SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

Odsjek za informacijske znanosti

Petar Kukulj

METODOLOGIJA IZGRADNJE PARALELNIH KORPUSA I EKSTRAKCIJE SPECIFIČNE TERMINOLOGIJE

Diplomski rad

Mentor: prof. dr. sc. Sanja Seljan

Zagreb, prosinac 2018.

Sadržaj

1. Uvod

Dostupnost različitih tekstova u elektroničkom obliku nikada nije bila veća nego danas. Zahvaljujući toj činjenici i razvoju tehnologije općenito, izgradnja i upotreba raznih tekstualnih podataka predstavlja važan resurs u analizi podataka. No dostup do velikih količina dvojezičnih paralelnih podataka predstavlja problem, osobito za manje raširene jezike. Takvi resursi vrlo su vrijedan izvor informacija primjenjivih za daljnja istraživanja u raznim područjima, a fokus ovog rada bit će na paralelnim korpusima. Prikazan je i razvoj područja koji je krenuo od prvotno jezične analize, preko korpusne analize, a koji se danas usmjerava prema analizi i rudarenju velikih količina podataka u cilju dohvaćanja informacija.

Na početku će se definirati pojam korpusa, prikazati faze razvoja te će se iznijeti njihova osnovna tipologija i potencijalna primjena. Prikupljanje tekstova i priprema korpusa za jezične analize važan su dio postupka njihove izgradnje te će se i te metode prezentirati u teorijskom dijelu rada, kao i pregled dosadašnjih istraživanja nad korpusima.

Praktični dio rada prikazat će konkretan proces prikupljanja tekstova za izgradnju paralelnog korpusa, njihovu obradu i stvaranje samoga korpusa koristeći online korpusni alat *Sketch Engine* koji će također biti ukratko opisan. Prikupljene tekstove čine razni sportski pravilnici na engleskom jeziku te njihovi prijevodi na hrvatski. Koristeći opcije koje *Sketch Engine* nudi, taj paralelni korpus bit će analiziran te će se izvući specifična dvojezična terminologija sadržanih sportova. Naposljetku, rezultati će biti prikazani i uspoređeni te interpretirani. Nakon istraživanja slijedi rasprava, zaključak i popis literature.

2. Razvoj korpusa

Korpus možemo definirati kao zbirku jezičnih tekstova u elektroničkom obliku, selektiranih prema eksternim kriterijima s ciljem da čine reprezentativan uzorak nekog jezika ili jezične varijante, a služi kao izvor podataka za različita istraživanja[1]. Ti kriteriji mogu biti eksterni ili interni[2]. Eksterni kriteriji odnose se na sudionike, prigodu te komunikativne funkcije jezičnog uzorka, a interni na obrasce unutar tog uzorka.

Danas se podrazumijeva da se korpusi izrađuju u strojno čitljivom obliku, ali to nije uvijek bio slučaj s obzirom na to da su se metodologije obrade jezika koristile i prije same pojave računala. Razvoj ove metodologije možemo podijeliti na tri faze: rana faza koja je trajala do 1950. godine, tridesetogodišnja stagnacija te moderna faza od 1980-ih godina naovamo[3].

U ranijoj fazi (30-ih godina 20. st.) istraživanja su uglavnom provodili korpusni lingvisti, dok se danas ovo područje usmjerava na pronalaženje informacija i obradu podataka. Iako 1930-ih godina nije postojao pojam korpusne lingvistike, terenski lingvisti i strukturalisti svoja istraživanja provodili su na prikupljenim i organiziranim uzorcima jezika u upotrebi te tu metodologiju zasigurno možemo opisati kao temeljenu na korpusima. Takvi „protokorpusi“ koristili su se u proučavanju usvajanja jezika, ortografskim istraživanjima, studijama podučavanja stranog jezika, komparativnoj lingvistici te sintaktičkim i semantičkim analizama.

Za pad popularnosti korpusne lingvistike krajem 1950-ih godina zaslužan je Noam Chomsky koji je u principu oživio stare nesuglasice između empirista i racionalista te uvodi princip

[1] Sinclair, J. Corpus and Text - Basic Principles. // Developing Linguistic Corpora: a Guide to Good Practice / uredio Martin Wynne. Oxford: Oxbow Books, 2005. Str. 1-16.
[2] EAGLES. Preliminary recommendations on Corpus Typology. 1996. URL: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html. (7. 12. 2017.)
[3] McEnery, A.; Wilson, A. Corpus Linguistics – An Introduction. 2nd ed. Edinburgh: Edinburgh University Press, 2005.

formalne obrade prirodnih jezika stvarajući time preduvjete za računalnu primjenu u obradi prirodnih jezika. Chomsky je zamjerao ranim korpusnim lingvistima pretpostavku da je jezik konačan skup te da se sve rečenice nekog jezika mogu prebrojati, a on je vjerovao da postoji samo konačan broj sintaktičkih pravila pomoću kojih možemo stvoriti beskonačan broj rečenica. Jezik gotovo zasigurno nije konačan skup rečenica, ali ni introspekcija nije objektivan pristup znanstvenom istraživanju. Istina je najvjerojatnije negdje između. Ne možemo zanemariti vrijednost kvantitativnih podataka dobivenih iz korpusa, ali ni ti podaci nemaju vrijednost ako ne znamo pravila po kojima se te jezične jedinice kombiniraju. Dakle, ove dvije teorije ne mogu pobijati jedna drugu, već se jedino mogu komplementirati. Unatoč velikom padu popularnosti istraživanja temeljenih na korpusima, ona nikada nisu u potpunosti napuštena te su se i tijekom ovog razdoblja koristila u fonetici i proučavanju usvajanja jezika. Dolazi i do raznih tehnoloških i metodoloških inovacija, a konkretno treba naglasiti radove Quirka, Svartvika i Leecha. U ovom razdoblju Francis i Kucera započinju i rad na Brown korpusu.

Zahvaljujući razvoju računala, analiza korpusa postala je jeftiniji, brži i precizniji proces. 1980-ih godina dolazi do naglog razvoja i rasta popularnosti korpusne lingvistike i računalne obrade jezika, a taj rast traje i danas. U tom razdoblju korpus postaje strojno čitljiv jezični resurs sa standardiziranom i dosljednom arhitekturom i metodologijom pogodnom za daljnje obrade. Današnja istraživanja obrade jezika usmjerena su na istraživanja velikih količina podataka primjenom dubinske analize teksta u području računalnih znanosti.

2.1 Vrste i primjena korpusa

Svi korpusi dizajnirani su s nekom svrhom na umu, a s obzirom na njihovu svrhu, možemo ih klasificirati na sljedeći način[4]:

- Specijalizirani korpusi: Ovi korpusi sadrže po jednu specifičnu vrstu tekstova, a čine reprezentativan uzorak tih specifičnih tekstova, na primjer geografskih udžbenika, predavanja, akademskih članaka iz određene znanstvene discipline, ležernih razgovora, studentskih eseja itd. Koriste se za istraživanje jednog specifičnog aspekta jezika, a osim vrste tekstova, mogu biti određeni nekim vremenskim razdobljem ili određenom temom itd.
- Generalni ili referentni korpusi: Sadrže tekstove različitih vrsta, bilo u pisanom ili izgovorenom obliku, te čine reprezentativan uzorak čitavog jezika. Obično su mnogo veći od specijaliziranih korpusa, a koriste se kao referentni materijali za učenje jezika ili prevođenje te kao uzorak općenitog jezika u odnosu na koji se istražuju razni specijalizirani diskursi.
- Usporedivi korpusi: Mogu biti višejezični ili mogu sadržavati različite varijetete istog jezika, a obično su sačinjeni od različitih vrsta tekstova koji su zastupljeni u sličnim omjerima. Usporedivi korpusi različitih varijeteta nekog jezika koriste se za istraživanja razlika između tih varijeteta, a dvojezični pomažu u pronalaženju razlika i sličnosti između tih jezika.
- Paralelni korpusi: Dvojezični ili višejezični korpusi koji sadrže izvorne oblike tekstova i njihove prijevode, a koriste se za razvoj sustava za strojno i strojno potpomognuto prevođenje[5], kontrastivna i terminološka istraživanja[6], glotodidaktiku te u dvojezičnoj

[4] Hunston, S. Corpora in Applied Linguistics. 3rd ed. Cambridge: Cambridge University Press, 2005.
[5] Brkić, M.; Seljan, S.; Bašić Mikulić, B. Using Translation Memory to Speed up Translation Process. // INFuture 2009 : Digital resources and knowledge sharing / uredili Stančić, H., Seljan, S., Bawden, D., Lasić-Lazić, J. & Slavić, A. Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. Str. 353-363.

i višejezičnoj leksikografiji[7]. Obično su sravnjeni na razini riječi ili rečenice kako bi se olakšalo pronalaženje prijevodnih ekvivalenata. Ovdje možemo i spomenuti nekoliko hrvatskih višejezičnih paralelnih korpusa kao što su *Srpsko – hrvatsko – engleski kontrastivni projekt* te hrvatsko-engleski prijevod Platonove *Republike* koji je bio samo jedan od više prijevodnih parova izdanih na TELRI CD-ROM-u[8]. *Hrvatsko-engleski paralelni korpus* jednosmjerni je paralelni korpus s hrvatskim kao izvornim jezikom, a tekstovi su pribavljeni iz tjednika *Croatia Weekly*[9]. *SETimes* sastavljen je od novinskih članaka objavljenih na portalu *Southeast European Times* koji su izvorno pisani engleskim jezikom te prevedeni na devet jezika, od kojih je jedan i hrvatski[10]. *Slovensko-hrvatski paralelni korpus* rezultat je akademske i istraživačke suradnje ovih dviju zemalja, a sastavljen je od različitih žanrova pribavljenih iz raznih izvora[11]. *HrEnWac* je hrvatsko-engleski paralelni web korpus prikupljen s hrvatskih domena[12], a valja spomenuti i resurse koji između ostalog sadrže *Pravnu stečevinu Europske unije* na 24 jezika[13] s više od 2 milijarde riječi[14].

- Učenički korpusi: Ovakvi korpusi zbirke su tekstova čiji su autori učenici nekog jezika, na primjer eseja. Cilj im je identificirati razlike između načina usvajanja jezika

---

[6] Seljan, S.; Gašpar, A. First Steps in Term and Collocation Extraction from English-Croatian Corpus. // Proceedings of 8th International Conference on Terminology and Artificial Intelligence. Toulouse, France: 2009. URL: http://ceur-ws.org/Vol-578/paper21.pdf (15. 5. 2018.)

[7] Simeon, I. Paralelni korpusi i višejezični rječnici. // Filologija. 38-39, 2002, str. 209-215.

[8] Tadić, M. Building the Croatian-English Parallel Corpus. // Second International Conference on Language Resources and Evaluation LREC2000 / uredili Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis S. Pariz, Atena: ELRA, 2000. Str. 523-530.

[9] ibid.

[10] Bekavac, B.; Seljan, S.; Simeon, I. Corpus-Based Comparison of Contemporary Croatian, Serbian and Bosnian. // Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages / uredili Marko Tadić, Mila Dimitrova-Vulchanova, Svetla Koeva. Zagreb: Croatian Language Technologies Society, 2008. Str. 33-39.

[11] Požgaj Hadži, V.; Tadić, M. Slovensko-hrvatski paralelni korpus. // Izazovi kontrastivne lingvistike (Izzivi kontrastivnega jezikoslovja) / Vesna Požgaj Hadži et al. Ljubljana: Znanstvena založba Filozofske Fakultete Univerze v Ljubljani, 2012. Str. 45-54.

[12] Ljubešić, N.; Esplà-Gomis, M.; Ortiz Rojas, S. et al. Croatian-English parallel corpus hrenWaC 2.0. // Slovenian language resource repository CLARIN.SI (2016). URL: http://hdl.handle.net/11356/1058. (6. 3. 2018.)

[13] Schlüter, P. Statistics on the DGT-Translation Memory (DGT-TM). 2018. URL: https://wt-public.emm4u.eu/Resources/DGT-TM_Statistics.pdf (13. 6. 2018.)

[14] ibid.

pojedinaca te razlike između materinjeg jezika i jezika koji uče kako bi se taj proces olakšao.

- Pedagoški korpusi: Sastoji se od uzorka jezika kojemu su učenici izloženi, na primjer udžbenika ili audio zapisa korištenih u nastavi kako bi se bolje upoznali s naučenim riječima ili frazama u raznim kontekstima.
- Povijesni ili dijakronijski korpusi: Korpusi tekstova skupljenih iz raznih vremenskih razdoblja koriste se za istraživanje promjena određenog aspekta jezika kroz neki vremenski period.
- Monitor korpusi: Dizajnirani su tako da prate suvremene jezične promjene što zahtijeva godišnju, mjesečnu ili čak dnevnu nadogradnju, no omjer vrsta tekstova mora uvijek biti isti radi konzistentne usporedbe tih različitih verzija korpusa.

Razni korpusi važan su resurs i pri rudarenju teksta, procesu u kojem se identifikacijom i istraživanjem interesantnih poveznica i činjenica unutar i između različitih tekstualnih dokumenata pokušava doći do korisnih informacija[15]. Metode rudarenja teksta primjenjuju se iznimno uspješno u raznim područjima kao što su analiza patenata, klasifikacija tekstova, bioinformatika pa i u filtriranju nepoželjne e-pošte[16]. Sam proces uglavnom počiva na načelima rudarenja podataka, s jednom bitnom razlikom, a to je činjenica da tekst sadrži nestrukturirane podatke koje je potrebno obraditi tako da budu prikladni za računalnu analizu[17]. Stoga je predobrada tekstovnih podataka iznimno važan korak te počiva na dostignućima drugih računalnih disciplina koje se bave obradom prirodnog jezika kao što su pronalaženje informacija, ekstrakcija informacija te korpusna lingvistika[18]. Iako ne toliko iscrpna, priprema tekstova za rudarenje uvelike se zasniva na principima predobrade tekstova

[15] Feldman, R.; Sanger, J. The Text Mining Handbook. 1st ed. Cambridge; New York: Cambridge University Press, 2007.
[16] Hotho, A.; Nürnberger, A.; Paaß, G. A brief survey of text mining. // Ldv Forum. 20, 1 (2005), str. 19-62.
[17] Witten, I. H. Text Mining. // Practical handbook of internet computing / uredio M. P. Singh. Boca Raton, FL: Chapman & Hall / CRC Press, 2006. Str. 314-342.
[18] Feldman, R.; Sanger, J. The Text Mining Handbook. 1st ed. Cambridge; New York: Cambridge University Press, 2007.

za korpusne analize[19]. No neka istraživanja upućuju na vrijednost detaljnije obrade teksta radi ekstrakcije višerječnih termina koji se mogu koristiti pri izradi taksonomija[20].

2.2 Priprema korpusa za analizu

Prije analize, podatke sadržane u korpusu potrebno je obraditi. Taj proces sadrži dva dijela: pripremu metapodataka, to jest zaglavlja te pripremu samog teksta[21].

Korpuse obilježavamo *mark-up* jezicima, a danas se najčešće koristi XML koji je nasljednik SGML-a te je najnoviji u nizu standarda u obilježavanju jezičnih resursa[22]. Neovisan je o operativnom sustavu i primjenjiv na sva pisma i jezike, a u kombinaciji s jezikom za oblikovanje (XSL) i jezikom za preoblikovanje (XSLT) omogućuje odabir, preobliku i fleksibilan prikaz podataka[23]. Nadalje, omogućuje obilježavanje kakvo sa SGML-om nije bilo moguće te je primjenjiv na različite formate zapisa dokumenata[24].

S obzirom na to da su korpusi zbirke različitih dokumenata, te dokumente potrebno je označiti metapodacima kako bismo mogli identificirati izvore pretraživanih riječi te opisati njihove karakteristike kao što su datum objavljivanja, autor, medij (pisani ili izgovoreni tekstovi), tematika teksta itd. Ti metapodaci sadržani su u zaglavljima dokumenata te, uz informacije o njima, pružaju korisnicima i mogućnost ograničavanja pretraživanja na određenu vrstu teksta te sastavljanje frekvencijskih popisa i pronalaženje ključnih riječi određenih tekstova itd[25].

---

[19] Hotho, A.; Nürnberger, A.; Paaß, G. A brief survey of text mining. // Ldv Forum. 20, 1 (2005), str. 19-62.
[20] Feldman et al. Text mining at the term level. // Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98) / uredili Jan M. Żytkow, Mohamed Quafafou. Nantes: Springer, 1998. Str. 65-73.
[21] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.
[22] Tadić, M. Uporaba XML-a u hrvatskim korpusima. // Upravljanje informacijama u gospodarstvu i znanosti (CroInfo 2000): zbornik. Zagreb: Nacionalna i sveučilišna knjižnica; Pliva, 2000. Str. 132-137.
[23] ibid.
[24] ibid.
[25] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.

```
<cesHeader>
     <fileDesc></fileDesc>
     <encodingDesc></encodingDesc>
     <profileDesc></profileDesc>
     <revisionDesc></revisionDesc>
</cesHeader>
```

Slika 1: Primjer *CES* zaglavlja[26]

Na slici 1 vidi se zaglavlje kodirano prema *CES* (*Corpus Encoding Standard*) standardu koji određuje minimalnu razinu deskriptivnog označavanja korpusa te njihovu generalnu arhitekturu. Element <*fileDesc*> sadrži potpun bibliografski opis korpusa ili teksta koji se u njemu nalazi, <*encodingDesc*> prikazuje poveznicu između teksta i njegovog izvora, <*profileDesc*> pruža dodatne informacije o tekstu kao što su jezik, datum objave itd., a element <*revisionDesc*> prikazuje tijek revizije teksta[27].

Prvi korak pripreme teksta određivanje je i usklađivanje kodiranja znakova nakon kojeg slijedi obilježavanje samog teksta koje se sastoji od označavanja odlomaka, paragrafa i rečenica, pojavnica i lema te označavanja vrsta riječi (*POS tagging*) i gramatičkih struktura[28]. Kod predobrade paralelnih korpusa potrebno je, uz ove korake, i sravniti tekst i to najčešće na razini rečenice.

Standardizirano kodiranje znakova nužno je za interoperabilnost i dosljednu manipulaciju dokumentima općenito, kao i u kontekstu jezičnih korpusa, a najčešći takav standard danas je UTF-8. Dokumente s nestandardno kodiranim znakovima moguće je konvertirati u standardni format, a nakon toga slijedi označavanje samog teksta.

[26] Corpus Encoding Standard (1996), Document CES 1, Version 1.4, listopad 1996, URL: http://www.cs.vassar.edu/CES/ (6. 8. 2018.)
[27] ibid.
[28] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.

Sljedeći korak je pronalaženje jedinica teksta, to jest riječi i rečenica. Identifikaciju riječi i punktuacija nazivamo i tokenizacija, a određivanje granica rečenica nazivamo segmentacija[29]. Segmentacija omogućuje prikaz tih rečenica, a tokenizacija utječe na pretraživanje te sortiranje i filtriranje rezultata itd[30].

```
<BODY>
 <DIV0 type="MAIN">
 <HEAD type="NA">
  <S>
   <W type="R">Outsideri</W>
   <W type="R">i</W>
   <W type="R">u</W>
   <W type="R">Zagrebu</W>
  </S>
 </HEAD>
 <HEAD type="PN">
  <S>
   <W type="R">Istodobno</W>
   <W type="R">s</W>
   <W type="R">petim</W>
   <W type="R">»Sajmom</W>
   <W type="R">outsider</W>
   <W type="R">umjetnosti«</W>
   <W type="R">u</W>
   <W type="R">New</W>
   <W type="R">Yorku</W>
   <W type="R">u</W>
   <W type="R">Muzeju</W>
   <W type="R">suvremene</W>
   <W type="R">umjetnosti</W>
   <W type="R">u</W>
   <W type="R">Zagrebu</W>
   <W type="R">postavljena</W>
   <W type="R">je</W>
   <W type="R">izložba</W>
   <W type="R">djela</W>
   <W type="R">hrvatskih</W>
   <W type="R">Outsidera</W>
  </S>
 </HEAD>
```

Slika 2: Primjer tokenizacije teksta[31]

Nakon odrađene segmentacije i tokenizacije, paralelne korpuse potrebno je sravniti. Ovaj korak vrlo je koristan te omogućuje prikazivanje prijevodnih ekvivalenata. Može se izvršiti na

[29] Erjavec. Compilation and Exploitation of Parallel Corpora. // Journal of Computing and Information Technology. 11, 2 (2003), str. 93-102.
[30] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.
[31] Tadić, M. Uporaba XML-a u hrvatskim korpusima. // Upravljanje informacijama u gospodarstvu i znanosti (CroInfo 2000): zbornik. Zagreb: Nacionalna i sveučilišna knjižnica; Pliva, 2000. Str. 132-137.

razini rečenice ili, rjeđe na razini riječi, a postupak otežava činjenica da se jedna rečenica iznimno rijetko prevodi jednom rečenicom[32]. Tekst je moguće sravniti ručno, što iziskuje mnogo vremena i napora, ili strojno koristeći razne metode kao što su upotreba dvojezičnih leksikona i strukture dokumenata ili raznih metoda neovisnih o jeziku[33].

```
</tu>
<tu>
    <tuv xml:lang="en">
        <seg>The goal lines must be of the same width as the goalposts and the crossbar.</seg>
    </tuv>
    <tuv xml:lang="hr">
        <seg>Poprečne linije moraju biti jednake širine kao stupovi vrata i greda.</seg>
    </tuv>
</tu>
```

Slika 3: Primjer sravnjenog teksta na razini rečenice (.TMX)

Lematizacija ili morfološka analiza proces je kojim identificiramo osnovne oblike riječi koje nazivamo lemama, a korisna je za pronalaženje konkordancija i prijevodnih ekvivalenata.

```
<seg lang="sl">
<w lemma="do">Do</w>
<w lemma="let letati leto">leta</w>
<w type=dig>2008</w>
<w lemma="se">se</w>
<w lemma="pričakovati">pričakuje</w>
<c>,</c>
<w lemma="da dati">da</w>
<w lemma="biti">bo</w>
<w lemma="odpravljen odpraviti">odpravljen</w>
<w lemma="bistven">bistveni</w>
<w lemma="del delo deti">del</w>
<w lemma="tradicionalen">tradicionalnih</w>
<w lemma="problem">problemov</w>
<w lemma="onesnaženost">onesnaženosti</w>
```

Slika 4: Primjer lematiziranog teksta[34]

[32] Erjavec (2003), Compilation and Exploitation of Parallel Corpora. // Journal of Computing and Information Technology. 11, 2 (2003), str. 93-102.
[33] ibid.
[34] Tadić, M. Introduction to Corpus Linguistics. Predavanja na ljetnoj školi Jadertina Summer School in Empirical and Compututional Linguistics (JSSECL). Zadar. 2006. URL: http://hnk.ffzg.hr/txts/mt4JSSECL/JSS2006_Corp-lin.htm (8. 12. 2017.)

POS (part-of-speech) tagging automatski je proces prepoznavanja i označavanja vrsta riječi, a služi kao osnova za daljnju sintaktičku ili semantičku obradu[35]. U paralelnim korpusima takve anotacije upravljaju automatskom dvojezičnom ekstrakcijom leksikona ili strojnim prevođenjem temeljenom na primjerima[36].

For_IF the_AT members_NN2 of_IO this_DD1 university_NN1 this_DD1 charter_NN1 enshrines_VVZ a_AT1 victorious_JJ principle_NN1 ;_; and_CC the_AT fruits_NN2 of_IO that_DD1 victory_NN1 can_VM immediately_RR be_VBI seen_VVN in_II the_AT international_JJ community_NNJ of_IO scholars_NN2 that_CST has_VHZ graduated_VVN here_RL today_RT ._.

NN	singular common noun
NNS	plural common noun
NP	singular proper noun
NP$	genitive proper noun
PP$	possesive pronoun
RP	adverbial particle
VBD	past tense form of lexical verb
VBN	past participle of lexical verb...

Slika 5: Primjer teksta označenog *POS tagovima*[37]

Naposljetku, parsiranjem označavamo sintaktičke strukture svih rečenica u korpusu[38].

```
[S
    [NP
        [N Klaudija N]
    NP]
    [VP
        [V sjedi V]
        [PP
            [P na P]
            [N stolcu]
        PP]
    VP]
S]

[S[NP[N Klaudija NP][VP [V sat V][PP [P na P][N stolcu N]PP]VP]S]
```

Slika 6: Primjer prikaza sintaktičke strukture rečenice[39]

[35] Erjavec (2003), Compilation and Exploitation of Parallel Corpora. // Journal of Computing and Information Technology. 11, 2 (2003), str. 93-102.
[36] ibid.
[37] Tadić, M. Introduction to Corpus Linguistics. Predavanja na ljetnoj školi Jadertina Summer School in Empirical and Compututional Linguistics (JSSECL). Zadar. 2006. URL: http://hnk.ffzg.hr/txts/mt4JSSECL/JSS2006_Corp-lin.htm (8. 12. 2017.)
[38] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.

2.3. Korpusni alati

Korpusni alati omogućuju različite analize tekstova sadržanih u korpusima. Pomoću njih možemo tražiti određene riječi, prebrojavati ih i računati njihove relativne frekvencije te prikazivati njihove pojavnice u kontekstu radi daljnjih istraživanja[40]. Traženje riječi ili nizova riječi moguće je izvoditi na svakom korpusu, a ako je on tokeniziran i parsiran, moguće je pretraživati i pojavnice određene vrste riječi te razne vrste rečenica ili jezičnih struktura[41]. Frekvencijski popisi prikazuju broj svih riječi u nekom korpusu.

Kod anotiranih ili obilježenih korpusa, to jest onih koji su tokenizirani i lematizirani, te liste sadrže mnogo detaljnije i relevantnije informacije, a tek u usporedbi s drugim korpusima možemo iskoristiti sav njihov potencijal i tako usporediti više jezika, jezičnih varijeteta ili vrsta tekstova.

Popisi ključnih riječi pružaju nam uvid u relativne frekvencije riječi nekog manjeg, specijaliziranog korpusa u usporedbi s većim referentnim korpusom koji predstavlja sveobuhvatan uzorak nekog jezika. Takvim usporedbama tražimo razlike između specijaliziranog i općeg diskursa, to jest specijaliziranu terminologiju koja je rangirana po stupnju odstupanja od svakodnevne upotrebe jezika[42].

Popisi kolokacija sadrže nizove riječi koji imaju tendenciju zajedničkog pojavljivanja u određenom rasponu prije i iza središnje riječi. Mogu biti rangirani po čistim frekvencijama, ali u tom slučaju oni sadrže velik broj gramatičkih riječi. Takve su riječi odraz određenog jezika, vrlo su česte, no nisu semantički relevantne u odnosu na središnju riječ. Taj fenomen moguće je kompenzirati raznim statističkim metodama te generirati popise relevantnih

[39] Tadić, M. Introduction to Corpus Linguistics. Predavanja na ljetnoj školi Jadertina Summer School in Empirical and Compututional Linguistics (JSSECL). Zadar. 2006. URL: http://hnk.ffzg.hr/txts/mt4JSSECL/JSS2006_Corp-lin.htm (8. 12. 2017.)
[40] Hunston, S. Corpus Linguistics. // Encyclopedia of Language & Linguistics / uredio Keith Brown. Boston: Elsevier, 2006. Str. 234-248.
[41] ibid.
[42] ibid.

kolokacija koje nadalje možemo rangirati po riječima koje su specifične ili po riječima koje su atipične u kontekstu određene kolokacije[43]. Naposljetku, konkordancije nam prikazuju sve pojavnice tražene riječi u kontekstu. Moguće ih je sortirati abecedno po riječima prije ili poslije središnje riječi što olakšava uočavanje čestih fraza koje sadržavaju te riječi te njihovih značenja i diskurzivnih funkcija.

Razvoj korpusnih alata možemo podijeliti na četiri generacije[44]:

- Alati prve generacije pojavljuju se 1960-ih i 1970-ih, a radili su na središnjim računalima. Obrađivali su samo ASCII znakove što je ograničilo njihovu upotrebu samo na engleske korpuse, a obično su imali samo jednu funkciju kao što je prebrojavanje riječi ili pregledavanje konkordancija.
- Druga generacija korpusnih alata pojavljuje se 1980-ih i 1990-ih. Kao i prvoj generaciji, njihova funkcionalnost bila je ograničena, ali bilo ih je moguće koristiti na osobnim računalima što je omogućilo razna manja istraživanja, ali i njihovu upotrebu u poučavanju jezika.
- Većina korpusnih alata koji se danas koriste alati su treće generacije koji su se pojavili krajem 1990-ih, a i dalje se razvijaju i nadograđuju[45]. Njihove glavne prednosti u odnosu na starije generacije korpusnih alata su multifunkcionalnost, integrirane opcije upotrebe raznih statističkih metoda, bolja funkcionalnost s većim korpusima, obrada znakova izvan ASCII skupa što je omogućilo rad s više jezika te grafičko sučelje prilagođeno korisnicima koji nemaju puno informatičkog znanja[46]. Najveći nedostaci treće generacije korpusnih alata poteškoće su u obradi korpusa većih od 100 milijuna riječi te sve izraženiji problemi po pitanju autorskih prava skupljenih tekstova.

[43] ibid.
[44] McEnery, T.; Hardie, A. Corpus linguistics: Method, theory and practice. 1st ed. Cambridge; New York : Cambridge University Press, 2012.
[45] ibid.
[46] ibid.

- Četvrta generacija korpusnih alata odgovor je na ova dva problema. Rad s iznimno velikim korpusima omogućen je spremanjem podataka na web servere te predindeksiranjem tih podataka, a koristeći korisničko sučelje koje onemogućuje pregledavanje čitavog korpusa umanjeni su problemi s autorskim pravima[47]. Ipak, alati ove generacije imaju nekoliko nedostataka kao što je obavezna predobrada podataka te njihovo postavljanje na server što može biti dugotrajno i skupo, a povezanost tih online korpusa s jedinstvenim korpusnim alatima može rezultirati pojavom raznih i različitih web alata koji se mogu koristiti samo s danim korpusom[48].

Korpusne alate koje danas koristimo možemo klasificirati koristeći sljedeću tipologiju[49]:

a) Samostalni i online alati: Samostalni alati koriste se na lokalnom računalu na kojem se nalazi i korpus (npr. *WordSmith Tools*, *MonoConc Pro* i *AntConc*), a online alatima možemo analizirati korpuse spremljene na internetu s bilo kojeg računala (npr. *Sketch Engine* i *KorpusDK*).
b) Alati vezani uz korpus i neovisni alati: Neki alati mogu se koristiti samo s određenim korpusima, a najčešće su razvijeni kao dio nekog korpusnog projekta ili za određenu instituciju (npr. *SARA* i *XAIRA* ili alat za pristup španjolskom referentnom korpusu). Posebna skupina ovakvih alata koristi se za pristup i analizu nekoliko različitih korpusa (npr. *KorpusDK*). Ostali alati neovisni su o korpusu te ih možemo koristiti za izgradnju ili analizu bilo kojeg korpusa (npr. *Sketch Engine*, *Corpus WorkBench*, *WordSmith Tools*, *MonoConc Pro* i *AntConc*).
c) Pripremljeni korpusi i web korpusi: Većina alata razvijena je za tradicionalne korpuse, ali ne može se zanemariti internet koji je potencijalno ogroman izvor jezičnih podataka. *Google* i ostale web tražilice nisu jezični alati, ali nam mogu pružiti pristup i

[47] ibid.
[48] ibid.
[49] Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // Electronic Lexicography / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.

uvid u taj „korpus“. S obzirom na navedeno, nije iznenađujuće da se počinju razvijati alati koji su u osnovi web tražilice, ali pružaju korisniku mogućnost pregledavanja rezultata u obliku konkordancija. Nazivamo ih web konkordancerima, a jedan od vodećih primjera takvih alata je *Webcorp*.

d) Jednostavni alati i napredni alati: Ovisno o broju i vrsti funkcija, korpusni alati mogu biti jednostavni ili napredni. Jednostavni alati omogućuju traženje kolokacija, konkordancija i ključnih riječi (npr. *AntConc* i *MonoConc Easy*), a napredni sadrže funkcije koje mogu biti od koristi naprednim korisnicima kao što su lingvisti i leksikografi (npr. *Sketch Engine*, *XAIRA* i *KorpusDK*).

3. Istraživanje

U nastavku slijedi prikaz istraživanja. Za potrebe ovoga rada prikupljen je korpus iz područja sporta na engleskom i hrvatskom jeziku. Prikupljen je korpus od ukupno 17 500 prijevodnih parova, odnosno ukupno 304 765 tokena za engleski i hrvatski jezik. Nakon postupka preformatiranja i čišćenja slijedi postupak sravnjivanja i zatim ekstrakcije specifične terminologije. Na kraju su prikazani rezultati istraživanja.

3.1. Materijali

Za potrebe rada prikupljeno je devet vrsta sportskih pravilnika na izvornom, to jest engleskom jeziku te njihovi prijevodi na hrvatski:

- Badminton
- Dvoranski hokej
- Futsal
- Hokej na travi
- Košarka
- Nogomet
- Rukomet
- Stolni tenis
- Vaterpolo

Svi su pravilnici preuzeti u *Portable Document Formatu* (PDF), a prvi korak u njihovoj obradi bio je njihova konverzija u *.doc* format kako bi se olakšala daljnja priprema tekstova čiji je cilj bio stvaranje *Translation Memory Exchange* datoteka (TMX). U tu svrhu bilo je

potrebno pročistiti pravilnike od pogrešaka nastalih prilikom konverzije iz PDF formata u *.doc* format, ukloniti razne slike i dijagrame te sravniti tekst na razini rečenice. Tako obrađene tekstove bilo je potrebno spremiti u *plain text* formatu kodiranom u UTF – 8 kodu.

Najčešće pogreške proizašle iz konverzije bili su viškovi ili nedostaci razmaka te razni problemi s dijakritičkim znakovima koji su ispravljeni, a slike i dijagrami su uklonjeni. Sravnjivanje na razini rečenice odrađeno je ručno, a postignuto na principu jedna rečenica – jedan redak, što znači da su i engleska i hrvatska verzija tekstova morale imati isti broj rečenica te isti broj redaka, radi potreba sravnjivanja (eng. alignment). Takve dvojezične baze predstavljaju osnovu za daljnja istraživanja[50],[51],[52].

Sljedeći korak bio je postavljanje engleskih i hrvatskih *plain text* datoteka na repozitorij prijevodnih memorija (http://concordia.vm.wmi.amu.edu.pl/tmrepository/)[53] kako bi se iz njih dobile TMX datoteke[54]. Sam proces je jednostavan te je moguće segmentirano stvaranje prijevodne memorije iz više datoteka.

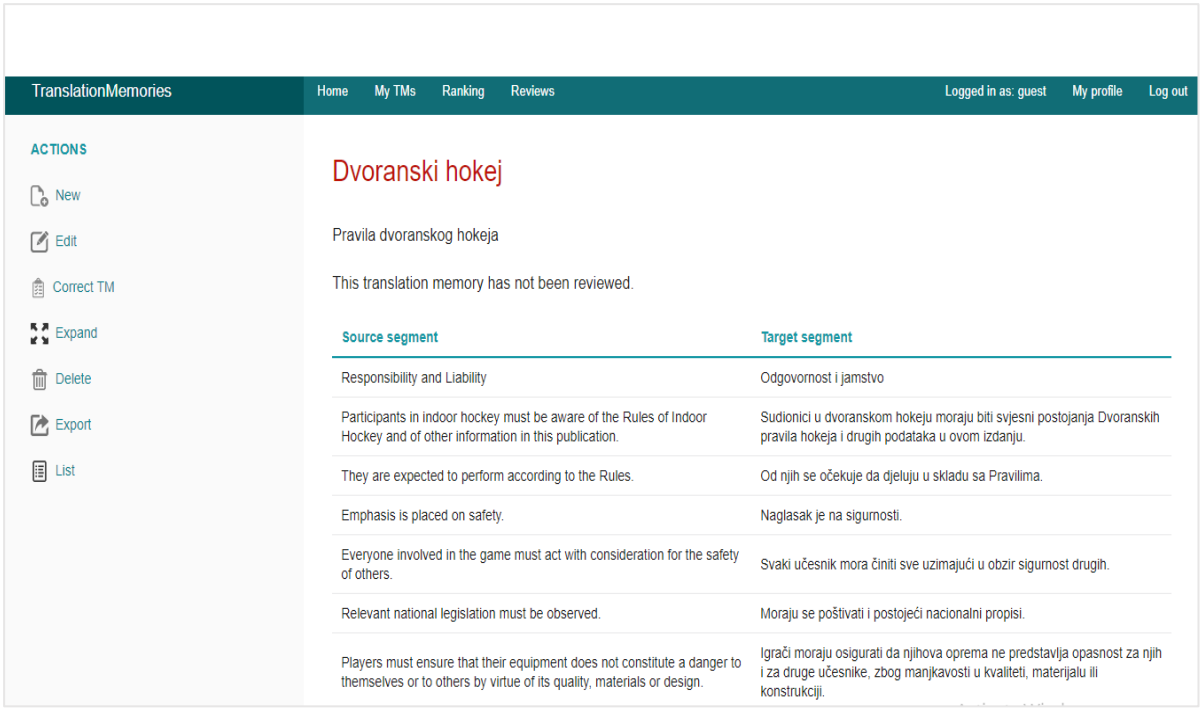
[50] Seljan, S.; Gašpar, A.; Pavuna, D. Sentence Alignment as the Basis For Translation Memory Database. // INFuture 2007 - Digital Information and Heritage / uredili Seljan, S., Stančić, H. Zagreb: Odsjek za Informacijske znanosti, Filozofski fakultet Zagreb, 2007. Str. 299-311.

[51] Brkić, M.; Matetić, M.; Seljan, S. Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus. // Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology ICCSIT 2011. Chengdu, China, 2011. Str. 1068-1070.

[52] Seljan, S.; Pavuna, D. Translation Memory Database in the Translation Process. // Proceedings of the 17th International Central European Conference on Information and Intelligent Systems IIS / uredili Aurer, B., Bača, M. Varaždin: FOI, 2006. Str. 327-332.

[53] Jaworski, R.; Dunđer, I.; Seljan, S. Usability Analysis of the Concordia Tool Applying Novel Concordance Searching. // Lecture Notes in Computer Science (LNCS), (in print).

[54] Jaworski, R.; Seljan, S.; Dunđer, I. Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. // Human Language Technologies as a Challenge for Computer Science and Linguistics / uredili Vetulani, Z. & Paroubek, P. Poznan: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2017. Str. 332-336.



Slika 7: Prikaz sravnjenog korpusa

Na slici 7 vidi se prikaz paralelnog korpusa na repozitoriju iz kojeg je moguće generiranje u .TMX formatu.

3.2. Translation Memory Exchange

Translation Memory Exchange format namijenjen je razmjeni prijevodnih memorija, a temeljen je na XML-u[55]. Razmjena se može vršiti na dvije razine. Prva razina sadrži samo tekst bez informacija o formatu dok druga razina sadrži i tekst i informacije o sadržajnom označavanju.[56] S obzirom na mogućnosti alata kojim su kreirane prijevodne memorije, to jest TM repozitorija, u ovom radu korištene su TMX datoteke prve razine.

[55] Seljan, S.; Tadić, M.; Agić, Ž.; Šnajder, J.; Dalbelo Bašić, B.; Osmann, V. Corpus Aligner (CorAl) Evaluation on English - Croatian Parallel Corpora. // Proceedings of Language Resources and Evaluation (LREC 2010) / uredili Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M. & Tapias, D. Valletta: European Language Resources Association, 2010. Str. 3481-3484.
[56] Raya, R. XML in localisation: Reuse translations with TM and TMX. XML in localisation. 2005. URL: https://www.ibm.com/developerworks/library/x-localis3/. (17. 1. 2018.)

Korijenski element ovakvih dokumenata je <*tmx*>, a glavni elementi koje sadrži su <*header*> i <*body*>. Element <*header*> sadrži metapodatke o samom dokumentu, a atributi koji nose te podatke mogu biti: *creationtool* i *creationtoolversion* koji nam govore kojim alatom i kojom verzijom alata je dokument kreiran, *datatype* definira vrstu podataka u dokumentu, *segtype* definira na koje je jedinice dokument segmentiran, *adminlang* naznačuje jezik administrativnih i informativnih elemenata, *srclang* označava jezik izvornog dokumenta, a *o-tmf* daje informacije o izvornom formatu prijevodne memorije.

Element <*body*> sačinjen je od skupa prijevodnih jedinica (<*tu*> elementi) koje se sadrže od varijanti prijevodnih jedinica (<*tuv*>) s atributom *xml:lang* koji označava njen jezik. Sam tekst sadržan je u <*seg*> elementima koji mogu imati i dodatne informacije o varijantama prijevodnih jedinica te elemente za označavanje sadržaja.

```
<tmx version="1.4">
    <header
        creationtool="TM repository" creationtoolversion="1.0"
        datatype="PlainText" segtype="sentence"
        adminlang="en-us" srclang="en"
        o-tmf="TM repository"/>
    <body>
        <tu>
            <tuv xml:lang="en">
                <seg>Futsal Laws of the Game 2010/2011</seg>
            </tuv>
            <tuv xml:lang="hr">
                <seg>Futsal Pravila igre 2010/2011</seg>
            </tuv>
        </tu>
        <tu>
            <tuv xml:lang="en">
                <seg>Authorised by the Sub-Committee of the International Football Association Board.</seg>
            </tuv>
            <tuv xml:lang="hr">
                <seg>Odobrena od strane Podkomisije International Football Association Board-a (IFAB).</seg>
            </tuv>
        </tu>
```

Slika 8: Primjer koda u .TMX formatu

3.3. Sketch Engine

Sam korpus sportskih pravilnika izgrađen je i analiziran koristeći online korpusni alat *Sketch Engine*. Prvenstveno je zamišljen kao leksikografski alat, ali našao je primjenu i u različitim znanstvenim istraživanjima iz područja jezičnih tehnologija koji se može zatim koristiti u razne svrhe, kao što su analiza podataka, analiza prijevodnih ekvivalenata, korpusne analize, izrada terminoloških baza, učenje jezika, itd.[57]

Sketch Engine sadrži razne korpuse koje održava njihov tim[58]:

- online općejezični korpusi koji sadrže velik uzorak suvremenog općeg jezika
- paralelni korpusi koji sadrže paralelne tekstove 300 jezičnih parova, a sastavljeni su većinom iz dva izvora: *EUROPARL* i *OPUS* projekata
- učenički korpusi sadrže tekstove korisne za učenje jezika, a na *Sketch Engineu* se može pristupiti učeničkim korpusima za slovenski, češki i engleski
- povijesni korpusi koji daju uvid u razvoj i promjene određenog jezika kroz neko vremensko razdoblje
- korpusi govornog jezika sastavljeni su od audio ili video snimki te transkripata tih snimki, a korisni su za proučavanje i učenje izgovora

Ostali zanimljivi korpusi na *Sketch Engineu* su[59]:

- *Oxford Children's Corpus* koji sadrži materijale namijenjene djeci te materijale koje su sama djeca stvarala, a koristan je za proučavanje procesa učenja čitanja i pisanja
- *Brown* korpus i *British National Corpus* koji se koriste kao referentni korpusi
- *London English* korpus namijenjen sociolingvističkim istraživanjima, to jest proučavanju jezičnih varijacija između raznih društvenih i dobnih skupina i zajednica

[57] Kilgarriff, A., Baisa, V., Bušta, J. et al., The Sketch Engine: ten years on. // Lexicography: Journal of ASIALEX. 1, 1 (2014), str. 7-36.
[58] ibid.
[59] ibid.

Uz navedene korpuse, *Sketch Engine* nudi korisnicima opciju stvaranja i analiziranja vlastitih jednojezičnih ili paralelnih korpusa.

To se može postići postavljanjem datoteka raznih formata s računala na web ili pomoću *WebBootCat* procedure koja se koristi za sakupljanje podataka dostupnih na internetu. Korisničke korpuse moguće je proširivati, brisati datoteke te dijeliti na potkorpuse, a potrebno ih je kompilirati prije analize pomoću raznih funkcija koje *Sketch Engine* sadrži.

Jedna od osnovnih funkcija je pregledavanje jednojezičnih ili paralelnih konkordancija temeljenih na različitim upitima kao što su leme, specifični oblici riječi, fraze itd. Moguće je navesti i željeni kontekstualni raspon oko središnje riječi te koristiti razne filtere kako bi dobili što preciznije rezultate. Te rezultate moguće je dodatno sortirati, uzorkovati te naknadno analizirati i prikazati distribuciju ključnih riječi pomoću grafova. Moguće je stvarati razne frekvencijske liste riječi, lema, termina ili kolokacija. Te liste mogu prikazivati sirove frekvencije nekog korpusa, a mogu se koristiti i za pronalaženje specifičnih izraza u usporedbi s referentnim korpusima.

Za podržane jezike omogućena je i izravna jednojezična ili višejezična ekstrakcija specifične terminologije ili ključnih riječi. Ekstrakcija višerječnih izraza zahtijeva specijalizirani i referentni korpus koji su obrađeni alatima za tokenizaciju, lematizaciju i *POS* označavanje te gramatiku izraza koja prepoznaje tipične strukture termina[60], a u vrijeme istraživanja ona nije postojala za hrvatski jezik. Jedna od funkcija specifičnih za *Sketch Engine* je *Word Sketch* koja ukratko prikazuje gramatičke i kolokacijske odlike neke riječi, to jest njene kolokacije i gramatičke odnose između njih.

[60] Kilgarriff, A. Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine. // Proceedings of ASLIB 35th Translating and the Computer Conference / London, UK: 2013. URL: https://www.sketchengine.eu//wp-content/uploads/2015/05/Terminology_finding_2013.pdf (11. 3. 2018.)

Koristeći *Word Sketch Difference*, moguće je usporediti takve razlike između dviju riječi na istom jeziku ili čak između riječi različitih jezika. Naposljetku, distribucijski tezaurus prikazuje riječi koje dijele najviše kolokacija, to jest one riječi koje se pojavljuju u sličnim kontekstima nekog korpusa.

3.4. Korpus sportskih pravilnika

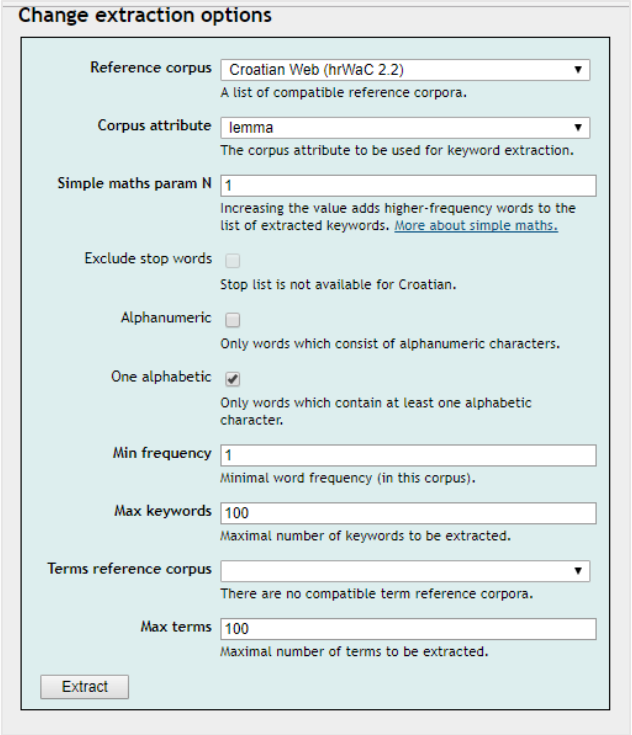
Kao što je navedeno, korpus kojim se bavimo u ovom radu izrađen je koristeći TMX datoteke koje su postavljene na *Sketch Engine*. Svaki dokument, to jest pravilnik određenog sporta sačinjava jedan potkorpus, dakle korpus sportskih pravilnika sastoji se od devet potkorpusa. Treba naglasiti da *Sketch Engine* paralelne korpuse prikazuje kao dva odvojena korpusa što znači da se naš korpus sastoji od korpusa engleskih pravilnika te od korpusa hrvatskih pravilnika, a oba su podijeljena na devet potkorpusa. Veličina korpusa i ostale detaljnije informacije bit će prikazane u rezultatima.

Cjelokupni korpus te pojedinačni potkorpusi analizirani su koristeći alate koje nudi *Sketch Engine* kako bi iz njih izvukli specifičnu terminologiju, to jest karakteristične riječi i višerječne izraze. Ovaj proces bio je poprilično jednostavan u analizi čitavog korpusa, pogotovo za engleski jezik. Naime, *Sketch Engine* sadrži funkciju pretraživanja ključnih riječi i višerječnih izraza na engleskom, dok je na hrvatskom moguće pretraživati ključne riječi, ali ne i višerječne izraze.

Moguće je izraditi frekvencijske liste specifičnih kolokacija pa su specifični višerječni izrazi identificirani pregledavanjem tih lista s obzirom na to da se riječi takvih izraza pojavljuju zajedno znatno češće nego obične riječi koje su u slučajnom suodnosu[61]. Posebna pozornost bila je usmjerena na kolokacije koje sadrže imenice. Za obje metode moguće je, ovisno o

[61] Manning, C.; Schütze, H. Foundations of statistical natural language processing. 2nd ed. Cambridge, MA, London: The MIT Press, 1999.

naravi istraživanja, pretraživati rijetke ili česte riječi koristeći *simple maths* parametar[62] čija najniža vrijednost prikazuje najrjeđe riječi dok najviša prikazuje najčešće. Na slici 9 mogu se vidjeti opcije ekstrakcije ključnih riječi za hrvatski jezik, a pretraživane su leme kako bi se dobio precizniji broj željenih pojmova.



Slika 9: Opcije ekstrakcije ključnih riječi za hrvatski jezik

S obzirom na to da je u ovom radu fokus istraživanja bila specifična terminologija, preferirani su rijetki izrazi koristeći *simple maths* parametar koji u tom slučaju treba sadržavati minimalnu vrijednost. Takvi izrazi pronalaze se u usporedbi sa širim uzorkom nekog jezika, to jest u usporedbi s nekim većim, općenitijim referentnim korpusom. Tako je engleski korpus sportskih pravilnika uspoređen s engleskim web korpusom (*enTenTen*), a za pronalaženje specifičnih višerječnih izraza korišten je *British National Corpus* (*BNC*). Svi specifični izrazi hrvatskog korpusa sportskih pravilnika pronađeni su u usporedbi s hrvatskim web korpusom (*hrWac 2.2*). Na razini pojedinačnih sportova, to jest potkorpusa, korišteni su isti referentni

[62] Kilgarriff, A. Getting to know your corpus. // Text, Speech and Dialogue / uredili Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. Berlin; Heidelberg: Springer, 2012. Str. 3-15.

korpusi kako bi se identificirala specifična terminologija u usporedbi s općenitim uzorkom jezika.

<< Back to corpus files

Use WebBootCaT with selected words

Single-word	Score	F	RefF
referee	1,048.20	1,096	121,609
goalkeeper	797.62	414	48,939
timekeeper	556.24	113	5,351
umpire	510.00	215	35,501
WP	412.13	198	43,637
goalkeeping	403.98	76	3,293
goalpost	338.62	71	6,278
unsportsmanlike	321.07	60	3,135
unsporting	296.53	51	1,086
crossbar	294.59	74	12,019
substitution	283.65	213	80,997
Futsal	272.64	51	3,173
FIH	269.69	46	897
foul	256.50	393	188,835
retake	250.24	91	27,553
offence	249.61	360	176,424
restart	236.97	280	140,454
thrower	232.05	63	14,841
infringement	220.54	207	106,924
ITTF	211.25	36	900
penalise	200.65	56	15,905
scoresheet	193.55	36	3,060
offside	173.96	47	14,692
nearest	166.46	74	38,767
penalty	159.64	800	669,082

Multi-word	Score	F	RefF
free kick	822.38	416	233
indirect free kick	810.64	138	4
goal line	743.69	143	19
free throw	723.48	121	2
opposing team	703.46	129	13
direct free kick	646.11	109	3
position of free kick	565.75	93	0
penalty area	552.01	208	145
playing court	541.46	89	0
penalty mark	541.46	89	0
penalty kick	533.34	104	21
penalty corner	466.78	91	21
goal area	433.04	73	3
team official	430.90	72	2
defending team	389.10	65	2
penalty stroke	388.40	70	11
field goal	384.05	67	7
old text	350.18	59	3
new text	319.93	60	16
game clock	298.56	49	0
unsporting behaviour	297.61	51	5
corner kick	289.55	53	13
third referee	280.34	46	0
team bench	278.11	46	1
playing time	254.38	54	33

Slika 10: Rezultati ekstrakcije ključnih riječi i višerječnih pojmova za engleski jezik

Sve frekvencijske liste naknadno su prekontrolirane, a izrazi koji su procijenjeni kao nespecifični za relevantni diskurs su uklonjeni. Takvi primjeri bili su rezultat strukture samih dokumenata ili su to bile riječi koje, unatoč njihovoj frekventnosti, imaju općenito značenje koje nije karakterizirano nekim od obrađenih sportova. Na slici 10 vidimo izraz *WP* (skraćenica od *water polo*) te višerječne fraze *old text* i *new text* koji su se na popisu našli zahvaljujući strukturi pravilnika u kojim se nalaze. Naime, svaki članak vaterpolskih pravila označen je oznakom *WP* i brojem članka (*WP 1, WP 2, WP 3* itd.), a sintagme *old text* i *new text* dio su pojašnjenja promjena pravila nogometne igre u kojima se uspoređuje stari tekst i novi tekst pravilnika. Na popisu se pronašao i prilog *nearest* za koji je procijenjeno da nije specifičan za taj diskurs te je uklonjen s konačne frekvencijske liste. Kratice raznih sportskih organizacija, saveza ili federacija također su uklonjene kao i izrazi koji se u hrvatskim pravilnicima koriste u izvornom, engleskom obliku (npr. *timeout*).

4. Rezultati

U ovom poglavlju predstavit će se rezultati istraživanja. U prvom dijelu bit će prikazane općenite informacije o korpusu, nakon čega slijede rezultati ekstrakcije specifične terminologije iz čitavog korpusa te pojedinačnih potkorpusa na engleskom i hrvatskom jeziku.

4.1. Informacije o korpusu

Prikupljen je korpus od ukupno 304 765 tokena, odnosno, 164 675 tokena za engleski i 140 090 tokena za hrvatski jezik. Za početak treba naglasiti da su *Sketch Engineovi* statistički podaci o korpusima aproksimativne naravi što se može iščitati iz činjenice da engleski korpus sadrži 8748 rečenica, a hrvatski 8844 što nije slučaj u stvarnosti jer su svi tekstovi sravnjeni na razini rečenice u odnosu 1:1. Unatoč tome, ovi podaci bi nam trebali pružiti dobar uvid u strukturu korpusa. Tablica 1 i tablica 2 sadrže osnovne informacije o engleskom dijelu korpusa sportskih pravilnika te o njegovim potkorpusima:

Tablica 1: Statistika korpusa (en)

Pojavnice	164 675
Riječi	142 499
Rečenice	8748
Dokumenti	9

Tablica 2: Statistika potkorpusa (en)

Potkorpus	**Pojavnice**	**Riječi**	**% korpusa**
Badminton	3987	~ 3450	2.42
Dvoranski hokej	13 532	~ 11 709	8.21
Futsal	39 950	~ 11 709	21.83
Hokej na travi	11 296	~ 9 774	6.85
Košarka	26 333	~ 22 786	15.99
Nogomet	25 516	~ 22 945	16.10
Rukomet	21 958	~ 19 001	13.33
Stolni tenis	12 793	~ 11 070	7.76
Vaterpolo	12 310	~ 10 652	7.47

Tablica 3 i tablica 4 sadrže informacije o hrvatskom dijelu korpusa i njegovim potkorpusima:

Tablica 3: Statistika korpusa (hr)

Pojavnice	140 090
Riječi	117 472
Rečenice	8844
Dokumenti	9

Tablica 4: Statistika potkorpusa (hr)

Potkorpus	**Pojavnice**	**Riječi**	**% korpusa**
Badminton	3331	~ 2793	2.37
Dvoranski hokej	11 189	~ 9382	7.98
Futsal	30 198	~ 25 322	21.55
Hokej na travi	9303	~ 7800	6.64
Košarka	21 963	~ 18 417	15.67
Nogomet	24 139	~ 20 241	17.23
Rukomet	18 626	~ 15 618	13.29
Stolni tenis	11 516	~ 9656	8.22
Vaterpolo	9825	~ 8238	7.01

U oba korpusa najveći su potkorpusi pravilnika za futsal, nogomet i košarku te sačinjavaju sličan postotak korpusa u kojima se nalaze.

4.2. Specifična terminologija

4.2.1. Cijeli korpus

Izvlačenje ključnih riječi i specifičnih višerječnih izraza za engleski jezik bio je jednostavan proces s obzirom na to da *Sketch Engine* sadrži sve potrebne alate za takve analize. Iako smo istraživanjem dobili puno veći broj rezultata, u radu će biti prikazano dvadeset najrelevantnijih.

Tablica 5: Prvih 20 ključnih riječi čitavog korpusa (en)

1	referee	**11**	Futsal
2	goalkeeper	**12**	foul
3	timekeeper	**13**	retake
4	umpire	**14**	offence
5	goalkeeping	**15**	restart
6	goalpost	**16**	thrower
7	unsportsmanlike	**17**	infringement
8	unsporting	**18**	penalise
9	crossbar	**19**	scoresheet
10	substitution	**20**	offside

Tablica 6: Prvih 20 specifičnih višerječnih izraza čitavog korpusa (en)

1	free kick	**11**	penalty kick
2	indirect free kick	**12**	penalty corner
3	goal line	**13**	goal area
4	free throw	**14**	team official
5	opposing team	**15**	defending team
6	direct free kick	**16**	penalty stroke
7	position of free kick	**17**	field goal
8	penalty area	**18**	game clock
9	playing court	**19**	unsporting behaviour
10	penalty mark	**20**	corner kick

Ključne riječi za hrvatski jezik izvučene su koristeći odgovarajući alat. Kao što je navedeno, specifični višerječni izrazi identificirani su pomoću specifičnih kolokacija jer ta funkcija nije podržana za hrvatski, a uspoređene su s općenitim uzorkom jezika (*hrWac 2.2*).

Tablica 7: Prvih 20 ključnih riječi čitavog korpusa (hr)

1	dosuđivati	**11**	prekršaj
2	mjeritelj	**12**	opomenuti
3	obrambeni	**13**	stativa
4	vratarev	**14**	dosuditi
5	nesportski	**15**	zapisnički
6	zapisničar	**16**	podbacivanje
7	ubacivanje	**17**	lopta
8	suparnički	**18**	zgoditak
9	neizravan	**19**	pomoćni
10	vratarski	**20**	postignut

Tablica 8: Prvih 20 specifičnih višerječnih izraza čitavog korpusa (hr)

1	slobodni udarac	**11**	suparnička momčad
2	kazneni udarac	**12**	izravni slobodni udarac
3	slobodno bacanje	**13**	mjeritelj vremena
4	kazneni prostor	**14**	nastavak igre
5	teren za igru	**15**	nesportsko ponašanje
6	zamjenski igrač	**16**	kazneni korner
7	posjed lopte	**17**	sat za igru
8	izvođenje kaznenog udarca	**18**	gol-aut crta
9	neizravni slobodni udarac	**19**	suparnički igrač
10	službena osoba	**20**	igrač s vratarskim pravima

Vidi se da engleski i hrvatski korpus nemaju puno zajedničkih ključnih riječi. Jedini pravi parovi su *goalpost* - *stativa* te *foul* - *prekršaj* dok *nesportski* ima dva pandana u engleskom korpusu – *unsportmanlike* i *unsporting*. *Timekeeper* i *mjeritelj* odražavaju razlike između engleskog i hrvatskog jezika. Naime, oba izraza označavaju isti koncept, ali u engleskom je on označen jednom riječju, a na hrvatskom frazom *mjeritelj vremena* koja se i našla na popisu specifičnih višerječnih izraza.

Popisi ključnih riječi, a pogotovo popis za hrvatski jezik, sadrže velik broj pridjeva što potvrđuje njihovu specifičnost, ali potencijalno indicira i na veću specifičnost i relevantnost višerječnih izraza.

Popisi višerječnih izraza dijele puno veći broj zajedničkih specifičnih fraza – njih 12, no treba naglasiti da je hrvatski prijevod za *penalty kick* (specifičan za nogomet i futsal) te *penalty stroke* (specifičan za dvoranski hokej i hokej na travi) jednak za sve navedene sportove – *kazneni udarac* što nije jedini takav slučaj.

4.2.2. Badminton

Specifična terminologija potkorpusa ekstrahirana je u usporedbi s općenitim uzorkom jezika (*enTenTen* i *hrWac 2.2*). Slijede rezultati najmanjeg potkorpusa.

Tablica 9: Prvih 20 ključnih riječi (en)

1	stringed	**11**	fault
2	racket	**12**	server
3	umpire	**13**	cord
4	badminton	**14**	net
5	shuttle	**15**	suspend
6	referee	**16**	opponent
7	receiver	**17**	boundary
8	feather	**18**	court
9	rally	**19**	stroke
10	offend	**20**	player

Tablica 10: Specifični višerječni izrazi (en)

1	service court
2	serving side
3	stringed area
4	offending side
5	receiving side

Hrvatski dio potkorpusa također sadrži samo pet specifičnih višerječnih izraza, a svi rezultati navedeni su u sljedećim tablicama.

Tablica 11: Prvih 20 ključnih riječi (hr)

1	ožičje	**11**	primatelj
2	serverov	**12**	suparnički
3	primateljski	**13**	servirati
4	strana	**14**	server
5	otezanje	**15**	igralište
6	serverski	**16**	pero
7	loptica	**17**	smetnja
8	reket	**18**	meč
9	servisni	**19**	servis
10	gem	**20**	suigrač

Tablica 12: Specifični višerječni izrazi (hr)

1	servisno polje
2	vrhovni sudac
3	primateljska strana
4	serverov reket
5	površina igrališta

Popisi ključnih riječi sadrže pet zajedničkih izraza, a najčešći specifični višerječni izraz isti je za oba jezika. U potkorpusu pravilnika za badminton vidimo još jedan primjer u kojem različite terminološke kategorije označavaju isti specifični pojam: *stringed* je naveden u popisu ključnih riječi, a dio je specifičnog višerječnog izraza *stringed area* koji je na hrvatski preveden jednom riječju – *ožičje*.

4.2.3. Dvoranski hokej

Engleski dio potkorpusa dvoranskog hokeja specifičan je jer sadrži velik broj ključnih riječi koje su složenice dviju samostalnih riječi, ali ortografski sačinjavaju samo jednu riječ.

Tablica 13: Prvih 20 ključnih riječi (en)

1	side-board	**11**	headgear
2	back-line	**12**	goalkeeper
3	goal-post	**13**	time-out
4	back-board	**14**	penalise
5	centre-line	**15**	rules
6	goalkeeping	**16**	offence
7	cross-bar	**17**	substitution
8	goal-line	**18**	half-time
9	re-start	**19**	deflect
10	umpire	**20**	penalty

Tablica 14: Specifični višerječni izrazi (en)

1	free push	**10**	field player
2	penalty stroke	**11**	protective headgear
3	full protective equipment	**12**	offending team
4	penalty corner	**13**	playing side
5	goalkeeper wearing full protective equipment	**14**	defending team
6	playing distance	**15**	protective equipment
7	centre pass	**16**	yellow card
8	colour shirt	**17**	face mask
9	indoor hockey		

Tablice 15 i 16 sadrže ključne riječi i specifične višerječne izraze hrvatskog potkorpusa dvoranskog hokeja:

Tablica 15: Prvih 20 ključnih riječi (hr)

1	gol-aut	**11**	palica
2	znak	**12**	korner
3	zagrađivanje	**13**	plosnat
4	vratarski	**14**	hokej
5	dosuđivati	**15**	drška
6	stativa	**16**	zgoditak
7	jednolik	**17**	brid
8	zviždaljka	**18**	prečka
9	štitnik	**19**	dvoranski
10	guranje	**20**	buli

Tablica 16: Specifični višerječni izrazi (hr)

1	kazneni korner	**9**	bočna greda
2	kazneni udarac	**10**	štitnik za noge
3	gol-aut crta	**11**	zaštitna oprema
4	igrač sa vratarskim pravima	**12**	nastavak igre
5	slobodno guranje	**13**	slobodni udarac
6	igrač u polju	**14**	središnja crta
7	izvođenje kaznenog kornera	**15**	gol crta
8	izvođenje kaznenog udarca		

Najfrekventnije ključne riječi zajedničke engleskom i hrvatskom dijelu potkorpusa poprilično su rijetke (*goal-post* - *stativa* te *goalkeeping* - *vratarski*), a ta dva zajednička primjera rezultat su navedenih ortografskih specifičnosti terminologije engleskog potkorpusa. Uistinu, četiri engleske ključne riječi prevedene su kao fraze te su se našle na popisu hrvatskih specifičnih višerječnih izraza (*back-line* i *gol-aut crta*, *goal-line* i *gol crta*, *centre-line* i *središnja crta*, *side-board* i *bočna greda*). Popisi specifičnih višerječnih izraza dijele četiri zajednička termina: *free push* - *slobodno guranje*, *penalty stroke* - *kazneni udarac*, *field player* - i*grač u polju* te *protective equipment* - *zaštitna oprema*.

4.2.4. Futsal

Ključne riječi i specifični višerječni izrazi engleskog dijela potkorpusa futsala prikazani su u tablicama 17 i 18:

Tablica 17: Prvih 20 ključnih riječi (en)

1	kick-in	**11**	referee
2	four-second	**12**	crossbar
3	unsporting	**13**	goalkeeper
4	sending-off	**14**	time-out
5	back-line	**15**	goal-line
6	futsal	**16**	retake
7	sent-off	**17**	chronometer
8	goalscoring	**18**	goalkeeping
9	timekeeper	**19**	infringement
10	goalpost	**20**	penalise

Tablica 18: Prvih 20 specifičnih višerječnih izraza (en)

1	position of free kick	**11**	four-second count
2	third referee	**12**	second penalty mark
3	indirect free kick	**13**	goalscoring opportunity
4	goal clearance	**14**	substitution zone
5	penalty mark	**15**	corner arc
6	penalty area line	**16**	double caution
7	direct free kick	**17**	defending goalkeeper
8	substitution procedure	**18**	second referee
9	unsporting behaviour	**19**	touch line
10	obvious goalscoring opportunity	**20**	restarting play

Zajedničke ključne riječi hrvatskog i engleskog potkorpusa su *unsporting* - *nesportski*, *chronometer* - *kronometar* te *futsal* - *futsal* koji su procijenjeni kao relevantni te nisu uklonjeni iz rezultata iako je termin identičan. Sve ključne riječi i specifični višerječni pojmovi sadržani su u tablicama 19 i 20.

Tablica 19: Prvih 20 ključnih riječi (hr)

1	obrambeni	**11**	suparnički
2	pomoćni	**12**	nesportski
3	mjeritelj	**13**	ubacivanje
4	znak	**14**	opomenuti
5	vratarev	**15**	akumuliran
6	dosuđivati	**16**	zgoditak
7	jednominutni	**17**	kronometar
8	neizravan	**18**	zviždaljka
9	futsal	**19**	nesmotren
10	gol-aut	**20**	zamjenski

Tablica 20: Prvih 20 specifičnih višerječnih izraza (hr)

1	slobodni udarac	**11**	ubacivanje nogom
2	kazneni udarac	**12**	treći sudac
3	kazneni prostor	**13**	mjeritelj vremena
4	zamjenski igrač	**14**	nesportsko ponašanje
5	izvođenje kaznenog udarca	**15**	izgledna prilika za postizanje zgoditka
6	suparnička momčad	**16**	poprečna crta
7	neizravni slobodni udarac	**17**	crta kaznenog prostora
8	izravni slobodni udarac	**18**	udarac iz kuta
9	akumulirani prekršaj	**19**	službena osoba
10	točka za izvođenje kaznenog udarca	**20**	spuštanje lopte

Timekeeper - mjeritelj vremena i *kick-in - ubacivanje nogom* termini su koji su se našli na engleskom popisu ključnih riječi i na popisu hrvatskih višerječnih pojmova. Zajedničke specifične višerječne fraze su: *third referee - treći sudac*, *indirect free kick - neizravni slobodni udarac, penalty mark - točka za izvođenje kaznenog udarca, penalty area line - crta kaznenog prostora, direct free kick - izravni slobodni udarac, unsporting behaviour - nesportsko ponašanje* te *obvious goalscoring opportunity - izgledna prilika za postizanje zgoditka*.

4.2.5. Hokej na travi

Ključne riječi i specifični višerječni izrazi pa i frekvencije termina pravilnika hokeja na travi vrlo su slični onima iz pravilnika dvoranskog hokeja te dijele sedamnaest ključnih riječi i petnaest specifičnih višerječnih izraza na engleskom jeziku:

Tablica 21: Prvih 20 ključnih riječi (en)

1	back-line	**11**	re-start
2	goal-post	**12**	umpire
3	flag-post	**13**	headgear
4	side-board	**14**	goalkeeper
5	back-board	**15**	penalise
6	centre-line	**16**	rules
7	goalkeeping	**17**	substitution
8	cross-bar	**18**	offence
9	side-line	**19**	deflect
10	goal-line	**20**	kicker

Tablica 22: Specifični višerječni izrazi (en)

1	free hit	**9**	field player
2	penalty stroke	**10**	protective headgear
3	full protective equipment	**11**	offending team
4	penalty corner	**12**	playing side
5	goalkeeper wearing full protective equipment	**13**	defending team
6	playing distance	**14**	protective equipment
7	centre pass	**15**	yellow card
8	colour shirt	**16**	face mask

S obzirom na to da engleski pravilnici dvoranskog hokeja i hokeja na travi uvelike dijele specifičnu terminologiju i ortografiju, za očekivati je i da hrvatski dio pokazuje iste osobine i specifičnosti prijevoda. Šesnaest ključnih riječi bilo je zajedničko ovim potkorpusima, a

zajedničkih specifičnih višerječnih izraza bilo je dvanaest s tim da je *bočna greda* iz pravilnika dvoranskog hokeja u ovom pravilniku prevedena kao *bočna daska*.

Tablica 23: Prvih 20 ključnih riječi (hr)

1	gol-aut	**11**	brid
2	znak	**12**	plosnat
3	zagrađivanje	**13**	drška
4	strana	**14**	korner
5	vratarski	**15**	hokej
6	dosuđivati	**16**	prečka
7	stativa	**17**	crta
8	jednolik	**18**	takmičenje
9	štitnik	**19**	zgoditak
10	palica	**20**	papučica

Tablica 24: Prvih 20 specifičnih višerječnih izraza (hr)

1	gol-aut crta	**11**	bočna daska
2	igrač sa pravima vratara	**12**	igranje loptom
3	kazneni korner	**13**	izvođenje kaznenog kornera
4	kazneni udarac	**14**	izvođenje kaznenog udarca
5	igrač u polju	**15**	nacionalni savezi
6	aut crta	**16**	vratar s punom zaštitnom opremom
7	slobodni udarac	**17**	stražnja daska
8	štitnici za noge	**18**	glava palice
9	zaštitna oprema	**19**	nastavak igre
10	udarac na vrata	**20**	postizanje zgoditka

Zajedničkih ključnih riječi unutar potkorpusa ipak ima malo više nego u pravilnicima za dvoranski hokej: *goal-post - stativa*, *goalkeeping - vratarski*, *cross-bar - prečka*, *kicker - papučica*. Ključne riječi prevedene kao višerječni pojmovi su: *back-line* i *gol-aut crta*, *side-board* i *bočna daska*, *back-board* i *stražnja daska*, *side-line* i *aut crta*. Zajednički specifični višerječni pojmovi su: *free hit - slobodni udarac*, *penalty stroke - kazneni udarac*, *penalty corner - kazneni korner*, *goalkeeper wearing full protective equipment - vratar s punom zaštitnom opremom*, *field player - igrač u polju* te *protective equipment - zaštitna oprema*.

4.2.6. Košarka

Zanimljiv primjer u engleskom dijelu ovog potkorpusa je različita ortografija istog termina. Naime, slobodno bacanje se u nekim situacijama koristilo kao jedna riječ - *free-throw*, a u drugim kao *free throw*, a pojam se našao na oba popisa.

Tablica 25: Prvih 20 ključnih riječi (en)

1	throw-in	**11**	semi-circle
2	endline	**12**	throw
3	non-scoring	**13**	backboard
4	scoresheet	**14**	backcourt
5	out-of-bound	**15**	correctable
6	free-throw	**16**	scorer
7	time-out	**17**	foul
8	no-charge	**18**	mid-point
9	unsportsmanlike	**19**	goaltending
10	frontcourt	**20**	infraction

Tablica 26: Prvih 20 specifičnih višerječnih izraza (en)

1	playing court	**11**	team bench area
2	free throw	**12**	jump ball
3	team bench	**13**	live ball
4	alternating possession	**14**	free-throw shooter
5	legal guarding position	**15**	initial legal guarding position
6	game clock signal	**16**	game clock
7	end of playing time	**17**	alternating possession arrow
8	no-charge semi-circle	**18**	disqualifying foul
9	shot clock signal	**19**	ball situation
10	field goal area	**20**	foul penalty

I hrvatski dio potkorpusa sadrži jednu zanimljivost. Iznimno rijetko smo imali primjer u kojem je engleski pojam nekog koncepta sadržavao više riječi, a hrvatski samo jednu. Ovdje je to izraz *jump ball* koji je na hrvatski preveden kao *podbacivanje*.

Tablica 27: Prvih 20 ključnih riječi (hr)

1	obrambeni	**11**	dosuđivati
2	pomoćni	**12**	doticanje
3	podbacivanje	**13**	ubacivanje
4	isključujući	**14**	čeon
5	zazviždati	**15**	nesportski
6	zapisničar	**16**	opunomoćenik
7	mjeritelj	**17**	posjed
8	zapisnički	**18**	bacanje
9	polukrug	**19**	nepropisan
10	stajni	**20**	protivnički

Tablica 28: Prvih 20 specifičnih višerječnih izraza (hr)

1	minuta odmora	**11**	klupa momčadi
2	posjed lopte	**12**	čeona crta
3	slobodno bacanje	**13**	mrtva lopta
4	sat za igru	**14**	obrambeni igrač
5	protivnički igrač	**15**	šut na koš iz igre
6	sat za napad	**16**	zapisnički stol
7	granična crta	**17**	ubacivanje lopte
8	pogodak iz igre	**18**	moment šuta
9	tehnička pogreška	**19**	zadnje polje
10	živa lopta	**20**	promjenjivi posjed

Zajedničke ključne riječi u košarkaškom diskursu su: *unsportsmanlike - nesportski, semi-circle – polukrug* te *throw - bacanje*. Engleske ključne riječi prevedene kao višerječni izrazi su: *throw-in* i *ubacivanje lopte, endline* i *čeona crta*, *free-throw* i *slobodno bacanje*, *time-out* i *minuta odmora*, *backcourt* i *zadnje polje*. Specifičnih višerječnih izraza zajedničkih obama dijelovima potkorpusa bilo je pet: *free throw - slobodno bacanje, team bench - klupa momčadi, alternating possession - promjenjivi posjed, live ball - živa lopta* te *game clock - sat za igru*.

4.2.7. Nogomet

Iako postoje sličnosti s terminologijom futsala one nisu tako izražene u engleskom dijelu pravilnika za nogomet kao u hrvatskom dijelu. Frekvencijski popisi engleskih potkorpusa futsala i nogometa dijele sedam ključnih riječi te sedam specifičnih višerječnih izraza.

Tablica 29: Prvih 20 ključnih riječi (en)

1	flagpost	**11**	referee
2	cautionable	**12**	pre-match
3	shinguard	**13**	restart
4	unsporting	**14**	goalkeeper
5	sending-off	**15**	crossbar
6	throw-in	**16**	team-mate
7	touchline	**17**	retake
8	goal-scoring	**18**	penalise
9	goalpost	**19**	feint
10	offside	**20**	kick-off

Tablica 30: Prvih 20 specifičnih višerječnih izraza (en)

1	indirect free kick	**11**	pre-match inspection
2	penalty mark	**12**	position of free kick
3	direct free kick	**13**	compulsory equipment
4	offside offence	**14**	team official
5	unsporting behaviour	**15**	play restarts
6	obvious goal-scoring opportunity	**16**	stop play
7	goal-scoring opportunity	**17**	assistant referee
8	serious foul play	**18**	corner flagpost
9	match official	**19**	sending-off offence
10	outside agent	**20**	next stoppage

Hrvatski dio potkorpusa pak pokazuje nešto više sličnosti s potkorpusom futsala. Sadrže devet zajedničkih ključnih riječi i četrnaest specifičnih višerječnih izraza s tim da je u futsalu korišten izraz *poprečna crta*, a u nogometu *poprečna linija*.

Tablica 31: Prvih 20 ključnih riječi (hr)

1	vratarev	**11**	suparnički
2	pomoćni	**12**	igračev
3	obrambeni	**13**	opomenuti
4	podhlačice	**14**	sučev
5	fintiranje	**15**	nesportski
6	postignut	**16**	nesmotren
7	dosuđivati	**17**	poprečan
8	kostobran	**18**	uzdužan
9	gol-crta	**19**	kutni
10	neizravan	**20**	jedanaesterac

Tablica 32: Prvih 20 specifičnih višerječnih izraza (hr)

1	teren za igru	**11**	udarac iz kuta
2	slobodni udarac	**12**	član sudačkog tima
3	kazneni udarac	**13**	pravila nogometne igre
4	zamjenski igrač	**14**	izgledna prilika za postizanje pogotka
5	kazneni prostor	**15**	suparnička momčad
6	neizravni slobodni udarac	**16**	spuštanje lopte
7	suparnički igrač	**17**	poprečna linija
8	nastavak igre	**18**	nesportsko ponašanje
9	izravni slobodni udarac	**19**	službena osoba
10	sudački tim	**20**	zamijenjeni igrač

Zanimljivo je da nijedna engleska ključna riječ nije završila na popisu najfrekventnijih hrvatskih višerječnih izraza što nipošto ne znači da takvi slučajevi ne postoje. Zajedničke ključne riječi bile su *shinguard - kostobran, unsporting - nesportski* te *feint - fintiranje,* a višerječni izrazi *indirect free kick - neizravni slobodni udarac, direct free kick - izravni slobodni udarac, unsporting behaviour - nesportsko ponašanje, obvious goal-scoring opportunity - izgledna prilika za postizanje pogotka, match official - član sudačkog tima* te *team official - službena osoba*.

4.2.8. Rukomet

Rukometni diskurs također sadrži dosta složenica koje su se našle na popisu engleskih ključnih riječi:

Tablica 33: Prvih 20 ključnih riječi (en)

1	goalkeeper-throw	**11**	goalpost
2	throw-off	**12**	disqualification
3	goal-area	**13**	referee
4	free-throw	**14**	goalkeeper
5	unsportsmanlike	**15**	substitution
6	scorekeeper	**16**	forewarn
7	timekeeper	**17**	infraction
8	throw-in	**18**	half-time
9	time-out	**19**	crossbar
10	thrower	**20**	whistle

Tablica 34: Prvih 20 specifičnih višerječnih izraza (en)

1	substitution area	**11**	7-meter throw
2	2-minute suspension	**12**	personal punishment
3	team time-out	**13**	passive play
4	whistle signal	**14**	playing court
5	goal-area line	**15**	substitution line
6	outer goal line	**16**	build-up phase
7	team official	**17**	guilty player
8	goal area	**18**	4-meter line
9	responsible team official	**19**	faulty substitution
10	forewarning signal	**20**	unsportsmanlike conduct

Hrvatske ključne riječi i specifični višerječni izrazi potkorpusa rukometa prikazani su u tablicama 35 i 36:

Tablica 35: Prvih 20 ključnih riječi (hr)

1	obrambeni	**11**	bacanje
2	vratarev	**12**	obračunavanje
3	mjeritelj	**13**	zvižduk
4	dosuđivati	**14**	progresivno
5	zapisničar	**15**	stativa
6	nesportski	**16**	isteći
7	zapisnički	**17**	protivnički
8	sedmerac	**18**	sučev
9	diskvalifikacija	**19**	isključenje
10	uzdužan	**20**	prekobrojan

Tablica 36: Prvih 20 specifičnih višerječnih izraza (hr)

1	slobodno bacanje	**11**	izvođenje bacanja
2	službena osoba	**12**	trajanje igre
3	mjeritelj vremena	**13**	protivnička momčad
4	prostor za zamjenu igrača	**14**	pasivna igra
5	vratarev prostor	**15**	izvođenje sedmeraca
6	nesportsko ponašanje	**16**	isključenje na 2 minute
7	početno bacanje	**17**	službeni predstavnik
8	posjed lopte	**18**	protivnički igrač
9	polje za igru	**19**	obrambeni igrač
10	uzdužna linija	**20**	jasna prigoda za postizanje zgoditka

Zajedničke ključne riječi bile su *unsportsmanlike - nesportski, scorekeeper - zapisničar, goalpost - stativa, disqualification – diskvalifikacija* te *whistle - zvižduk*. Engleske ključne riječi prevedene kao višerječni izrazi bile su *throw-off* i *početno bacanje*, *free-throw* i *slobodno bacanje*, *timekeeper* i *mjeritelj vremena* te su tipični primjeri takvih slučajeva. Zajedničkih višerječnih izraza bilo je sedam: *substitution area - prostor za zamjenu igrača*, *2-minute suspension - isključenje na 2 minute*, *team official - službeni predstavnik*, *goal area - vratarev prostor, passive play - pasivna igra*, *playing court - polje za igru* te *unsportsmanlike conduct - nesportsko ponašanje*. Bio je i jedan slučaj engleskog višerječnog izraza koji je preveden kao jedna riječ: *7-meter throw* i *sedmerac*.

4.2.9. Stolni tenis

Ključne riječi i specifični višerječni izrazi engleskog potkorpusa stolnog tenisa prikazani su u tablicama 37 i 38:

Tablica 37: Prvih 20 ključnih riječi (en)

1	pimpled	**11**	continental
2	umpire	**12**	disqualify
3	racket	**13**	disciplinary
4	time-out	**14**	advertisement
5	paralympic	**15**	receiver
6	referee	**16**	adhesive
7	authorise	**17**	adviser
8	expedite	**18**	covering
9	laws	**19**	offence
10	regulations	**20**	rally

Tablica 38: Specifični višerječni izrazi (en)

1	assistant umpire	**10**	playing surface
2	racket control	**11**	end line
3	paralympic title	**12**	team event
4	individual match	**13**	management committee
5	pimpled rubber	**14**	physical disability
6	net assembly	**15**	red card
7	playing area	**16**	total area
8	team match	**17**	free hand
9	individual event		

Ključne riječi i specifični višerječni izrazi hrvatskog potkorpusa stolnog tenisa prikazani su u tablicama 39 i 40:

Tablica 39: Prvih 20 ključnih riječi (hr)

1	ekspeditivan	**11**	momčadski
2	pravila	**12**	primatelj
3	ždrijebati	**13**	redoslijed
4	nazubljena	**14**	ždrijeb
5	reket	**15**	opomenuti
6	igračev	**16**	dodir
7	pričvršćen	**17**	kontinentalan
8	paraolimpijski	**18**	predsjedavajući
9	serviranje	**19**	rangirati
10	propagandni	**20**	stegovni

Tablica 40: Prvih 20 specifičnih višerječnih izraza (hr)

1	vrhovni sudac	**11**	momčadski susret
2	prostor za igru	**12**	pojedinačno natjecanje
3	propagandna poruka	**13**	izvršni odbor
4	površina igranja	**14**	službena osoba
5	pojedinačni susret	**15**	pomoćni sudac
6	rang lista	**16**	kontrola reketa
7	igra parova	**17**	komplet mreže
8	međunarodno natjecanje	**18**	otvorena međunarodna prvenstva
9	prijavljeni igrač	**19**	momčadsko natjecanje
10	ekspeditivni sistem	**20**	ploha reketa

Zajedničke ključne riječi ovog potkorpusa bile su *pimpled - nazubljena, racket - reket, paralympic - paraolimpijski, expedite - ekspeditivan, laws - pravila, continental - kontinentalan, disciplinary – stegovni* te *receiver - primatelj*. *Referee* i *advertisement* ključne su riječi koje su prevedene višerječnim izrazima *vrhovni sudac* i *propagandna poruka*. Bilo je devet višerječnih izraza koji su se našli na oba frekvencijska popisa: *assistant umpire - pomoćni sudac, racket control - kontrola reketa, individual match - pojedinačni susret, net assembly - komplet mreže, playing area - površina igranja, team match - momčadski susret, individual event - pojedinačno natjecanje, playing surface - površina igranja* te *team event - momčadsko natjecanje*.

4.2.10. Vaterpolo

Ključne riječi i specifični višerječni izrazi engleskog potkorpusa vaterpola prikazani su u tablicama 41 i 42:

Tablica 41: Prvih 20 ključnih riječi (en)

1	timekeeper	**11**	stoppage
2	throw	**12**	opposing
3	re-entry	**13**	retake
4	goalkeeper	**14**	exclusion
5	crossbar	**15**	brutality
6	referee	**16**	elapse
7	timeout	**17**	dribble
8	thrower	**18**	rules
9	re-enter	**19**	substitute
10	foul	**20**	offence

Tablica 42: Prvih 20 specifičnih višerječnih izraza (en)

1	penalty throw	**11**	offending player
2	re-entry area	**12**	defending player
3	5 metre area	**13**	defending goalkeeper
4	goal throw	**14**	ordinary foul
5	half distance line	**15**	penalty foul
6	neutral throw	**16**	goal line
7	corner throw	**17**	substitute goalkeeper
8	third personal foul	**18**	5 metre line
9	probable goal	**19**	official table
10	improper re-entry	**20**	goal judge

Ključne riječi i specifični višerječni izrazi hrvatskog potkorpusa vaterpola prikazani su u tablicama 43 i 44:

Tablica 43: Prvih 20 ključnih riječi (hr)

1	obrambeni	**11**	kapica
2	postignut	**12**	bacanje
3	mjeritelj	**13**	prekršitelj
4	dosuđivanje	**14**	dosuditi
5	zapisničar	**15**	ometanje
6	vratarski	**16**	odugovlačenje
7	stativa	**17**	prekršaj
8	zapisnički	**18**	isključenje
9	zviždaljka	**19**	protivnički
10	peterac	**20**	nepravilan

Tablica 44: Prvih 20 specifičnih višerječnih izraza (hr)

1	posjed lopte	**11**	gol bacanje
2	slobodno bacanje	**12**	obrambena ekipa
3	gol crta	**13**	protivnička ekipa
4	ponovni ulazak	**14**	polovina igrališta
5	kazneno bacanje	**15**	obrambeni igrač
6	kazneni udarac	**16**	izvođenje peteraca
7	gol sudac	**17**	kraj utakmice
8	mjeritelj vremena	**18**	osobna pogreška
9	prostor za ponovni ulazak	**19**	polovica igrališta
10	čista igra	**20**	prekršaj za isključenje

Zajedničke ključne riječi engleskog i hrvatskog dijela potkorpusa bile su *throw - bacanje, foul - prekršaj, opposing - protivnički, exclusion - isključenje, offence - prekršaj*, a *timekeeper* i *re-entry* prevedeni su kao *mjeritelj vremena* i *ponovni ulazak* te su se našli na hrvatskom popisu specifičnih višerječnih izraza. Na tim popisima u oba potkorpusa našli su se *penalty throw - kazneni udarac* / *kazneno bacanje, re-entry area - prostor za ponovni ulazak, goal throw - gol bacanje, half distance line - polovica igrališta, defending player - obrambeni igrač, penalty foul - prekršaj za isključenje, goal line - gol crta* te *goal judge - gol sudac.*

4.3. Rasprava

Broj zajedničkih višerječnih izraza bio je veći od broja zajedničkih ključnih riječi u svim potkorpusima osim badmintona. Treba naglasiti i da je iz tog potkorpusa ekstrahirano samo pet specifičnih višerječnih izraza na oba jezika što je najvjerojatnije odraz činjenice da je to daleko najmanji potkorpus u usporedbi s ostatkom korpusa sportskih pravilnika. Kao što je navedeno, tako je bilo i na razini čitavog korpusa, a zajedničkih višerječnih izraza bilo je četiri puta više – njih dvanaest u odnosu na tri zajedničke ključne riječi, s obzirom na to da su dvije ključne riječi i dva višerječna izraza prevedena po jednim terminom.

Ovi rezultati potvrđuju ulogu višerječnih termina u okviru specijalizirane terminologije. Naime, višerječni izrazi čine više od 70 % specijaliziranih terminoloških leksikona[63], a na konkretnom primjeru korpusa sažetaka znanstvenih tekstova iz područja biomedicine takvi izrazi činili su od 50 % do 80 % specifičnih izraza[64]. Nadalje, na uzorku (11.3 %) englesko – hrvatskog dijela *Englesko – hrvatsko i hrvatsko – engleskog rječnika elektronike*, višerječni izrazi sačinjavali su 72.6 % natuknica na engleskom jeziku ili njihovih prijevodnih

[63] Krieger, M. G.; Finatto, M. J. B. Introdução à Terminologia: teoria & prática. 2nd ed. São Paulo: Contexto, 2004.
[64] Ramisch, C. E. Multi-word terminology extraction for domain-specific documents. Magistarski rad. Grenoble: Grenoble INP – Institut National Polytechnique de Grenoble, 2009.

ekvivalenata[65]. U istom istraživanju otkriveno je da natuknice koje su jednorječne na oba jezika čine samo 2.8 % engleskog te 4.6 % hrvatskog dijela rječnika[66].

Iako korpus pravilnika sadrži terminologiju različitih sportova od kojih su neki izrazi prevedeni istim terminom, možemo pretpostaviti da se odnose na iste koncepte u različitom kontekstu te na temelju te pretpostavke predstaviti u kojem omjeru hrvatski i engleski dio dijele ključne riječi i višerječne izraze. Prema provedenom istraživanju engleski dio korpusa sportskih pravilnika sadrži 20 % ključnih riječi zajedničkih s hrvatskim dijelom i 65 % zajedničkih specifičnih višerječnih izraza. Hrvatski dio sadrži 15 % zajedničkih jednorječnih i 60 % višerječnih izraza.

U svim potkorpusima relativne brojke bile su veće u korist višerječnih izraza pa tako i u pravilnicima za badminton. Najviše zajedničkih jednorječnih izraza bilo je u pravilnicima stolnog tenisa – njih 40 %, a višerječnih u engleskom dijelu istog potkorpusa – njih 53 %, s tim da je hrvatski dio sadržavao 45 % takvih slučajeva, a razlika u postocima rezultat je različitog broja identificiranih fraza na frekvencijskim popisima potkorpusa.

Najmanje zajedničkih ključnih riječi bilo je u pravilnicima dvoranskog hokeja (10 %) što je najvjerojatnije odraz velikog broja složenica koje su na engleskom jeziku ortografski označene kao jedna riječ. Najmanje zajedničkih specifičnih višerječnih izraza bilo je opet u istom potkorpusu: njih 23.5 % u engleskom dijelu i 26.7 % u hrvatskom.

Jedino su frekvencijski popisi zajedničkih višerječnih izraza pravilnika košarke imali manje zajedničkih jedinica od hrvatskog dijela potkorpusa dvoranskog hokeja – njih 25 %. S obzirom na to da su i pravilnici košarke obilježeni ortografskim razlikama između hrvatskog i engleskog jezika, možemo zaključiti da je to glavni uzrok ovakvih rezultata. Te razlike

[65] Štambuk, A. Multi-word lexical units in English and Croatian terminology of electronics. // Studia Romanica et Anglica Zagrabiensia. 42, 2 (1997), str. 373-390.
[66] ibid.

možemo iščitati iz broja ključnih riječi koje su prevedene kao višerječni izrazi. Potkorpus košarke sadrži pet takvih slučajeva, a potkorpusi dvoranskog hokeja i hokeja na travi po četiri. U prilog ovoj pretpostavci ide i činjenica da engleski dio potkorpusa s najvećim brojem zajedničkih ključnih riječi i višerječnih izraza (stolni tenis) sadrži samo jednu ovakvu riječ. Valja napomenuti da je obrnutih slučajeva bilo iznimno malo – samo dva i to po jedan u potkorpusu košarke (*jump ball* i *podbacivanje*) i rukometa (*7-meter throw* i *sedmerac*).

S obzirom na to da pridjevi opisuju imenice, njihova prisutnost na frekvencijskim popisima ključnih riječi najvjerojatnije je još jedan odraz važnosti višerječnih izraza u sportskoj terminologiji. Takvih slučajeva bilo je dosta više u hrvatskom dijelu korpusa, a glavni uzrok tome su sintaktičke razlike između jezika i način formiranja višerječnih izraza. Dok engleski dopušta spajanje dviju imenica u nominativu, hrvatski kao flektivni jezik ne može samo spojiti dvije imenice u jednu semantičku jedinicu bez da izrazi odnos između njih pomoću prepozicijskih fraza ili upotrebe padeža, najčešće imenice u genitivu.

U istraživanju terminologije elektronike, fraze s imenicom u nominativu i imenicom u genitivu činile su samo 12 % hrvatskih višerječnih izraza dok su kombinacije pridjeva i imenica sačinjavale 52 % uzorka. U engleskom dijelu kombinacije pridjeva i imenica činile su 45 %, a kombinacije dviju imenica 29 % uzorka[67]. Na temelju ovih podataka možemo zaključiti da prisutnost pridjeva na frekvencijskim popisima uistinu indicira na još veću važnost višerječnih izraza, konkretno u sportskoj terminologiji te potencijalno u terminologijama općenito.

Na popisu ključnih riječi čitavog korpusa sportskih pravilnika bila su tri pridjeva u engleskom dijelu te sedam u hrvatskom dijelu, a na popisima specifičnih višerječnih izraza našao se jedan od tih engleskih pridjeva te čak pet hrvatskih. Najviše takvih slučajeva na razini potkorpusa bilo je u pravilnicima nogometa: pet pridjeva na engleskom te trinaest pridjeva na

[67] ibid.

hrvatskom popisu ključnih riječi, a svi engleski te pet hrvatskih pojavilo se kao dio specifičnog višerječnog izraza. Svaki potkorpus sadržavao je barem po jedan pridjev na popisu ključnih riječi, a skoro svaki je taj pridjev sadržavao kao dio specifičnog višerječnog izraza (osim engleskog dijela potkorpusa dvoranskog hokeja i hokeja na travi).

Valja napomenuti i činjenicu da je za engleski jezik podržano automatsko prepoznavanje specifičnih višerječnih izraza dok ta funkcija nije podržana za hrvatski jezik što je moglo imati utjecaj na broj pridjeva na frekvencijskim listama ovih jezika. No s obzirom na to da se velik broj engleskih pridjeva s popisa ključnih riječi pojavio kao dio specifičnih višerječnih izraza, možemo zaključiti da utjecaj tehnološke potpore na rezultate i nije toliko značajan te da je veći broj pridjeva na frekvencijskim listama hrvatskog korpusa i njegovih potkorpusa uistinu posljedica razlika između ovih jezika.

Naposljetku, ove rezultate možemo analizirati i iz kvalitativne perspektive. Pretpostavku da discipline koje su na neki način povezane uvelike dijele i terminologiju[68] možemo pokušati ispitati i na području sporta te potražiti zajedničke termine kojih bi trebalo biti više u srodnim sportovima. Logično je očekivati da timski sportovi imaju više zajedničkih specifičnih termina nego na primjer timski i pojedinačni sportovi ili konkretnije vaterpolo i stolni tenis. Dakle, dvoranski hokej i hokej na travi, nogomet i futsal te badminton i stolni tenis bi trebali dijeliti relativno velik broj izraza. Rukomet i košarka su timski sportovi koji se igraju rukama te bi i oni trebali imati sličnu terminologiju, za razliku od gore navedene kombinacije vaterpola i stolnog tenisa.

Najveće sličnosti pokazuju potkorpusi dvoranskog hokeja i hokeja na travi. Oni dijele 85 % ključnih riječi na engleskom te 80 % ključnih riječi na hrvatskom jeziku. Postotak zajedničkih

[68] Kováříková, D. Exploring Corpus Data Using Data Mining: A Project of Automatic Term Recognition. // Corpus Linguistics 2017 Conference, University of Birmingham, 25-28 July 2017 (2017). URL: http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper102.pdf. (13. 5. 2018.)

višerječnih izraza bio je različit u ova dva potkorpusa zbog različitog broja izraza na frekvencijskim popisima, a za engleski dio iznosio je 88.23 % u pravilniku dvoranskog hokeja te 93.75 % u pravilniku hokeja na travi. Na hrvatskom jeziku pravilnik dvoranskog hokeja dijelio je 80 % višerječnih izraza s pravilnikom hokeja na travi što je činilo 60 % specifičnih višerječnih izraza ovog potkorpusa.

Slijede futsal i nogomet koji su dijelili 35 % ključnih riječi i 35 % višerječnih izraza na engleskom te 45 % ključnih riječi i 70 % višerječnih izraza na hrvatskom jeziku.

Engleski pravilnici košarke i rukometa imali su 25 % zajedničkih ključnih riječi i 5 % zajedničkih višerječnih izraza, a hrvatski 40 % ključnih riječi i 20 % višerječnih izraza. Pomalo iznenađujuće, pravilnici badmintona i stolnog tenisa imali su manje zajedničkih termina od potkorpusa košarke i rukometa i to 25 % ključnih riječi i nijedan višerječni izraz na engleskom jeziku.

U hrvatskom dijelu potkorpusa bilo je 10 % zajedničkih ključnih riječi i 20 % višerječnih izraza u pravilniku badmintona te 5 % višerječnih izraza u pravilniku stolnog tenisa. Na kraju, kao što je očekivano, najmanje termina dijelili su pravilnici vaterpola i stolnog tenisa, i to 15 % ključnih riječi i nijedan višerječni izraz na engleskom, a na hrvatskom jeziku nije bilo ni zajedničkih ključnih riječi niti zajedničkih višerječnih izraza. Iako su potkorpusi badmintona i stolnog tenisa pokazali manje sličnosti od očekivanih, možemo zaključiti da je pretpostavka o zajedničkim terminima u sličnim disciplinama točna jer je većina srodnih sportova ipak poprilično terminološki bliska, a sportovi za koje je procijenjeno da dijele najmanje karakteristika uistinu dijele i najmanje termina.

5. Zaključak

U radu je, uz teoretski okvir razvoja analize tekstualnih podataka i korpusne lingvistike prikazan pregled vrsta korpusa i alata te koraci u postupku anotacije korpusa. U dijelu istraživanja, prikazan je konkretan primjer izrade jednog paralelnog korpusa te njegova potencijalna primjena. Prikazane su faze prikupljanja i predobrade korpusa, a kroz mogućnosti obilježavanja korpusa uočena je važnost dosljedne predobrade i anotacije korpusa što omogućuje razne analize te stvaranje paralelnih tekstova koji se koriste u prijevodnim memorijama. Paralelni tekstovi dobiveni stvaranjem prijevodnih memorija korišteni su u stvaranju paralelnog korpusa. Daljnja analiza izvršena je pomoću *Sketch Enginea* čijim je alatima ekstrahirana specifična dvojezična terminologija čitavog korpusa sportskih pravilnika te njegovih pojedinačnih potkorpusa.

Terminološke analize pokazale su da hrvatski i engleski dijelovi korpusa dijele više specifičnih višerječnih izraza nego ključnih riječi što je potvrdilo već primijećenu ulogu takvih izraza u kontekstu specijaliziranih terminologija. Prisutnost pridjeva na popisima ključnih riječi potvrđuju ove zaključke jer oni modificiraju imenice te sami rijetko označavaju neki konkretan koncept, a pogotovo su česti u hrvatskim višerječnim izrazima. Kvalitativnom usporedbom terminologije potkorpusa potvrđena je i pretpostavka o većem broju zajedničkih termina srodnih disciplina koju smo primijetili između sportova koji dijele određene sličnosti. Izrada ovakvih višejezičnih terminoloških baza može poslužiti u daljnjoj primjeni kroz višejezično pretraživanje, strojno prevođenje, izradu terminoloških baza i analizi informacija sadržanih u korpusu.

6. Literatura

Bekavac, B.; Seljan, S.; Simeon, I. Corpus-Based Comparison of Contemporary Croatian, Serbian and Bosnian. // *Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages* / uredili Marko Tadić, Mila Dimitrova-Vulchanova, Svetla Koeva. Zagreb: Croatian Language Technologies Society, 2008. Str. 33-39.

Brkić, M.; Matetić, M.; Seljan, S. Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus. // *Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology ICCSIT 2011*. Chengdu, China, 2011. Str. 1068-1070.

Brkić, M.; Seljan, S.; Bašić Mikulić, B. Using Translation Memory to Speed up Translation Process. // *INFuture 2009 : Digital resources and knowledge sharing* / uredili Stančić, H., Seljan, S., Bawden, D., Lasić-Lazić, J. & Slavić, A. Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009. Str. 353-363.

Corpus Encoding Standard. Document CES 1, Version 1.4, listopad 1996. 1996. URL: http://www.cs.vassar.edu/CES/ (6. 8. 2018.)

EAGLES. Preliminary recommendations on Corpus Typology. 1996. URL: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html. (7. 12. 2017.)

Erjavec. Compilation and Exploitation of Parallel Corpora. // *Journal of Computing and Information Technology*. 11, 2 (2003), str. 93-102.

Feldman, R.; Sanger, J. The Text Mining Handbook. 1st ed. Cambridge; New York: Cambridge University Press, 2007.

Feldman et al. Text mining at the term level. // *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)* / uredili Jan M. Żytkow, Mohamed Quafafou. Nantes: Springer, 1998. Str. 65-73.

Hotho, A.; Nürnberger, A.; Paaß, G. A brief survey of text mining. // *Ldv Forum*. 20, 1 (2005), str. 19-62.

Hunston, S. Corpora in Applied Linguistics. 3rd ed. Cambridge: Cambridge University Press, 2005.

Hunston, S. Corpus Linguistics. // *Encyclopedia of Language & Linguistics* / uredio Keith Brown. Boston: Elsevier, 2006. Str. 234-248.

Jaworski, R.; Seljan, S.; Dunđer, I. Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. // *Human Language Technologies as a Challenge for Computer Science and Linguistics* / uredili Vetulani, Z. & Paroubek, P. Poznan: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2017. Str. 332-336.

Jaworski, R.; Dunđer, I.; Seljan, S. Usability Analysis of the Concordia Tool Applying Novel Concordance Searching. // *Lecture Notes in Computer Science (LNCS)*, (in print).

Kilgarriff, A. Getting to know your corpus. // *Text, Speech and Dialogue* / uredili Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. Berlin; Heidelberg: Springer, 2012. Str. 3-15.

Kilgarriff, A. Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine. // *Proceedings of ASLIB 35th Translating and the Computer Conference* /

London, UK: 2013. URL: https://www.sketchengine.eu//wp-content/uploads/2015/05/Terminology_finding_2013.pdf (11. 3. 2018.)

Kilgarriff, A.; Kosem, I. Corpus tools for lexicographers. // *Electronic Lexicography* / uredile Sylviane Granger, Magali Paquot. Oxford: Oxford University Press, 2012. Str. 31-55.

Kilgarriff, A., Baisa, V., Bušta, J. et al., The Sketch Engine: ten years on. // *Lexicography: Journal of ASIALEX*. 1, 1 (2014), str. 7-36.

Kováříková, D. Exploring Corpus Data Using Data Mining: A Project of Automatic Term Recognition. // *Corpus Linguistics 2017 Conference, University of Birmingham, 25-28 July 2017* (2017). URL: http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper102.pdf. (13. 5. 2018.)

Krieger, M. G.; Finatto, M. J. B. Introdução à Terminologia: teoria & prática. 2nd ed. São Paulo: Contexto, 2004.

Ljubešić, N.; Esplà-Gomis, M.; Ortiz Rojas, S. et al. Croatian-English parallel corpus hrenWaC 2.0. // *Slovenian language resource repository CLARIN.SI* (2016). URL: http://hdl.handle.net/11356/1058. (6. 3. 2018.)

Manning, C.; Schütze, H. Foundations of statistical natural language processing. 2nd ed. Cambridge, MA, London: The MIT Press, 1999.

McEnery, T.; Hardie, A. Corpus linguistics: Method, theory and practice. 1st ed. Cambridge; New York : Cambridge University Press, 2012.

McEnery, A.; Wilson, A. Corpus Linguistics – An Introduction. 2nd ed. Edinburgh: Edinburgh University Press, 2005.

Požgaj Hadži, V.; Tadić, M. Slovensko-hrvatski paralelni korpus. // Izazovi kontrastivne lingvistike (Izzivi kontrastivnega jezikoslovja) / Vesna Požgaj Hadži et al. Ljubljana: Znanstvena založba Filozofske Fakultete Univerze v Ljubljani, 2012. Str. 45-54.

Ramisch, C. E. Multi-word terminology extraction for domain-specific documents. Magistarski rad. Grenoble: Grenoble INP – Institut National Polytechnique de Grenoble, 2009.

Raya, R. XML in localisation: Reuse translations with TM and TMX. XML in localisation. 2005. URL: https://www.ibm.com/developerworks/library/x-localis3/. (17. 1. 2018.)

Schlüter, P. Statistics on the DGT-Translation Memory (DGT-TM). 2018. URL: https://wt-public.emm4u.eu/Resources/DGT-TM_Statistics.pdf (13. 6. 2018.)

Seljan, S.; Gašpar, A.; Pavuna, D. Sentence Alignment as the Basis For Translation Memory Database. // *INFuture 2007 - Digital Information and Heritage* / uredili Seljan, S., Stančić, H. Zagreb: Odsjek za Informacijske znanosti, Filozofski fakultet Zagreb, 2007. Str. 299-311.

Seljan, S.; Pavuna, D. Translation Memory Database in the Translation Process. // *Proceedings of the 17th International Central European Conference on Information and Intelligent Systems IIS* / uredili Aurer, B., Bača, M. Varaždin: FOI, 2006. Str. 327-332.

Seljan, S.; Tadić, M.; Agić, Ž.; Šnajder, J.; Dalbelo Bašić, B.; Osmann, V. Corpus Aligner (CorAl) Evaluation on English - Croatian Parallel Corpora. // *Proceedings of Language Resources and Evaluation (LREC 2010)* / uredili Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M. & Tapias, D. Valletta: European Language Resources Association, 2010. Str. 3481-3484.

Seljan, S.; Gašpar, A. First Steps in Term and Collocation Extraction from English-Croatian Corpus. // *Proceedings of 8th International Conference on Terminology and Artificial Intelligence*. Toulouse, France: 2009. URL: http://ceur-ws.org/Vol-578/paper21.pdf (15. 5. 2018.)

Simeon, I. Paralelni korpusi i višejezični rječnici. // *Filologija*. 38-39, 2002, str. 209-215.

Sinclair, J. Corpus and Text - Basic Principles. // *Developing Linguistic Corpora: a Guide to Good Practice* / uredio Martin Wynne. Oxford: Oxbow Books, 2005. Str. 1-16.

Štambuk, A. Multi-word lexical units in English and Croatian terminology of electronics. // *Studia Romanica et Anglica Zagrabiensia*. 42, 2 (1997), str. 373-390.

Tadić, M. Building the Croatian-English Parallel Corpus. // *Second International Conference on Language Resources and Evaluation LREC2000* / uredili Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis S. Pariz, Atena: ELRA, 2000. Str. 523-530.

Tadić, M. Introduction to Corpus Linguistics. Predavanja na ljetnoj školi Jadertina Summer School in Empirical and Computitional Linguistics (JSSECL). Zadar. 2006. URL: http://hnk.ffzg.hr/txts/mt4JSSECL/JSS2006_Corp-lin.htm (8. 12. 2017.)

Tadić, M. Uporaba XML-a u hrvatskim korpusima. // *Upravljanje informacijama u gospodarstvu i znanosti (CroInfo 2000): zbornik*. Zagreb: Nacionalna i sveučilišna knjižnica; Pliva, 2000. Str. 132-137.

Witten, I. H. Text Mining. // *Practical handbook of internet computing* / uredio M. P. Singh. Boca Raton, FL: Chapman & Hall / CRC Press, 2006. Str. 314-342.

Sažetak

Sažetak:

U radu su prikazane metodologije izgradnje paralelnih korpusa i ekstrakcije specifične terminologije. Prvi dio pruža teorijsku pozadinu te opisuje razvoj korpusa i korpusnih alata, kao i proces pripreme korpusa za analizu. Zatim slijedi opis istraživanja u kojem su prikupljeni sportski pravilnici na engleskom i hrvatskom jeziku sravnjeni te iskorišteni za izgradnju paralelnog korpusa koristeći online korpusni alat *Sketch Engine*. Isti alat korišten je i za ekstrakciju specifične terminologije čiji su rezultati analizirani na kraju rada.

Ključne riječi: paralelni korpus, engleski jezik, hrvatski jezik, specifična terminologija, sport

Abstract:

This paper demonstrates the methodologies of corpus building and terminology extraction. The first part provides a theoretical background and describes the development of corpora and corpus analysis tools, as well as the process of preparing a corpus for analysis. A description of the carried out research follows in which the collected English and Croatian sports rulebooks were aligned and used to build a parallel corpus using *Sketch Engine*, an online corpus tool. The same tool was used for the extraction of specific terminology, the results of which are discussed at the end of the paper.

Keywords: parallel corpus, english language, croatian language, domain-specific terminology, sport