ka juhul, kui sõnal ei saagi lõppu olla (nt. hüüdsõnal), pannakse sõnale lõpp - nn. null-lõpp.
- <morfoloogiline info> on üks variantidest, mis on kõik esitatud morfoloogiliste kategooriate tabelis[17].[12]
- <süntaktilised märgendid> on eesti keele kitsenduste grammatika süntaktilised märgendid

Kui on tegemist liitsõna või tuletisega, siis:

[17] http://www.cl.ut.ee/korpused/morfliides/seletus

- Tüvi on eristatud eelnevast komponendist '_' märgiga;
- Lõpp on eristatud eelnevast komponendist '+' märgiga; nn. null-lõpp ongi '+0'
- Sufiks on eristatud eelnevast komponendist '=' märgiga. Sufiksite märkimine ei ole järjekindel: märgitakse ainult teatud hulka produktiivseid sufikseid.
- Lemmatüvi leitakse ainult viimase parempoolse komponendi alusel

Mitmesõnalised nimed on sellisel kujul:

New Yorgis New York+s //_S_ prop sg in // [12]

Järgnevalt on toodud ära ülevaade morfoloogilistest märgenditest [13]:

- Sõnaliik
    - Käändsõnad: nimisõnad ehk substantiivid (_S_), omadussõnad ehk adjektiivid (_A_), arvsõnad ehk numeraalid (_N_), asesõnad ehk pronoomenid (_P_) ja lühendid ning akronüümid (_Y_).
    - Pöördsõnad: tegusõnad ehk verbid (_V_).
    - Muutumatud sõnad: määrsõnad ehk adverbid (_D_), kaassõnad ehk adpositsioonid (_K_), sidesõnad (_J_), hüüdsõnad (_I_), ainult verbidega koos esinevad sõnad (_X_).
- Kääne: nimetav ehk nominatiiv (nom), omastav ehk genitiiv (gen), osastav ehk partitiiv (part), lühike sisseütlev ehk aditiiv (adit), sisseütlev ehk illatiiv (ill), seesütlev ehk inessiiv (in), seestütlev ehk elatiiv (el), alaleütlev ehk allatiiv (all), alalütlev ehk adessiiv (ad), alaltütlev ehk ablatiiv (abl), saav ehk translatiiv (tr), rajav ehk terminatiiv (ter), olev ehk essiiv (es), ilmaütlev ehk abessiiv (ab), kaasaütlev ehk komitatiiv (kom).
- Arv: ainsus ehk singular (sg), mitmus ehk pluural (pl).
- Komparatsioon: algvõrre (pos), keskvõrre (comp), ülivõrre (super).
- Isik: esimene (ps1), teine (ps2), kolmas (ps3), neid kombineeritakse arvuga.
- Aeg: olevik (pres), lihtminevik (impf). Täis- ja enneminevik moodustatakse *olema* verbi ja mineviku partitsiibi abil (partic past).
- Tegumood: isikuline (ps), umbisikuline (imps).
- Kõneviis: kindel (indic), tingiv (cond), käskiv (imper), kaudne (quot).

- Kõneliik: jaatav (af) ja eitav (neg).

Lisaks neile eristatakse põhi- ja järgarvsõnu (card ja ord), ees- ja tagasõnu (prep ja post), põhi-, modaal ja abiverbe (main, mod, aux) ning eraldi märgendatakse verbi käändelisi vorme: da-infinitiiv (inf) ja supiin (sup), mis omakorda jaguneb: a) ma-supiin (ill), b) mas-supiin (in), c) mast-supiin (el), d) maks-supiin (tr), e) mata-supiin (abes); des-vorm ehk gerundiiv (ger) ja kesksõnad ehk partitsiibid, mis jagunevad: a) v-kesksõna (partic pres ps), tav-kesksõna (partic pres imps), nud-kesksõna (partic past ps), tud-kesksõna (partic past imps). Asesõnadest eristatakse personaal-, demonstratiiv-, indefiniit-, possessiiv-, interrogatiiv-, relatiiv-, refleksiiv-, retsiprook- ja determinatiivpronoomeneid (vastavalt pers, dem, indef, pos, inter, rel, rec, det). Nende eestikeelsed vasted on isikuline, näitav, umbmäärane, omastav, küsiv, siduv, vastastikune ja määratlev asesõna.[13]

Süntaktilised märgendid, mida korpuses on kasutatud, on järgmised [14]:

Öeldise märgendid

@+FMV - finiitne öeldis

@-FMV - infiniitne öeldis

@+FCV - *olema* liitaegades ning modaalverbid ahelverbides, finiitne vorm

@-FCV - *olema* liitaegades ning modaalverbid ahelverbides, infiniitne vorm

@NEG - verbi eitus

Põhja märgendid

@SUBJ - alus ehk subjekt

@OBJ - sihitis ehk objekt

@PRD - öeldistäide ehk predikatiiv

@ADVL - määrus ehk adverbiaal, ka fraasiadverbiaal

Laiendite märgendid

@AN> - omadus- ja järgarvsõna eestäiendina

@<AN - omadus- ja järgarvsõna järeltäiendina

@AD> - määrsõna eestäiendina

@<PN - kaassõna järeltäiendina

@NN> - nimi-, ase- ja põhiarvsõna eestäiendina

@<NN - nimi-, ase- ja põhiarvsõna järeltäiendina

@VN> - partitsiip eestäiendina

@<VN - partitsiip järeltäiendina

@<INF_N - verbi infinitiitne vorm järeltäiendina

@<P - eessõna laiend

@P> - tagasõna laiend

@<Q - kvantori järellaiend

Muud

@J - sidend

@I - hüüatus

6 Kokkuvõte

Jalgpall on spordiala, millel on oma kindel terminoloogia. Seal on kasutuses palju sõnu, mida me teistes valdkondades mõistame mingi muu tähenduse all. Näiteks kui me mõtleme sõnast tulistama, siis see võib tähendada nii püssist laskmist, kui ka jalgpallis pealelööki väravale. Kicktionary annab meile vajaliku konteksti, et näidata sõna tähendust nii, nagu see kehtib just jalgpalli puhul.

Kictionary koostamisel on võetud aluseks FrameNet, et koostada sarnaselt suurt freimisemantikal põhinevat andmebaasi. Kicktionary pakub freimi semantilist informatsiooni, mis võimaldab kasutajal mõista sõnade tähendusi, mida kasutatakse jalgpalli kirjeldamiseks. Kicktionary on saadaval inglise, prantsuse ja saksa keeles, kuid käesoleva töö raames tõlgiti see ka eesti keelde. Selleks koostati freimileksikon, mis sisaldab eestikeelseid Kicktionary freime.

Et ka koostatud freimidest kasu oleks, on vaja eestikeelseid tekste, mida saaks nende freimidega märgendada. Kuna freimid on mõeldud just jalgpalli-alastele tekstidele, siis tekkis vajadus jalgpalli tekstikorpuse järele. Varasemalt saadaolevad korpused ei sisaldanud kas üldse või väga vähesele määral jalgpallitemaatilisi tekste. Praktilise töö teise tulemusena tekkis üks selline eestikeelne jalgpalli-alane tekstikorpus. Selleks sai kogutud erinevaid jalgpallitekste mitmetest uudisteportaalidest. Pärast seda ühestati tekstid morfoloogiliselt ja süntaktiliselt, et neid oleks võimalik hiljem ka freimidega märgendada. Selleks kasutati Tartu Ülikoolis väljatöötatud süntaksianalüsaatorit. Morfoloogilise ühestamise käigus lisati igale sõnale kõik tema morfoloogilised teisendused ja seejärel valiti välja antud konteksti sobiv sõnavorm. Süntaktilise ühestamise käigus lisati sõnavormile kõikvõimalikud süntaktilised märgendid ja pärast seda eemaldatati sõnavormilt konteksti põhjal kõik lubamatud süntaktilised märgendid.

Töö tulemusteks on freimileksikon ja tekstikorpused. Freimileksikon koosneb 103-st eesti keelde tõlgitud Kicktionary freimist. Tekstikorpusi on kaks: korpus tavatekstina ning ühestatud korpus. Korpus tavatekstina sisaldab ühestamata kujul jalgpallitekste erinevatest uudisteportaalidest. Ühestatud korpus sisaldab neid samu tekste morfoloogiliselt ja süntaktiliselt ühestatud kujul. Korpused koosnevad 21-st uudistekstist, mis on kogutud erinevatest allikatest.

Valminud tulemuste abil on võimalik jalgpalli-alaseid tekste analüüsida. Töö tulemusi on juba kasutatud bakalaureusetöö „Eestikeelse jalgpalli-alase tekstikorpuse automaatne märgendamine jalgpalli-alase Frameneti abil“ koostamisel.

Materjale on võimalik veel edasi arendada. Freime on võimalik lisada eraldi stseenidesse, et valmiks stseenide ja freimide hierarhia. Leksikaalseid üksuseid on võimalik paikutada sünonüümide hulkadesse st grupeerida sõnad, millel on sarnane või samasugune tähendus. Sünonüümide hulkadest saaks ehitada mõistete hierarhiad. Samuti on võimalik koostada lisaks olemasolevatele freimedele juurde uusi freime ning lisada freimidele veel leksikaalseid üksuseid. Korpusesse on võimalik juurde lisada uusi uute jalgpalli-alaseid tekste.

7 Tsiteeritud teosed

[1] Mare Koit, Tiit Roosmaa (2011). "Tehisintellekt". (12.05.2014)

http://dspace.utlib.ee/dspace/bitstream/handle/10062/28296/tehisintellekt.pdf?sequence=2

[2] Ilona Tragel, Piret Piiroja. E-kursuse "Tuntud keeleteadlasi ja koolkondi" materjalid.

http://www.e-ope.ee/_download/euni_repository/file/1388/oppematerjalide_avaleht.docx (12.05.2014)

[3] Wikipedia

http://en.wikipedia.org/wiki/FrameNet (09.05.2014)

[4] FrameNet - *Glossary*

https://framenet.icsi.berkeley.edu/fndrupal/glossary

[5] Kadri Muischnek, Heili Orav, Heiki-Jaan Kaalep, Haldur Õim. "Eesti keele tehnoloogilised ressursid ja vahendid".

http://www.hm.ee/index.php?popup=download&id=3993 (12.05.2014)

[6] FrameNet

https://framenet.icsi.berkeley.edu/fndrupal/about (09.05.2014)

[7] Kicktionary

http://kicktionary.de/index.html (09.05.2014)

[8] Kicktionary - *Background*

http://kicktionary.de/background.html (09.05.2014)

[9] *International Computer Science Institute*

https://www.icsi.berkeley.edu/icsi/news/2006/07/thomas-schmidt-kicktionary (09.05.2014)

[10] Kaili Müürisep. Analüüsi etapid. [Online].

http://www.cs.ut.ee/~kaili/parser/demo/overview.html (11.05.2014)

[11] Kaili Müürisep. Eesti keele süntaksianalüsaatori märgenditest.

http://www.cs.ut.ee/~kaili/papers/myyriseprakling03final.pdf (11.05.2014)

[12] Tartu Ülikooli arvutilingvistika uurimisrühm. Morfoloogiliselt ühestatud korpus. http://www.cl.ut.ee/korpused/morfkorpus/ (11.05.2014)

[13] Ülevaade morfoloogilistest märgenditest

http://math.ut.ee/~kaili/thesis/pt3_2.html (12.05.2014)

[14] Eesti keele kintsenduste grammatika süntaktilised märgendid

http://math.ut.ee/~kaili/thesis/pt3_4.html (13.05.2014)

Lisad

I. Freimid

II. Korpus tavatekstina

III. Morfoloogiliselt ja süntaktiliselt ühestatud korpus

IV. Freimi elementide eestikeelsed tõlked

GOAL_PART = värava osa

SHOOTER = lööja

SHOT = löök

SOURCE = allikas

MOVING_BALL = liikuv pall

PART_OF_BODY = kehaosa

MOVE = käik

TARGET = sihtmärk

PATH = teekond

SECOND_SHOOTER = teine lööja

SECOND_SHOT = teine löök

GOALKEEPER = väravavaht

DISTANCE = kaugus

BALL = pall

SHOOTER_TEAM = lööja meeskond

INTERVENING_PLAYER = sekkuv mängija

INTERVENTION_RESULT = sekkumise tulemus

INTERVENTION_TARGET = sekkumise sihtmärk

INTERVENTION_LOCATION = sekkumise asukoht

INTEVENTION = sekkumine

OPPONENT_TEAM = vastasmeeskond

RECIPIENT = vastuvõtja

PASSER = söötja

AREA = piirkond

MARKER = katja

PASS = sööt

SECOND_RECIPIENT = teine vastuvõtja

DIRECTION = suund

FLICK_ON_TAGET = edasilükkamise sihtmärk

INTERCEPTOR = vahelesekkuja

PASSERS = söötjad

GOAL = värav

REFEREE = kohtunik

IRREGULARITY = reeglipäratus

SCORER = väravalööja

TEAM_MATE = meeskonnakaaslane

SCORER_TEAM = lööja meeskond

CONCEDING_TEAM = sisselaskev meeskond

RESULTING_SCORE = saadud tulemus

SET_PIECE = standardolukord

PREPARING_EVENT = ettevalmistav sündmus

PREVIOUS_SCORE = eelnev tulemus

ASSISTER = väravasöödu andja

OPPONENT_PLAYER = vastasmängija

PLAYER_WITH_BALL = palliga mängija

AREA = piirkond

ACTION = tegevus

CHALLENGE = võitlus

PLAYERS = mängijad

OPPONENT_PLAYERS = vastasmängijad

TARGET = sihtmärk

ATTACK = rünnak

TEAM_WITH_BALL = palliga meeskond

PLAYER1 = mängija1

PLAYER2 = mängija2

REFEREE = kohtunik

OFFENSE = süüdistus

OFFENDED_TEAM = kannatanud meeskond

COMPENSATION = kompensatsioon

OFFENDER = põhjustaja

OFFENDED_PLAYER = kannatanud mängija

COACH = treener

DECISION = otsus

CARD = kaart

OFFENDER_TEAM = põhjustaja meeskond

TEAM = meeskond

EXECUTING_PLAYER = lahtimängiv mängija

PENALTY = trahvilöök

PLAYER = mängija

OPPORTUNITY = võimalus

PREVIOUS_EVENT = eelnev sündmus

SUBSTITUTED_PLAYER = vahetatud mängija

SUBSTITUTE = asendaja

LINEUP = rivistus

LEADING_PLAYER = juhtiv mängija

ATTACKING_TEAM = ründav meeskond

PART_OF_TEAM = meeskonna osa

OCCASION = sündmus

LEADER = liider

SCORE = skoor

MARGIN = vahe

DOMINATING_TEAM = domineerivam meeskond

MATCH_PERIOOD = mänguperiood

TEAMS = meeskonnad

SPECTATORS = pealtvaatajad

SPECTATOR_ACTIVITY = pealtvaataja tegevus

TRAILER = mahajääja

VISITOR = külaline

MATCH_LOCATION = mängu asukoht

HOST = võõrustaja

TIME = AEG

COMPETITION_STAGE = võistlusfaas

LOSER = kaotaja

FINAL_SCORE = lõppskoor

WINNER = võitja

MATCH = mäng

COMPETITION = võistlus

DECISIVE_EVENT = otsustav sündmus

REASON = põhjus

FIRST_OR_SECOND = esimene või teine

NEXT_ROUND = järgmine ring

SPECIFIER = täpsustaja

NOT_PLAYER = mitte mängija

NOT_PLAYING_REASON = mittemängimise põhjus

SANCTION = sanktsioon

AUTHORITY = autoriteet

ORIENTATION = orientatsioon

POSITION_SPECIFICATION = positsiooni täpsustus

GOAL_TARGET = värava sihtmärk

DISTANCE_TO_BALL = kaugus pallini

UP_DOWN_ORIENTATION = üles alla orientatsioon

V. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina **Andreas Orlo** (sünnikuupäev: 04.02.1992)
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Jalgpalli Frameneti tõlkimine eesti keelde ja jalgpalli tekstikorpuse koostamine,
(*lõputöö pealkiri*)

mille juhendaja on Neeme Kahusk,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **14.05.2014**