

This work is licensed under a Creative Commons Attribution 4.0 International License.


Ovaj rad dostupan je za upotrebu pod međunarodnom licencom Creative Commons Attribution 4.0.



<https://doi.org/10.31820/f.37.2.3>

Mirjana Borucinsky

OD N-GRAMA DO VIŠERJEČNIH IZRAZA: O ODNOSU KORPUSNE LINGVISTIKE I LEKSIKOGRAMATIKE*

dr. sc. Mirjana Borucinsky, Sveučilište u Rijeci, Pomorski fakultet
mirjana.borucinsky@pfri.uniri.hr  orcid.org/0000-0002-1132-9720

izvorni znanstveni rad

UDK 811.163.42'374'324

811.111'374'324

811.163.42'373'36

811.111'373'36

rukopis primljen: 3. ožujka 2025; prihvaćen za tisak: 29. listopada 2025.

U ovom radu istražuje se odnos leksika i gramatike kroz n-grame ili učestale leksičke spojeve (ULS) u engleskom i hrvatskom jeziku. Kriteriji utvrđivanja ULS-ova iz n-grama uključuju frekvenciju, rekurentnost, semantičku neovisnost te gramatičku strukturu. ULS-ovi su vrsta višerječnih izraza koji predstavljaju nepotpune strukturne jedinice koje nadilaze gramatičke strukture (npr. at the same time, on the basis of, s obzirom na to, u skladu s člankom itd). Analizom podataka iz korpusa engleskog i hrvatskog jezika pomoću računalnojezikoslovnog alata Sketch Engine (Kilgarriff i dr. 2004) promatra se kako su navedene gramatičke strukture instancirane u tekstu. Za teorijsku osnovu odabran je okvir systemske funkcionalne gramatike (engl. Systemic Functional Grammar, SFG, Halliday 1994; Halliday i Matthiessen 2004) zbog primjenjivosti i usredotočenosti na jezik u uporabi. U radu se istražuje i sintaktička sinonimija u vidu dvojnih gramatičkih struktura poput in case of + IMENSKA SKUPINA, in case + SUREČENICA; in the event of + IMENSKA SKUPINA, in the event +

* Ovaj rad financiralo je Sveučilište u Rijeci projektom uniri-iz-25-63.

SUREČENICA i njihovih hrvatskih istovrijednica: u slučaju + IMENSKA SKUPINA i u slučaju + SUREČENICA. Kontrastivnom analizom pokazuje se odnos između gramatičkih i leksičkih struktura te kako je njihova funkcionalnost u tekstu motivirana značenjem. Naglašava se važnost kvantitativne analize podataka dobivenih iz korpusa stavljajući u suodnos korpusnu lingvistiku i leksikogramatiku. Dok postoje brojna istraživanja o ULS-ovima u engleskom jeziku, oni gotovo da i nisu istraživani u hrvatskome. Rad donosi nove podatke o uporabi leksičko-gramatičkih sveza u hrvatskome jeziku. Rezultati istraživanja mogu se primijeniti u poučavanju jezika, kontrastivnim istraživanjima koja povezuju leksikogramatičke obrasce različitih jezika te u unapređenju korpusne metodologije, osobito u pogledu preciznijeg označavanja imenskih skupina i njihovih postmodifikatora.

Ključne riječi: korpusna lingvistika; leksikogramatika; n-gram; učestali leksički spoj (ULS)

1. Uvod

U ovom radu stavljaju se u suodnos korpusna lingvistika i leksikogramatika. Korpusna lingvistika, točnije podaci dobiveni iz korpusa, prikladni su za pojašnjavanje jezičnih pojava iz perspektive funkcionalne lingvistike jer pružaju empirijske podatke iz pomno sastavljenog i odabranog jezičnog uzorka. Stoga se mnoga jezična istraživanja u današnje vrijeme temelje na podacima dobivenima iz korpusa, i to kroz liste riječi, konkordancije, skice riječi, n-grame, ključne riječi i sl. Međutim, valja naglasiti ono što su davne 1998. godine ustanovili Precht i dr., a to je da svrha korpusne lingvistike nije u tome da pruži isključivo kvantitativne podatke, već da istakne ili razjasni jezične pojave uz pomoć analize jezičnih podataka dobivenih iz korpusa, a to vrijedi i danas. Korpusna lingvistika vrijedan je alat za jezikoslovna istraživanja, ali ga valja kombinirati s drugim metodama, npr. kvantitativnom analizom, kontrastivnom analizom i introspekcijom. Vrijednost korpusne lingvistike kao alata očituje se prije svega u velikoj količini podataka koji su dostupni za jezičnu analizu i mogućnosti ponavljanja istraživanja. Korpusno utemeljeno istraživanje ne podrazumijeva, dakle, samo brojanje jezičnih jedinica ili njihovih obilježja, nego uvijek uključuje i kvalitativnu analizu odnosno interpretaciju dobivenih podataka.

U ovom radu istražuju se n-grami ili učestali leksički spojevi (ULS) u engleskom i hrvatskom jeziku kako bismo odgovorili na pitanje koji je odnos između leksika i gramatike i kako su gramatičke strukture oprimjerene (engl. *instantiated*) u tekstu. Za teorijsku osnovu odabran je okvir Sistem-

ske funkcionalne gramatike (SFG, engl. *Systemic Functional Grammar*, SFG, Halliday 1994; Halliday i Matthiessen 2004), te se polazi od pretpostavke da postoji sistematičnost između paradigmatskog slaganja riječi i formalnih svojstava tih riječi. Poseban naglasak stavlja se na pojavu sintaktičke sinonimije, odnosno gramatičkih struktura koje izražavaju slično značenje (npr. *in case of* + *IMENSKA SKUPINA*; *in case* + *SUREČENICA*; *in the event of* + *IMENSKA SKUPINA*; *in the event* + *SUREČENICA*; *u slučaju* + *IMENSKA SKUPINA*; *u slučaju* + *SUREČENICA*) čime se nastoji pokazati kako su gramatičke varijante motivirane značenjem te kako granica između leksičkih i gramatičkih obrazaca nije strogo definirana, već je fluidna.

2. Teorijska pozadina

2.1. Leksikogramatika i korpusna lingvistika

Leksikogramatika je naziv koje je uveden u *Sistemsjoj funkcionalnoj gramatici* (SFG) kako bi se naglasila povezanost (kontinuitet) između leksika i gramatike (Halliday 1994, Matthiessen i sur., 2010). Halliday (1996) leksikogramatiku definira kao spoj gramatike i vokabulara, koji su međusobno neodvojivi i čine jedinstvenu cjelinu. Hasan (1987, 1996) također iznosi pretpostavke o povezanosti leksika i gramatike, navodeći da krajnje točke sustava izbora nisu leksičke opcije nego samo jedna leksička opcija, pri čemu leksik predstavlja gramatičke izbore (Hunston i Francis 2002). Hallidayeva leksikogramatika jedan je od važnijih pojmova za razvoj funkcionalne gramatike, a Halliday jedini među tradicijskim funkcionalistima ističe neodvojivost leksika i gramatike. Takvo shvaćanje također čini okosnicu netradicijskog funkcionalizma, tj. kognitivnih gramatika (Belaj i Tanacković Faletar 2014: 25). Leksikogramatika je konstrukt za izražavanje. Halliday (1996) često za isti taj pojam koristi i skraćeni naziv gramatika. Takva gramatika (tj. leksikogramatika) slojevita je i obuhvaća semantiku, leksikogramatiku i fonologiju (Halliday 1994, usp. također Borucinsky 2015). Svaki sloj paradigmatski je sustav koji se sastoji od niza izbora s ulaznim uvjetom. Izbori su ostvareni kao sintagmatske strukture. Te strukture sačinjavaju funkcije ili snopove funkcija. Pri tome funkcije nisu proizvoljne, već su motivirane mogućnostima koje ostvaruju.

Jedna od prednosti SFG-a u odnosu na druge funkcionalne gramatike kao što je primjerice gramatika uloga i referenci (engl. *Role and Reference Grammar*, RRG, Foley i Van Valin 1984, Van Valin 1993) jest fokusiranost

na jezik u uporabi i primjenjivost. To, dakako, ne znači da je SFG jedini gramatički model koji je usmjeren na jezik u uporabi. SFG kao teorijski okvir odabran je jer se radi o tzv. *top down* pristupu u kojem se leksički elementi ne biraju na početku procesa generiranja, već su rezultat izbora u mreži sustava, a manje delikatni izbori izražavaju se gramatički. SFG daje prednost paradigmatskim odnosima, u smislu da se izbori smatraju generativnim središtem gramatike, tj. strukture koje se realizacijom izvode iz skupova sistemskih značajki procesa.

Leksik i gramatika tradicionalno su smatrani odvojenim, zasebnim pojavama. Stoga se u tradicionalnim modelima jezika previše važnosti nije davalo višerječnim izrazima, koji se sastoje i od leksičkih i gramatičkih jedinica (usp. Stefanowitsch i Gries 2003: 210). Nadalje, korpusna lingvistika prvobitno se nije bavila gramatikom, izuzmemo li koligacije (tj. preferencije i ograničenja između određenih leksičkih jedinica i vrsta riječi koje dolaze uz njih). Međutim, u novije vrijeme se nastoji dati sveobuhvatniji prikaz jezika pa postoje brojne teorije koje pokazuju da su višerječni izrazi važna spona između leksika i gramatike, a sve je češća i uporaba korpusa kao metodologije u istraživanju takvih struktura. Jedan od istaknutijih i recentnijih primjera korpusno-utemeljenog opisa gramatike engleskoga jezika jest *Grammar of Spoken and Written English* (Biber i dr. 2021). No, valja napomenuti da ne postoji korpusno-utemeljen opis gramatike hrvatskoga jezika. Drugim riječima, ne postoji opis gramatike koji se temelji na strukturnim obrascima pronađenima u korpusu.

Naše poimanje korpusne lingvistike kao metodologije naslanja se na Stefanowitscha (2020: 1) koji korpusnu lingvistiku definira kao vrstu lingvističkog propitivanja koje se temelji na podacima dobivenim iz korpusa. To je, dakle, način pronalaženja odgovora na jezikoslovna istraživačka pitanja utemeljen na cjelovitoj i sustavnoj analizi načina na koji su jezične pojave distribuirane u jezičnom korpusu (Stefanowitsch 2020: 55). Korpusna lingvistika i leksikogramatika povezane su na nekoliko načina, a možemo reći i da su komplementarne na način da korpusna lingvistika pruža podatke i alat za leksikogramatičku analizu jezika. Primjerice kroz konkordancije dobivamo uvid u to kako se riječi rabe u kontekstu; kolokacije i skice riječi omogućuju nam uočavanje karakterističnih jezičnih obrazaca; frekvencije pokazuju koje jezične strukture su učestale. N-grami pokazuju koje se riječi pojavljuju u nizu jedna do druge, a kvalitativnom analizom utvrđuje se u kojem se funkcionalnom stilu javljaju pojedini ULS-ovi i koje funkcije imaju. Biber i dr. (2004) primjerice navode da se u znanstvenim

člancima vrlo često upotrebljavaju imenski i prijedložni ULS-ovi. Korpusni podaci mogu poslužiti i za gramatičku analizu, tj. kroz njih dobivamo uvid u to kako su gramatičke strukture oprimjerene u tekstu, te koji je odnos između leksika i gramatike. S druge pak strane, teorijske postavke SFG-a mogu pridonijeti boljem razumijevanju jezika, a time i načinu na koji se korpus označava. Primjerice, Halliday i Matthiessen (2006) pokazali su kako se SFG može primijeniti u računalnojezikoslovnom sustavu za generiranje jezika navodeći za primjere vremenska izvješća i recepte. Borucinsky i Kegalj (2019) prikazale su kako teorijske postavke SFG-a mogu pomoći u interpretaciji imenske skupine, a s obzirom na to da oko sintaktičkog označavanja korpusa još uvijek ne postoji konsenzus, postavke i okvir SFG-a mogu biti jedna od teorija koja će pridonijeti boljem razumijevanju imenskih skupina, a sukladno tome i njihovu označavanju u korpusu.

2.2. Odnos leksika i gramatike

Jedna od pretpostavki SFG-a jest da leksik predstavlja gramatičke izbore. Tucker (1998: 98) navodi da leksički odnosi djeluju i na sintaktičku strukturu u kojoj se pojavljuju ili kojom se realiziraju. Primjerice u engleskoj sintagmi *How X is he?* neoznačeni leksički izbor koji će popuniti mjesto X dio je antonima koji izražavaju usporedbu (npr. *tall rather than short*), čime zaključuje da leksičko-semantički odnosi kao što su sinonimija i antonimija stoje u korelaciji sa strukturalnim razlikama koje ovise o njihovu supojavljanju. Drugim riječima, kako navodi Tucker (ibid.), sustav mreža za leksički ostvarene izbore motiviran je i leksičkim i gramatičkim obrascima supojavljanja.

Prema postavkama SFG-a, u gramatici pojedinoga jezika ostvaruju se tri funkcionalne sastavnice poznatije kao metafunkcije, a to su: ideacijska (engl. *ideational*), interpersonalna (engl. *interpersonal*) i tekstna (engl. *textual*) sastavnica. Ideacijska metafunkcija pokazuje na koji način su ideje i iskustva reprezentirani u jeziku. Interpersonalna metafunkcija pokazuje na koji način smo u interakciji sa svijetom, dok se tekstna metafunkcija odnosi na način na koji su informacije strukturirane. Prva se dijeli u dvije podskupine, a to su iskustvena (engl. *experiential*) i logička (engl. *logical*). „Sve tri funkcije preslikane su i na skupinu i na surečenicu (engl. *clause*¹). No, svaka

¹ Engleski naziv *clause* odnosi se na rečenicu (npr. *The man is reading a book*) i surečenicu (npr. ... *who is reading a book*).

sastavnica u surečenici pridonosi zajedničkoj strukturi tako da se u surečenici sve tri sastavnice kombiniraju u jednu, dok unutar skupine ove tri sastavnice nisu odvojene strukture već svaka jednim svojim dijelom pridonosi zajedničkoj strukturi. Dakle, razlika između skupine i surečenice je u stupnju organizacije“ (Halliday 1994: 179). Sve tri metafunkcije preslikane su dakle i na razinu skupine i na razinu surečenice. Međutim, način na koji se te funkcije ostvaruju razlikuje se s obzirom na stupanj organizacije. U surečenici su tri funkcije strukturno razdvojene, a zajedno tvore jedinstvenu hijerarhijsku cjelinu, dok su u skupini one isprepletene i djeluju unutar jedne, kompaktnije strukture. Drugim riječima, u surečenici se svaka sastavnica (proces, sudionik, okolnost) ostvaruje kao zaseban element, dok unutar skupine svaka sastavnica doprinosi zajedničkoj strukturi samo jednim svojim dijelom. To se može prikazati na sljedećim primjerima. U rečenici *The old man is reading a book* tri metafunkcije jasno su razdvojene. Ideacijska funkcija prikazuje proces čitanja (*is reading*) i sudionike (*the old man, a book*), interpersonalna funkcija izražava stav govornika kroz izjavnu rečenicu, dok tekstna funkcija određuje organizaciju rečenice, pri čemu je *the old man* tema, a *is reading a book* rema. U primjeru skupine *the very old man*, sve tri funkcije djeluju unutar jedne strukture, pa tako ideacijska funkcija opisuje pojam i njegove osobine (*old man*), interpersonalna se očituje u stupnju procjene (*very*), a tekstna u organizaciji elemenata unutar skupine (*the* kao determinator imenske skupine). Razlika je između skupine i surečenice u stupnju organizacije. Dok je surečenica hijerarhijski strukturirana jedinica u kojoj su funkcije jasno odvojene, skupina je kompaktna jedinica u kojoj se funkcije međusobno preklapaju i istodobno ostvaruju. Upravo ta razina organizacije čini osnovu Hallidayeve distinkcije između skupine i (su)rečenice.

Skupina (engl. *group*) je proširena riječ² ili skup riječi (Halliday, 1994). Riječi čine skupinu na način da su glave ili modifikatori skupine. Slično skupinama, i fraze (engl. *phrase*) se nalaze na središnjem položaju između riječi i rečenice. Međutim, za razliku od skupina, fraze nisu logički strukturirane skupine riječi, već umanjene rečenice (v. također Borucinsky 2015). S obzirom na interpersonalnu funkciju, imenska skupina može biti subjekt ili

² Proširena riječ se u SFG-u odnosi na jezičnu jedinicu koja uključuje osnovnu riječ (glavu) i dodatne riječi koji proširuju ili dopunjuju njezino značenje (modifikatore). Proširena riječ može uključivati modifikatore (npr. *very old*) koji su izborni i omogućuju izražavanje specifičnog značenja referenta (*man*).

dopuna, dok prema iskustvenoj strukturi imensku skupinu čini jedna od sastavnica – deiktik, numerativ, epitet, klasifikator, Stvar³, i/ili kvalifikator. Njena logička struktura sastoji se od tri sastavnice, a to su premodifikator, glava i postmodifikator. Prema tekstnoj funkciji, imenska skupina predstavlja i kontekstualizira diskursne referente. Skupine se sastoje od riječi koje su logički strukturirane. Fraze, s druge strane, nemaju logičku strukturu, već samo iskustvenu strukturu. Upravo iz razloga što su skupine spojevi riječi s logičkom strukturom, bilo ih je potrebno uvesti kao zasebnu jedinicu koja zauzima položaj između fraze i surečenice (Matthiessen i dr. 2010: 112). Ovo je jedna od bitnih razlika između SFG-a i ostalih funkcionalnih teorija jezika.

U Hallidayevoj (1994: 212) teoriji razlikuje se prijedložna skupina (engl. *preposition group*, *Pg*) od prijedložne fraze (engl. *prepositional phrase*, *PP*). Prijedložna skupina odnosi se na nizove riječi koji funkcioniraju kao jedan prijedlog, te poput drugih skupina mogu, uz glavu, imati i modifikatore (npr. *right behind* pri čemu *right* funkcionira kao modifikator, a *behind* kao glava). Prijedložna fraza, s druge pak strane, ne smatra se skupinom jer ne predstavlja proširenu glavu koja ima modifikatore, nego djeluje kao surečenica i ponaša se kao predikator, a sastoji se od prijedloga kojemu je imenska skupina dopuna (npr. *in the house*, pri čemu je imenska skupina *the house* dopuna prijedlogu *in*), usp. Halliday i Matthiessen (2004).

2.3. N-grami i učestali leksički spojevi (ULS)

N-grami predstavljaju nizove koji mogu biti: 1. riječi, npr. *one of the, a lot of, a type of, itd; da će se, kao što je*, 2. znakovi, ili tzv. *character grams / char-grams*, npr. *-ing, -ness; -ment; -ost, -telj, -lac, -ica*, te 3. glasovi, npr. *kæt*. Računalnojezikoslovni alati kao što je *Sketch Engine* (Kilgarriff i dr. 2004) izlistavaju n-grame u obliku riječi, a ne slova. Za n-grame koji nisu riječi već slova, postoji alat koji je razvio Fletcher (2012). Takav alat može biti koristan za uočavanje stranih riječi u nekome jeziku, no za hrvatski ne postoji sličan alat. U ovom istraživanju isključivo se bavimo n-gramima koji predstavljaju riječi, točnije nizove riječi. N-grami su, dakle, vrsta višerječnih izraza (engl. *multi-word expressions*), koji su u anglističkoj tradiciji poznati i kao učestali leksički spojevi (ULS, engl. *lexical bundles*, Biber i dr. 1999), *clu-*

³ Prema Hallidayevoj leksikogramatici središnji dio iskustvene strukture imenske skupine jest *Stvar* (engl. *Thing*) pa se stoga označava i velikim početnim slovom.

sters (Scott 2008), *chains* (Stubbs, 2001; Stubbs i Barth 2003), *recurrent sequences* (De Cock 2004), *recurrent word combinations* (Altenberg 1998), i koji imaju važnu ulogu u oblikovanju diskursa i razlikovanja funkcionalnih stilova. Generiranje ULS-ova iz korpusa ovisi o tri čimbenika: duljini, frekvenciji i disperziji. Rasprave se vode o tome koja je granica za određivanje ULS-a, a najčešće se promatraju n-grami koji se sastoje od 3 ili 4 riječi (tj. trigrama ili kvadrigrama) koji se pojavljuju barem 10 puta u korpusu (Biber i dr. 1999, Cortes 2004, usp. također Borucinsky 2022) i koji predstavljaju proširene kolokacije. Hyland (2008) je utvrdio da kvadrigama daju bolji uvid u strukturu i funkcije u odnosu na trigrame. Osim učestalosti pojavljivanja riječi, za određivanje je li neki N-gram ujedno i ULS važna je i disperzija, stoga se uvodi i mjera poznata kao srednja reducirana frekvencija (SRF, engl. *average reduced frequency*, Hlaváčová 2006). Može se, primjerice, dogoditi da se neka riječ često pojavljuje u jednome tekstu, ali da nije (jednako) zastupljena u svim tekstovima. Stoga je uvedena ova dodatna mjera, a to je položaj na kojem se riječ nalazi u korpusu. Uzmimo da se riječ r1 pojavljuje 5 puta u jednomilijunskome korpusu, a riječ r2 također pet puta u jednomilijunskome korpusu. Obje dakle imaju istu apsolutnu i relativnu frekvenciju, no prva riječ r1 pojavljuje se samo u jednome tekstu, dok je druga r2 ravnomjerno raspoređena u svim tekstovima. Druga riječ prema tome ime veći SRF i bit će značajnija pri donošenju odluke o njezinu uvrštavanju u rječnik, enciklopediju i sl., kako navodi Hlaváčová (2006). Ovom se mjerom ukida neravnomjernost riječi u korpusu koja može negativno utjecati na rezultat (usp. također Borucinsky 2023).

O ULS-ovima se dosta pisalo, posebice u engleskom jeziku (npr. Biber i dr. 1999, Biber, Conrad i Cortes 2004, Cortes 2004, Biber i Barbieri 2007, Hyland 2008, i dr., Biber i dr. 2021), dok u hrvatskom gotovo da i nema radova o ovoj temi. Jedno istraživanje proveli su Borucinsky i Pritchard (2022) u kojemu se bave ULS-ovima u specijaliziranom korpusu. Römer (2009) navodi da postoje i druge primjene pristupa ULS-ovima osim tradicionalne žanrovske analize. Naime, s obzirom na to da ULS-ovi predstavljaju strukturne jedinice koje nadilaze gramatičke strukture (Biber, Conrad i Cortes 2004), smatramo da nam mogu dati nove uvide u odnos između leksika i gramatike, a na toj se pretpostavci temelji ovo istraživanje.

ULS-ovi su jednostavno nizovi riječi koje se često supojavljaju u prirodnom govoru ili pismu. U većini slučajeva ti izrazi nisu strukturne jedinice i govornici ih ne prepoznaju kao frazeme ili druge fiksne leksičke sveze.

Jedno od osnovnih obilježja ULS-ova jest njihova gramatička struktura. Iako ULS-ovi u pravilu ne predstavljaju potpunu strukturnu jedinicu, pokazuje se da se mogu svrstati u nekoliko osnovnih strukturnih tipova kao što su: imenički (npr. *a wide variety of*; *činjenica je da je*), glagolski (npr. *will have to be*; *najavio je da će*), prijedložni (*at the same time*; *s obzirom na to*), i rečenični (npr. *it is important to*; *očekuje se da će*), dok prema funkciji mogu biti usmjereni na istraživanje, (engl. *research-oriented*, npr. *the size of the*; *način na koji je*), tekst (*text-oriented*, npr. *in case of an*; *u slučaju koji je*) ili sudionika (engl. *participant-oriented*, npr. *should be able to*; *kako nam je rekao*). Razlika između glagolskih i rečeničnih ULS-ova ogleda se u njihovoj gramatičkoj strukturi i funkciji. Glagolski ULS-ovi sastavljeni su od glagolskih skupina i najčešće služe za izražavanje radnje, stanja ili događaja, a u pravilu uključuju pomoćne ili modalne glagole (npr. *will have to be*). S druge strane, rečenični ULS-ovi pokrivaju složenije strukture koje čine sintaktičke jedinice više razine (npr. *očekuje se da će – it is expected that*). Za razliku od engleskoga, u kojemu se razlikovanje glagolskih i rečeničnih ULS-ova temelji na jasnim sintaktičkim obilježjima, u hrvatskome je takva klasifikacija znatno složenija. Razlog tomu leži u morfo-sintaktičkoj složenosti jezika, slobodnom redu riječi i mogućnosti ispuštanja subjekta. Posljedično, mnoge sveze koje na površinskoj razini izgledaju kao glagolske zapravo funkcioniraju kao cijele rečenice. Na primjer, ULS *očekuje se da će* sadrži finitni glagol i veznik koji uvodi zavisnu surečenicu, pa se stoga klasificira kao rečenični ULS. S druge strane, ULS-ovi poput *može se vidjeti* ili *je temeljeno na* predstavljaju glagolske ULS-ove jer čine samo dio predikata i ne uvode zavisnu surečenicu. U ovom istraživanju bavimo se imeničkim i prijedložnim ULS-ovima, prije svega njihovom strukturom koju dovodimo u svezi s metafunkcijama. Pod imeničke ULS-ove svrstat ćemo i izraze poput *one of the most*, *the best of the* (usp. također Biber i dr. 2021: 1008).

Iako se ULS-ovi često koriste u akademskim tekstovima, njihovi strukturni obrasci znatno se razlikuju od onih u drugim funkcionalnim stilovima. Tako su primjerice u svakodnevnom govoru češći rečenični (npr. *I don't know why*) i glagolski (npr. *are you talking about*) ULS-ovi, dok su u akademskim tekstovima češći prijedložni (npr. *in the next chapter*) i imenički (npr. *the nature of the*) ULS-ovi.

ULS-ovi najčešće prelaze iz jedne gramatičke strukture u drugu. Primjeri koje Biber i dr. (2021: 992) među ostalima navode za engleski jezik su *in the case of* i *at the same time*. Na primjeru *in the case of* može se ilustrirati tvrdnja da ULS-ovi prelaze iz jedne gramatičke strukture u drugu. Dakle, struktura *in*

the case of sastoji se od prijedloga *in* i imenske skupine *the case of*. Međutim, ta imenska skupina nije potpuna jer se sastoji od determinatora *the*, glave *case* i samo je djelomično uveden postmodifikator *of*, sa sljedećim mogućim izborima: *in the case of a*, *in the case of the*, *in the case of any*, *in the case of such*, *in the case of double genitive*, itd. U slučaju kada ULS predstavlja potpunu gramatičku strukturu, u akademskim tekstovima je to najčešće prijedložna fraza (npr. *in the first place*, *in the next chapter*). Biber i dr. (2021: 988) navode da se 15 % ULS-ova u razgovornom jeziku smatra potpunim strukturnim jedinicama, dok je taj postotak u akademskim tekstovima i manji, gdje svega 5 % ULS-ova predstavlja potpune strukturne jedinice. To je stoga što ULS-ovi u ovim vrstama teksta uvode referente (npr. *the nature of the* može se proširiti na *the nature of the problem*) ili imaju tekstnu funkciju (npr. *in the next chapter*), odnosno služe za oblikovanje diskursa. Valja naglasiti da ULS-ovi predstavljaju sustavne obrasce uporabe koji se mogu prepoznati pomoću opsežnih korpusnih istraživanja. Upravo nam korpusi omogućuju uvid u dio jezika koji se često zanemaruje, a to jest da gramatika nije samo proučavanje apstraktnih kategorija i struktura, već specifičnih funkcija koje riječi imaju unutar kategorija i struktura (Biber i dr. 2021: 982).

3. Cilj rada i metodologija

Cilj je ovoga rada istražiti odnos između leksika i gramatike u engleskom i hrvatskom jeziku kroz analizu imeničkih i prijedložnih ULS-ova. Poseban naglasak stavlja se na pojavu sintaktičke sinonimije, odnosno gramatičkih struktura koje izražavaju slično značenje (npr. *in case of* + *IMENSKA SKUPINA*; *in case* + *SUREČENICA*; *in the event of* + *IMENSKA SKUPINA*; *in the event* + *SUREČENICA*; *u slučaju* + *IMENSKA SKUPINA*; *u slučaju* + *SUREČENICA*) čime se nastoji pokazati kako su gramatičke varijante motivirane značenjem te da je granica između leksičkih i gramatičkih obrazaca fluidna. Pri tome polazimo od korpusne analize koja je nužna za razumijevanje (leksiko)gramatičkih obrazaca. U radu se ne opisuju samo gramatički aspekti ULS-ova u dvama jezicima, već ih se promatra u kontekstu, tj. uzima se u obzir kako se govornici koriste ULS-ovima te se sagledavaju njihove varijacije u odnosu na druge (slične) strukture kao i faktori koji utječu na uporabu određenog ULS-a.

U radu nastojimo odgovoriti na sljedeće pitanje: Koje podatke o svezi leksika i gramatike možemo otkriti pomoću korpusno-utemeljenog istraživanja proučavajući ULS-ove? Polazište analize jest engleski jezik, a kon-

trastivnom se analizom nastoji dati uvid u ULS-ove za hrvatski jezik, u kojemu su ovi višerječni izrazi gotovo neistraženi.

Resursi korišteni u ovom istraživanju su mrežni korpusi *enTenTen21* (61 585 997 113 pojavnica) za engleski jezik te *MaCoCu Cro 20-21* (2 728 381 677 pojavnica) za hrvatski jezik. Mrežni korpusi odabrani su zbog količine jezičnih podataka kako bi se dobio što veći jezični uzorak. ULS-ovi koji se analiziraju i interpretiraju u ovome radu dobiveni su pomoću funkcionalnosti *N-gram* u računalnojezikoslovnom alatu *Sketch Engine* (Kilgarriff i dr. 2004).

Način sastavljanja i označavanja korpusa (Pan, Reppen i Biber 2020) te odabir kriterija kao što su duljina, frekvencija i disperzija bitno utječu na popis ULS-ova u nekom jeziku ili njegovom funkcionalnom stilu, posebice u kontrastivnim istraživanjima. Međutim, s obzirom na to da cilj ovoga rada nije dobivanje cjelokupnog popisa ULS-ova u engleskom i hrvatskom jeziku, već promatramo njihovu leksikogramatičku strukturu i način na koji su instancirani u tekstu, smatramo da su ti čimbenici manje ograničavajući za ovo istraživanje.

Na temelju popisa n-grama dobivenih iz navedenih resursa, odredili smo četveročlane ULS-ove na temelju sljedećih kriterija (v. također Tablicu 1):

- I. minimalna čestota pojavljivanja – najmanje 10 pojava u korpusu (usp. Biber i dr. 1999);
- II. srednja reducirana frekvencija (SRF, Hlaváčová 2006)
- III. kriteriji koji su poslužili za uklanjanje šuma iz podataka dobivenih korpusnom pretragom (usp. Salazar 2014) prikazani u Tablici 1.

Tablica 1. Kriteriji na temelju kojih je uklonjen šum iz korpusnih podataka

Kriterij	Primjer
vlastita imena, toponimi, imena institucija	<i>in the United Kingdom; Akademija znanosti i umjetnosti</i>
n-grami koji sadrže 3+ leksičke riječi tj. ključne riječi	<i>the hundred shares index; energija iz obnovljivih izvora</i>
šum	<i>mm mm mm mm</i>

Kao što je navedeno, u ovome radu razmatraju se imenički i prijedložni ULS-ovi u engleskom i hrvatskom jeziku (npr. *at the same time, at the end of, a member of the; bez obzira na to, iz godine u godinu* itd.).

Također su od posebnog interesa dvojne gramatičke strukture u oba jezika kao što su:

- (1) a. *in case of* + IMENSKA SKUPINA (npr. *in case of legal dispute*), *in case* + SUREČENICA (npr. *in cases that could cause conflict*)
 b. *in the event of* + IMENSKA SKUPINA (npr. *in the event of an emergency*), *in the event* + SUREČENICA (npr. *in the event that no personal representative has been appointed*)
- (2) *u slučaju* + IMENSKA SKUPINA (npr. *u slučaju pogibljii*) i *u slučaju* + SUREČENICA (npr. *u slučaju da članica ne odgovori na upozorenje; u slučaju kada sud donese rješenje*).

Ove strukture generirane su iz korpusa pomoću jezika za postavljanje upita korpusu (engl. *corpus query language*, CQL) te pregledavane u konkordancijama uz pomoć funkcionalnosti *Slučajni uzorak*.⁴

4. Rezultati i rasprava

Popis n-grama koji možemo dobiti iz računalnojezikoslovnoga alata prikazuje frekvencijsku listu niza pojavnica. Pitanje koje se postavlja jest je li svaki n-gram ujedno i ULS. Kao što svaki n-gram nije kolokacija, tako ni svaki n-gram nije ULS. Kriteriji utvrđivanja koji nizovi riječi se smatraju ULS-ovima su frekvencija (veća čestota indicira da je n-gram ULS), rekurzivnost (to su obrasci koji su karakteristični za određeni jezični varijetet), semantička neovisnost (ULS funkcionira kao jedna cjelina), te gramatička struktura (iako ULS-ovi ne predstavljaju potpune strukturne jedinice ipak se mogu svrstati u nekoliko osnovnih strukturnih kategorija kao što su imenička, glagolska, rečenična, prijedložna).

Na temelju primjera iz korpusa engleskoga jezika prikazanih u Tablicama 2 i 3 uočavamo da se ULS-ovi sastoje od gramatičkih i leksičkih riječi. ULS-ovi prikazani u Tablicama 2 i 3 ručno su pročišćeni prema kriterijima navedenima u prethodnom poglavlju te su isključena vlastita imena i oni

⁴ [word="in"] [word="the"]? [lemma="case"] [word="of"]
 [word="in"] [word="the"]? [lemma="case"] [word="that|which|when|where"] [tag="Vmp.*"]
 [word="u"] [lemma="slučaj"] [tag="Nc.*"]
 [word="u"] [lemma="slučaj"] [word="koji|ako|da|kada"]
 [word="u"] [lemma="slučaj"] [tag="Vmp.*"]

višerječni izrazi koji sadrže tri ili više leksičke riječi. Naime takve višerječne izraze obično definiramo ključnim riječima. Popis n-grama koji dobijemo iz računalnojezikoslovnog alata katkada izlista i ključne riječi (npr. *the hundred shares index; energija iz obnovljivih izvora*). Međutim, način na koji se generiraju n-grami i ključne riječi potpuno je drugačiji. N-grami se generiraju na temelju frekvencije pojavnica, dok se ključne riječi dobivaju na temelju ključnosti, tj. omjera koji pokazuje da je tražena riječ ili višerječni izraz češći u fokusnom korpusu u odnosu na zadani, referentni korpus (usp. Borucinsky 2023). Ključne riječi mogu biti jednorječne (engl. *single word lexical unit*) ili višerječne (engl. *multi-word lexical unit*, MWLU). Potojnje, osim uvjeta da su češće u fokusnom korpusu u odnosu na referentni korpus, moraju zadovoljiti i uvjet da njihova struktura odgovara strukturi naziva koji su postavljeni za neki jezik, odnosno tzv. *gramatici za crpljenje naziva* (engl. *term grammar*). Traženje ključnih riječi u korpusu od posebne je važnosti za specijalizirane korpuse. Za razliku od ULS-ova, ključne riječi ne nadilaze strukturu skupine, a najčešće su sastavljene od punoznačnih riječi (uglavnom imenica, nešto rjeđe glagola) koje nose glavni dio značenja. Za razliku od ključnih riječi, ULS-ovi nadilaze razinu skupine. ULS-ovi se razlikuju i od kolokacija, a smatra se da predstavljaju neku vrstu proširenih kolokacija, te se i sami mogu dalje proširivati. Primjerice iz kvadrigrama *at the end of* može se dobiti n-gram od 5, 6 ili 7 riječi (*at the end of the, at the end of the academic, at the end of the academic year*).

U nastavku su prikazani najfrekventniji imenički i prijedložni ULS-ovi dobiveni iz korpusa *enTenTen21* i *MaCoCu* za engleski i hrvatski jezik.

Tablica 2. *Popis najfrekventnijih imeničkih ULS-ova iz korpusa enTenTen21 s apsolutnom frekvencijom (AF) i srednjom reduciranom frekvencijom (SRF)*

ULS	AF	SRF
<i>the end of the</i>	64853	11312
<i>one of the most</i>	52740	920
<i>a member of the</i>	45100	787
<i>the rest of the</i>	35104	612
<i>a wide range of</i>	32562	568
<i>the beginning of the</i>	17866	312
<i>a wide variety of</i>	16217	283
<i>the top of the</i>	14975	262
<i>the fact that the</i>	14286	250

ULS	AF	SRF
<i>a part of the</i>	14071	246
<i>one of the best</i>	13309	233
<i>some of the most</i>	12994	227
<i>the start of the</i>	12794	223
<i>a result of the</i>	12664	221
<i>the name of the</i>	11971	209
<i>the development of the</i>	11054	193
<i>the time of the</i>	10752	188
<i>the course of the</i>	10660	186
<i>the history of the</i>	10647	186
<i>the first time in</i>	10611	185

Tablica 3. Popis najfrekventnijih prijedložnih ULS-ova iz korpusa *enTenTen21* s apsolutnom frekvencijom (AF) i srednjom reduciranom frekvencijom (SRF)

ULS	AF	SRF
<i>at the end of</i>	59366	1036
<i>at the same time</i>	55659	972
<i>for the first time</i>	35688	623
<i>on the other hand</i>	29690	519
<i>in the form of</i>	23590	412
<i>on the basis of</i>	22825	398
<i>in the context of</i>	21472	375
<i>at the time of</i>	19078	333
<i>in the field of</i>	17940	313
<i>in the middle of</i>	16681	291
<i>by the end of</i>	16561	289
<i>in the case of</i>	16201	283
<i>in a way that</i>	16076	281
<i>in addition to the</i>	15978	279
<i>in a variety of</i>	15325	268
<i>in accordance with the</i>	13969	244
<i>on behalf of the</i>	13879	242
<i>at the beginning of</i>	13269	232
<i>for more information on</i>	12349	216
<i>in the face of</i>	12322	215

U analizi učestalosti ULS-ova korištena su dva pokazatelja: apsolutna frekvencija (AF) i srednja reducirana frekvencija (SRF). Apsolutna frekvencija označava ukupan broj pojavljivanja određene jedinice u korpusu i daje osnovni uvid u njezinu čestoću. Međutim, sama po sebi ne govori dovoljno o tipičnosti izraza u jeziku jer visoka vrijednost može biti rezultat pojavljivanja u manjem broju dokumenata ili određenom funkcionalnom stilu. Stoga se kao pouzdaniji kriterij koristi reducirana frekvencija (Hlavačová 2006), odnosno njezina prosječna vrijednost (SRF), koja uzima u obzir distribuciju izraza kroz korpus. Visok SRF upućuje na to da je jedinica ne samo česta, već i ravnomjerno rasprostranjena u različitim tekstovima i funkcionalnim stilovima, što je čini tipičnom za jezik u cjelini. Suprotno tome, jedinice s visokom apsolutnom frekvencijom, a niskom srednjom reduciranom frekvencijom najčešće su ograničene na određene funkcionalne stilove.

Primjerice, u engleskom jeziku ULS *the end of the* bilježi AF od 64 853 i vrlo visok SRF (11 312), što pokazuje da je riječ o izrazito tipičnom obrascu prisutnom u različitim funkcionalnim stilovima. Slično tome, u hrvatskom jeziku izraz *obzirom na to da* (AF = 33 009, SRF = 10 142) potvrđuje se kao važan element u oblikovanju diskursa. Nasuprot tome, izraz *udaljenost i kilometraža između* u hrvatskom korpusu ima visoku apsolutnu frekvenciju (AF = 24 923), ali nisku srednju reduciranu frekvenciju (SRF = 171), što znači da je, iako učestao, u većoj mjeri ograničen na specifične kontekste, primjerice tehničke ili administrativne tekstove. Ovakvi kontrasti potvrđuju da kombinacija AF-a i SRF-a omogućuje preciznije razlikovanje specifičnih obrazaca.

Rezultati za hrvatski jezik pokazuju da je s obzirom na prirodu jezika teže odrediti ULS-ove. ULS-ovi za hrvatski jezik sadržavali su više leksičkih riječi (npr. *na konferenciji za novinare, s teškoćama u razvoju*) pa bi se mogli smatrati ključnim riječima.

Tablica 4. *Popis najfrekventnijih imeničkih ULS-ova iz korpusa MaCoCu s apsolutnom frekvencijom (AF) i srednjom reduciranom frekvencijom (SRF)*

ULS	AF	SRF
<i>obzirom na to da</i>	33009	10142
<i>udaljenost i kilometraža između</i>	24923	171
<i>adresa je zaštićena od</i>	8042	1880
<i>ova e-mail adresa je</i>	7910	1835
<i>posto u odnosu na</i>	6516	1533
<i>za osobe s invaliditetom</i>	6268	1627
<i>podatci su ažurirani u</i>	4336	100
<i>način na koji se</i>	4234	1451
<i>odnosu na isto razdoblje</i>	3589	873
<i>notne sveske i na</i>	3587	8
<i>sveske i na umu</i>	3587	8
<i>stan se sastoji od</i>	3580	348
<i>godine s početkom u</i>	3242	911
<i>sve od sebe da</i>	3074	1013
<i>način na koji je</i>	2675	924
<i>nešto nije u redu</i>	2675	906
<i>kuća se sastoji od</i>	2583	353
<i>stroj se nalazi u</i>	2569	8
<i>taj stroj se nalazi</i>	2568	7
<i>druge strojeve iz kategorije</i>	2568	7

Tablica 5. *Popis najfrekventnijih prijedložnih ULS-ova iz korpusa MaCoCu s apsolutnom frekvencijom (AF) i srednjom reduciranom frekvencijom (SRF)*

ULS	AF	SRF
<i>s obzirom na to</i>	19533	6100
<i>s obzirom da je</i>	19208	6530
<i>na to da je</i>	14175	4672
<i>bez obzira na to</i>	12062	4084
<i>iz dana u dan</i>	9759	3348
<i>iz godine u godinu</i>	9725	3262
<i>na način da se</i>	8743	2846
<i>na bilo koji način</i>	7850	2682

ULS	AF	SRF
<i>s teškoćama u razvoju</i>	7823	1806
<i>na to da se</i>	7751	2613
<i>u skladu s člankom</i>	7528	635
<i>o izmjenama i dopunama</i>	6092	1515
<i>za sve one koji</i>	5838	2010
<i>o tome kako je</i>	5592	1920
<i>na to da su</i>	5062	1704
<i>na najbolji mogući način</i>	5008	1706
<i>s vremena na vrijeme</i>	4894	1664
<i>među kojima je i</i>	4823	1595
<i>s obzirom da se</i>	4681	1620
<i>u bilo kojem trenutku</i>	4628	1477

Analiza najfrekventnijih imeničkih i prijedložnih ULS-ova u engleskom i hrvatskom jeziku otkriva bitne razlike u njihovoj distribuciji i funkcijama koje ostvaraju u tekstu. U engleskom jeziku najčešći imenički ULS-ovi (npr. *the end of the, one of the most, a member of the*) jasno pokazuju tendenciju prema strukturiranju diskursa putem čvrsto ustaljenih okvira koji označuju pripadnost, kategorizaciju ili evaluaciju. Slično tome, prijedložni ULS-ovi (npr. *at the end of, at the same time, on the other hand*) u pravilu funkcioniraju kao vremenski, kontrastivni ili kontekstualni markeri, što je u skladu s prepoznatim obrascima u engleskom akademskom i javnom registru. Suprotno tome, rezultati za hrvatski jezik ukazuju na drukčiju prirodu ULS-ova. Među najfrekventnijim imeničkim ULS-ovima ističu se obrasci uvjetovani formalno-administrativnim i tehničkim kontekstima (*udaljenost i kilometražna između, adresa je zaštićena od, za osobe s invaliditetom*), kao i izraze vezane uz svakodnevni govor (*nešto nije u redu, sve od sebe da*). U prijedložnim ULS-ovima dominiraju koncesivni i uvodni izrazi (*s obzirom na to, bez obzira na to, na način da se*), kao i idiomatske vremenske izraze (*iz dana u dan, iz godine u godinu, s vremena na vrijeme*).

U usporedbi s engleskim, gdje su česte imenske skupine uvedene prijedlogom *of* (npr. *the beginning of the, in the context of the*), hrvatski podaci pokazuju veću sklonost prema sintaktički složenijim, ali leksički varijabilnim obrascima (*u skladu s člankom, o izmjenama i dopunama, za sve one koji*). Valja, naravno, uzeti u obzir prirodu mrežnih korpusa koji nam omogućuju uvid u jezik kojemu se ne može pristupiti izravno (Borucinsky 2023), ali

imaju puno nedostataka. Za analizu funkcionalnih stilova trebalo bi promatrati uravnoteženi korpus poput BNC-a, međutim za hrvatski ne postoji istovjetan korpus. No, u ovom se radu ne bavimo žanrovskom analizom, već pomoću ULS-ova nastojimo razumjeti suodnos leksika i gramatike.

Dosadašnja analiza pokazala je frekvencijske obrasce imeničkih i prijedložnih ULS-ova u engleskome i hrvatskome jeziku. Međutim, čestoća ne daje uvid u njihove funkcije. Stoga u nastavku izdvajamo primjere konkordancija iz gore navedenih resursa za engleski i hrvatski jezik te ih promatramo kroz prizmu metafunkcija Hallidayeve systemske funkcionalne gramatike.

(3) *At the end of the module, students will be expected to submit their final report in journal article format.*

Prijedložni ULS *at the end of* + *IMENSKA SKUPINA* prikazan u primjeru (3) vremensko značenje izražava sažetim imeničkim izrazom. Funkcionalno ostvaruje tekstnu metafunkciju jer strukturira diskurs u vremenu, ali i ideacijsku metafunkciju s obzirom na to pozicionira događaje.

(4) *On the other hand, critics claim that these programs hamper the fragile job recovery in the U.S.*

Prijedložni ULS *on the other hand* prikazan u primjeru (4) gotovo je u potpunosti gramatikaliziran kao diskursni marker. Iako je formalno prijedložna fraza, funkcionalno služi da bi se istaknuo kontrast i time ostvaruje tekstnu metafunkciju.

(5) *S obzirom na to da je istraživanje započelo prije izbijanja pandemije koronavirusa, u studiju nisu bili uključeni pacijenti s covidom-19.*

Prijedložni ULS *s obzirom na to* uvodi zavisnu surečenicu. Funkcionalno označava kauzalno-ekspanzivne odnose i tipično se javlja u administrativnim i argumentacijskim tekstovima.

(6) *Iz godine u godinu povećava se incidencija T1D kako u svijetu tako i u nas, a bolest se sve češće javlja u mlađoj životnoj dobi.*

Prijedložni ULS prikazan u primjeru (6) izražava temporalnu iteraciju i tipičan je za akademske i publicističke tekstove. Funkcionalno signalizira kontinuitet i postupnu promjenu. Poput izraza *on the other hand*, pokazuje kako višerječni izrazi s vremenom poprimaju diskurzivne funkcije koje nadilaze doslovno značenje.

U nastavku rada fokusiramo se na dvojne gramatičke strukture, ili sintaktičku sinonimiju u smislu kako se definira u Silić i Pranjeković (2007: 209): „Ta mogućnost da se isto ili slično značenje, odnosno funkcija izrazi različitim tipovima gramatičke veze među tagmemima, tj. da se sintagme mi tipa *plavooka djevojka* mogu zamjenjivati sintagmemima tipa *djevojka plavih očiju*, jedno je od glavnih izvorišta gramatičke sinonimije na razini sintagmema”.

Slijede kvantitativna i kvalitativna kontrastivna analiza kroz sinonimne leksičke i gramatičke strukture kao što su *in (the) case(s) of + IMENSKA SKUPINA/ SUREČENICA*; *in (the) event(s) of + IMENSKA SKUPINA/ SUREČENICA* i *u slučaju + IMENSKA SKUPINA/SUREČENICA*.

Iz Tablice 6 razvidno je da je struktura *in (the) case(s) of + IMENSKA SKUPINA* češća od strukture *in (the) case(s) + SUREČENICA*. U slučaju sinonimnog para *in the event* također je češća postmodifikacija imenskom skupinom u odnosu na postmodifikaciju surečenicom (Tablica 7). Sinonimija postaje očita i kroz kontrastivnu analizu, tj. kada primjere prevedemo na hrvatski jezik u kojemu postoji samo jedna istovrijednica za lekseme *case* i *event*. Postmodifikacija imenskom skupinom također je češća od postmodifikacije surečenicom u hrvatskome jeziku (Tablica 8).

Tablica 6. Učestalost pojavljivanja ULS-a *in (the) case(s) of + Ng* i *in (the) case(s) + SUREČENICA*

	<i>in (the) case(s) of + Ng</i>	<i>in (the) case(s) + SUREČENICA</i>
<i>enTenTen21</i>	21,3	0,53

Tablica 7. Učestalost pojavljivanja ULS-a *in (the) event(s) of + Ng*, *in (the) event(s) + SUREČENICA*

	<i>in (the) event(s) of + Ng</i>	<i>in (the) event(s) + SUREČENICA</i>
<i>enTenTen21</i>	4,93	1,58

Tablica 8. Učestalost pojavljivanja skupine *u slučaju + Ng*, *u slučaju + SUREČENICA*

	<i>u slučaju + Ng</i>	<i>u slučaju + SUREČENICA</i>
<i>MaCoCu Cro</i>	37,24	19,72

Korpusni podaci prikazani do sada daju nam uvid u to koje strukture su učestalije, što nam govori o tome koje strukture korisnici jezika preferiraju, tj. kojim jezičnim sredstvima oblikuju izričaj. U nastavku slijedi kvantitativna analiza.

Biber i dr. (2021: 630) navode da je prijedlog *of* vrlo produktivan kada se radi o postmodifikaciji prijedložnom skupinom, i to zbog toga što se njime ostvaruje niz funkcija, te se iskazuje prislan semantički odnos između glave skupine i imenske skupine koja ju modificira.

Primjeri navedeni pod (7), (8) i (9) odabrani su za ilustraciju iz korpusa pomoću funkcionalnosti *Slučajni uzorak*, dok su primjeri (10-19) osmišljeni na temelju primjera iz korpusa, tj. u ovom slučaju koristili smo se tzv. korpusno oprimjerenim pristupom (engl. *corpus-illustrated approach*). U primjeru (9) uz hrvatske primjere navode se i engleske istovrijednice u svrhu kontrastivne analize.

(7)

<i>in (the) case(s) of + Ng</i>	<i>in (the) cases + SUREČENICA</i>
<i>in case of ambiguity</i>	<i>in case where a solution will serve a large and varied number of customers</i>
<i>in case of dispute</i>	<i>in cases that could cause deaths</i>
<i>in case of fraud</i>	<i>in case when goods and services are supplied</i>
<i>in case of injury</i>	<i>in the case which is subject of this appeal</i>
<i>in the cases of interest</i>	<i>in the case which has been referred to</i>

(8)

<i>in (the) event(s) of + Ng</i>	<i>in (the) event(s) + SUREČENICA</i>
<i>in the event of failure</i>	<i>in the event that the primary communications line is not operating</i>
<i>in the event of an emergency</i>	<i>in the events which led to the deaths</i>
<i>in the event of continued dispute</i>	<i>in the event that an internal review is sought</i>
<i>in the event of an injury</i>	<i>in the event when the customer has any doubts</i>
<i>in the event of leakage</i>	<i>in the event that an internal review is sought</i>

(9)

u slučaju + Ng

u slučaju pobjede

in case of victory

u slučaju potrebe

in case of need

u slučaju prijetnji

in case of threat

*u slučaju privremene
spriječenosti*

*in the event of temporary
inability*

u slučaju opiranja

*in case of resistance/ in case
Ø there is resistance*

u slučaju + SUREČENICA

u slučaju da se dokaže odgovornost

in the event that liability is proven

*u slučajevima propisanim ovim
Zakonom*

in cases stipulated in this Act

u slučaju kada sud prihvati prijedlog

*in cases when the Court accepts a
proposal*

u slučaju koji navodi čitatelj

*in the case Ø referred to by the
reader*

u slučaju ako to ne bude moguće

in case Ø this is not possible

Analiza višerječnih izraza prikazanih u primjerima (7), (8) i (9) pokazala je kako leksički izbori aktiviraju različite mreže gramatičkih izbora te da se engleski i hrvatski razlikuju u stupnju semantičke diferencijacije i formalnoj distribuciji. Primjeri (11–19) otkrivaju na koji se način isti leksički izbor (*case/sučaj*) ostvaruje kroz različite postmodifikacijske obrasce, uključujući prijedložne fraze, zavisne surečenice i sl. Na taj se način može precizno pratiti kako se teorijske postavke leksikogramatike, posebice koncept mijenjanja položaja (engl. *rankshifting*, Halliday 2002), realizira u konkretnoj jezičnoj upotrebi.

U engleskom jeziku izrazi *in (the) case of* i *in (the) event of* ilustriraju leksikogramatičku varijaciju u smislu Hallidayeve (1994) definicije leksikogramatike kao jedinstvenog spoja leksičkih i gramatičkih izbora. Oba se izraza mogu realizirati imenskom skupinom (*in case of dispute, in the event of failure*) ili zavisnom surečenicom (*in cases when goods are supplied, in the event that liability is proven*), pri čemu izbor između *case* i *event* nije samo leksički, nego i funkcionalno motiviran. Dok se imenica *case* tipično povezuje s pravnim i administrativnim funkcionalnim stilom u kojem označava

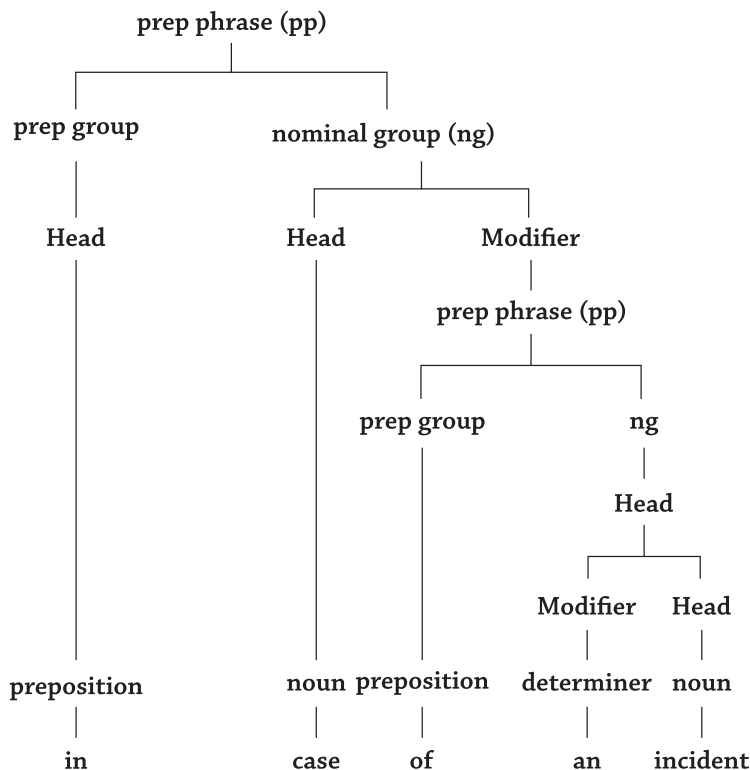
pojedinačan ili apstraktan slučaj, imenica *event* prevladava u tehničkim i sigurnosnim kontekstima, gdje upućuje na događaj ili njegov ishod. Suprotno tome, hrvatski jezik sustavno koristi izraz *u slučaju*, koji se jednako pojavljuje uz imensku skupinu (*u slučaju prijetnji*, *u slučaju pobjede*) i uz zavisnu surečenicu (*u slučaju da se dokaže odgovornost*, *u slučaju kada sud prihvati prijedlog*), čime se neutralizira distinkcija između engleskih izraza *case* i *event*.

Iz primjera (7) i (8) razvidno je da postoje pogodbeni i nepogodbeni načini uporabe struktura *in case* i *in the event*. To postaje vidljivo kada strukturu skupine/fraze *in case of injury* proširimo u strukturu surečenice *in case if there is an injury*, pri čemu se u ovom primjeru izražava pogodbeni način, dok u primjeru *in the case of discussion*, koji se može proširiti na *in the case which is under discussion*, nije izražena pogodba. Nepogodbeni načini uporabe imaju sintaktičku funkciju atributa koji proširuje imenicu. Drugim riječima, oni funkcioniraju kao postmodifikatori koji specificiraju okolnosti povezane s imenicom kojoj se pridružuju. U primjerima poput *in case of ambiguity* i *in the event of failure*, skupine su neovisni atributi koji proširuju značenje imenice te nisu povezani s glavnom rečenicom. Takvo razlikovanje između pogodbenih i nepogodbenih načina moguće je objasniti i u okviru tradicionalne hrvatske gramatike (Silić i Pranjković, 2007), prema kojoj atribut ima funkciju dodatka imenici, može biti izražen prijedložnim skupinama (npr. *u slučaju nesreće*, *bez opasnosti od kvara*), te se ne veže sintaktički na glagolsku jezgru rečenice. Pogodbeni način, naprotiv, ostvaruje se zavisnim rečenicama pogodbenog značenja koje uvode veznici poput *ako*, *kad* ili *u slučaju da*, što odgovara engleskim istovrijednicama *if* i *in case (that)*.

Usporedi li se ova analiza s prethodnim uvidima o učestalosti i funkcionalnosti ULS-ova u engleskom jeziku, može se zaključiti da skupine poput *in case of* i *in the event of* predstavljaju primjer frazeologizacije i gramatikalizacije unutar prijedložnih skupina. One se pojavljuju visokom frekvencijom u korpusu i postupno prelaze iz slobodnih sintaktičkih kombinacija u ustaljene leksičko-gramatičke obrasce koji prenose kompleksna značenja. U tom procesu dolazi do semantičke kondenzacije pri čemu višerječne jedinice preuzimaju funkciju jedne gramatičke kategorije, što je u skladu s Hallidayevim (1994) shvaćanjem jezika kao sustava izbora kroz koji se značenje gramatikalizira. Nepogodbene varijante poput *in the case of X* funkcioniraju kao nominalne leksičke cjeline (usp. Silić i Pranjković 2007), dok pogodbeni oblici (*in case if...*, *in the event that...*) ostvaruju punu propozicijsku strukturu koja povezuje dvije situacije. Takva transformacija od sintaktičke prema frazeološkoj stabilnosti pokazuje da su ULS-ovi ključni nositelji funkcionalnih značenja u jeziku.

Ono što je zanimljivo jest da govornici jezika vrlo spretno biraju odgovarajuću strukturu za dani kontekst, dok analiza i interpretacija istih predstavlja izazov za analitičara. Kao što je ilustrirano primjerom (10), unutar jedne prijedložne fraze (*in case of an incident*) dolazi do mijenjanja položaja (engl. *rankshifting*, Halliday 2002: 43) pri čemu se jedinice više razine spuštaju na nižu razinu i zauzimaju ulogu jedinica niže razine. U primjeru (10) prijedložna fraza *of an incident* spušta se na razinu riječi te modificira glavu *case* u skupini *case of an incident*, koja se potom također spušta na nižu razinu i modificira prijedlog *in* u prijedložnoj frazi *in case of an incident*. Za razliku od glave skupine, premodifikatori i postmodifikatori nisu obavezni, no u ovom je slučaju prijedložna fraza obavezna jer je rečenica **Notify us immediately in case* negramatična. I surečenice se također spuštaju na nižu razinu te djeluju kao modifikatori glave (npr. *when an incident arises* modificira glavu *case* fraze *in case when an incident arises*).

(10)



Struktura prikazana u primjeru (10) može se pojednostavljeno prikazati na sljedeći način:

(11)

<i>in</i>	<i>case</i>	[<i>of an incident</i>]
	glava	Postmodifikator
	Stvar	kvalifikator
	imenica	prijedložna fraza

Kao što je razvidno iz primjera (11), te kao što Halliday (1994: 213) navodi, *of* nije predikator već strukturna oznaka u imenskoj skupini, slično kao što je to riječ *to* u glagolskoj skupini. Nadalje, osim fraza koje mogu postmodificirati glavu imenske skupine, postmodifikatori mogu biti i surečenice. U primjerima (12–15) navedeni su mogući oblici postmodifikatora unutar imenske skupine u kojoj je glava imenica *case*. U primjeru (12) glavu skupine pobliže određuje odnosna ili atributna surečenica uvedena odnosnom zamjenicom *that*, dok u primjeru (13) imamo istu skupinu u kojoj glavu također postmodificira odnosna surečenica bez odnosne zamjenice, u primjerima (14) i (15) imamo odnosnu surečenicu s prilozima mjesta ili vremena:

(12) *in case* [[**that** an incident arises]]

(13) *in case* [[\emptyset an incident arises]]

(14) *in case* [[**when** an incident arises]]

(15) *in case* [[**where** an incident arises]]

Prema Hallidayu (1994), samo su restriktivne odnosne surečenice (*cases arising in incidents – cases which arise in incidents*) umetnute i imaju ulogu postmodifikatora ili kvalifikatora unutar imenske skupine, dok nerestriktivne odnosne surečenice (npr. *just in case that an incident arises*) nisu umetnute već su strukture hipotakse, tj. smatraju se zasebnim jedinicama rečenične strukture.

Usporedba struktura **in case arising in an incident* i *in cases arising in incidents*, te *?in a case arising in an incident* i *in the case arising in an incident*⁵

⁵ Znak * označava negramatičnu strukturu, dok ? označava gramatične strukture, koje se, međutim, ne preferiraju u uporabi.

pokazuje da je sustav mreže za leksički realizirane opcije motiviran obrascima supojavljivanja. Sustav mreža za leksički ostvarene izbore motiviran je i leksičkim i gramatičkim obrascima supojavljivanja. Primjerice, kada glavu *case* postmodificira nerestriktivna nefinitna surečenica (*arising in incidents*), ona mora sadržavati (određeni) član (deiktik).

Višerječni izraz *in case* u hrvatskome ima sljedeće istovrijednice:

(16) *u slučaju [nesreće]*

(17) *u slučaju [[da dođe do nesreće]]*

(18) *u slučaju [[kada dođe do nesreće]]*

(19) *u slučaju [[ako dođe do nesreće]]*

U primjerima (16–19) prikazuje se prijevodna istovrijednica *u slučaju* koju može postmodificirati imenska skupina (*u slučaju nesreće*) ili zavisna surečenica uvedena veznikom *da*, *kada* ili *ako* (*u slučaju da/kada/ako dođe do nesreće*). I ovdje dolazi do spuštanja gramatičkih struktura na niži položaj, budući da se surečenica spušta na razinu imenske skupine i preuzima funkciju kvalifikatora glave. Međutim, za razliku od engleskog, hrvatski se oslanja na bogatiji sustav veznika, a ne na odnosne zamjenice ili priložne izraze, čime se ostvaruje šira funkcionalna fleksibilnost. Sam leksički izbor (*slučaj*) time aktivira drugačiju mrežu gramatičkih izbora u odnosu na englesku imenicu *case*, što potvrđuje Hallidayevu tezu da leksik predstavlja gramatičke izbore.

Valja istaknuti da oba jezika potvrđuju osnovno polazište leksikogramatike – da su leksički i gramatički sloj nerazdvojivi i da se ostvaruju kroz paradigmatske izbore unutar mreže sustava. U engleskom se leksički izbor *case* aktivira kroz postmodifikaciju odnosnim surečenicama, dok se u hrvatskom isti izbor aktivira kroz mrežu vezničkih zavisnih surečenica. Govornici oba jezika intuitivno biraju odgovarajuću strukturu prema kontekstu (pravni, administrativni, tehnički), dok analitičaru ostaje izazov identificirati granicu između leksičkog i gramatičkog sloja, budući da se neprestano preklapaju. Ovi primjeri potvrđuju da je sustav mreža leksički realiziranih izbora motiviran i leksičkim i gramatičkim obrascima supojavljivanja (usp. Tucker 1998), a razlikovni potencijal leži u sredstvima koja su specifična za neki jezik. Navedeni primjeri pokazuju i da se značenje ne ostvaruje kroz pojedinačne riječi, već kroz obrasce leksičko-gramatičkih izbora koji odražavaju društvene, komunikacijske i funkcionalne potrebe jezika. Unutar tog

sustava, ULS-ovi funkcioniraju kao gramatikalizirani obrasci konceptualizacije iskustva, u kojima se leksički i gramatički elementi spajaju u stabilne strukture koje prenose značenje s visokom predvidljivošću. Analizirani primjeri potvrđuju i da su ULS-ovi više od frekvencijske pojave. Naime, pomoću njih ostvaruju se različite metafunkcije jezika i otkrivaju tipološke razlike među jezicima, u ovom slučaju engleskom i hrvatskom. Engleski se oslanja na nominalne i prijedložne obrasce koji kondenziraju značenje, dok hrvatski češće koristi zavisne rečenice kako bi izrazio iste odnose. Te razlike potvrđuju Hallidayevu tezu da su jezični sustavi različiti po načinu na koji organiziraju metafunkcionalne resurse za ostvarivanje značenja u specifičnim društvenim i komunikacijskim kontekstima (Halliday i Matthiessen 2004).

Istraživanje je također pokazalo da su leksički i gramatički obrasci u dinamičkoj interakciji te da izbor jezičnih struktura nije slučajan, već motiviran značenjem te jezičnim i izvanjezičnim čimbenicima. Naposljetku, pokazalo se da je varijacija u promatranim strukturama vrlo sistematična, što znači govornici jezika biraju leksiko-gramatičke obrasce ovisno o jezičnim i izvanjezičnim čimbenicima (v. također Biber i dr. 2021: 5).

5. Zaključak

Ovim radom nastojalo se prikazati komplementarnost leksikogramatike i korpusne lingvistike. U radu se također nastojalo pokazati da su leksički i gramatički obrasci u interakciji na sistematičan način, te da je gramatička varijacija inherentno obilježje jezika. Na temelju rezultata dobivenih iz korpusa, određeni su kriteriji za utvrđivanje učestalih leksičkih spojeva, koji u hrvatskome u odnosu na engleski sadrže više leksičkih riječi, a s obzirom na prirodu jezika, tj. padežne oblike teže ih je odrediti. Rezultati istraživanja pokazuju da postoje razlike u broju i vrsti ULS-ova koji se rabe u engleskome i hrvatskome. Znanstveni doprinos rada očituje se na trima razinama. Kao prvo, rad donosi nove podatke o leksičko-gramatičkim svezama u hrvatskome jeziku. Kao drugo, potvrđuje se da je struktura tih sveza motivirana značenjem. Naposljetku, ova saznanja mogu biti korisna u primjeni korpusne lingvistike za analizu jezičnih obrazaca u različitim jezicima. Rezultati istraživanja mogu se primijeniti u poučavanju jezika, pri čemu podizanje svijesti o učestalim obrascima može doprinijeti razvoju jezične kompetencije, u kontrastivnim istraživanjima koja povezuju leksikogramatičke obrasce različitih jezika te u unapređenju korpusne metodologije, osobito u pogledu preciznog označavanja imenskih skupina i njihovih postmodifikatora.

Pitanje gramatičke sinonimije vrlo je zanimljivo. Na temelju korpusnih podataka jesmo dobili neke preferencije i obrasce u kojima se dvojne strukture preferiraju, no valjalo bi ih pomnije istražiti. Stoga se u budućim istraživanjima planira provesti kolostrukcijska analiza nad strukturama kao što su *in case / in the event*.

Literatura

- Altenberg, Bengt (1998) „On the Phraseology of Spoken English: The Evidence of Current Word Combinations“, ur. Cowie, Anthony P., *Phraseology: Theory, Analysis and Applications*, str. 101–122, Oxford, Oxford University Press.
- Belaj, Branimir; Tanacković Faletar, Goran (2014) *Kognitivna gramatika hrvatskoga jezika. Knjiga prva: Imenska sintagma i sintaksa padeža*, Zagreb, Disput.
- Biber, Douglas; Barbieri, Federica (2007) „Lexical Bundles in University Spoken and Written Registers“, *English for Specific Purposes* 26, str. 263–286.
- Biber, Douglas; Johansson, Stig; Leech, Geoffrey, N.; Conrad, Susan; Finegan, Edward (2021) *Grammar of Spoken and Written English*, Amsterdam/Philadelphia, John Benjamins Publishing Company.
- Biber, Douglas; Johansson, Stig; Leech, Geoffrey, N.; Conrad, Susan; Finegan, Edward (1999) *Longman Grammar of Spoken and Written English*, Harlow, Pearson.
- Biber, Douglas; Conrad, Susan; Cortes, Viviana (2004) „If You Look at...: Lexical Bundles in University Teaching and Textbooks“, *Applied Linguistics*, 25, str. 371–405.
- Borucinsky, Mirjana (2023) *Primjena metoda korpusne lingvistike u jezikoslovnim istraživanjima*, Rijeka, Sveučilište u Rijeci, Pomorski fakultet.
- Borucinsky, Mirjana; Pritchard, Boris (2022) Lexical bundles in maritime texts, *ICAME Journal*, 46, 1, str. 5–17.
- Borucinsky, Mirjana; Kegalj, Jana (2019) Syntactic ambiguity of (complex) nominal groups in technical English, *International journal of English Studies*, 19, 2, str. 83–102.
- Borucinsky, Mirjana (2015) *Modifikacija u imenskim skupinama u engleskome i hrvatskome jeziku*, doktorska disertacija, Zagreb, Sveučilište u Zagrebu, Filozofski fakultet.

- Cortes, Viviana (2004) „Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology“, *English for Specific Purposes*, 23, str. 397–423.
- De Cock, Sylvie (2004) „Preferred Sequences of Words in NS and NNS Speech“, *Belgian Journal of English Language and Literature*, 2, str. 225–246.
- Fletcher, William H. (2012) *Phrases in English*. URL: <http://phrasesinenglish.org/>.
- Foley, William A.; Van Valin, Robert D. Jr. (1984) *Functional Syntax and Universal Grammar*, Cambridge, Cambridge University Press.
- Halliday, Michael Alexander Kirkwood (1994) *An Introduction to Functional Grammar*, 2. izd., London, Edward Arnold.
- Halliday, Michael Alexander Kirkwood (1996) „On Grammar and Grammatics“, u: Hasan, Ruqaiya; Cloran, Carmel Cloran; Butt, David (ur.), *Functional Descriptions*, str. 1–38. Amsterdam, John Benjamins. doi:10.1075/cilt.121
- Halliday, Michael Alexander Kirkwood; Matthiessen, Christian Mathias Ingemar Martin (2004) *An Introduction to Functional Grammar*, 3. izd., London, Edward Arnold.
- Halliday, Michael Alexander Kirkwood; Matthiessen, Christian Mathias Ingemar Martin (2006) *Construing Experience through Meaning: A Language-based Approach to Cognition*, Bloomsbury Academic.
- Hasan, Ruqaiya (1987) „The Grammarian’s Dream: Lexis as Most Delicate Grammar“, u: Halliday, Michael Alexander Kirkwood; Fawcett, Robin P. (ur.), *New Developments in Systemic Linguistics*, vol. 1, str. 184–211, London, Pinter.
- Hasan, Ruqaiya (1996) „Semantic Networks: A Tool for the Analysis of Meaning“, u: Cloran, Carmel; Butt, David Butt; Williams, Geoffrey (ur.), *Ways of Saying: Ways of Meaning*, str. 104–131, London, Cassel.
- Hlaváčová, Jaroslava (2006) „New Approach to Frequency Dictionaries - Czech Example“, u: Calzolari, Nicoletta; Choukri, Khalid; Gangemi, Aldo; Maegaard, Bente; Mariani, Joseph; Odijk, Jan; Tapias, Daniel (ur.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, str. 373–378. Genoa, Computer Science Bibliography.

- Hunston, Susan; Francis, Gill (2002) *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*, Amsterdam, John Benjamins.
- Hyland, Ken (2008) „As Can Be Seen: Lexical Bundles and Disciplinary Variation“, *English for Specific Purposes*, 27, str. 4–21.
- Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel; Tugwell, David (2004) Itri-04-08 *The Sketch Engine. Information Technology*. URL: https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2004.pdf.
- Matthiessen, Christian Mathias Ingemar Martin; Teruya, Kazuhiro; Lam, Man Kit Marvin (2010) *Key Terms in Systemic Functional Linguistics*, London, Continuum.
- Pan, Fan; Reppen, Randi; Biber, Douglas (2020) „Methodological issues in contrastive lexical bundle research. The influence of corpus design on bundle identification.“ *International Journal of Corpus Linguistics*, 25, 2, str. 216–230, doi: 10.1075/ijcl.19063.pan
- Precht, Knut; Biber, Douglas; Conrad, Susan; Reppen, Randi (1998) „Corpus Linguistics: Investigating Language Structure and Use“, *TESOL Quarterly*, 32(4), doi:10.2307/3588017
- Römer, Ute (2009) „The Inseparability of Lexis and Grammar“, *Corpus Linguistics Perspectives, Annual Review of Cognitive Linguistics*, 7, str. 141–163. doi:10.1075/arcl.7.06rom
- Salazar, Danica (ur.) (2014) *Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching*, Amsterdam, John Benjamins. doi:10.1075/scl.65
- Scott, Michael (2008) *WordSmith Tools Version 5*. Liverpool, Lexical Analysis Software.
- Silić, Josip; Pranjković, Ivo (2007) *Gramatika hrvatskoga jezika - Za gimnazije i visoka učilišta*, Zagreb, Školska knjiga.
- Stefanowitsch, Anatol; Gries, Stephan Th. (2003) „Collostructions: Investigating the interaction of words and constructions“. *International Journal of Corpus Linguistics* 8(2), str. 209-243. doi:10.1075/ijcl.8.2.03ste
- Stefanowitsch, Anatol (2020) *Corpus Linguistics: A Guide to the Methodology*. Berlin, Language Science Press.
- Stubbs, Michael; Barth, Isabel (2003) „Using Recurrent Phrases as Text-Type Discriminators: A Quantitative Method and Some Findings“, *Functions of Language* 10(1), str. 61–104. doi:10.1075/fol.10.1.04stu

- Stubbs, Michael (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, Blackwell.
- Tucker, Gordon H. (1998) *The Lexicogrammar of Adjectives: A Systemic Functional Approach to Lexis*, London, Cassel.
- Van Valin, Robert D. Jr. (1992) „A Synopsis of Role and Reference Grammar“, u: Van Valin, Robert D. Jr. (ur.), *Advances in Role and Reference Grammar*, str. 1–164. Amsterdam, John Benjamins.

Popis kratica: **AF** – apsolutna frekvencija; **CQL** – jezik za postavljanje upita korpusu (engl. *corpus query language*); **Ng** – imenska skupina (engl. *nominal group*); **Pg** – prijedložna skupina (engl. *preposition group*); **PP** – prijedložna fraza (engl. *prepositional phrase*); **RRG** – Gramatika uloga i referenci (engl. *Role and Reference Grammar*); **SFG** – Sistemska funkcionalna gramatika (engl. *Systemic Functional Grammar*); **SkE** – *Sketch Engine*, **SRF** – srednja reducirana frekvencija; **ULS** – učestali leksički spoj (engl. *lexical bundle*).

SUMMARY

Mirjana Borucinsky

EXPLORING THE INTERPLAY OF LEXIS AND GRAMMAR THROUGH N-GRAMS IN ENGLISH AND CROATIAN

This paper looks at the intricate relationship between lexis and grammar by studying N-grams in English and Croatian. N-grams, i.e. sequences of N words extracted from corpora, are also referred to as lexical bundles (Biber et al. 1999), clusters (Scott 2008), chains (Stubbs 2001; Stubbs & Barth, 2003), recurrent sequences (De Cock 2004), and recurrent word combinations (Altenberg 1998). Lexical bundles (e.g. as well as, in order to, in case of, in terms of) are formulaic sequences that provide “the building blocks of coherent discourse” (Hyland 2008: 6). The lexical bundle approach has focused mainly on register and text-type differences (Hyland 2008), trying to answer the question whether there is a discipline specific lexical repertoire or a core vocabulary. One such attempt is described in Borucinsky and Pritchard (2022). However, since lexical bundles present incomplete structural units that cross grammatical structures (Biber, Conrad and Cortes 2004), there is room for further applications of this approach, both in linguistics and in language teaching (e.g. in cross-linguistic analyses), as suggested by Römer (2009). One such application could be the understanding of the interplay between lexis and grammar. Hence, this paper aims at answering the following question: What can we uncover about the interplay of lexis and grammar through corpus-based research by studying MWEs, and in particular lexical bundles? The starting point of the analysis is the English language,

while a contrastive analysis aims to provide insight into MWEs in Croatian, where these multi-word expressions are almost entirely unexplored. We use Sketch Engine (Kilgarriff et al. 2004) to extract N-grams from English and Croatian corpora (enTenTen21 and MaCoCu), and to identify four-word lexical bundles based on the following criteria: (1) a minimum cut-off frequency of at least ten occurrences (cf. Biber et al. 1999); (2) the average reduced frequency (Hlaváčová 2006) and (3) exclusion criteria (cf. Salazar 2014) to eliminate noise resulting from corpus processing. We focus only on NP- and PP-based lexical bundles, and study them cross-linguistically. Furthermore, special attention is paid to grammatical variation or syntactic synonyms such as in case of + Ng (e.g. in case of legal dispute), in case + CLAUSE (e.g. in cases that could cause conflict); in the event of + Ng (e.g. in the event of an emergency), in the event + CLAUSE (e.g. in the event that no personal representative has been appointed). These forms show that grammatical structures are motivated by meaning and that the boundary between lexis and grammar is fluid. This is even further brought to light through contrastive analysis, as illustrated by the following examples: u slučaju + Ng (e.g. u slučaju pogibljivi lit. 'in case of distress') and u slučaju + CLAUSE (e.g. u slučaju da članica ne odgovori na upozorenje lit. 'in case that the member does not respond to the warning'; u slučaju kada sud donese rješenje 'in case when the court passes a decision'). The contribution of this paper is threefold. Empirically, it provides new data on the use of lexico-grammatical constructions in Croatian. Theoretically, it confirms that the structure of these constructions is motivated by meaning and that it interacts with grammar in different ways in the two languages. Methodologically, the study demonstrates the effectiveness of combining frequency-based and functional analysis in contrastive lexico-grammar. The results may have multiple applications: in language teaching, where raising awareness of frequent patterns can contribute to the development of language competence, in contrastive research that links lexico-grammatical patterns across different languages, and in advancing corpus methodology, particularly with regard to more precise annotation of nominal groups and their postmodifiers.

Keywords: *lexicogrammar; corpus linguistics; N-grams; lexical bundles; English; Croatian*