SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

AK. GOD. 2016./2017.

Selmir Hasanić

Alati i tražilice za pretraživanje korpusa

diplomski rad

Mentorica:

dr. sc. Nives Mikelić Preradović, izv. prof.

Zagreb, 2017.

Sadržaj

1. Uvod

Dati odgovor na pitanje „što je jezik?“ nije jednostavan i lak poduhvat. Mnoštvo je jezičara i jezičnih „škola misli“ pokušalo dati odgovor i diskusije su proizašle iz nekoliko postulata; neki tvrde da je jezik ljudsko svojstvo s kojim se rađamo, neki tvrde da je mehanizam kojeg u potpunosti moramo naučiti, a neki, pak, da je to fenomen sačinjen od ukupnog broja jezičnih tvorevina unutar određene jezične zajednice. U svakom slučaju, jezik je živa tvorevina, pojava s mnoštvom slojeva i dimenzija kojih većina njegovih korisnika nije ni svjesna. Učenje jezika provodi se pomoću raznoraznih metoda i pomagala. Rječnici su jezična pomagala koja nam daju uvid u leksik pojedinog jezika. Više ih je vrsta i svakako su jedna od polazišnih točaka u proučavanju jezika kao cjeline. No obrazovni učinak rječnika bilo koje vrste ima određeni doseg – i upravo je ovdje vrlo važna uloga korpusa. Kao jezična pomagala, korpusi su od iznimne važnosti, budući da pružaju kontekstualan okvir uporabe pojedinih riječi, fraza, pa i rečenica. Samim time, korpusi nam daju tzv. *grassroots* pristup jezičnim informacijama i korisniku pružaju bolji uvid u središnje i tipične aspekte pojedinog jezika, te ono najznačajnije u učenju i proučavanju jezika - značenje. U današnje vrijeme korpusna lingvistika pronašla je svoje mjesto u mnogim lingvističkim područjima kao neizostavan alat u istraživanju.

Od 1960-ih godina, odnosno od početka razvoja korpusne lingvistike i tehnologije do danas, napravljeni su nevjerojatni pomaci u smislu strukture i upravljanja korpusima, te su istaknute mogućnosti i prednosti korištenja korpusima u svrhu jezičnog obrazovanja i istraživanja. Ovaj rad služi kao prikaz nekih od dostupnih arhitektura, odnosno alata i tražilica za pretraživanje korpusa. U prvom dijelu rada prikazana je teorijska pozadina korpusa, njihova općenita definicija, podjela i povijesni razvoj. Zatim je dan detaljan pregled odabranih alata i tražilica za pretraživanje korpusa, kao i primjeri korpusa koji se njima koriste. U ovom dijelu navedeni su i neki od hrvatskih korpusa i alata kojima služe kao operativna podloga. U trećem dijelu rada opisano je anketno istraživanje o uporabi korpusa u obrazovne ili istraživačke svrhe od strane korisnika aktivnih u nekima od jezičnih područja.

2. Korpusna lingvistika

2.1. Što je korpusna lingvistika?

Prije prelaska na samu definiciju korpusa i pobliže opisivanje njegovih vrsta i funkcija u današnjem jezikoslovlju, valja najprije izraditi jasni pregled lingvističke grane koja se bavi njegovom izradom i uporabom – korpusne lingvistike.

Jedna od najjednostavnijih definicija korpusne lingvistike jest da je korpusna lingvistika proučavanje jezika na temelju primjera iz uporabe jezika u „stvarnom životu“[1], odnosno, da je njezina zadaća proučavati jezik na temelju diskursa.[2] U knjizi „Corpus Linguistics: An Introduction“, lingvist Wolfgang Teubert predstavio je temeljnu svrhu i povijesni razvoj korpusne lingvistike, kao i njezinu jukstapoziciju s nekim drugima lingvističkim smjerovima kao što je „chomskijanska lingvistika“, odnosno generativna gramatika, o čemu će više govora biti u nastavku rada. Svrha korpusne lingvistike jest stjecanje dubljeg uvida u jezik i njegovu uporabu istraživanjem korpusa govornog ili pisanog jezika, gdje se, u načelu, istražuju tekstovi različitih opsega, a ne izolirane rečenice.[3] Iz ovoga proizlazi kako je fokus interesa korpusne lingvistike, za razliku od standardne i chomskijanske lingvistike, upravo značenje koje pojedine jezične jedinice nose. Važno je napomenuti da je značenje zapravo, poput jezika, društveni fenomen i temelj je rasprave članova diskursne zajednice.[4] Budući da činjenice postaju činjenice tek onda kada uđu u diskurs, upravo je na članovima diskursne zajednice da u diskurs uvedu ono što smatraju činjenicama, i tako ostvaruju svojevrsno značenje.[5] Dakle, izvan diskursa može postojati mnoštvo činjenica, no one za nas ne postoje ako nisu u diskursu, odnosno prirodnom jeziku, i on je jedini izvor s kojim se možemo konzultirati želimo li saznati značenje neke riječi, kolokacije ili druge konstrukcije.

Korištenje korpusom svakako se može smatrati empirijskim pristupom budući da su početna točka, kao u svim vrstama znanstvenog istraživanja, autentični podaci.[6] Postupci kojima se prikupljeni podaci opisuju su induktivni budući da se svaki teoretski zaključak ili pretpostavka dobivaju od aktualnih primjera.[7] U svojem znanstvenom članku „Corpus Linguistics – The Basic Form of Linguistic Analysis“, Nikola Dobrić naveo je lingvistička

[1] McEnery; Wilson, 2001.
[2] Teubert; Čermáková, 2007.
[3] Klobučar Srbić, 2008.
[4] Teubert; Čermáková, 2007.
[5] Ibid.
[6] Tognini-Bonelli, 2001.
[7] Ibid.

područja kojima empirijski pristup temeljen na korpusu (engl. *corpus-based approach*) može uvelike koristiti:

a) *Leksikografija:* omogućava pregled lingvističkih i ne-lingvističkih povezanosti određenih riječi i uvelike može pomoći leksikografima u slaganju rječnika;
b) Sociolingvistika: korpusne tehnike omogućuju istraživanje dijalekata i registara koje se prethodno nisu mogle provoditi;
c) Analiza diskursa: osigurava uzorak jezika koji je dovoljno velik za donošenje zadovoljavajućih generalizacija o karakteristikama jezika koje sežu izvan njegove strukture;
d) Morfologija: rezultati istraživanja provedenog primjenom ovog pristupa mogu nam mnogo toga reći o čestoti, distribuciji i funkciji varijanti riječi;
e) Fonologija: može nam dati kvalitetan uvid u različite načine fonetske distribucije;
f) Semantika: ne postoji ijedan drugi pristup čijom se primjenom mogu dobiti tako cjeloviti dokazi koji pružaju uvid u značenje riječi;
g) Sintaksa: istraživanje strukture jezika na ovaj način može polučiti empirijskim dokazima o načinu na koji strukturiramo rečenice i izražavamo se pomoću jezika;
h) Komparativna i kontrastivna lingvistika: može nam prikazati sličnosti i razlike u uporabi različitih jezika, uz uvjet da postoje paralelni korpusi;
i) Edukativna lingvistika: ovako velika istraživanja mogu pomoći pri osmišljavanju učinkovitih edukativnih materijala i aktivnosti za učenike određenih jezika.[8]

Dobrić je ovdje pristup temeljen na korpusu iskoristio kao koncept koji obuhvaća sve aktivnosti vezane uz korištenje korpusom u lingvističke svrhe. Međutim, talijanska lingvistkinja Elena Tognini-Bonelli uvela je jednu posebnu, i ključnu, distinkciju između dvije vrste pristupa uporabe korpusa za jezičnu analizu. Jedan je, već ranije spomenuti, pristup temeljen na korpusu (engl. *corpus-based approach*), a kojeg je Tognini-Bonelli, pak, definirala kao pristup unutar kojeg se korpusom koristi uglavnom kako bi se izložila, ispitala, opovrgnula ili potvrdila teorija.[9] Drugi je pristup opisala kao pristup vođen korpusom (engl. *corpus-driven approach*), gdje se korpusom služi za pronalaženje lingvističkih otkrića i formiranje teorija pomoću metodologije korpusne lingvistike, koja se zatim intelektualno obrađuju i pretvaraju u rezultate.[10] Teubert navodi da bi korpusna lingvistika svakako trebala upotrebljavati pristup

[8] Dobrić, 2009.
[9] Tognini-Bonelli, 2001.
[10] Teubert; Čermáková, 2007.

vođen korpusom kako bi mogla komplementirati standardnu lingvistiku, a ne samo biti njezin dodatak.[11]

2.2. Klasifikacija korpusne lingvistike

Ovom se distinkcijom dotičemo jedne od dilema u lingvističkim krugovima vezanih uz poimanje korpusne lingvistike, odnosno kako klasificirati korpusna lingvistiku - kao metodologiju, teoriju, zasebnu lingvističku disciplinu? Pristup temeljen na korpusu podupire stajalište da je korpusna lingvistika metodologija, budući da se njome služi za provjeru nečega postojećeg, primjerice, nadopunjavanje teorijskih tvrdnja i intuitivnih pretpostavki.[12] Pristup vođen korpusom, pak, podupire suprotno stajalište – da se korpusna lingvistika, odnosno korpus, može smatrati jezičnom teorijom. Važno je napomenuti da su lingvisti oko ovih klasifikacija vrlo neodlučni. Oni koji se u svojem radu koriste korpusima korpusnu lingvistiku doživljavaju kao nešto značajnije od puke metodologije. Halliday je istaknuo kako korpusna lingvistika objedinjuje postupke prikupljanja podataka i teoretiziranja, što dovodi do kvalitativnih promjena u našem shvaćanju jezika.[13]

Iako je dio primijenjene lingvistike, Tognini-Bonelli ponudila je za korpusnu lingvistiku naziv metodologija pred-primjene (engl. *pre-application methodology*). Budući da se metodologija može definirati kao uporaba određenih pravila i informacija u određenoj situaciji, „metodologija pred-primjene“ znači da, za razliku od ostalih primjena koje započinju prihvaćanjem određenih zadanih činjenica, korpusna lingvistika može definirati svoja pravila i informacije prije nego što se uopće primijene; stoga, korpusna lingvistika može značajno doprinositi ostalim primjenama i, kao takva, ima status teorije.[14]

Tony McEnery i Andrew Wilson u svojoj su knjizi „Corpus Linguistics: An Introduction“ ponudili dogovor na pitanje može li se korpusna lingvistika klasificirati kao zasebna disciplina unutar lingvistike. Zaključili su kako korpusna lingvistika istovremeno jest i nije disciplina unutar lingvistike.[15] Ostale utemeljene lingvističke discipline predstavljaju specifične aspekte uporabe jezika. Gledajući na predmetnu dilemu iz navedene perspektive, korpusna lingvistika može se smatrati metodologijom primjenjivom u nizu lingvističkih istraživanja. Iako nije

[11] Teubert; Čermáková, 2007.
[12] Klobučar Srbić, 2008.
[13] Tognini-Bonelli, 2001.
[14] Ibid.
[15] McEnery; Wilson, 2001.

samostalna disciplina, svakako pruža uvid u pristup istraživanju jezika i time na svojevrstan način pobliže opisuje jedno ili više lingvističkih područja.[16]

Ipak, vidljivo je kako stručnjaci naginju više ka opisu korpusne lingvistike kao metodologije, iako je nedvojbeno da je, kao takva, praktički nadišla granice te definicije svojom, gotovo univerzalnom, primjenjivošću i izuzetno učinkovitim pristupom istraživanju jezika.

2.3. Chomskijanska i korpusna lingvistika

Dakle, korpusna lingvistika ima empirijski temelj, odnosno temelji se na induktivnoj metodi. Usprkos prednostima koje nudi i sve rastućem ugledu unutar akademske i istraživačke zajednice, jedan od najvećih lingvista 20. stoljeća, Noam Chomsky, bio je gorljivi protivnik njezine uporabe. Chomskyjeve su teorije bile iznimno cijenjene i u ovom je kontekstu vrlo važno opisati stajalište lingvista čiji je lingvistički pravac dominirao teorijskom lingvistikom sredinom prošlog stoljeća, budući da je konsenzus oko toga što sačinjava lingvističke podatke uvelike definiran u kontekstu rasprava koje je započeo Chomsky.[17] Nadalje, upravo je Chomsky taj koji je odgovoran za dugogodišnju raspravu između racionalista i empirista.[18] Ranih 1960-ih, njegove se publikacije pokrenule novu struju i preusmjerile su fokus lingvističkog istraživanja od empirizma ka racionalizmu, odnosno od jezične upotrebe (engl. *performance*) ka jezičnoj kompetenciji (engl. *competence*).[19]

Noam Chomsky imao je naturalistički pristup učenju jezika. Smatrao je kako je zadaća lingvistike proučavanje ljudske mogućnosti stvaranja neograničenog broja raznovrsnih gramatičkih rečenica, a srž njegove lingvističke revolucije bila je generativna moć pravila.[20] U svom djelu „Aspects of the Theory of Syntax“, Chomsky opisuje pojam *generativne gramatike*, koja označava sustav pravila koja na eksplicitan i definiran način dodjeljuju strukturalne opise rečenicama.[21] Takva gramatika, naglašava, većinom je okrenuta kognitivnim procesima kojih govornik nije svjestan i pokušava opisati ono što govornik zaista zna, a ne što može reći o svome znanju.[22] Gramatika nije adekvatna ako objašnjava konačni skup promatranih podataka,

[16] McEnery; Wilson, 2001.
[17] Seuren, 1998.
[18] McEnery; Wilson, 2001.
[19] Dobrić, 2009.
[20] Teubert; Čermáková, 2007.
[21] Chomsky, 1970.
[22] Ibid.

već ako proizvodi beskonačni skup gramatičkih rečenica i njezina je zadaća pritom objasniti i određene intuitivne uvide koje govornik ima o svojem jeziku, primjerice, prosudbe o tome je li neka za njega nova rečenica gramatična ili nije, shvaćanje dvosmislenosti pojedinih rečenica, uočavanje jednake interpretacije različitih rečenica (parafraziranje) i slično[23]. Generativni lingvisti vjeruju u generativnu moć gramatike, odnosno da uporabom gramatike idealnog izvornog govornika možemo proizvesti neograničen skup rečenica.[24]

Chomsky je promijenio cilj lingvističkog istraživanja; umjesto apstraktnih opisa jezika, proučavaju se teorije koje odražavaju psihološku stvarnost, odnosno kognitivno vjerojatne modele jezika.[25] Već spomenuti koncept jezične kompetencije ključan je pojam u chomskijanskoj lingvistici. Iz toga proizlazi da je zadaća lingvista modelirati jezičnu kompetenciju, a ne upotrebu.[26] Chomsky je terminima „kompetencija" i „upotreba" pridodao novu interpretaciju de Saussureovim terminima „langue" i „parole", pri čemu kompetencija ne predstavlja popis elemenata poput termina „langue", nego sustav generativnih pravila.[27] Temelj Chomskyjeve „revolucije" bila je teorija da svi ljudi dijele jednu jedinstvenu značajku – svatko od nas ima urođenu sposobnost, odnosno mehanizam za brzo usvajanje jezika i kreativno korištenje njime (engl. *LAD – Language Acquisition Device*). Prema Chomskyju, upravo ovaj mehanizam regulira već spomenutu gramatiku.[28] Chomsky je u ovom kontekstu upotrijebio i pojam „univerzalna gramatika" (engl. *universal grammar*), a označava temeljnu jednakost svih jezika na biološkoj razini.[29]

Kako je ranije spomenuto, Chomsky je iznimno važna figura u razvoju korpusne lingvistike. Tvrdio je kako je korpusna lingvistika beskorisna te da je njezina uloga krajnje beznačajna, budući da je jezik sam po sebi produktivan.[30] Tvrdio je kako se uporabom korpusa ne može doći do novih spoznaja, budući da se radi o analizi već izrečenih rečenica. Jezična uporaba je u potpunosti sporedna u njegovoj teoriji – jezična kompetencija je aspekt koji objašnjava i karakterizira govornikovo znanje jezika.[31] Međutim, zadaća korpusne lingvistike nešto je drugačija od zadaće chomskijanske lingvistike. Gramatika i postupak generiranja novih rečenica uz pomoć ograničenog vokabulara i pravila primarni su ciljevi generativne, odnosno

[23] Hrvatska enciklopedija. URL: http://www.enciklopedija.hr/natuknica.aspx?id=21594 (20.3.2017.)
[24] Teubert; Čermáková, 2007.
[25] McEnery; Wilson, 2001
[26] Ibid.
[27] Hrvatska enciklopedija. URL: http://www.enciklopedija.hr/natuknica.aspx?id=21594 (20.3.2017.)
[28] Teubert; Čermáková, 2007.
[29] Ibid.
[30] Ibid.
[31] McEnery; Wilson, 2001

chomskijanske lingvistike. U središtu korpusne lingvistike jest značenje, odnosno semantička promjena, kao i pragmatika.[32] Korpus, dakle, pruža uvid u uporabu prirodnoga jezika; odnosno, u značenjske „finese“ raznoraznih izraza, parafraza, idioma te njihovih konotacija. Pruža iznimno važan društveni kontekst pojedinih jedinica značenja i time važan uvid u mogućnost primjene navedenih elemenata u određenim situacijama. Korpusna lingvistika počiva na tvrdnji da nam definicije u rječniku ne mogu pružiti sva značenja određenih riječi. Također, dovodi u upit stajalište da je riječ temeljna jedinica značenja.[33] Najbolji su primjer toga idiomi, koji se sastoje od najmanje dvije riječi no čije je značenje različito od zasebnih „rječničkih“ definicija tih riječi. Korpusna lingvistika počiva na diskursu čiji članovi određuju značenje određenih riječi. Drugim riječima, ako nešto u diskursu nosi određeno značenje, ono se uzima kao istinito, budući da je enciklopedijsko znanje prisutno u našem diskursu kao i u definicijama u rječniku.[34] Korpusna lingvistika iz diskursa izvlači sve što možemo znati o značenju i jedinstven je pokazatelj koliko je zapravo naše poimanje svijeta crno-bijelo.[35] Jedan od najjednostavnijih primjera kontekstualnog značaja korpusa su kolokacije, naročito u prevođenju. Primjerice, u hrvatskom jeziku uvriježena je kolokacija „teška ozljeda“. Ako nismo iskusni poznavatelj engleskog jezika i htjeli bismo prevesti ovu kolokaciju na engleski, služeći se rječnikom možemo doći do rezultata poput „difficult injury“ ili „heavy injury“. Upišemo li te rezultate u korpus COCA (*Corpus of Contemporary American English*), korpusni sustav izbacuje obavijest da nije pronašao odgovarajuću kolokaciju. No iskoristimo li mogućnost korpusa da prema imenici „injury“ pronađemo najčešće pridjeve koji se upotrebljavaju u tom kontekstu, dobit ćemo pridjeve poput „serious“, „major“, „severe“, „significant“ i slično. Iako postoji mnoštvo rječnika u kojima određene kolokacije i jesu navedene, korpusi, između ostalog, pružaju i statistički pregled čestote uporabe pojedinog pridjeva, tekstnu okolinu unutar koje se kolokacija upotrebljava, pregled ostalih imenica koji čine kolokacije s navedenim pridjevom te jednostavnost pretraživanja.

Dakle, u 20. stoljeću razvila su se dva komplementarna pogleda na jezik, onaj faktografski koji se vodi generalizacijama iz stvarnog jezičnog fenomena te onaj koji je utemeljen na univerzalijama i potkrepljuje ih lingvističkim konstruktima.[36] Iako je kroz pregled očigledno da su pristupi različiti, Maja Bratanić pokušala je ova dva pogleda na jezik ponešto i približiti.

[32] Teubert; Čermáková, 2007.
[33] Ibid.
[34] Ibid.
[35] Ibid..
[36] Bratanić, 1998.

Tako je za korpuse navela kako u „izvrnutoj i ponešto isforsiranoj“ perspektivi u određenom smislu istovremeno simuliraju lingvističku kompetenciju idealnoga izvornoga govornika u rasponu koji može nadići pretpostavke bilo kojega konkretnoga modela.[37]

2.4. Povijest korpusne lingvistike

Budući da je njezina popularnost tada uvelike opala, opće je mišljenje kako je korpusna lingvistika bila praktički napuštena 1950-ih te da je ponovno, i sasvim slučajno, oživljena ranih 1980-ih; no pioniri u tom području obrađivali su i analizirali korpusne podatke neprestano nakon njezina nastanka.[38]

Začeci korpusne lingvistike vezani su uz tekstove religijskog ili kulturnog značaja. U Indiji su određeni dokumenti pod nazivom *Prātiśākhya* opisivali zvučne obrasce Sanksrta iz Veda, a gramatika klasičnog Sanskrta koju je sastavio drevni gramatičar Pāṇini djelomično se temeljila na analizi istog korpusa.[39] Na sličan su se način arapski gramatičari bavili jezikom Kurana, a u zapadnoj Europi učenjaci su proučavali jezik Biblije i drugih kanonskih tekstova.[40]

Dakle, moderna korpusna lingvistika pojavila se u razdoblju kada je Noam Chomsky svojim publikacijama protresao temelje lingvistike. Iako je lingvistička zajednica bila zaintrigirana pojmom univerzalnosti jezičnog aparata, 1950-ih godina pojavila su se određena empirijska pitanja i problemi koje nije mogla riješiti Chomskyjeva introspekcija; bili su potrebni konkretni jezični podaci.[41] Poznati britanski lingvist Randolph Quirk 1959. godine pokrenuo je prvi veći projekt skupljanja podataka za empirijsko proučavanje gramatike pod nazivom „Survey of English Usage“ (SEU). Upravo se taj projekt smatra začetkom moderne korpusne lingvistike jer je kao takav bio početna točka svima onima koji su bili zainteresirani za empirijsko proučavanje jezika.[42] Jednim od najvažnijih djela u najranijim stadijima razvitka korpusne lingvistike smatra se publikacija lingvista Henryja Kučere i Williama Nelsona Francisa pod nazivom „Computational Analysis of Present-Day American English“ izdana 1967. godine, koja se temeljila na analizi milijunskog korpusa Brown (engl. *Brown Corpus*) i u kojoj su lingvisti predstavili statističke podatke vezane uz korpus .[43] Upravo su ova dvojica

[37] Bratanić, 1998.
[38] McEnery; Wilson, 2001.
[39] Wikipedia. Corpus Linguistics. URL: https://en.wikipedia.org/wiki/Corpus_linguistics (21.3.2017.)
[40] Ibid.
[41] Teubert; Čermáková, 2007.
[42] Teubert; Čermáková, 2007.
[43] Wikipedia. Corpus Linguistics. URL: https://en.wikipedia.org/wiki/Corpus_linguistics (21.3.2017.)

lingvista odgovorna i za kompiliranje korpusa Brown od 1961. do 1964. godine, koji je ime dobio prema sveučilištu Brown u Providenceu, u saveznoj državi Rhode Island. Korpus Brown bio je prvi sustavno organizirani računalni korpus suvremenog američkog pisanog jezika koji se sastojao od jednog milijuna pojavnica, odnosno 500 tekstova prikupljenih iz raznovrsnih izvora.[44] Ovaj je korpus bio pažljivo organiziran, jednostavan za korištenje te korigiran i lektoriran da gotovo nije ni sadržavao jezične greške.[45] Jedan od poznatijih korpusa nastalih na temeljima SEU-a jest korpus London-Lund (engl. *London-Lund Corpus*), čiju je izradu predvodio Jan Svartvik. Svartvik je kompjuterizirao SEU, a računalo je 1970-ih godina postalo neizostavno pomagalo u korpusnoj lingvistici.[46] Korpus Brown također je utjecao na razvoj mnogih korpusa, poput korpusa LOB (engl. *Lancaster-Oslo-Bergen Corpus*), korpusa Kolhapur, korpusa Wellington, Australskog korpusa engleskog jezika (engl. *Australian Corpus of English*), korpusa Frown i korpusa FLOB (engl. *Freiburg-LOB Corpus*).[47]

Najvažniji korpusni projekt u ranijim stadijima razvoja, pod vodstvom lingvista Johna M. Sinclaira, bio je svakako „English Lexical Studies" započet 1963. godine u Edinburgu i dovršen u Birminghamu.[48] U to je vrijeme Sinclair pomno razmišljao o dilemi je li riječ temeljna jedinica značenja i ovim je projektom započeo istraživanje u tom smjeru. Sinclair je ujedno i prvi lingvist koji se korpusom služio u svrhu leksičkog istraživanja, a projekt se provodio nad elektroničkim uzorkom govornog i pisanog jezika kako bi se proučilo značenje koncepta „leksička jedinica" (engl. *lexical items*) koji je uključivao i, tada novi koncept, kolokacije.[49] Na temelju projekta nastalo je vrlo utjecajno djelo u kontekstu razvoja korpusne lingvistike – izvještaj OSTI (engl. *Report to the Office for Scientific and Technical Information*, *OSTI-Report*). Istraživanje na temelju kojeg je izvještaj napisan postavilo je temelje za razvoj teorije korpusne lingvistike uspostavom protokola i metodologija.[50]

Na temelju izvještaja OSTI nastao je istraživački projekt pod nazivom COBUILD. Najpoznatiji proizvod ovog istraživanja bio je rječnik koji je izrađen na temelju dubinske analize korpusa na računalu, pod nazivom „Collins COBUILD English Language Dictionary", veliki projekt izrade rječnika na temelju korpusa. Dizajniran je sredinom 1970-ih, a završen je 1987. godine pod vodstvom Johna Sinclaira na sveučilištu u Birminghamu. Korpus koji je

[44] Malmkjaer, 2004.
[45] Teubert; Čermáková, 2007.
[46] McEnery; Wilson, 2001.
[47] Wikipedia. Corpus Linguistics. URL: https://en.wikipedia.org/wiki/Corpus_linguistics (25.3.2017.)
[48] Ibid.
[49] Ibid.
[50] Williams, 2005.

služio kao baza riječi za rječnik sastojao se od 18,3 milijuna riječi i u to je vrijeme bio najveći korpus općeg jezika na svijetu.[51] Nažalost, upravo je njegova veličina bila ograničavajući čimbenik u vrsti njegove uporabe. Leksikografi su se rijetko njime koristili kako bi odredili sva moguća značenja riječi ovisno o njezinoj uporabi; njegova je primarna zadaća bila da služi kao validacijski alat za odluke leksikografa te da pruži potrebne primjere.[52]

Korpusna lingvistika obuhvatila je i neke od drevnih jezika. Korpus klasičnog arapskog jezika pod nazivom „Quranic Arabic Corpus" računalni je korpus koji se sastoji od teksta iz Kurana, islamske svete knjige. Kuran, odnosno korpus, sastoji se od 77.430 riječi i ima nekoliko slojeva obilježavanja.[53] Lingvisti Andersen i Forbes razvili su obilježeni korpus drevnog hebrejskog jezika koji se sastoji od oko pola milijuna riječi. Ovaj fond riječi čine Hebrejska Biblija, koja se sastoji od preko 300.000 riječi, suvremene inskripcije koje su pronašli arheolozi i dodatni kanonski tekstovi poput Svitaka s Mrtvog mora.[54]

Kroz svoj razvoj, korpusna lingvistika suočavala se s mnogim problemima i samim je time njezin je napredak u određenim razdobljima bio usporen. Kao u svim područjima gdje se primjenjuje računalna tehnologija, u pitanju je bilo mnoštvo tehničkih poteškoća vezanih uz izradu korpusa, no počele su se izražavati određene sumnje i postavljati ključna pitanja. Između ostalog, stručnjaci su se pitali može li uistinu korpus predstaviti diskurs zbog svoje ograničenosti i doveli su u pitanje mogućnost standardizacije te načine označavanja i kodiranja.[55] No unatoč problemima i uz pomoć računalne tehnologije, korpusna lingvistika postepeno se razvijala i postajala sve složenija. Danas se više ne smatra marginaliziranim pristupom kojim se koristi isključivo u lingvistici engleskog jezika, već neizostavnim pristupom u mnoštvu lingvističkih istraživanja i područja kojim se proučavaju mnogi jezici i njihovi dijalekti diljem svijeta.[56]

3. Korpus

3.1. Što je korpus?

Prije daljnje analize, izrazito je bitno definirati temeljni pojam ovog rada. Iako se definicije korpusa u većini slučaja uvelike podudaraju, u najužem smislu riječi, korpus

[51] Teubert; Čermáková, 2007.
[52] Ibid.
[53] Dukes; Atwell; Habash, 2011.
[54] Andersen; Forbes, 2012.
[55] Teubert; Čermáková, 2007.
[56] McEnery; Wilson, 2001.

(lat. *corpus:* tijelo) jest, kao što i etimologija termina sugerira, određeno tijelo tj. cjelovita zbirka podataka, dokumenata ili građe za neku disciplinu.[57] Neke od općenitijih definicija korpus opisuju kao zbirku tekstualne građe, što je, ipak, ponešto uža definicija budući da postoje i tzv. govorni korpusi, odnosno zbirke zvučnih zapisa prirodnog jezika, o čemu će više biti rečeno u nastavku rada. U svojem znanstvenom članku „Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika", renomirani hrvatski lingvist Marko Tadić korpus definira kao zbirku jezičnih odsječaka odabranih i skupljenih prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak i pri tome napominje kako korpus ne moraju sačinjavati isključivo cjeloviti tekstovi, nego i dijelovi tekstova koji su dovoljno veliki da čine korpusni uzorak. [58] Kako je spomenuto ranije u radu, korpusna lingvistika i računalna tehnologija postali su nerazdvojni. Stoga se u današnje vrijeme pod pojmom korpus gotovo uvijek misli na računalni korpus, odnosno korpus kodiran na standardan i dosljedan način s nakanom da bude otvoren za računalno pretraživanje.[59]

3.2. Struktura korpusa

Vrlo je bitno naglasiti kako ne postoji strukturalno uniformirana verzija korpusa. Različiti korpusi služe različitim svrhama i, u skladu s tim, dizajnirani su i implementirani na drugačije načine. Korpusi mogu biti različitih veličina, iako je poželjno da im je obujam što veći, kako bi se analizom dobili što točniji i iscrpniji jezični podaci. Njihovo tijelo tekstova može se sastojati od odsječaka ili od potpunih formalnih, neformalnih, književnih ili općih zvučnih i tekstualnih zapisa. Mnogi jezični stručnjaci vjeruju kako je govorni oblik jezika bolji pokazatelj temeljne strukture i organizacije jezika od pisanog oblika, a sam Sinclair tvrdi kako je improvizirani govor najpouzdaniji.[60] Nadalje, veliki je broj korpusa određen specifičnim vremenskim razdobljem, tj. sadržava zapise koji se mogu kategorizirati u isto razdoblje. Tako, primjerice, postoji Korpus suvremenog američkog engleskog jezika (engl. *Corpus of Contemporary American English, COCA*), koji se sastoji od suvremene varijante engleskog jezika koja se govori u SAD-u. No i tu postoji određeni problem razgraničavanja. Određenom je književnom djelu potrebno možda dulje vrijeme da se etablira, a njegov utjecaj može biti prisutan vrlo dugo;

[57] Sveučilišni računski centar. Hrvatski jezični portal. URL: http://hjp.znanje.hr/index.php?show=search_by_id&id=elllXRQ%3D (1.4.2017.)
[58] Tadić, 1998.
[59] Ibid.
[60] Sinclair, 1991.

primjerice, frazeologija W. Shakespearea i Biblija kralja Jakova i danas imaju utjecaj na upotrebu suvremenog engleskog jezika.[61]

Za računalne je korpuse iznimno važan pojam standardizacije, odnosno standardizirani načini prikazivanja teksta putem računala. Između ostalog, vrlo je bitno istaknuti inicijativu kodiranja teksta TEI (engl. *Text Encoding Initiative*). TEI je međunarodni projekt pokrenut 1988. godine čija je svrha razvoj generičkih smjernica za standardnu shemu kodiranja tekstova.[62] TEI smjernice definiraju format XML, a specificiraju potpune bibliografske informacije, omogućuju odvajanje čistog teksta od ostalih oblika kodiranja standardiziranom metodom i sustavno kategoriziraju ostale oblike kodiranja poput, primjerice, onih za ekstralingvističke metapodatke.[63]

Tekst u korpusu najčešće je jednostavnog formata, u obliku linearnog niza pismena (engl. *characters*) koje čine slova, razmaci i interpunkcijski znakovi.[64] Kako je navedeno u kratkom rječniku korpusne lingvistike sastavljenom na Odsjeku za lingvistiku Filozofskog fakulteta u Zagrebu, slova sačinjavaju oblike riječi, odnosno pojavnice (engl. *token*) koje se definiraju kao nizovi pismena omeđeni graničnicima, dok leme (engl. *lemma*) predstavljaju temeljne oblike pojavnica koji su navedeni u rječniku. Računalu je izrazito važno osigurati uputu, odnosno metodu lematizacije (engl. *lemmatisation*) kojom ono može različite pojavnice svesti na zajedničku lemu. Tekst za računalnu obradu najčešće se i obilježava (engl. *annotation*) čime se dodaju eksplicitne informacije ondje gdje su implicitno prisutne čitatelju. Tako se delimitiranim jezičnim jedinicama pomoću oznaka (engl. *tag*) opisuje struktura ili određena osobina koja pripada toj jedinici. Najčešći je primjer toga POS označavanje (engl. *POS tagging*) kojim se označavaju vrste riječi u rečenici. Prije samog POS označavanja, potrebno je provesti tokenizaciju (engl. *tokenisation*), odnosno identificiranje i eksplicitno obilježavanje pojavnica. Važno je i spomenuti još jedan postupak prisutan u izgradnji većine korpusa, tzv. *parsing*, kojim se određuje sintaktička struktura rečenica na način da se rečenični dijelovi odvajaju te se pobliže opisuju njihovi međusobni odnosi.[65]

Većina korpusa vrlo je korisna utoliko što rezultate prikazuju u listama čestote (engl. *frequency list*) koje se, pak, mogu postaviti po kriterijima prikaza; primjerice, po prvom

[61] Sinclair, 1991.
[62] Text Encoding Initiative. What is the TEI?. URL: http://www.tei-c.org/Vault/SC/J31/WHAT.htm (2.4.2017.)
[63] Sinclair, 1991.
[64] Ibid.
[65] Hrvatski nacionalni korpus. Rječnik korpusne lingvistike. URL: www.hnk.ffzg.hr/bb/definicijekl.doc (2.4.2017.)

pojavljivanju riječi, abecednim redoslijedom, čestoti riječi i slično.[66] Veliki broj korpusa također ima i opciju pronalaska odnosno prikaza kolokacija (engl. *collocation*) prema traženoj pojavnici ili vrsti riječi. Naročito koristan način prikazivanja pojavnica su konkordancije (engl. *concordance*), prikaz traženih pojavnica s okolnim ko-tekstom, tj. pojavnica u njihovoj tekstnoj okolini.[67] Glavna riječ unutar konkordancije naziva se stožernica (engl. *headword*), a konkordancija se može prikazati u dva oblika: KWAL (engl. *Key-Word And Line*) gdje je dopušteno nekoliko redaka konteksta s lijeve i desne strane okoline, i KWIC (engl. *Key-Word In Context*) gdje su stožernice prikazane unutar unaprijed definirane lijeve i desne okoline.[68]

Slika 1. *Prikaz KWAL (gore) i KWIC (dolje) konkordancija za pojavnicu „corpus“ u korpusu COCA*[69]

Postoji još mnogo funkcionalnosti vezanih uz korpuse, no predstavljene su neke od značajki koje su više-manje prisutne kod većine današnjih računalnih korpusa i samim time predstavljaju reprezentativnu skupinu korpusnih značajki.

3.3. Vrste korpusa

Kako postoji mnoštvo značajki koje korpus može posjedovati, tako postoje i različite vrste korpusa dizajnirane u određenu svrhu. Isto tako, klasifikacije korpusa ovise o parametru klasifikacije, primjerice, jeziku, opsegu specijalizacije, obliku govora i slično. Na taj način jedan korpus može biti klasificiran u više kategorija istovremeno.

[66] Sinclair, 1991.
[67] Ibid.
[68] Hrvatski nacionalni korpus. Rječnik korpusne lingvistike. URL: www.hnk.ffzg.hr/bb/definicijekl.doc (2.4.2017.)
[69] Mark Davies. Corpus of American Contemporary English. URL: http://corpus.byu.edu/coca/ (2.4.2017.)

Prva klasifikacija korpusa učinjena je s obzirom na parametar opsega specijalizacije. Na taj način razlikujemo opći korpus (engl. *general corpus*) i specijalizirani korpus (engl. *specialized corpus*). Opći korpusi su korpusi koji nisu ograničeni na određeni tip teksta, tematsko područje ili registar, odnosno njihov sadržaj u jednakoj mjeri predstavlja područja i žanrove reprezentativne za jezik.[70,71] Najpoznatiji primjer općeg korpusa je Britanski nacionalni korpus (engl. *British National Corpus)* čija je izgradnja trajala od 1990. do 1994. godine, a koji se sastoji od 100 milijuna riječi, od čega je 90% jezika u pisanom, a 10% u govornom obliku.[72] Ostali poznati opći korpusi su Američki nacionalni korpus (engl. *American National Corpus*), Korpus suvremenog američkog engleskog jezika (engl. *Corpus of Contemporary American English*), Ruski nacionalni korpus (engl. *Russian National Corpus*), Hrvatski nacionalni korpus (engl. *Croatian National Corpus*), Mađarski nacionalni korpus (engl. *Hungarian National Corpus*) i mnogi drugi. Specijalizirani korpusi su korpusi čiji su tekstovi ograničeni na određeni tip teksta, domenu ili žanr, dakle predstavljaju određenu varijantu jezika, a ne jezik u općenitim pojmovima.[73] Mogu biti raznih veličina, a neki od poznatijih su korpus CHILDES, koji sadrži jezik generiran od strane djece, korpus MICASE (engl. *Michigan Corpus of Academic Spoken English*) i drugi.[74]

Sljedeća klasifikacija korpusa jest prema jezičnom obliku, gdje razlikujemo korpus pisanog jezika (engl. *written corpus*) i korpus govornog jezika (engl. *spoken corpus*). Budući da je već spomenuto mnogo korpusa pisanog jezika, valja navesti neke od korpusa govornog jezika, primjerice, Korpus Santa Barbara govornog američkog engleskog jezika (engl. *Santa Barbara Corpus of Spoken American English*), korpus CANCODE (engl. *Cambridge and Nottingham Corpus of Discourse in English*) i drugi.

Korpusi se klasificiraju i po vremenskom razdoblju koje pokrivaju, pa tako postoje sinkronijski korpus (engl. *synchronic corpus*) i dijakronijski korpus (engl. *diachronic corpus*). Sinkronijski korpus je korpus koji sadrži jezične podatke iz jednog vremenskog razdoblja, što znači da je pogodan za uspoređivanje jezičnih varijeteta. Najpoznatiji takav korpus jest Međunarodni korpus engleskog jezika (engl. *International Corpus of English*) kojem je primarni cilj usporedba varijeteta engleskog jezika u državama u kojima je engleski službeni ili

[70] University of Essex. Corpus Linguistics. URL: http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction2.html (5.4.2017.)
[71] Megyesi, Beata B.. Corpus usage. Uppsala University. URL: http://stp.lingfil.uu.se/~bea/uv/uv09/dokverktyg/02-corpususage.pdf (5.4.2017.)
[72] Aston; Burnard, 1998.
[73] Bennett, 2010.
[74] Ibid.

većinski jezik.[75] Dijakronijski korpus sadrži jezične podatke iz različitih vremenskih razdoblja, što znači da je pogodan za proučavanje tijeka mijenjanja pojedinog jezika. Primjer dijakronijskog korpusa je Helsinški dijakronijski korpus engleskih tekstova (engl. *Helsinki Diachronic Corpus of English Texts*) koji se sastoji od 1,5 milijuna riječi iz 400 tekstnih uzoraka.[76]

Još jedna klasifikacija korpusa jest po broju jezika koje sadrži. Tako razlikujemo jednojezični korpus (engl. *monolingual corpus*) i višejezični korpus (engl. *multilingual corpus*). Višejezični korpus dalje se grana na paralelni korpus (engl. *parallel corpus*) koji većinom sadrži dva paralelna jezika, odnosno tekst jednog jezika prijevod je teksta drugog jezika. Sličan paralelnom korpusu je komparativan korpus (engl. *comparable corpus*), koji sadrži komponente prikupljene prema zajedničkom kriteriju uzorkovanja te slične uravnoteženosti i reprezentativnosti;[77] drugim riječima, tekstovi imaju jednak sadržaj no nisu izravni prijevodi.

Prema režimu prikupljanja podataka razlikujemo monitor korpuse (engl. *monitor corpus*) i uravnotežene korpuse (engl. *balanced corpus*, *sample corpus*). Monitor korpusi specifični su jer se neprestano ažuriraju i nadopunjuju kako bi pratili jezične promjene, a najpoznatiji je takav korpus Bank of English, započet 1991. godine u okviru projekta COBUILD.[78] Uravnoteženi korpus je statičan, njegov je sadržaj iz određenog vremenskog razdoblja i ne nadopunjuje se.[79]

Postoje i referentni korpusi (engl. *reference corpus*) i ciljni korpusi (engl. *target korpus*), koji su u međusobnom odnosu na način da se referentni korpus upotrebljava kao korpus s kojim se drugi, ciljni korpusi uspoređuju, najčešće vezano uz analizu statističkih podataka. Postoji još klasifikacija po raznoraznim parametrima, poput učeničkog korpusa, pedagoškog korpusa i slično, no navedene su klasifikacije najistaknutije i najspominjanije u stručnoj literaturi.

3.4. Korpusi u Hrvatskoj

U usporedbi s mnogoljudnim jezicima poput engleskog, hrvatska korpusna lingvistika nije toliko razvijena i njezina povijest mnogo je kraća. Kod mnogoljudnih jezika postoji

[75] Megyesi, Beata B. Corpus usage. Uppsala University. URL: http://stp.lingfil.uu.se/~bea/uv/uv09/dokverktyg/02-corpususage.pdf (5.4.2017.)
[76] Ibid.
[77] McEnery, 2003.
[78] Xiao, 2008.
[79] Ibid.

zainteresiranost velikog tržišta i relativno je lako naći privredne i neprivredne subjekte, a korpusni su projekti većinski financirani upravo iz takvih izvora.[80] Hrvatski je jezik od oko pet milijuna govornika, te se za, primjerice, projekt nacionalnog korpusa samim time nameće isključivo rješenje u obliku pomoći jezične zajednice, tj. države koja bi morala pokrivati većinu troškova.[81] Treća verzija Hrvatskog nacionalnog korpusa (HNK 3.0) sadrži 101,3 milijuna pojavnica i predstavlja ususstavljenu zbirku odabranih tekstova uglavnom suvremenoga hrvatskog jezika koji pokrivaju raznovrsne medije, žanrove, stilove, područja i tematiku i javno je i slobodno dostupan u svrhu istraživanja, obrazovanja i ostalih nekomercijalnih uporabi. [82,83] HNK nastao je na temelju predložene strukture iz nekoliko članaka Marka Tadića, a njegova prva verzija sadržavala je samo 30 milijuna pojavnica.

Još jedan projekt je Hrvatska jezična riznica, a razvio ga je Institut za hrvatski jezik i jezikoslovlje. Projekt je zapravo korpus hrvatskog jezika koji se prikuplja od odabranih tekstova hrvatskoga jezika svih struka i funkcionalnih usmjerenja, uglavnom iz razdoblja konačnog oblikovanja hrvatske standardojezične norme u drugoj polovici 19. st. do danas.[84] U Hrvatskoj se razvija i prvi hrvatski učenički korpus. Učenički su korpusi elektronske zbirke tekstova koje su generirane od strane neizvornih govornika, odnosno učenika jezika.[85] Prvi učenički korpus hrvatskog pisanog jezika zove se Hrvatski učenički korpus (engl. *Croatian Learner Corpus*), a sastoji se od dva potkorpusa – korpusa pisanog jezika (engl. *CROatian Learner TExt Corpus CROLTEC*) i korpusa govornog jezika (engl. *CROatian Learner SpEech Corpus, CROLSEC*), građu mu sačinjavaju tekstovi i zvučni zapisi stranih učenika koji pohađaju Croaticum – Centar za hrvatski kao drugi i strani, a svrha mu je omogućiti dubinsku analizu jezika učenika te ga opisati, kao i odstupanje od standardnog jezika.[86] Još jedan iznimno zanimljiv projekt zove se hrWaC. hrWac je *web* korpus, odnosno korpus koji se sastoji od tekstova preuzetih s internetskih stranica s vršnom domenom Hrvatske, odnosno oznakom „.hr“. Korpus je razvila skupina za obradu prirodnog jezika na Odsjeku za informacijske i komunikacijske znanosti

[80] Tadić, 1997.
[81] Ibid.
[82] Mikelić Preradović; Berać; Boras, 2015.
[83] Hrvatski nacionalni korpus. URL: http://www.hnk.ffzg.hr/default.htm (7.4.2017.)
[84] Institut za hrvatski jezik i jezikoslovlje. Hrvatska jezična riznica. URL: http://riznica.ihjj.hr/dokumentacija/index.hr.html (7.4.2017.)
[85] McEnery; Hardie, 2011.
[86] Mikelić Preradović; Berać; Boras, 2015

Filozofskog fakulteta u Zagrebu.[87] hrWaC je osvanuo i u svojoj drugoj verziji, koja sadrži 1,9 milijardi pojavnica, a razvili su ga Nikola Ljubešić i Filip Klubička.[88]

4. Alati i tražilice za pretraživanje korpusa

Na početku ovog poglavlja važno je napomenuti kako svi alati i tražilice predstavljene u nastavku nisu prikazani na isti način ili u jednakom opsegu zbog dostupnosti, odnosno nedostupnosti, i vrste dokumentacije i literature.

4.1. Corpuscle

Corpuscle se razvija na multidisciplinarnom institutu Uni Research u Bergenu u Norveškoj, točnije na odjelu za računalstvo Uni Research Computing, koji je jedan od šest odjela na institutu. Na institutu se vode projekti istraživanja u područjima biotehnologije, zdravlja, okoliša, klime, energetike i društvenih znanosti.[89] Razvoj ovog alata financirali su infrastruktura CLARINO, zaklada Meltzer i Norveško istraživačko vijeće.[90]

Corpuscle je tražilica za pretraživanje korpusa i sustav za upravljanje korpusom koji je se počeo razvijati na spomenutom institutu 2009. godine. Jedan od autora alata, Paul Meurer, u svojem je znanstvenom članku „Corpuscle – a new corpus management platform for annotated corpora“ predstavio detaljan prikaz strukture alata i naveo kako je Corpuscle primarno dizajniran s namjerom da može osigurati platformu za korpuse velikih opsega, naglasivši i zahtjeve pri njegovu dizajnu:

1) Potpuna podrška za Unicode standard za prikaz slova;
2) Podrška za hijerarhijski strukturirane podatke (npr. XML);
3) Jednostavna implementacija atributa s više vrijednosti i atributa s određenim vrijednostima:
4) Podrška za korpuse iznimno velikih opsega (2 milijarde pojavnica i više);
5) Brzina izvršenja upita jednaka onoj alata IMS Open Corpus Workbench ili bolja;
6) Moćna sintaksa upita koja se može usporediti sa CQL-om (engl. *Corpus Query Language*) prema kojem je i dizajnirana;

[87] Natural Language Processing group. People. URL: http://nlp.ffzg.hr/people/ (12.4.2017.)
[88] Ljubešić; Klubička, 2014.
[89] Uni Research. URL: http://uni.no/en/about-uni-research/ (15.4.2017.)
[90] Uni Computing. Corpuscle. URL: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page# (15.4.2017.)

7) Jednostavno integriranje ručnog obilježavanja i uređivanja, uz izvođenje upita u realnom vremenu, barem za korpuse manjeg opsega;
8) Visoka funkcionalnost za generiranje konkordancija, kolokacija, statistike distribucije i slično;
9) Kvalitetno definirano aplikacijsko programsko sučelje (engl. *Application Programming Interface, API*);
10) Integrirano sučelje za *web*.[91]

Korpus se, dakle, sastoji od skupa korpusnih pozicija kojima su pridružene pojavnice, tj. riječi ili interpunkcijski znakovi, koje predstavljaju vrijednosti atributa *riječ*.[92] Prema članku, Corpuscle je alat koji zahtijeva da se obilježeni dokumenti vertikaliziraju prije nego što se uvezu radi obrade, slično kao u Sketch Engineu ili CWBU-u, gdje svaki red u vertikalnoj datoteci (engl. *vertical file*) predstavlja korpusnu poziciju i gdje su pozicijski atributi odvojeni tabom. Takva se datoteka može automatski generirati iz XML dokumenata, no važno je napomenuti da će takav korpus sadržavati samo strukturalne pozicije i osnovni atribut *riječi* koji odgovaraju sekvenci pojavnica, dok će lingvističko obilježavanje biti obavljeno zasebno, za što je potrebna predobrada podataka.[93]

word	pos
<NP>	
the	Det
girl	N
with	PP
<NP>	
the	Det
telescope	N
</NP>	
</NP>	

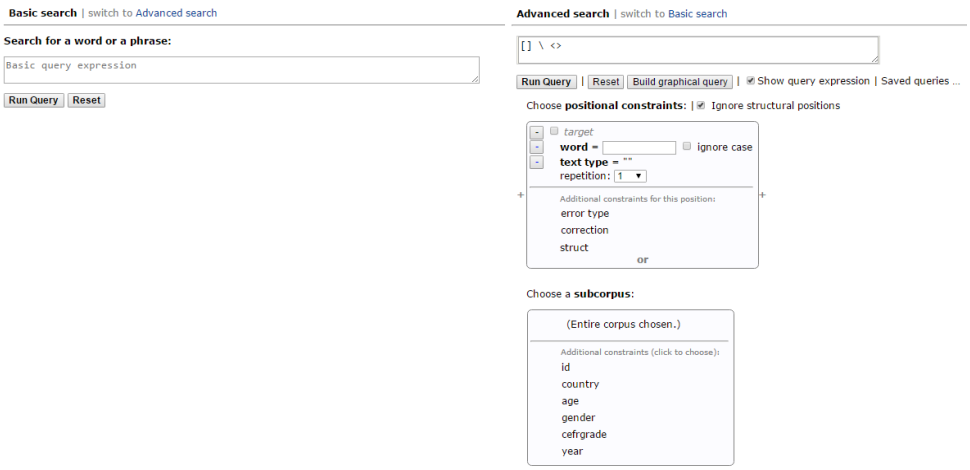
Slika 2. *Primjer vertikaliziranog teksta s oznakama POS*[94]

Meurer dalje navodi kako se, radi učinkovitosti, pretraživanje korpusa izvršava uz pomoć indeksa, točnije obrnutog indeksa (engl. *inverted index*) koji postoji za svaki korpusni atribut, a obilježeni je tekst korpusa kodiran kao skupina sekvenci numeričkih vrijednosti, tj. ID-ova gdje se jedna sekvenca dodjeljuje jednom atributu. Također je važno napomenuti kako

[91] Meurer, 2012.
[92] Uni Computing. Corpuscle: Getting started. URL: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-getting-started (15.4.2017.)
[93] Meurer, 2012.
[94] Ibid.

je autor iskoristio metodu sufiksnog niza (engl. *suffix array*) za implementaciju leksikona, odnosno za indeksiranje podnizova znakova prema pripadajućim nizovima znakova, te metodu strukturnog stabla (engl. *structure tree*) kojom se kodira hijerarhijska struktura korpusa.[95]

Kako je već navedeno ranije, autori Corpusclea stvorili su jezik, donosno sintaksu upita koja se uvelike temelji na onoj alata IMS Open Corpus Workbench. Imena Corpuscle Query Language (CQL), jezik je vrlo složen i predstavlja najizravniji način pretraživanja pomoću tražilice Corpuscle, iako, naravno, postoji i opcija jednostavnijeg pretraživanja pomoću jednostavnije varijante jezika ili pomoću sastava upita koji se mogu odabrati putem grafičkog sučelja .[96]



Slika 3. *Osnovne (lijevo) i napredne (desno) opcije pretraživanja korpusa u Corpuscleu*[97]

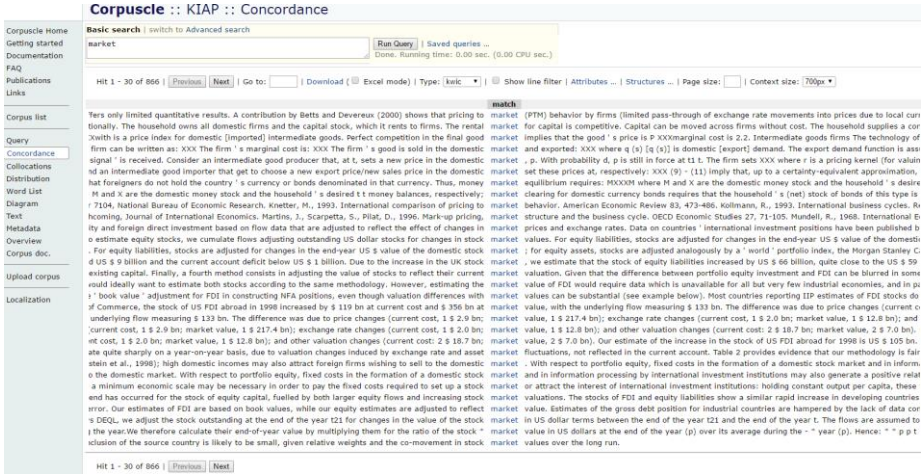
Meurer je opisao i jezik upita, koji se sastoji od pozicijskih ograničenja (engl. *positional constraint*) koja odgovaraju pojedinim korpusnim pozicijama i pišu se unutar uglatih zagrada ([…]), a unutar njih mogu biti sadržana atributna ograničenja (engl. *attribute constraint*) koja ograničenje povezuju s korpusnom pozicijom koja sadržava definirani atribut u sljedećem formatu „*atribut* = '*vrijednost*'“, a koja mogu biti atomska ili strukturna. Za vrijednosti u atributnim ograničenjima također je moguće i upotrebljavati regularne izraze (engl. *regular expression*), a pozicijska se ograničenja pomoću sekvencijskog operatora i operatora *, +, *{n,m},?* i | mogu uključiti u regularne izraze. Još jedna korisna značajka je mogućnost postavljanja upita za strukturalne pozicije, odnosno XML oznake, što se izvodi na vrlo praktičan način budući da se upotrebljavaju standardni XML elementi <…> i </…>.[98] Na

[95] Meurer, 2012.
[96] Uni Computing. Corpuscle: URL: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-getting-started (15.4.2017.)
[97] Uni Computing. Corpuscle: URL: http://clarino.uib.no/korpuskel/simple-query (16.4.2017.)
[98] Meurer, 2012.

internetskoj stranici Corpusclea nalazi se opširna dokumentacija vezana uz korištenje jezikom Corpuscle Query Language s detaljnim i jasnim uputama.

Što se tiče aplikacijskog programskog sučelja, alat je napisan u programu Lisp, točnije u njegovoj inačici Common Lisp, i pohranjen je na internetskom poslužitelju s kojim se komunicira pomoću određenih HTTP zahtjeva *get* i *post*, a zauzvrat se dobivaju odgovori u obliku stranice u XML-u i JSON-u.[99] Ovo je sučelje jedini vanjski API koji je implementiran i samim je time vrlo pogodno za izradu internetskih aplikacija povrh temeljne funkcionalnosti alata.[100]

Internetsko sučelje Corpusclea izrađeno je vrlo jednostavno i minimalistički. Corpuscle na svojoj internetskoj stranici nudi pristup 33 različita korpusa različitih jezika – škotskog, norveškog, španjolskog, engleskog, bugarskog, slovenskog, njemačkog, francuskog i staronordijskog jezika. Pristupiti se ne može svim korpusima. Nekoliko je korpusa dostupno i bez prijave na stranicu, neki imaju ograničen pristup i potrebna je prijava kako bi im se pristupilo, a neki su isključivo dostupni članovima norveške infrastrukture CLARINO. Nekoliko je mogućih načina prijave. Ukoliko je institucija pri kojoj korisnik ima račun član mreža CLARIN Service Provider ili eduGAIN, te ukoliko korisnik ima CLARIN IdP ili OpenIdP račun, odobrava mu se pristup korpusima s javnim licencama i, ovisno o pripadnosti, akademskim licencama.[101] Prije ulaska u svaki korpus potrebno je pristati na uvjete licence pod kojom se izdaje pojedini korpus.

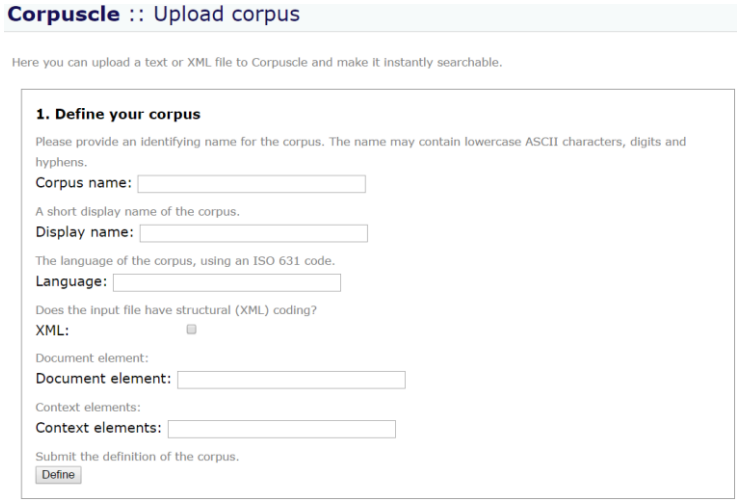


Slika 4. *Prikaz korpusnih pozicija za upit „market" u formatu KWIC unutar korpusa KIAP*[102]

[99] Meurer, 2012.
[100] Ibid.
[101] Uni Computing. Corpuscle: URL: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-getting-started (16.4.2017.)
[102] Uni Computing. Corpuscle: URL: http://clarino.uib.no/korpuskel/concordance (16.4.2017.)

Na primjeru korpusa KIAP, korpusa engleskih i francuskih članaka iz područja ekonomije, lingvistike i medicine, prikazan je KWIC rezultat za upit „market“. Unutar konkordancija, korisnik može prilagoditi prikaz informacija ali i veličinu prikaza. Korisnik može odlučiti želi li prikaz konkordancija kao KWIC ili unutar konteksta definiranog strukturnim atributom, npr. rečenicom ili paragrafom.[103] Kada su konkordancije prikazane u KWIC formatu, moguće je i odabrati određene atribute (npr. autor, izvor i sl.) te strukture (npr. <body>, <references> i sl.). Korisnik također može i preuzeti prikazane rezultate u .txt formatu. Na Slici 4. s desne su strane vidljive određene opcije postavljanja upita. Primjerice, u odjeljku „Word List“ zajedno su prikazani rezultati upita uz njihov broj pojavljivanja te je moguće i podesiti da se dobiju odgovarajući rezultati za druge atribute osim atributa *riječ*, a za naprednije statistike vezane uz učestalosti i distribuciju tu je i odjeljak „Distribution“.[104] Osim toga, korisnik može dobiti uvid u izvorni tekst u pitanju, metapodatke korpusa, njegovu dokumentaciju i pregled korpusa.

Internetsko sučelje je općenito vrlo praktično i intuitivno, a što se tiče korisničke podrške, stranica sadrži izrazito korisne poveznice, publikacije te odjeljak „FAQ“ u kojemu su pruženi odgovori na najčešće postavljena pitanja. Još jedna korisna i podosta pojednostavljena opcija je opcija za prenošenje vlastitog korpusa na Corpuscle – potrebno je ispuniti nekoliko polja vezano uz budući korpus i prenijeti dokument koji će biti pretraživ.



Slika 5. *Prikaz sučelja za unos podataka o korpusu kojeg korisnik želi prenijeti na Corpuscle*[105]

[103] Muerer, 2012.
[104] Ibid.
[105] Uni Computing. Corpuscle. URL: http://clarino.uib.no/korpuskel/upload-corpus?session-id=242651015837346 (17.4.2017.)

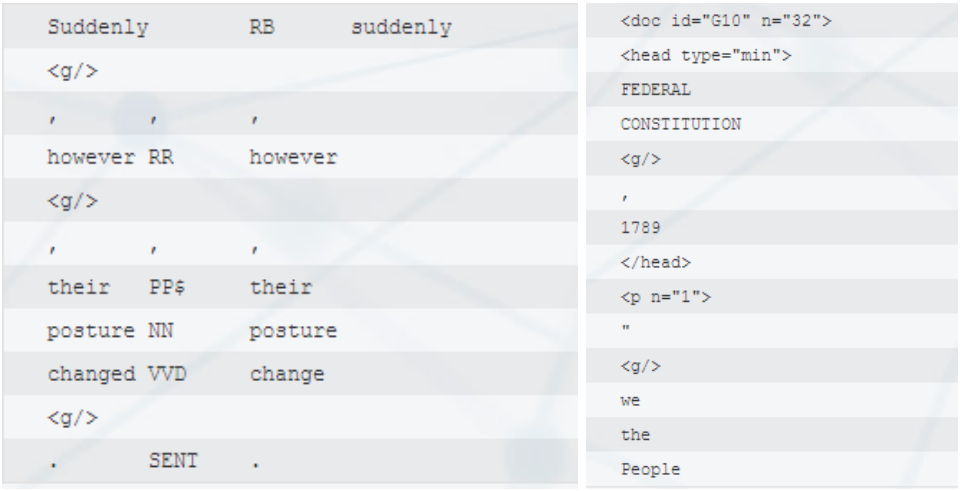
4.2. Sketch Engine

Sketch Engine softver je za upravljanje korpusima i njihovu analizu kojeg razvija tvrtka Lexical Computing Ltd. od 2003., a njegova je prva inačica izdana 23. lipnja 2003. Tvrtku je osnovao britanski lingvist i leksikograf Adam Kilgarriff s ciljem pružanja usluga i proizvoda na području korpusne obrade.[106] Važno je napomenuti kako je Sketch Engine nastao suradnjom s Pavelom Rychlýjem, informatičarom iz Centra za obradu prirodnog jezika pri Masarykovom sveučilištu u Brnu, kojeg je Kilgarriff imao prilike upoznati nakon što je Rychlý razvio sustav za korpusne upite.[107] Nakon upoznavanja, Rychlý je u svoj sustav odlučio inkorporirati Kilgarriffov pojam nacrta riječi (engl. *word sketch),* te je razvio komponente Manatee i Bonito, o kojima će više biti riječi u nastavku.[108] Od svog začetka do danas, softver je doživio mnoge promjene i, s vremenom, dobivao dodatne značajke i poboljšane funkcionalnosti. Budući da je jedan od najpoznatijih alata u svom području, Sketch Engine ima široku uporabu unutar lingvistike i leksikografije.

Sketch Engine ima tri temeljna komponenta, vlastiti sustav upravljanja pod nazivom *Manatee*, internetsko sučelje pod nazivom *Bonito* te modul *Corpus Architect* koji je odgovoran za izradu korisničkih korpusa i upravljanje njima.[109] Dakle, alat Sketch Engine čine softverski paket te internetski servis koji, uz temeljni softver, uključuje mnoštvo gotovih korpusa kojima se korisnici mogu koristiti i koje održava tim Sketch Enginea, kao i mogućnost stvaranja i instaliranja vlastitog korpusa te upravljanja njime pomoću alata WebBootCaT.[110] Korisnik ima dvije opcije u pogledu stvaranja vlastitog korpusa - može prenijeti i instalirati već postojeći korpus, a može ga i početi graditi ispočetka. Korpusna građa, odnosno dokumenti, mogu se prenositi u standardnim formatima kao što su *.pdf, .doc, .txt, .html* i *.tmx,* a mogu biti i komprimirani u formatima *.zip, .tar, .gz,* i *.bz2*, no nakon što se prenesu, procesom konvertiranja pretvaraju se u običan tekst, odnosno format *.txt*.[111] Kao i Corpuscle, Sketch Engine zahtijeva da se tekst vertikalizira, odnosno da bude u formatu w*ord-per-line (WPL)*, budući da je to ulazni format alata. Dakle, u svakom je redu po jedna pojavnica, koja zauzima korpusnu poziciju. Općenito, izvorna se građa sastoji od: a) već spomenutih korpusnih pozicija, kojima su pridruženi određeni atributi; b) struktura, tj. korpusnih segmenata koji imaju početne i završne pozicije, a koji najčešće označavaju rečenice, paragrafe ili dokumente (XML oznake) i c)

[106] Adam Kilgarriff. URL: http://kilgarriff.co.uk/cv.htm (19.4.2017.)
[107] Sketch Engine. URL: https://www.sketchengine.co.uk/adam-kilgarriff-structured-bibliography/ (19.4.2017.)
[108] Ibid.
[109] Kilgarriff; Baisa; Bušta et al., 2014.
[110] Ibid.
[111] Ibid.

strukturnih atributa, tj. atributa pojedinačnih struktura koji sadrže metapodatke o tim strukturama kao što je ime autora.[112] Ako je tekst lematiziran i POS-označen, vertikaliziranom se tekstu dodaju dva stupca koji su odvojeni tabom. Važno je naglasiti kako riječi mogu imati mnoštvo atributa, a ne samo uobičajene leme i POS oznake.

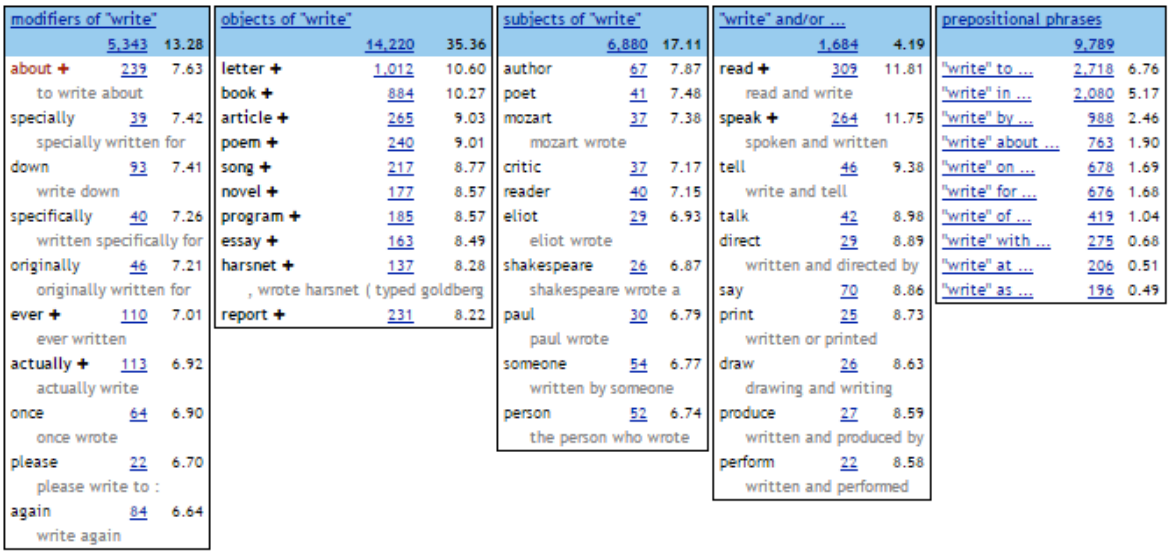


Slika 6. *Prikaz vertikaliziranog teksta, lematiziranog i s POS oznakama (lijevo) i prikaz vertikaliziranog teksta sa strukturalnim obilježjima, odnosno XML oznakama (desno)*[113]

Jedna od temeljnih funkcionalnosti alata Sketch Engine je pojam kojeg je uveo Kilgarriff i po kojem je alat i dobio ime, a to je nacrt riječi (engl. *word sketch*). Nacrti riječi izrazito su inovativni i korisni načini prikazivanja rezultata pretrage korpusa. Nacrti riječi su sažeci gramatičkog i kolokacijskog ponašanja pojedine riječi.[114] Pri pretraživanju nacrta riječi, moguće je odabrati vrstu riječi koja se pretražuje i mnoštvo naprednih postavki kao što su minimalna učestalost, način sortiranja kolokacija, mogućnost strukturiranja nacrta riječi po gramatičkim odnosima (engl. *gramrel*) i slično. Nadalje, u nekim je korpusima također moguće služiti se značajkom *Sketch difference* koja generira nacrte riječi za dvije riječi radi lakše usporedbe, a izrazito je korisna za bliske sinonime i antonime.[115] Nacrti riječi, ovisno o odabiru vrste riječi, detaljan su prikaz interakcije tražene riječi s okolnim gramatičkim elementima i vrlo su korisni za leksikografe, ali i učenike pojedinih jezika. U prikazu nacrta riječi naveden je i broj koji se odnosi na ukupan broj puta koliko se riječ pojavila uz drugi gramatički element. U slučaju da korisnik želi uvid u kontekst pojedine pojavnice riječi, klikom na taj broj sučelje

[112] Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/preparing-a-text-corpus-for-the-sketch-engine-overview/ (19.4.2017.)
[113] Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/preparing-corpus-text/ (19.4.2017.)
[114] Kilgarriff; Baisa; Bušta et al., 2014.
[115] Sketch Engine. URL: https://www.sketchengine.co.uk/quick-start-guide/word-sketch-difference-lesson-2/ (19.4.2017.)

preusmjerava korisnika na stranicu s prikazom konkordancija koje, kao što je već navedeno, čine osnovni alat za sve one koji rade s korpusima.[116]



modifiers of "write"		
	5,343	13.28
about ✚	239	7.63
to write about		
specially	39	7.42
specially written for		
down	93	7.41
write down		
specifically	40	7.26
written specifically for		
originally	46	7.21
originally written for		
ever ✚	110	7.01
ever written		
actually ✚	113	6.92
actually write		
once	64	6.90
once wrote		
please	22	6.70
please write to :		
again	84	6.64
write again		

objects of "write"		
	14,220	35.36
letter ✚	1,012	10.60
book ✚	884	10.27
article ✚	265	9.03
poem ✚	240	9.01
song ✚	217	8.77
novel ✚	177	8.57
program ✚	185	8.57
essay ✚	163	8.49
harsnet ✚	137	8.28
, wrote harsnet (typed goldberg		
report ✚	231	8.22

subjects of "write"		
	6,880	17.11
author	67	7.87
poet	41	7.48
mozart	37	7.38
mozart wrote		
critic	37	7.17
reader	40	7.15
eliot	29	6.93
eliot wrote		
shakespeare	26	6.87
shakespeare wrote a		
paul	30	6.79
paul wrote		
someone	54	6.77
written by someone		
person	52	6.74
the person who wrote		

"write" and/or ...		
	1,684	4.19
read ✚	309	11.81
read and write		
speak ✚	264	11.75
spoken and written		
tell	46	9.38
write and tell		
talk	42	8.98
direct	29	8.89
written and directed by		
say	70	8.86
print	25	8.73
written or printed		
draw	26	8.63
drawing and writing		
produce	27	8.59
written and produced by		
perform	22	8.58
written and performed		

prepositional phrases		
	9,789	
"write" to ...	2,718	6.76
"write" in ...	2,080	5.17
"write" by ...	988	2.46
"write" about ...	763	1.90
"write" on ...	678	1.69
"write" for ...	676	1.68
"write" of ...	419	1.04
"write" with ...	275	0.68
"write" at ...	206	0.51
"write" as ...	196	0.49

Slika 7. *Nacrt riječi za riječ „write" iz BNC-a*[117]

Naravno, do konkordancija se može doći i putem osnovnog pretraživanja korpusa, odnosno izvođenja jednostavnog upita. Ako korisnik odabere ovu metodu prikaza, može dodatno podesiti tip upita, strukturu kontekstualnog prikaza te vrstu tekstova koje se pretražuje. Korisne značajke pri jednostavnom pretraživanju korpusa:

a) Jednostavni upiti nisu osjetljivi na velika i mala slova;
b) Upite od više elemenata, koji su odvojeni razmaknicom, interpretira kao sekvence;
c) Pretražuje temeljni oblik riječi, tj. lemu, ili određeni oblik riječi.[118]

Slično kao u alatu IMS Open Corpus Workbench, jezik upita kojim se koristi u Sketch Engineu jest proširena verzija CQL-a (engl. *Corpus Query Language*).[119] CQL se upotrebljava isključivo u konkordancijskom pretraživanju, a važno je napomenuti kako se unutar Sketch Enginea mogu upotrebljavati CQL, regularni izrazi i tzv. divlje karte (engl. *wild cards*).[120] Svaka opcija pretraživanja ima svoju posebnu svrhu. Kako je navedeno na internetskoj stranici Sketch Enginea, CQL se upotrebljava za postavljanje kriterija za pozicije ili pojavnice, tj. riječi, leme,

[116] Sketch Engine. URL: https://www.sketchengine.co.uk/quick-start-guide/word-sketch-difference-lesson-2/ (20.4.2017.)
[117] Sketch Engine. URL: https://the.sketchengine.co.uk/ (20.4.2017.)
[118] Kilgarriff; Baisa; Bušta et al., 2014.
[119] Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/corpus-querying/ (20.4.2017.)
[120] Ibid.

oznake itd., regularni izrazi postavljaju kriterije za nizove znakova i mogu se upotrebljavati unutar CQL-a ili za filtriranje lista riječi, a divlje karte predstavljaju jednostavan običaj uporabe upitnog znaka (?) za bilo koji neodređeni znak te zvjezdice (*) za bilo koji broj neodređenih znakova.[121] Na internetskoj stranici Sketch Enginea nalaze se upute za korištenje navedenim metodama pretraživanja.

Još jedna značajka Sketch Enginea jest njegov distribucijski tezaurus, pod odjeljkom „Thesaurus“. Ovaj tezaurus generira se po načelu zajedničke kolokacije, odnosno, ako dvije riječi imaju mnogo zajedničkim kolokacija, pojavit će se jedna drugoj u tezaurusu.[122] Pri velikim se izračunima za sve parove riječi računa koliko kolokata dijele - oni koji ih najviše dijele nakon procesa normalizacije su ti koji se pojavljuju u tezaurusu pojedine riječi.[123] Pri pretraživanju pojedine riječi, postoji mogućnost odabira vrste riječi, ali i napredne mogućnosti poput maksimalnog broja rezultata, podešavanja izračuna podudarnosti kolokata i slično.

tea *(noun)*
British National Corpus (BNC) freq = 8,357 (74.38 per million)

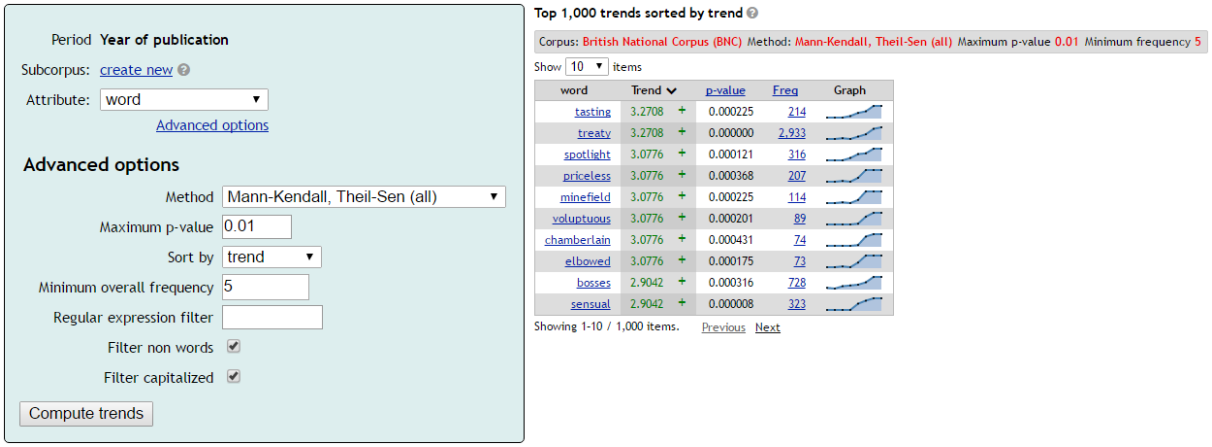
Lemma	Score	Freq		Score	Freq
coffee	0.419	6,447	glass	0.146	11,658
drink	0.311	7,112	biscuit	0.144	1,356
lunch	0.253	5,163	clothes	0.142	6,927
wine	0.253	7,069	evening	0.142	14,576
dinner	0.252	6,285	flower	0.139	7,225
milk	0.228	4,602	champagne	0.137	1,690
beer	0.227	3,616	one	0.136	12,413
meal	0.225	6,486	cos	0.134	16,731
breakfast	0.221	4,305	pie	0.133	1,533
fruit	0.208	4,937	pudding	0.133	1,050
bread	0.203	3,659	potato	0.130	2,428
cake	0.194	3,610	break	0.129	4,534
cheese	0.183	2,891	cigarette	0.129	3,332
food	0.183	20,597	vegetable	0.129	2,687
supper	0.180	1,374	bed	0.127	15,727
sandwich	0.177	1,854	salad	0.127	1,389
meat	0.177	3,705	christmas	0.126	8,557
stuff	0.170	6,667	morning	0.126	20,776
soup	0.164	1,402	everything	0.125	17,745
cream	0.160	3,236	bar	0.125	9,643
cup	0.160	13,100	oil	0.121	10,904
bottle	0.158	5,621	bath	0.121	4,443
chocolate	0.157	2,220	card	0.121	9,327
juice	0.156	1,884	pot	0.120	2,944
egg	0.155	6,003	round	0.120	22,772
sugar	0.155	3,861	sauce	0.119	1,584
dish	0.155	2,872	thing	0.119	74,810
whisky	0.150	1,847	holiday	0.119	10,163
butter	0.149	2,032	water	0.119	36,495
fish	0.148	10,496	apple	0.118	3,425

Slika 8. *Distribucijski tezaurus riječi „tea“ u BNC-u*[124]

Jedna iznimno korisna značajka je i mogućnost uvida u riječi kod kojih dolazi do promjena u učestalosti uporabe tijekom vremena. Odjeljak „Trends“ pruža dijakronijsku analizu riječi, a riječi se sortiraju prema sve učestalijoj ili sve rjeđoj uporabi.[125] Ova je značajka iznimno korisna za leksikografe koji se njome mogu služiti kako bi identificirali neologizme te za povjesničare koji se njome mogu služiti kako bi odredili vrijeme kada se određena riječ

[121] Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/corpus-querying/ (21.4.2017.)
[122] Kilgarriff; Baisa; Bušta et al., 2014.
[123] Ibid.
[124] Sketch Engine. URL: https://the.sketchengine.co.uk/ (21.4.2017.)
[125] Sketch Engine. URL: https://www.sketchengine.co.uk/user-guide/user-manual/trends/ (21.4.2017.)

počela ili prestala upotrebljavati ili kada se određena riječ iznimno učestalo upotrebljavala.[126] Nadalje, važno je napomenuti kako se navedena značajka ne može upotrebljavati za svaki korpus, nego isključivo za one korpuse koji su obilježeni za određeno vremensko razdoblje, primjerice British National Corpus, Feed Corpus, Early English Books Online Corpus i slično.[127] Odjeljak „Trends“ pojavit će se, dakle, isključivo za takve korpuse.



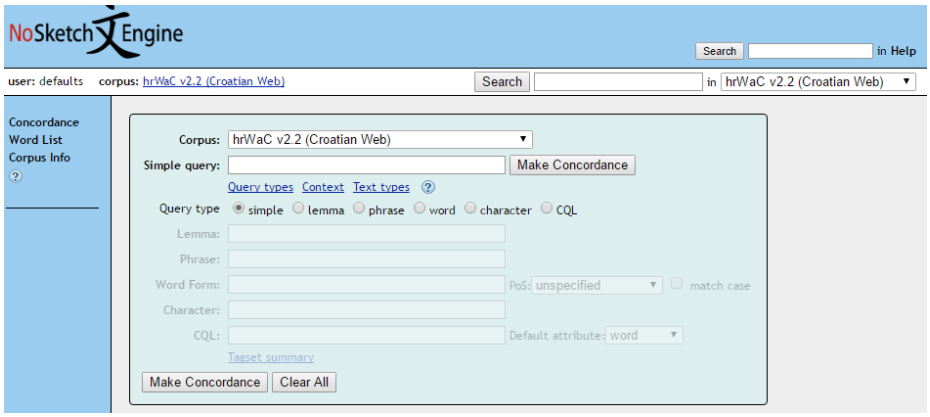
Slika 9. *Postavke za upit (lijevo) i rezultati upita (desno) za BNC*[128]

Neke od dodatnih značajki su sustav za procjenu prikladnosti rečenica da služe kao primjeri u rječniku pod nazivom GDEX (engl. *Good Dictionary Examples*), dvojezični nacrti, usporedba ključnih riječi i korpusa, pretraživanje terminologije unutar određenog područja te lokalizacija sučelja.[129] API Sketch Enginea u jednostavnom je JSON formatu, što omogućava programima pristup nacrtima riječi, kolokacijama, unosima u tezaurus te mogućnost pretraživanja terminologije u određenom dokumentu.[130]

Internetsko sučelje Sketch Enginea slično je onomu Corpusclea, vrlo je minimalistički dizajnirano i vrlo je intuitivno. Kako bi se mogli koristiti osnovnim funkcionalnostima, potrebno je imati korisnički račun koji je potrebno platiti, no postoji i mogućnost stvaranja probnog računa koji je aktivan 30 dana. S probnim je računom razina pristupa ograničena na mnogo manji broj korpusa. Testament kvaliteti i raširenosti Sketch Enginea kao alata jest i taj da je moguće odabrati 80 jezika kao parametar za daljnji odabir određenog korpusa. Važno je napomenuti kako ima i mnoštvo raznovrsnih korpusa – općih, paralelnih, učeničkih, povijesnih, korpusa dječjeg jezika te referentnih korpusa. Vrijedi spomenuti kako postoji i besplatna inačica

[126] Sketch Engine. URL: https://www.sketchengine.co.uk/user-guide/user-manual/trends/ (23.4.2017.)
[127] Ibid.
[128] Sketch Engine. URL: https://the.sketchengine.co.uk (23.4.2017.)
[129] Kilgarriff; Baisa; Bušta et al., 2014
[130] Ibid.

softvera koja ima manju funkcionalnost i otvorenog je izvornog koda, a naziva se NoSketch Engine. Razvija se u Centru za obradu prirodnog jezika pri Masarykovom sveučilištu u Brnu. Upravo je NoSketch Engine (Bonito) softverska podloga za Hrvatski nacionalni korpus (HNK) te hrWac. hrvatski *web* korpus koji je ujedno i najveći računalni korpus hrvatskoga jezika.



Slika 10. *Sučelje korpusa hrWac*[131]

4.3. BlackLab

BlackLab je tražilica za pretraživanje, otvorenog izvornog koda, koju razvija Institut za nizozemsku leksikologiju (engl. *Institute of Dutch Lexicology*) s ciljem stvaranja brzog i značajkama bogatog sučelja za pretraživanje korpusa povijesnih i suvremenih tekstova nizozemskog jezika.[132] Izdan je pod licencom Apache 2.0 2012. godine i otada mu je broj korisnika i suradnika rastao.[133]

Dakle, BlackLab je tražilica za pretraživanje dizajnirana na temelju softverske biblioteke za dohvaćanje informacija Apache Lucene, a primaran joj je cilj omogućiti složene upite i brzu obradu nad velikim tijelima teksta. Primarna svrha BlackLaba jest da služi kao pomagalo lingvistima koji tragaju za uzorcima u velikim tijelima teksta koji je obilježen lingvističkim elementima kao što su POS, paragrafi, rečenice i slično; također, upotrebljava se u povijesnim istraživanjima te u razvoju umjetne inteligencije.[134]

Prema službenoj internetskoj stranici tražilice, njezine su glavne značajke sljedeće:

1) Tekst s indeksiranim oznakama, za pretraživanje određenih lema ili vrsta riječi (POS);

[131] NoSketch Engine. hrWaC. URL: http://nl.ijs.si/noske/all.cgi/first_form?corpname=hrwac;align= (25.4.2017.)
[132] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/index.html (25.4.2017.)
[133] Ibid.
[134] Ibid.

2) Brzina i skalabilnost;
3) Pretraživanje pomoću regularnih izraza;
4) Višestruki ulazni formati, u BlackLabov je indeks jednostavno uvesti mnoštvo formata, poput formata TEI, Alto, FoLiA ili Sketch Engine. Također je jednostavno dodati podršku za određeni format, ukoliko ga tražilica ne podržava;
5) Višestruki jezici upita, kao što su Corpus Query Language i parser upita Lucene, a postoji i vrlo osnovna podrška za jezik upita SRU Common Query Language. Uz to, postoji jednostavna opcija dodavanja dodatnog jezika upita;
6) Jednostavan za uporabu, API je dizajniran prema „načelu najmanjeg iznenađenja“;
7) Pretraživanje unutar XML oznaka koje se pojavljuju u tekstu:
8) Brzo grupiranje i sortiranje velikih setova rezultata prema nekoliko kriterija, uključujući kontekst (traženi tekst, lijevi i desni kontekst);
9) Precizno isticanje (engl. *highlighting*) rezultata u dokumentu i brzi pregled rezultata u formatu KWIC;
10) Aktivan je projekt s otvorenim izvornim kodom, napisan u programskom jeziku Java i dizajniran na temelju softverske biblioteke za dohvaćanje informacija Apache Lucene.[135]

Za verziju BlackLab 2.0 planira se implementirati niz dodatnih značajki, poput integracije s pretraživačkom platformom Apache Solr i/ili tražilicom Elasticsearch, dodavanje podrške za distribuirano pretraživanje i slično.[136] Pri izgradnji vlastitog sustava temeljenog na BlackLabu, moguće je odabrati između dvije mogućnosti, BlackLab Server i BlackLab Core. BlackLab Server, kao što mu i samo ime govori, predstavlja internetski servis koji se može upotrebljavati putem bilo kojeg programskog jezika, a koji nudi jednostavno sučelje u REST formatu; iako je, takoreći, najfleksibilnija Javina biblioteka, to znači da se mora upotrebljavati programski jezik koji se može interpretirati i kompajlirati u Java virtualnom stroju (engl. *Java virtual machine*).[137] Priručnik na internetskoj stranici prvenstveno sadrži upute za instalaciju inačice BlackLab Core preuzimanjem datoteka sa središnjeg repozitorija Maven, preuzimanjem unaprijed izgrađenih binarnih datoteka s navedene stranice te izgradnja sustava uporabom izvornog koda koji se može preuzeti s hosting servisa GitHub.

[135] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/features.html (25.4.2017.)
[136] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/roadmap.html (25.4.2017.)
[137] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/getting-started.html#server-or-core (25.4.2017.)

Kako bi se građa, odnosno podaci mogli pretraživati pomoću BlackLaba, moraju biti u podržanom formatu te ih BlackLab mora indeksirati.[138] Kako je već ranije navedeno, BlackLab podržava niz formata, od kojih u najpoznatiji XML formati – TEI i FoLiA (engl. *Format for Linguistic Annotation*), a slučaju da korisnik posjeduje podatke u nepodržanom formatu, oni se mogu pretvoriti u drukčiji format pomoću alata OpenConvert koji je dostupan putem instalacijskog paketa ili kao internetsko sučelje.[139] Važno je napomenuti kako je potreban CLARIN račun za korištenje internetskim sučeljem. Korisnik gradi svoj indeks pokretanjem elementa IndexTool i potrebno je navesti direktorij, ulazne datoteke i format ulaznih datoteka, a iznimno je korisna i značajka QueryTool kojom se ispituje indeks.[140]

Što se tiče jezika upita, BlackLab podržava već spomenuti jezik Corpus Query Language koji je prvi put uveden projektom IMS Open Corpus Workbench budući da predstavlja već standardan, ali i iznimno učinkovit način pretraživanja korpusa.[141] Iako se Corpus Query Language upotrebljava kod alata IMS Open Corpus Workbench i Sketch Engine, važno je napomenuti kako ipak postoje određene razlike u jeziku za sva tri projekta.[142] Na internetskoj stranici projekta BlackLab postoje detaljne upute za uporabu navedenog jezika koje korisnika vode korak po korak kroz različite upite.

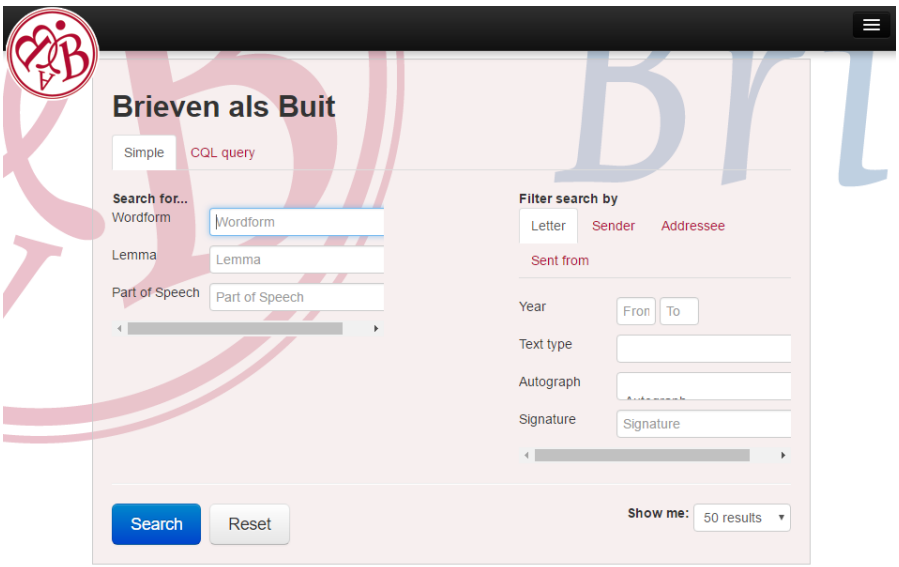
BlackLabom se koriste mnoge renomirane institucije. Naravno, institut na kojem se razvija koristi se njime za vlastite aplikacije pretraživanja korpusa nizozemskog jezika, poput korpusa Brieven als Buit, Gysseling i Hedendaags Nederlands.[143] Nadalje, BlackLab je također u uporabi, između ostalih, na Sveučilištu u Tilburgu za korpus OpenSonar, Institutu Meertens za korpus Fesli, Institutu Allen za umjetnu inteligenciju koji se njime koristi u alatu za ekstrakciju znanja IKE, Sveučilištu u Radboudu za korpuse Basilex i OpenChechen te Virtualnom institutu za afrikaans za korpus Korpusportaal.[144] Kako bi pristup pojedinim korpusima bio moguć, potrebno je imati CLARIN račun, a za neke je potrebno platiti kako bi se stvorio korisnički račun.

Jedan od korpusa kojima je pristup omogućen jest korpus Brieven als Buit, koji sadrži zbirku povijesnih pisama koje su slali i primali pomorci u razdoblju od 17. do 19. stoljeća.[145]

---

[138] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/getting-started.html (25.4.2017.)
[139] Ibid.
[140] Ibid.
[141] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/corpus-query-language.html (25.4.2017.)
[142] Ibid.
[143] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/who-uses-blacklab.html (25.4.2017.)
[144] Ibid.
[145] GitHub. BlackLab. URL: http://inl.github.io/BlackLab/blacklab-in-action.html (25.4.2017.)

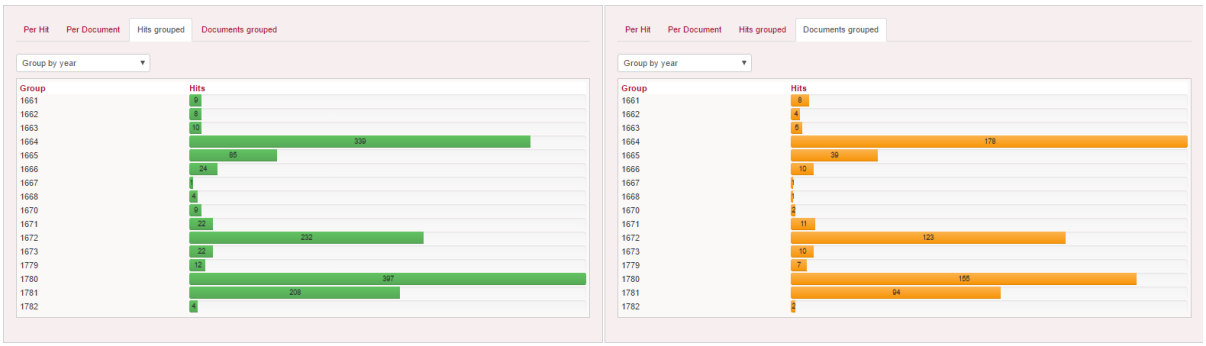
Sučelje korpusa vizualno je vrlo jednostavno i intuitivno. Postoje dvije glavne mogućnosti pretraživanja korpusa – jednostavna, pomoću grafičkog sučelja i ona složenija, pomoću upita formuliranih u CQL-u. Kod jednostavnog pretraživanja, korisnik ima nekoliko mogućnosti. Korpus se tako može pretraživati po pojavnici, lemi, ali i vrsti riječi (POS). Ako korisnik odabere opciju pretraživanja pomoću CQL-a, potrebno je u prazno polje unijeti željeni upit. Na desnoj se strani sučelja nalazi dodatna mogućnost filtriranja pretrage. Četiri su glavna filtra: a) pismo, gdje se može specificirati razdoblje unutar kojeg je poslano, vrsta pisma, autogram i potpis; b) pošiljatelj, gdje se može specificirati ime pošiljatelja, spol pošiljatelja, klasa pošiljatelja, dob pošiljatelja, njegova regija prebivališta te vrsta odnosa s adresatom; c) adresat, gdje se može specificirati ime adresata, mjesto adresata, država adresata, regija adresata te brod adresata i d) lokacija s koje je pismo poslano, gdje je moguće specificirati mjesto, državu, regiju i brod. Nadalje, moguće je odabrati i količinu prikazanih rezultata na jednoj stranici.



Slika 11. *Početna stranica korpusa Brieven als Buit*[146]

Rezultati pretrage mogu se prikazati prema odgovarajućem pronalasku u obliku konkordancija i prema dokumentu u kojem se pronađeni rezultat nalazi. Jedna iznimno korisna značajka je i prikaz grupiranih rezultata, gdje se mogu grupirati odgovarajući pronalasci, ali i sami dokumenti. Pronalasci se mogu grupirati prema 9 kriterija, primjerice, naslovu dokumenta, godini, lemi i slično, a dokumenti se mogu grupirati prema 4 kriterija, broju pronalazaka, godini, desetljeću i autoru.

[146] Brieven als Buit. URL: http://brievenalsbuit.inl.nl/zeebrieven/page/search (25.4.2017.)



Slika 12. *Grupiranje pronalazaka (lijevo) i dokumenata (desno) za lemu „schip“*[147]

4.4. IMS Open Corpus Workbench (CWB)

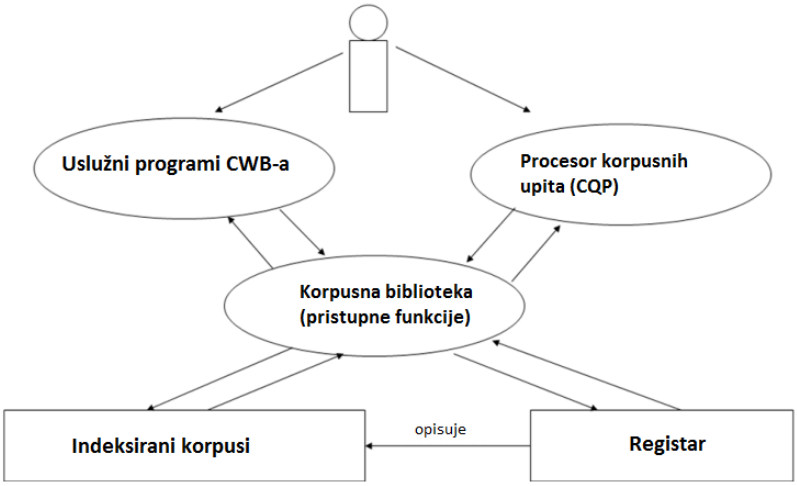
Ovaj je alat razvijen na Institutu za obradu prirodnog jezika pri Sveučilištu u Stuttgartu. Nastao je kao dio projekta pod nazivom TC Project (engl. *Text Corpora and Tools for Their Extraction*) prvenstveno kako bi se pružila softverska platforma za rad u područjima leksikografije i terminologije.[148] Projekt je započeo u ranim 1990-ima i, unatoč njegovu opsegu, u potpunosti je razvijen na već spomenutom institutu, što je još u 1970-ima uspostavljeno kao najbolja praksa za moćne korpusne alate koji pripadaju tzv. prvoj generaciji konkordancijskog softvera, naziv koji su skovali lingvisti McEnery i Hardie u svojoj knjizi „Corpus Linguistics: Method, Theory & Practice“.[149] Alat je isprva bio primarno namijenjen internoj uporabi na institutu i nije bio otvorenog izvornog koda, no kasnije je razvoj prešao na model otvorenog izvornog koda i postao je dostupan većem broju korisnika pod licencom GNU General Public Licence.[150]

Dakle, IMS Open Corpus Workbench (CWB) zbirka je alata otvorenog izvornog koda koji služe za upravljanje i pretraživanje velikih korpusa tekstova, opsega od 10 milijuna do 2 milijarde riječi, s lingvističkim obilježjima.[151] Dakle, ovaj je moćni i fleksibilni sustav za indeksiranje i pretraživanje korpusnih podataka dizajniran za uporabu s obilježenim podacima odnosno korpusima s višestrukim lingvističkim obilježjima na razini riječi.[152] Alatom se koristi u lingvistici temeljenoj na podacima za ekstrakciju lingvističkog znanja iz tekstualnih izvora ili provjeravanje lingvističkih teorija pomoću velikih tekstova, u leksikografiji za pribavljanje

[147] Brieven als Buit. URL: http://brievenalsbuit.inl.nl/zeebrieven/page/search (25.4.2017.)
[148] IMS. IMS Corpus Workbench (CWB). URL: http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html (28.4.2017.)
[149] Evert; Hardie, 2011.
[150] Ibid.
[151] Sourceforge. IMS Open Corpus Workbench (CWB). URL: http://cwb.sourceforge.net/index.php (28.4.2017.)
[152] Evert; Hardie, 2011.

dokaza iz korpusa za leksičke opise te u terminologiji za ekstrakciju termina i iskorištavanje terminoloških izvora.[153] U početku, CWB je zbog hardverskih zahtjeva bio dostupan isključivo na komercijalnoj inačici operativnog sustava Unix, no to se promijenilo početkom 21. stoljeća kad su osobna računala s različitim operativnim sustavima postala hardverski dovoljno moćna za pokretanje alata; stoga, počela se razvijati kompatibilnost sa svim sustavima, a glavne razvojne platforme za alat su Linux i Mac OS X.[154]

U znanstvenom članku „Twenty-first century Corpus Workbench: Updating a query architecture for the new millenium“, skupina je autora predstavila detaljan pregled svakog aspekta alata, uključujući i njegovu arhitekturu. Prema članku, temeljna jedinica podataka u CWB-u je korpus, a arhitektura mu počiva na skupini kodiranih i indeksiranih korpusa te dvije skupine metoda kojima se pristupa navedenima korpusima. Unutar sustava, korpus čini skupina binarnih datoteka koje predstavljaju indeksirane tekstove, lingvistička obilježja i strukturalne oznake, a koje su pohranjene na disku, dok su sve upravljačke informacije (engl. *housekeeping information*) pohranjene u jednoj registarskoj datoteci (engl. *registry file*).[155]



Slika 13. *Arhitektura CWB-a*[156]

Nadalje, članak navodi kako su registarske datoteke za sve indeksirane korpuse pohranjene u korpusnom registru (engl. *corpus registry*). Pristup niske razine podacima indeksiranog korpusa omogućuje skupina funkcija pod nazivom korpusna biblioteka (engl. *Corpus library – CL*), koja

[153] IMS. IMS Corpus Workbench (CWB). URL: http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html (28.4.2017.)
[154] Evert; Hardie, 2011.
[155] Ibid.
[156] Ibid.

se može upotrebljavati neovisno o ostatku CWB-a kako bi se ostale aplikacije mogle koristiti njezinim API-jem. Funkcijama korpusne biblioteke obično se pristupa putem uslužnih programa CWB-a (engl. *CWB utilities*) ili procesora korpusnih upita (engl. *Corpus Query Processor – CQP).* Uslužni programi CWB-a skupine su programa kojima se upravlja putem naredbenog retka radi upravljanja indeksiranim korpusima, dok je procesor korpusnih upita zapravo alat za konkordanciju koji na temelju korpusne biblioteke pretražuje indeksirani korpus pomoću jezika upita. Arhitektura je temeljno slična u slučaju da se CWB upotrebljava kao pozadinski sustav (engl. *back-end system*), no u tom slučaju pristupni softver (engl. *front-end software*) komunicira s uslužnim programima CWB-a i korpusnim procesorom upita, a krajnji korisnik komunicira s pristupnim softverom.[157]

Nadalje, u znanstvenom je članku također predstavljen proces obrade ulaznog formata. Slično kao u Sketch Engineu ili Corpuscleu, format korpusnih datoteka također mora biti vertikaliziran. To je ključno kako bi navedeni uslužni programi mogli indeksirati korpusne datoteke. Radi se o vertikaliziranim datotekama čiji su stupci odvojeni tabulatorom, gdje svaki stupac predstavlja određeno obilježje na razini riječi, uključujući i sami oblik riječi, a svaki red predstavlja jednu pojavnicu, bilo riječ ili interpunkcijski znak.[158]

```
<sentence>
It          PP          it
was         VBD         be
<np head="elephant">
an          DT          a
elephant    NN          elephant
</np>
.           SENT        .
</sentence>
```

Slika 14. *Primjer ulaznog formata korpusne datoteke u CWB-u*[159]

Pojavnice su zapravo temeljne jedinice unutar CWB-a, budući da CWB može obrađivati, odnosno indeksirati tekst koji je minimalno obilježen – tokeniziran, a pri indeksiranju se svakoj pojavnici dodjeljuje broj korpusne pozicije, počevši od 0. XML oznake poput <*sentence*> ne smatraju se pojavnicama. Upravo se numerizacijom korpusnih položaja ovaj sustav razlikuje od onih temeljenih na relacijskim bazama podataka, gdje su pojavnice, u pravilu, neodređene

[157] Evert; Hardie, 2011.
[158] Ibid.
[159] Ibid.

redoslijedom. Nadalje, numerirane korpusne pozicije igraju veliku ulogu u načinu kako CWB modelira indeksirane korpuse, budući da se korpusi indeksiraju kao skupine atributa vezanih za brojeve korpusnih pozicija. Postoji više vrsta atributa. Najosnovnija vrsta atributa, koja je zapravo obilježje na razini pojavnice, naziva se pozicijski atribut, ili p-atribut (engl. *positional attribute, p-attribute*); dakle, ona predstavlja bilo koji element korpusa gdje je svaka korpusna pozicija povezana s jednom vrijednošću, dakle, i same pojavnice mogu biti p-atributi. Osim p-atributa, tu su strukturalni atributi ili s-atributi (engl. *structural attribute, s-attribute*) i atributi usklađivanja ili a-atributi (engl. *alignment attribute, a-attribute*). Strukturalni atributi povezani su s neprekinutim rasponima pojavnica (engl. *token range*) koje su unutar sustava zapravo parovi korpusnih pozicija s početnom i završnom točkom; drugim riječima, to su XML oznake. Namjena atributa usklađivanja jest usklađivanje prevedenoga teksta na razini rečenice na način da se kodira međuodnos između područja teksta u dva korpusa koja predstavljaju istu vrstu podataka u dva različita jezika.[160]

Što se tiče jezika upita, CWB unutar korpusnog procesora upita koristi se već spomenutim jezikom upita Corpus Query Language (CQL). Međutim, autori članka „Twenty-first century Corpus Workbench: Updating a query architecture for the new millenium“ navode kako se, preferira termin CQP Query Language ili CQP-syntax budući da se naziv CQL dugo vremena upotrebljava za standard jezika upita koji nije namijenjen korpusima, ali i kao oznaka za temeljni jezik upita softvera SARA/Xaira. Naravno, jezik upita je prilagođen samom sustavu i sadržava značajke svojstvene isključivo njemu. Jedna od glavnih značajki jest da ta da upotrebljava uzorke regularnih izraza na dvije razine; na razini nizova znakova za oblike riječi i vrijednosti obilježavanja te na razini sekvenci pojavnica.[161] Na internetskoj se stranici CWB-a može preuzeti opširan priručnik o uporabi jezika upita, a dio toga opisan je i u navedenom članku.

U članku „Character encoding in corpus construction” navode kako je općeprihvaćeno upotrebljavati Unicode standard za kodiranje znakova budući da se njime može pokriti široki dijapazon ne-zapadnjačkih jezika, ali i zbog toga što se njegovom inačicom može ostvariti kompatibilnost unatrag sa standardom ASCII koji se dugo vremena upotrebljavao za alfabetične jezike poput engleskog.[162] Inačica Unicodea, UTF-8, danas je standard za kodiranje korpusnih podataka zbog svoje, takoreći, prilagodljivosti. Kada je CWB razvijen, niti Unicode niti ASCII

[160] Evert; Hardie, 2011.
[161] Ibid.
[162] McEnery; Xiao, 2005.

nisu bili standardi koji su bili vrlo rašireni, no od verzije 3.2 i nadalje, CWB ima potpunu kompatibilnost sa standardom UTF-8, zahvaljujući projektu Textométrie, sustavu za analizu korpusa.[163]

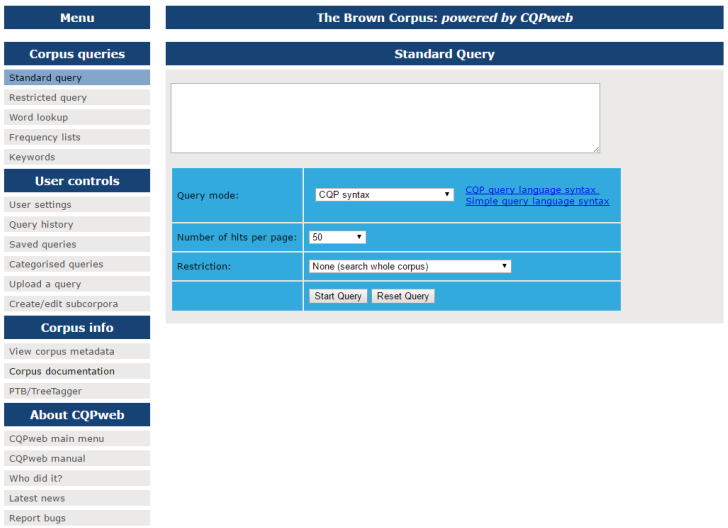
Vrlo je važno napomenuti kako CWB nije namijenjen početnicima, već korisnicima koji su iskusni na području računalne korpusne lingvistike budući da moraju biti upućeni u način rada sustava i upravljanje putem naredbenih redaka.[164] CWB se zapravo sastoji od tri različita programska paketa, a to su sljedeći:

1) CWB core, uključuje korpusnu biblioteku (CL) niske razine, uslužne programe CWB-a i korpusni procesor upita (CQP). CWB core vrlo je moćan alat za analizu korpusa, dugotrajan i iznimno utjecajan u području računalne korpusne lingvistike. Primjerice, ulazni format i jezik CWB-a upotrebljava već spomenuti pozadinski softver Sketch Enginea, Manatee, a sličnu verziju jezika upita upotrebljava sustav Poliqarp;
2) Sučelje CWB/Perl, koje se dalje grana na zasebne pakete u Perlu – CWB, CWB-CL i CWB-Web;
3) CQPweb.[165]

Nadalje, potrebno je naglasiti i prilagodljivost CWB-a. Osim što može funkcionirati kao pozadinski sustav za arhitekturu klijent/server, može se i upotrebljavati za korpuse manjeg opsega te je tako dostupan korisnicima za preuzimanje i instalaciju na osobnim računalima gdje, pak, postoji mogućnost odabira između naredbenog sučelja te grafičkog sučelja temeljenog na HTML-u.[166] Također, CWB je najviše korisnicima dostupan putem internetskih sučelja koja su korisnicima nerijetko preferirani način pristupa korpusima budući da su prilagođeniji korisnicima, a neka od sučelja su BNCweb, Leeds CQP, IntelliText te sučelje projekta VISL.[167]

U verziji CWB 3.0, u softverski je paket dodan već navedeni komponent CWB/Perl čija je namjena bila olakšati stvaranje internetskih sučelja ili sučelja s naredbenim retkom po želji, a najvažniji je dio komponenta tzv. *Common Elementary Query Language (CEQL)* – pojednostavljeni jezik upita koji omogućuje pristup najkorištenijim značajkama CQP-a, no u obliku koji je mnogo pristupačniji i razumljiviji početnicima.[168,169]

[163] McEnery; Xiao, 2005.
[164] Ibid.
[165] Ibid.
[166] Ibid.
[167] Ibid.
[168] Sourceforge. IMS Open Corpus Workbench (CWB). URL: http://cwb.sourceforge.net/index.php (30.4.2017.)
[169] Evert; Hardie, 2011.



Slika 15. *Sučelje za skupinu korpusa Brown*[170]

Standardno sučelje za CWB zove se CQPweb; neovisno je o korpusu i dizajnirano je na temelju sučelja BNCweb, s dodatnim značajkama implementiranim putem relacijske baze podataka MySQL, a također podržava i korpuse čiji su tekstovi kodirani standardom UTF-8.[171] Primjer toga je sučelje za skupinu korpusa Brown koje nudi pregršt značajki. Dizajn sučelja je, kao i u prijašnjim alatima, vrlo intuitivan i minimalistički. Upiti se mogu postavljati na standardan i ograničen način, i moguće je odabrati opciju pretraživanja s pomoću jezika upita ili jednostavnog upita. Vrlo je korisno što se pokraj polja za unos upita nalaze dvije poveznice na priručnike o uporabi jezika CQP Query Language i pojednostavljenog jezika upita. Prema zadanim postavkama, rezultati pretraživanja prikazuju se u KWIC formatu, no to je moguće promijeniti u postavkama korisnika na način da se prikazuju kao istaknute riječi u rečenicama. Korisna značajka je i „Word lookup“, koja korisniku omogućuje da pretragom određenog oblika riječi dobije prikaz broja pojavljivanja te riječi, riječi koje sadrže traženi oblik, riječi koje počinju traženim oblikom te riječi koje završavaju traženim oblikom. Primjerice, odabere li se opcija da se prikažu riječi koje započinju traženim oblikom, pretragom riječi „house“ korisnik dobije rezultate poput „houses“, „households“ i „housekeeper“. Rezultati se mogu preuzeti u formatu .txt. Također je moguće dobiti liste čestote s raznim mogućnostima postavki, poput odabira temeljne jedinice pretrage (oblik riječi, lema, vrsta riječi), načina sortiranja liste, broja rezultata na stranici i slično. Odjeljak „Keywords“ omogućuje korisniku prikaz liste ključnih riječi koja se generira usporedbom frekvencijskih listi stvorenih za neki drugi potkorpus.

[170] Brown Corpus. URL: https://corpling.uis.georgetown.edu/cqp/brown/ (30.4.2017.)
[171] Evert; Hardie, 2011.

No.	Search result	No. of occurrences	Percent
1	House_NN	396	46.81%
2	House_NP	186	21.99%
3	houses_NNS	81	9.57%
4	household_NN	32	3.78%
5	Housed_VVN	12	1.42%
6	schoolhouse_NN	11	1.3%
7	house_VV	10	1.18%
8	farmhouse_NN	8	0.95%
9	housekeeping_NN	6	0.71%
10	Parkhouse_NP	5	0.59%
11	Housewives_NNS	5	0.59%
12	housewife_NN	4	0.47%
13	clubhouse_NN	4	0.47%
14	warehouse_NN	4	0.47%

Slika 16. *Prikaz rezultata za riječ „house" s postavkom da rezultati sadrže traženi oblik riječi (odjeljak Word lookup)*[172]

Vrlo korisna značajka sučelja jesu i korisničke kontrole, odnosno „User controls". U odjeljku „User settings" moguće je podesiti mogućnosti prikaza, mogućnosti kolokacija, mogućnosti preuzimanja, definiranje makronaredbi i definiranje profila korisnika. Nadalje, moguće je pogledati povijest upita korisnika, pregledati pohranjene i kategorizirane upite, prenijeti datoteku od koje se može pohraniti upit te stvoriti i uređivati potkorpus. Naravno, dostupne su i informacije o korpusu poput metapodataka i dokumentacija, a korisnik može i nešto više saznati o samom CQPwebu – načinu korištenja, popis korpusa koji se njime koristi, tko ga je dizajnirao i slično.

4.5. Xaira

Xaira (engl. *XML-Aware Indexing and Retrieval Architecture*) je alat otvorenog izvornog koda za analizu i pretraživanje korpusa koji se temelji na poznatom alatu razvijenom na Odjelu za računalstvo pri Sveučilištu u Oxfordu – SARA (engl. *SGML-Aware Retrieval Application*), koji je primarno razvijen za Britanski nacionalni korpus (BNC).[173] Xairu su razvili Lou Burnard i Tony Dodd i distribuira se pod licencom GNU General Public Licence.[174] Ono po čemu se Xaira razlikuje od SARA-e jest da nije isključivo vezana uz BNC, već predstavlja

[172] Brown Corpus. URL: https://corpling.uis.georgetown.edu/cqp/brown/index.php?thisQ=lookup&uT=y (30.4.2017.)
[173] Xiao; Hu, 2015.
[174] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X01 (3.5.2017.)

tražilicu koja se može primijeniti na bilo koji korpus sačinjen od ispravno oblikovanih XML dokumenata.[175]

Xaira je analitički alat napisan u programskom jeziku C++, a namijenjen je za indeksirane korpuse s nekoliko vrlo korisnih značajki. Kako je već navedeno, Xaira je u potpunosti zasnovana na XML-u. Dizajnirana je prema alatu SARA i samim time ima i mogućnost prebacivanja SGML-a u XML. Ona indeksira riječi i XML strukturu, što znači da može pretraživati obilježeni tekst i XML oznake bilo koje zbirke dokumenata u XML-u, a postoji i mogućnost dodavanja oznaka, ukoliko je to potrebno.[176] Bitno je napomenuti kako je najoptimalnija vrsta za ovaj alat upravo TEI XML, a neispravno oblikovane XML ili SGML dokumente Xaira odbacuje. Xaira se XML oznakama koristi u različite svrhe:

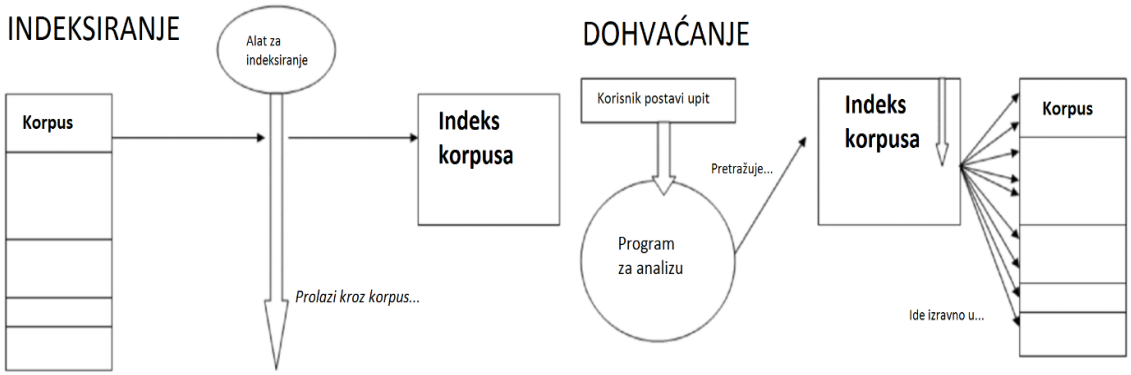
1) Za tokenizaciju teksta;
2) Za određivanje opsega pretraga;
3) Za određivanje formata prikaza rezultata;
4) Kao objekt pretrage, primjerice, za prebrojavanje elemenata ili oznaka određene vrste;
5) Za unutarnje upravljačke datoteke (engl. *internal control files*);
6) Pri pohranjivanju rezultata u vanjske datoteke za ponovnu uporabu od strane drugih aplikacija.[177]

Jedna od važnijih značajki alata jest i potpuna podrška standarda Unicode, što znači da se njime može koristiti za pretraživanje i prikaz teksta na bilo kojem jeziku, uz uvjet da korisnik u sustavu ima instalirane odgovarajuće fontove.[178] Kao što je već navedeno, Xaira je besplatan alat otvorenog izvornog koda koji se može preuzeti s hosting servisa SourceFourge (*http://xaira.sourceforge.net)* i ima otvorenu modularnu arhitekturu koja je dizajnirana kako bi se upotrebljavala kao internetski servis, ali i kao samostalna aplikacija na operativnom sustavu.[179]

Oxfordova internetska stranica Odjela za računalstvo navodi kako je arhitektura Xaire doživjela određene preinake tijekom vremena, ponajviše u pogledu dizajniranja modela pod nazivom „Xaira Object Model“, koji definira niz objekata i metoda za predstavljanje i pretraživanje velike količine lingvističkih podataka, a kojega implementira poslužiteljski

[175] Xiao; Hu, 2015.
[176] TEI Wiki. Xaira. URL: https://wiki.tei-c.org/index.php/Xaira (3.5.2017.)
[177] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X01 (3.5.2017.)
[178] Xiao; Hu, 2015.
[179] TEI Wiki. Xaira. URL: https://wiki.tei-c.org/index.php/Xaira (3.5.2017.)

program Xaire (engl. *server program*). Ovaj je model uveden kako bi se Xaira mogla upotrebljavati na većem broju operativnih platformi i putem klijentskih programa koji imaju različite funkcionalnosti. Program za indeksiranje (engl. *indexer program*) stvara indekse neovisne o operativnoj platformi iz zbirke XML dokumenata, koje poslužiteljski program može upotrebljavati. Klijentski programi mogu pristupiti poslužitelju Xaire putem API-ja kojeg upotrebljava klijentski program za operativni sustav Windows, ili putem protokola XMLRPC ili SOAP.[180] Klijentski program Xaire služi za kreiranje pretrage, dok poslužitelj služi za pretragu indeksa i pronalazak rješenja, no klijentski se program može upotrebljavati i kao običan alat za konkordanciju budući da korisnik nikada ne komunicira izravno s poslužiteljem.[181] Korisničko sučelje Xaire vrlo je jednostavno za dizajniranje korpusa i uporabu alata za indeksiranje, a njezin je klijent sofisticirani sustav za analizu korpusa u kojemu je moguće prikazivati liste riječi, konkordancije, kolokacije i slično.[182] Alat za indeksiranje i server dizajnirani su tako da se mogu pokrenuti na svim operativnim platformama, dakle Linuxu, MacOS-u i Windowsu.[183] Važno je napomenuti kako je primarna svrha trenutne inačice alata Xaira rad s korpusima koji su spremljeni na korisnikovom računalu, no alat je dizajniran i za pristup korpusima spremljenim na mrežnom poslužitelju i zato se Xaira ponekad naziva klijentskim programom.[184]

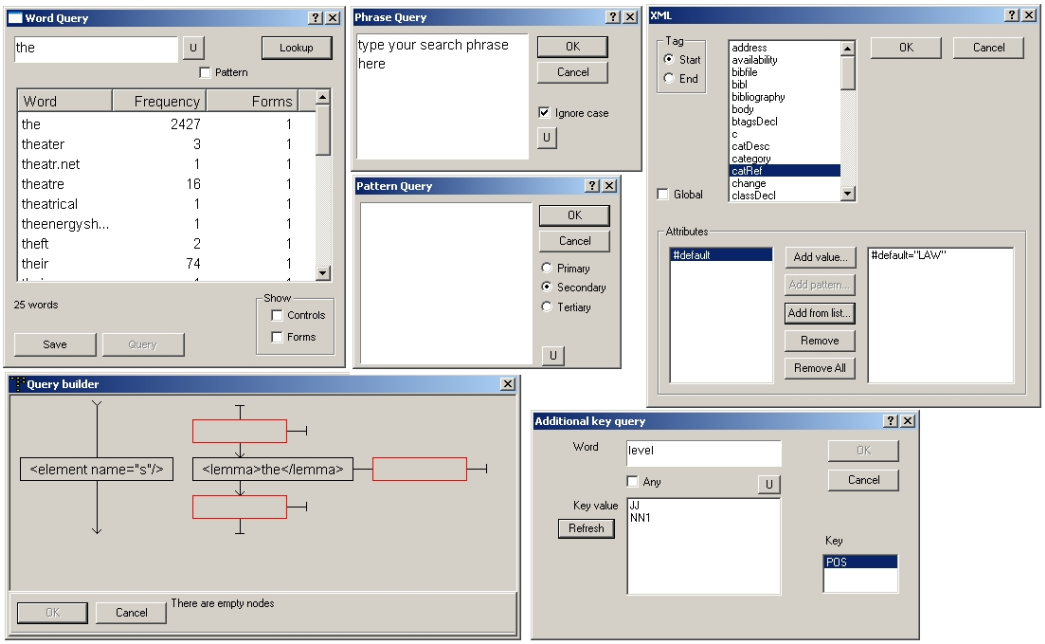


Slika 17. *Prikaz procesa indeksiranja i dohvaćanja informacija u alatu Xaira*[185]

[180] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/ (4.5.2017.)
[181] Hardie. Introduction to Xaira Part One: All about Xaira. URL: http://slideplayer.com/slide/793318/ (4.5.2017.)
[182] Ibid.
[183] TEI Wiki. Xaira. URL: https://wiki.tei-c.org/index.php/Xaira (4.5.2017.)
[184] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X01 (4.5.2017.)
[185] Hardie. Introduction to Xaira Part One: All about Xaira. URL: http://slideplayer.com/slide/793318/ (4.5.2017.)

Jezik upita kojeg upotrebljava Xaira naziva se XML Query Language (CQL), i s pomoću njega izvode se i izražavaju svi upiti u alatu. Xaira ima bogatu ponudu u pogledu pretraživanja, pa je tako moguće izvesti sljedeće pretrage:

1) Najlakši način pretrage jest upisati riječ ili frazu u polje „Quick Query“, što je jednako uporabi dijaloškog okvira „Phrase Query“;
2) Pretraživanje različitih oblika riječi u korpusu izvodi se putem dijaloškog okvira „Word Query“;
3) Pretraživanje riječi s dodatnim ključevima, poput POS kodova izvodi se putem „Addkey Query“;
4) Pretraživanje uzoraka riječi izvodi se pomoću dijaloškog okvira „Pattern Query“;
5) Pretraživanje početnih ili završnih XML oznaka izvodi se putem dijaloškog okvira „XML Query“;
6) Složeni upiti s različitim vrstama pretrage izvode se putem vizualnog sučelja „Query Builder“;
7) Pretraživanje se može vršiti i upisivanjem naredbi u CQL-u, putem dijaloškog okvira „CQL Query“.[186]

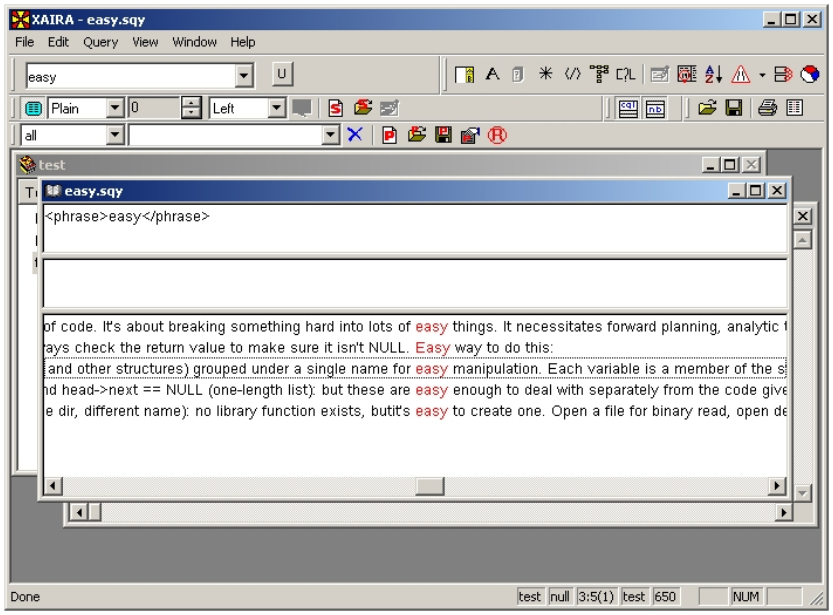


Slika 18. *Prikaz različitih dijaloških okvira mogućnosti pretraživanja u alatu Xaira*[187]

[186] University of Oxford IT Services. Xaira. URL http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X01 (6.5.2017.)
[187] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X02#simple (6.5.2017.)

Klijentski program alata Xaira ima korisniku razumljivo i intuitivno sučelje. Na primjeru operativnog sustava Windows, glavni se prozor ne razlikuje od standardnih prozora programa ove vrste. Ovaj se glavni prozor pojavljuje se pri otvaranju programa i on predstavlja početnu točku za sve daljnje aktivnosti korisnika. Dakle, prozor ima standardan izgled – naslovnu traku, traku izbornika i nekoliko alatnih traka. Kao što je vidljivo na Slici 19., glavni prozor može sadržavati unutar sebe više manjih prozora. Ukupno četiri vrste prozora mogu se pojaviti unutar glavnog prozora:

1) Korpusni prozor (engl. *corpus window*) koji sadrži informacije o tekstovima u korpusu koji se pretražuje;
2) Prozor za upite (engl. *query window*) koji sadrži rješenje za upit koji je postavljen;
3) Prozor za pregledavanje (engl. *browser window*) koji sadrži tekst datoteke iz korisnikova korpusa i tako omogućuje korisniku da pogleda originalni tekst i XML oznake;
4) Prozor za skriptu (engl. *script window*) koji sadrže područje u koje se može upisivati skripta za Xairu.[188]



Slika 19. *Prikaz glavnog prozora u klijentskom programu Xaire na OS Windows*[189]

Na internetskoj stranici http://projects.oucs.ox.ac.uk/xaira/ nalazi se opširan opis uporabe programa i jezika upita.

[188] University of Oxford IT Services. Xaira. URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X-01 (6.5.2017.)
[189] Ibid.

4.6. Poliqarp

Poliqarp (engl. *POLinterpretation Indexing Query and Retrieval Processor*) je softver za upravljanje korpusima koji je razvijen na Institutu za računalstvo pri Poljskoj akademiji znanosti u sklopu projekta kojeg je financirao Državni odbor za znanstvena istraživanja, a njegova se funkcionalnost djelomično temelji na sustavu CQP, odnosno IMS Open Corpus Workbench, s određenim dodatnim značajkama.[190] Alat se distribuira pod licencom GNU General Public Licence, a njegova prva inačica objavljena je 2006. godine.[191] Cilj navedenog projekta bio je stvoriti javno dostupan korpus poljskog jezika s morfosintaktičkim obilježjima.[192] Dakle, Poliqarp je softver, odnosno tražilica za upravljanje velikim korpusima i njihovo pretraživanje koja ima nekoliko korisnih značajki. Ovaj je alat univerzalan jer upotrebljava eksterno definirani skup oznaka i interno kodiranje u standardu UTF-8, što znači da korpusi nisu ograničeni jezikom.[193] Alat je moćnih performansi, ima svoj jezik upita, interpretira regularne izraze na razini znakova u riječima te riječima u rečenicama ili paragrafima, kompaktno prikazuje korpuse i podržava ga veći broj operativnih sustava.[194]

U znanstvenom članku „POLIQARP 1.0: Some technical aspects of a linguistic search engine for large corpora“, autori softvera Daniel Janus i Adam Przepiórkowski detaljno su predstavili arhitekturu i način funkcioniranja softvera. Prema članku, temeljni je format korpusne građe kodiran u standardu XCES (engl. *XML Corpus Encoding Standard*), no navedeni je format previše robustan u pogledu mjesta kojeg zauzima na disku, stoga se najprije građa mora pretvoriti u binarni oblik pomoću alata *bp* (engl. *build Poliqarp representation*). Svaki dokument u korpusu mora sadržavati datoteku zaglavlja imena *header.xml*, i *bp* također može dohvaćati metapodatke iz XML datoteka zaglavlja (engl. *XML header file*). Binarni se korpusi sastoje od nekoliko zbirki datoteka koje sadrže informacije vezane uz isti element korpusa, a zovu se *backendovi*. Oni se, pak, sastoje od dvije strukture za pohranu podataka - vektora (engl. *vector*) za podatke nepromjenjive veličine i rječnika (engl. *dictionary*) za podatke promjenjive veličine. Nadalje, postoje i *backendovi* koji sadrže strukturalne informacije o dokumentima koji sačinjavaju korpusnu građu, metapodatke, skupine oznaka i slično. Glavni

[190] Przepiórkowski; Krynicki; Dębowski; Woliński; Janus; Bański, 2004.
[191] Przepiórkowski; Krynicki; Dębowski; Woliński; Janus; Bański, 2004.
[192] Ibid.
[193] Ibid.
[194] Wikipedia. Poliqarp. URL: https://en.wikipedia.org/wiki/Corpus_linguistics (9.5.2017.)

backend korpusa upravo je vektor sačinjen od struktura veličine 8 bajta koje predstavljaju segmente i koje se sastoje od sljedećih polja:

1) Zastavice koja označava ima li mjesta ispred pojedinog segmenta;
2) Pomak na rječnik ortografskih riječi;
3) Pomak na rječnik skupina razjašnjenih interpretacija, tj. one koje su označene kao ispravne u danom kontekstu;
4) Pomak na rječnik skupina nerazjašnjenih interpretacija, tj. one koje je dodijelio alat za morfološku analizu, bez obzira jesu li označene kao ispravne u prethodnim fazama označavanja.

Nakon što se korpusna građa obradi pomoću alata *bp,* za dobiveni se binarni korpus s pomoću alata za indeksiranje (engl. *corpus indexer*) stvaraju obrnuti indeksi (engl. *inverted index*) koji služe ubrzavanju izvedbe upita. Budući da se u korpusu nalazi mnogo više segmenata nego što je specifičnih elemenata u sekvencama ortografskih riječi i obje sekvence skupina interpretacija, obrnuti indeksi su prijeko potrebni kako bi se tim specifičnim elementima pripisala lista pojavljivanja u korpusu. Svaka od sekvenci ima svoj indeks, no za neke je moguće narediti alatu za indeksiranje da preskoči izradu indeksa. Obrnuti indeksi mogu zauzimati mnogo mjesta ako se pohranjuju naivno, no to se rješava dijeljenjem korpusa u komade (engl. *chunk*) i specifičnim mehanizmom kompresije korpusa pod nazivom Golombovo kodiranje.[195]

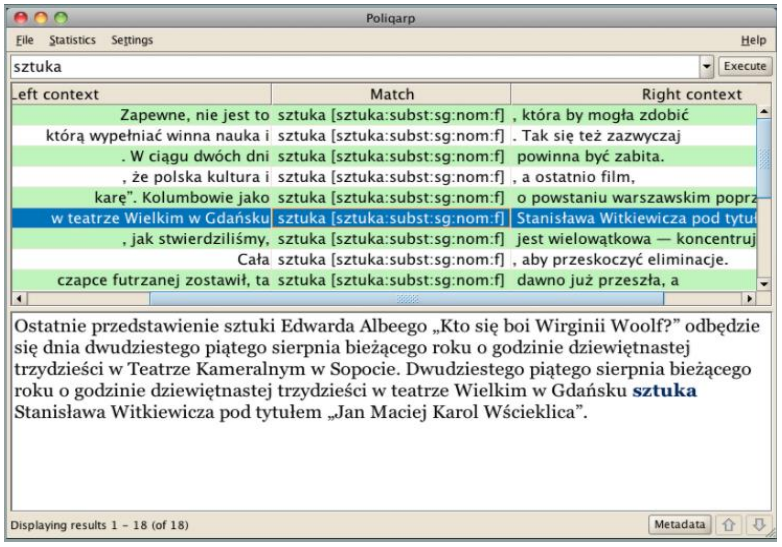
Nadalje, u članku opisuju i specifičnu arhitekturu klijent-poslužitelj. Slično kao i u Xairi, alat za konkordanciju podijeljen je na dva dijela – poslužitelja koji je napisan u programskom jeziku C i koji izvodi konkordanciranje, te klijenta koji prikazuje rezultate korisnikova pretraživanja. Protokol kojim se provodi komunikacija između klijenta i poslužitelja omogućuje dvosmjernu komunikaciju i putem njega se zahtjevi šalju poslužitelju. Na svaki upit, poslužitelj odgovara gotovo trenutno, a može pružiti i odgovore koji nisu vezani uz upit, već služe kao svojevrsne obavijesti korisniku. Klijent-poslužitelj korisna je arhitektura jer su unutar sustava odvojeni logički modul i modul sučelja, no ova vrsta arhitekture ima još prednosti:

1) Poslužitelj podržava višestruku povezanost, što znači da se na jedan poslužitelj mogu istovremeno spojiti dvije instance klijenta ili dva različita klijenta, bez udvostručavanja sustavnih resursa;

[195] Janus; Przepiórkowski, 2005.

2) Protokol podržava koncept sesije, što znači da su omogućene višestruke povezanosti, dakle, moguće je spojiti se na poslužitelj na kraće razdoblje kako bi se postavio upit i/ili kako bi korisnik provjerio postoje li novi rezultati za već postavljeni upit. Ovo je naročito korisno za internetske klijente;
3) Cjelokupna funkcionalnost korpusa nalazi se u C biblioteci s jasno definiranim sučeljem i stoga je moguće stvoriti dodatnu biblioteku koja je kompatibilna s istim sučeljem i upotrijebiti je s ostatkom sustava. [196]

Jezik upita kojim se koristi Poliqarp temelji se na već spomenutom jeziku upita IMS Open Corpus Workbencha, CQP-u, a regularni se izrazi mogu formulirati za korpusne pozicije.[197] Upiti za vrste riječi i morfosintaktičke kategorije mogu se postavljati zasebno, a jedinstvena je značajka Poliqarpa činjenica da se njime mogu pretraživati korpusi koji, uz razjašnjene interpretacije, sadrže i informacije o svim mogućim morfosintaktičkim interpretacijama koje je generirao alat za morfološku analizu.[198] U ranim je fazama Poliqarp dizajniran kao alat za korpuse koji su lingvistički obilježeni na razini riječi, no poslije su dodana proširenja i Poliqarp je postao alat za indeksiranje i pretraživanje određenih banaka stabala (engl. *treebank*).[199]

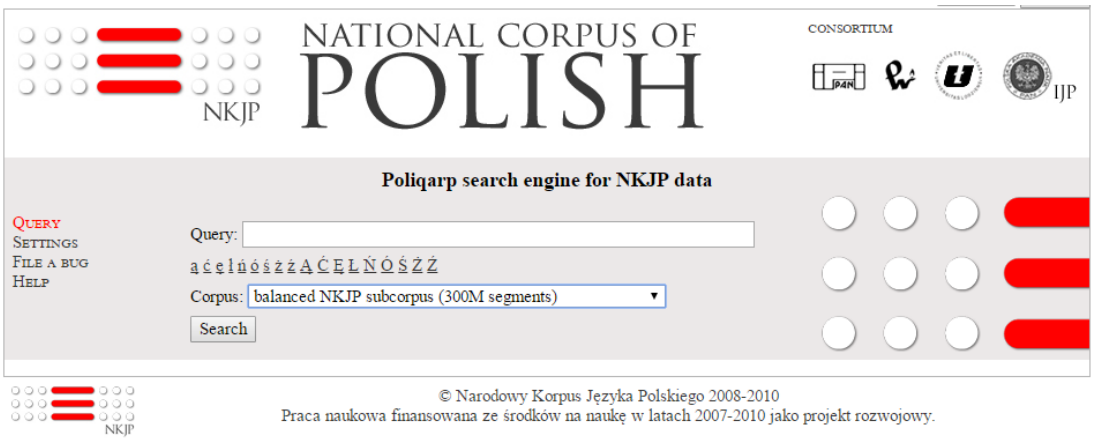


Slika 20. *Prikaz sučelja klijentskog programa za Poliqarp (MacOS)*[200]

[196] Janus; Przepiórkowski, 2005.
[197] Ibid.
[198] Ibid.
[199] Ibid.
[200] Daniel Janus. Smyrna: An easy Polish concordancer in Clojure. URL: http://danieljanus.pl/talks/reveal.js/2014-lambdadays.html#/7 (15.5.2017.)

Dostupne su dvije vrste sučelja za softver Poliqarp. Oba prikazuju rezultate u KWIC formatu, upotrebljavaju isti za alat obradu korpusa i podržavaju isti jezik upita - jedno je sučelje samostalni program napisan u Javi koji se može instalirati na više operativnih platformi, dok se drugo sučelje sastoji od PHP skripti i dizajnirano je za pretraživanje korpusa putem interneta, no nema toliko značajki kao samostalno sučelje.[201] Takvo se sučelje upotrebljava za korpus IPI PAN te Nacionalni korpus poljskog jezika.

Sučelje Nacionalnog korpusa poljskog jezika jednostavno je i intuitivno. Na početnoj stranici korpusa korisnik odmah može upisati upit služeći se jezikom upita ili jednostavno upisati traženi oblik riječi. Postoji i mogućnost odabira određenog potkorpusa. S lijeve se strane sučelja nalaze odjeljci s mogućnostima. Tako se tu nalazi i odjeljak „Settings“, gdje je moguće izvršiti raznorazne postavke, primjerice, broj prikaza na jednoj stranici, sortiranje rezultata prema raznim kriterijima, broj elemenata prikazanih u konkordancijama, širinu tekstne okoline u konkordancijama i slično. Nadalje, razvojnom je timu moguće prijaviti određene pogreške pod odjeljkom „File A Bug“, te je tu i odjeljak „Help“ koji služi kao svojevrstan priručnik za uporabu korpusa.



Slika 21. *Prikaz početne stranice Nacionalnog korpusa poljskog jezika*[202]

[201] Janus; Przepiórkowski, 2005.
[202] National Corpus of Polish. URL: http://nkjp.pl/poliqarp (15.5.2017.)

4.7. PhiloLogic

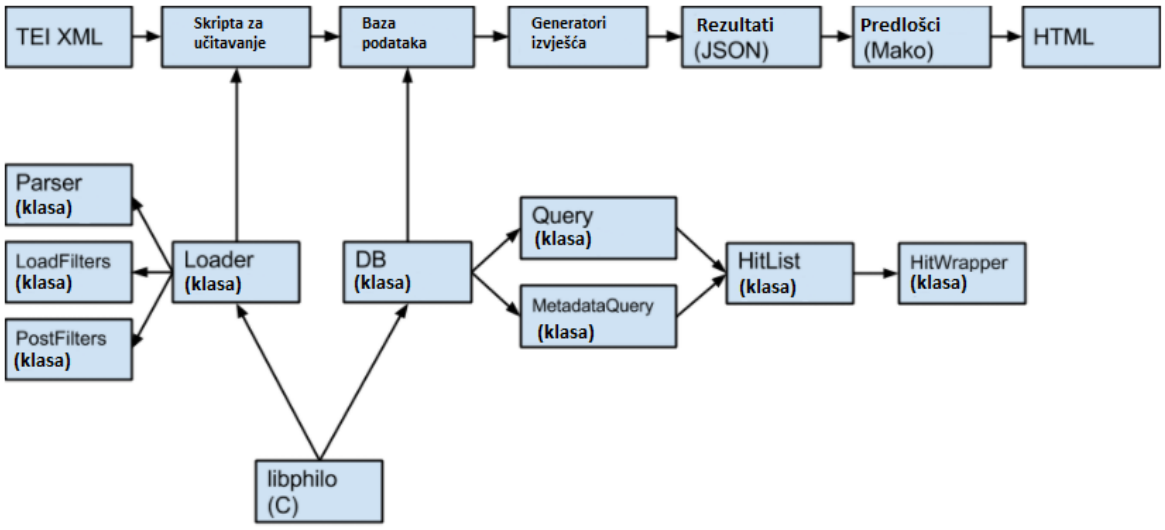
PhiloLogic je alat otvorenog izvornog koda za pretraživanje, dohvaćanje i analizu sadržaja koji je razvijen od strane projekta ARTFL i Razvojnog centra za digitalnu knjižnicu (engl. *Digital Library Development Center*) pri Sveučilištu u Chicagu. PhiloLogic je izrazito jednostavan za korištenje, no unatoč tome, vrlo je moćan sustav za potpuno pretraživanje teksta, dohvaćanje i obavještavanje za velike multimedijalne baze, s mogućnošću obrade složenih tekstualnih struktura s mnoštvom metapodataka.[203] PhiloLogic je dizajniran u svrhu akademskog istraživanja baza književnih, religijskih, filozofskih i povijesnih tekstova, a isto je tako pogodan i za povijesne enciklopedije i rječnike.[204] Trenutno je najnovija važeća inačica PhiloLogic 4.6 i ovo će se poglavlje isključivo koncentrirati na nju. PhiloLogic obiluje korisnim značajkama; ne zauzima mnogo procesorskih resursa, no unatoč tomu vrlo je brz i robustan. Koristi se već dugi niz godina, naročito u akademskoj zajednici, što znači da je i rigorozno ispitan. Njegova je instalacija neovisna o drugom softveru, i u softverskom su paketu ugrađene mogućnosti raznolikog konfiguriranja.[205]

Na početku je odmah važno napomenuti kako je PhiloLogic modularan sustav. Prema službenoj stranici na GitHubu, najnoviju je inačicu isključivo moguće instalirati na operativnim sustavima koji se temelje na Unixu, kao što su Linux distribucije, no ne postoji podrška za operativni sustav MacOS. Moguće ju je samo pokrenuti na internetskom poslužitelju Apache, potrebno je instalirati Python 3 kako bi se pokrenula, a njezinoj se internetskoj aplikaciji može pristupiti samo putem modernih internetskih preglednika.[206]

Na službenom blogu projekta ARTFL, jedan od programera koji rade na razvijanju PhiloLogica, Richard Whaling, pružio je uvid u arhitekturu najnovije inačice alata, PhiloLogic4, u objavi pod nazivom „PhiloLogic4: The Big Picture“. Građa koja se pretražuje pomoću alata PhiloLogic mora biti kodirana u ranije navedenom XML formatu pod nazivom TEI, kojega ima više vrsta. Upravo je ta raznovrsnost mogućih načina kodiranja TEI dokumenata bila razlog za redizajn alata. Whaling navodi kako je temelj svih usluga ovog alata skupina C funkcija koje se nalaze u biblioteci pod nazivom *libphilo* i upravo su te funkcije odgovorne za visokokvalitetno komprimiranje, indeksiranje i algoritme pretraživanja. Upravo

[203] University of Chicago Library. PhiloLogic. URL: https://www.lib.uchicago.edu/efts/ARTFL/philologic/ (18.5.2017.)
[204] ARTFL Project. PhiloLogic. URL: https://sites.google.com/site/philologic3/manual (18.5.2017.)
[205] ARTFL Project. PhiloLogic. URL: https://sites.google.com/site/philologic3/home (18.5.2017.)
[206] GitHub. PhiloLogic4. URL: https://github.com/ARTFL-Project/PhiloLogic4/blob/master/README.md (20.5.2017.)

ta biblioteka predstavlja temelj na koji se nadograđuju sve Python klase, od kojih su najvažnije klasa *Loader* koja upravlja parsiranjem i indeksiranjem TEI XML datoteka, te klasa *DB* koja upravlja pristupom bazi podataka PhiloLogica. Vrlo je važno napomenuti da se te klase služe drugim klasama te kako ne dijele gotovo nijedan element ili način rada, a ta je odvojenost ono po čemu se PhiloLogic razlikuje od ostalih sustava te vrste. Drugim riječima, skupina elemenata koje stvaraju bazu podataka razlikuje se od skupine elemenata koje vrše upite nad postojećom bazom podataka. Novi je parser napisan u Pythonu i zapravo je alat za evaluaciju XPathova kojim se upravlja specificiranjem objektnih XPathova (engl. *object XPaths*) i XPathova metapodataka (engl. *metadata XPaths*), te je mnogo jednostavniji u pogledu proširivanja, prilagodbe i održavanja od prošlog parsera koji je bio skripta napisana u Perlu.[207] Kako je ranije navedeno, budući da TEI dokumenti nisu unificirani pojavila se potreba za alatom koji će imati općenitiju biblioteku za obradu TEI dokumenata, kojoj je bilo potrebno dodati dva dodatna elementa, opću skriptu za prihvat dokumenta (engl. *general-purpose document-ingesting script*) za obradu pogrešaka i neodređenosti, te gotovu internetsku aplikaciju koja se može upotrijebiti u većinu svrha te prilagoditi.[208]



Slika 22. *Prikaz arhitekture PhiloLogica*[209]

Whaling dalje navodi kako nova skripta za učitavanje (engl. *load script*) u načelu funkcionira kao i ona prijašnja, no kraća je i jednostavnija za razumijevanje i prilagodbu, a njezine su tri glavne zadaće da: a) zaprima argumente u obliku naredbenih redaka i sve datoteke

[207] Allen; Gladstone; Whaling, 2013.
[208] ARTFL Project Research Blog. PhiloLogic4: The Big Picture. URL: https://artfl.blogspot.hr/2014/12/philologic4-big-picture.html (20.5.2017.)
[209] Ibid.

prebacuje u klasu *Loader* uz još nekoliko parametara; b) pohranjuje sve parametre specifične za sustave i c) pohranjuje sve konfiguracijske parametre specifične za tekst. U konačnici, skripta učitanu bazu podataka premješta na odgovarajuću lokaciju na poslužitelju i oko nje stvara internetsku aplikaciju. Unutar baze podataka, odnosno aplikacije, upiti prolaze kroz generatore izvješća (engl. *report generator*) čija je zadaća interpretirati upit i sukladno tome pristupiti bazi podataka, nakon čega se stvara rezultatski objekt u Pythonu koji se mapira blizu objekta u JSON-u. Taj objekt premješta se na datoteku predloška Mako koja ga može prebaciti u HTML format koji se može pregledavati putem internetskog preglednika.[210] Internetski element, odnosno internetska aplikacija PhiloLogica odvojena je od C biblioteke, no ta su dva dijelu međusobno ovisna.[211]

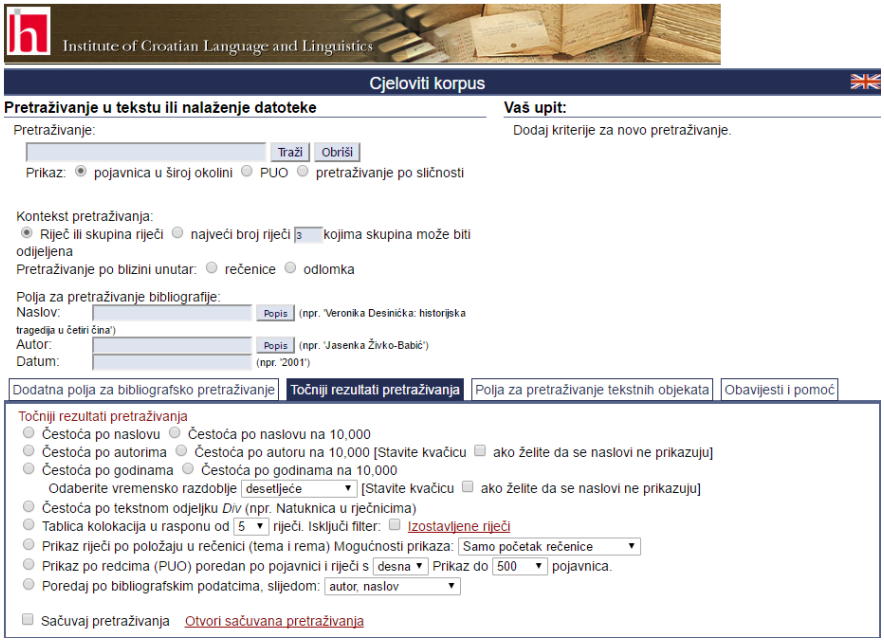
Pretraživanje riječi i metapodataka koristi se jednakom sintaksom, no ona se interpretira drugačije, što je detaljnije opisano na GitHub stranici PhiloLogica. Općenita sintaksa upita PhiloLogica sastoji se od 5 temeljnih operatora:

1) Obična pojavnica, odnosno bilo koja riječ odvojena razmaknicama, npr. *pojavnica*;
2) Citirana pojavnica, odnosno niz znakova unutar navodnika koji može sadržavati razmaknicu, npr. *„pojavnica“;*
3) Raspon, odnosno dvije pojavnice odvojene crticom, npr. *a-f* ;
4) Booleov OR, koji se prikazuje kao simbol „|“, npr. *pojavnica* | *riječ*;
5) Booleov NOT, koji se prikazuje kao „NOT“, npr. *pojavnica.NOT pojavnice*.[212]

PhiloLogic je sustav koji se koristi i u Hrvatskoj. Jedan od poznatijih projekata koji se koristi PhiloLogicom jest Hrvatski jezični korpus razvijen u sklopu projekta Hrvatska jezična riznica na čelu s Dunjom Brozović Rončević s Instituta za hrvatski jezik i jezikoslovlje. Projekt se služi starijom inačicom, PhiloLogic3. Ovaj se korpus sastoji od različitih izvora teksta, a neki od njih su internetske novine, knjige i članci, tiskane i objavljene knjige i druge tiskovine, digitalne datoteke tiskanih knjiga koje su dostupne putem izdavača te transkripcije prikupljenih podataka i snimaka.[213] Sučelje potkorpusa, odnosno korpusa, nešto je drugačije po vizualnom rasporedu elemenata od većine internetskih sučelja pregledanih u ovom radu. Opcije pretraživanja nisu poredane s lijeve strane u stupcu, nego se većina opcija, tj. oblika filtriranja,

[210] ARTFL Project Research Blog. PhiloLogic4: The Big Picture. URL: https://artfl.blogspot.hr/2014/12/philologic4-big-picture.html (20.5.2017.)
[211] ARTFL Project Research Blog. General Overview Of PhiloLogic4'S Web Architecture. URL: https://artfl.blogspot.hr/2014/12/general-overview-of-philologic4s-web.html (23.5.2017.)
[212] GitHub. PhiloLogic4. URL: https://github.com/ARTFL-Project/PhiloLogic4/blob/master/docs/query_syntax.md (23.5.2017.)
[213] Ćavar; Brozović Rončević, 2012.

nalaze ispod samog polja za pretraživanja. Također su dostupne i tri opcije za sužavanje pretraživanja, a nalaze se u obliku kartica na donjem dijelu sučelja. Kartica „Dodatna polja za bibliografsko pretraživanje“ sadrži polja kao što su „Nakladnik, „Datum Izdavanja“, „Jezik“ i slično, a uz svako se polje nalazi poveznica koja vodi na popis opcija koje je moguće unijeti u to polje. Kartica „Točniji rezultati pretraživanja“ pruža mogućnosti podešavanja sortiranja čestoće po raznim parametrima, oblik prikaza i još mnogo mogućnosti koje su vidljive na Slici 23. Kartica „Polja za pretraživanje tekstnih objekata“ pruža mogućnost pretraživanja odjeljaka karakterističnih za TEI dokumente, *Div* i *SubDiv,* a pokraj polja u odjeljcima nalaze se isto tako poveznice na popis opcija koje je moguće unijeti. Također, dostupna je i kartica pod nazivom „Obavijesti i pomoć“ koja sadrži nekoliko smjernica za upotrebu sustava PhiloLogic.



Slika 23. *Prikaz sučelja Hrvatskog jezičnog korpusa s otvorenom Karticom „Točniji rezultati pretraživanja“*[214]

Još jedan projekt koji se služi sustavom PhiloLogic kao softverskom platformom jest CroALa, odnosno Croatiae Auctores Latini – znanstvena i slobodno dostupna zbirka recenziranih latinskih tekstova hrvatskih autora i autora povezanih s Hrvatskom koju sačinjavaju djela od srednjeg vijeka do današnjeg doba.[215] Dio tekstova u zbirci plod su jednostavnih postupaka digitalizacije, skeniranja ili prijepisa starijih izdanja, a dio je rezultat modernog filološkog rada i digitalne inačice suvremenih kritičkih izdanja.[216] Koliko je

[214] Institut za hrvatski jezik i jezikoslovlje. Hrvatska jezična riznica. URL: http://riznica.ihjj.hr/philologic/Cijeli.whizbang.form.hr.html# (25.5.2017.)
[215] CroALa. URL: http://croala.ffzg.unizg.hr/ (25.4.2017.)
[216] Ibid.

dostupno iz literature, zbirka upotrebljava inačicu PhiloLogic3, iako je na blogu projekta Croatica et Tyrolensia navedeno kako je tim 2015. godine uspješno pokrenuo najnoviju inačicu PhiloLogica na lokalnom glavnom računalu.[217] Vezano uz zbirku CroALa, nastao je ranije spomenuti projekt Croatica et Tyrolensia čiji je cilj bio napraviti digitalnu usporedbu hrvatske i tirolske latinističke književnosti od tri baze podataka: CroALa, Latinitas Tyrolensis (LatTy) i Croatiae auctorum Latinorum bibliographia.[218] Postoji nekoliko sučelja za pretraživanje zbirke. Na početnoj stranici zbirke (http://croala.ffzg.unizg.hr/) moguće je pretraživati zbirku unosom jedne ili više riječi u jedno polje, nakon čega je moguće podesiti prikaz rezultata, tj. korisnik može odlučiti želi li da rezultati budu prikazani u standardnom formatu konkordancija ili u KWIC formatu. Međutim, postoje i dvije dodatne mogućnosti. Korisnik može odabrati sučelje za pretraživanje na engleskom jeziku koje je, vizualno gledajući, vrlo minimalno i jednostavno, no ne i intuitivno. Druga mogućnost je sučelje na latinskom jeziku prikazano na Slici 24. Po mogućnostima pretraživanja, sučelja su gotovo identična sučelju projekta Riznica.

Slika 24. *Prikaz sučelja na latinskom jeziku za pretraživanje zbirke CroALa*[219]

[217] Croatica et Tyrolensia. Technical note: installing PhiloLohic 4 on localhost. URL: http://crotyr.hypotheses.org/99 (28.5.2017.)
[218] Salopek, Ž. Pet godina zbirke Croatiae auctores Latini i prva godina projekta Croatica et Tyrolensia. URL: http://dfest.nsk.hr/wp-content/themes/boilerplate/2015/prezentacije/Salopek_Zeljka.pdf (28.5.2017.)
[219] CroALa. URL: http://croala.ffzg.unizg.hr/croala-form.html (28.5.2017.)

4.8. TEITOK

TEITOK je internetski alat za pregledavanje, izgradnju i uređivanje korpusa s opširnim tekstualnim oznakama i lingvističkim obilježjima kojeg razvija Centar za lingvistiku pri Sveučilištu u Lisabonu (port. *Centro de Linguística da Universidade de Lisboa, CLUL*). Autor alata je Maarten Janssen, a TEITOK je privatni projekt dostupan na internetskom repozitoriju GitLab.[220] Primarni cilj pri dizajniranju TEITOK-a bio je stvoriti alat koji u sebi objedinjuje dvije korpusne značajke važne dvjema različitim ciljnim publikama. TEITOK cilja zadovoljiti potrebe korisnika koji više naginju ka filološkom pristupu, ali i onim koji imaju primarno lingvističke interese pri pretraživanju korpusa. Tako za, primjerice, korpusnu građu koja se sastoji od paleografskih transkripcija povijesnih dokumenata TEITOK istovremeno omogućuje dodavanje tekstualnih oznaka (engl. *textual markup*) poput onih za prijelome redaka, slovoslagarske informacije, promjenu rukopisa, obrisane dijelove i slično, ali i lingvističkih obilježja (engl. *linguistic annotation*) poput POS oznaka, lema, normalizirane ortografije, semantičkih klasifikacija i slično, budući da mnogo alata koji imaju opciju lingvističkog obilježavanja nemaju opciju tekstualnog označavanja, i obratno.[221] S takvom se građom često previđa korist obje značajke. Uporaba lingvističkih obilježja bez tekstualnih oznaka često vodi do zaključaka koji nemaju veze sa samim dokumentom, nego s osobom koja ga je transkribirala, a sama lingvistička obilježja nisu korisna samo za gramatičke i statističke analize, nego i za bogatije opcije pretraživanja dokumenata.[222]

TEITOK je alat namijenjen prvenstveno korpusima na kojima je potrebno mnogo rada. To se, u prvom redu, odnosi na povijesne korpuse, učeničke korpuse te korpuse govornog jezika. Izrazito je prilagodljiv u smislu da je moguće promijeniti postojeće skripte ili postojećem modularnom sustavu dodavati funkcionalnosti u obliku novih skripti, kompatibilan je s većinom alata za generiranje TEI datoteka i može se koristiti za objavljivanje i pretraživanje korpusa obrađenog ranije spomenutim CQP-om, pa se samim time može usporediti i sa CQPwebom.[223]

Autor TEITOK-a Maarten Janssen u znanstvenom je članku pod nazivom „TEITOK: Text-Faithful Annotated Corpora“ detaljno je predstavio arhitekturu i značajke alata. Dakle, TEITOK je internetski alat koji služi za izgradnju, obilježavanje, održavanje i objavljivanje

[220] CLUL. TEITOK. URL: http://alfclul.clul.ul.pt/teitok/site/index.php?action=home (2.6.2017.)
[221] CLUL. TEITOK. URL: http://alfclul.clul.ul.pt/teitok/site/index.php?action=about (2.6.2017.)
[222] Janssen, 2016.
[223] Ibid.

korpusa, a većinom je napisan u kombinaciji programskih jezika PHP i Java. Korpusna građa sastoji se od XML datoteka kodiranih u formatu TEI s prilagođenim sustavom tokenizacije koji omogućuje jednostavan prikaz XML datoteka, uređivanje metapodataka i pojedinačnih pojavnica te pretraživanje korpusa. Tokenizacija se dodaje unutar samog TEI dokumenta, pri čemu se pojavnice označavaju elementom <*tok*> i sva se lingvistička obilježja predstavljaju kao atributi nad tim elementima koji okružuju svaku riječ. U tom slučaju, tekstualno označen korpus je sve osim elemenata <*tok*>, a lingvistički je korpus sekvenca elemenata <*tok*> koji se mogu prebaciti, u slučaju nužde, u vertikalizirani format. Za razliku od standardnih TEI smjernica koje za različite načine kodiranja ortografskog oblika riječi propisuju uporabu elementa <*choice*>, TEITOK upotrebljava elemente <*tok*> koji mogu imati višestruke ortografske oblike za jednu riječ, a koji su modelirani kao atributi. Jedna od prednosti ovog sustava jest da se na element <*tok*> može dodati koliko je god ortografskih atributa potrebno, dok element <*choice*> ima ograničene opcije, što je vrlo korisno za manuskripte. Uređivanje lingvističkih obilježja vrlo je jednostavno budući da se u načinu rada za prikaz teksta (engl. *text-view*) sva obilježja prikažu postavljanjem kursora miša na pojedinu pojavnicu. Odabirom pojavnice pojavljuje se HTML oblik sadržaja pojavnice koji se uređuje i nakon pohranjivanja se ažurira XML datoteka. Strukturalno uređivanje vrši se u vertikaliziranom formatu tablice, a za znatnije promjene preferabilno je definirati koje su pojavnice u pitanju i urediti ih pomoću jezika upita za CQP. TEITOK ima iznimno složen sustav metapodataka, no korisnik može odrediti skupinu relevantnih metapodataka i od njih napraviti tablicu za uređivanje (engl. *edit table*) u formatu HTML koja opisuje relevantna polja, dok se uz polja nalaze definicije XPatha koje specificiraju određena polja u TEI zaglavlju (engl. *TEI header*). Od tablice se generira jednostavan HTML oblik koji zamjenjuje XPathove s HTML poljima za unos i XML dokument se pregledava kako bi se pronašla odgovarajuća vrijednost, te korisnik može dodati ili uređivati informacije. Nakon pohranjivanja, informacije se upisuju na odgovarajuću lokaciju u XML datoteci i stvaraju se čvorovi (engl. *nodes*). Za dodavanje oznaka POS, TEITOK se služi alatom za označavanje NeoTag koji je neovisan o jeziku i koji se koristi internom strukturom riječi za označavanje riječi izvan vokabulara (engl. *out-of-vocabulary, OOV*). Ovaj je alat vrlo specifičan jer, uz primarnu funkciju označavanja teksta, može upotrijebiti određeni korpus kao „korpus za trening“ kako bi se izgradio alat za označavanje koji je specijaliziran za vrstu tekstova u korpusu i koji se poboljšava kako korpus raste. Označavanje se vrši izravno u XML datoteku.[224]

[224] Janssen, 2016.

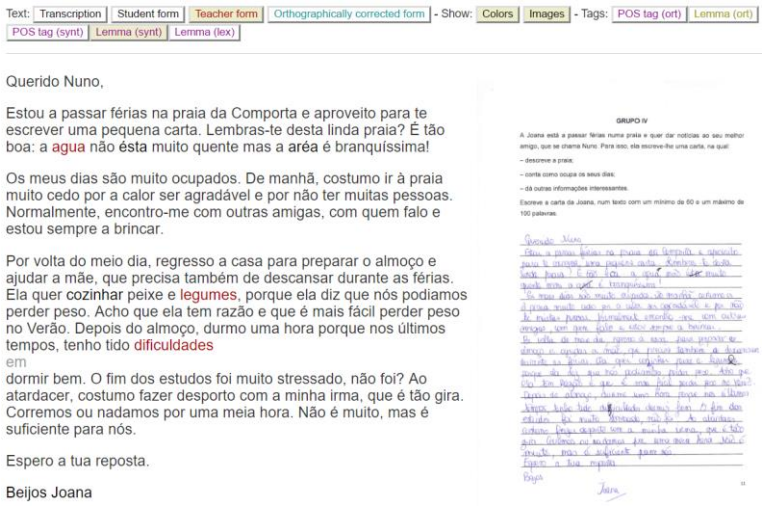
Nadalje, u članku se navodi i jezik upita, odnosno način pretraživanja korpusa pomoću alata TEITOK. CQP verzija korpusa gradi se automatski i moguće ju je pretražiti putem internetskog sučelja pomoću CQL-a. Upit prolazi kroz CQP i rezultati se prikazuju putem internetskog preglednika u formatu KWIC. CQP korpus gradi se pomoću posebno alata *tt-cwb-encode* koji gradi CQP korpusne datoteke izravno iz XML datoteka, pri tome zadržavajući posebnu datoteku koja ukazuje na poziciju pomaka pojedine pojavnice u XML datoteci kako bi se mogle brzo pronaći. Uz to, sučelje ne omogućuje samo lingvističke upite, već je moguće i pretraživati dokumente, a rezultati se prikazuju kao popis dokumenata s odgovarajućim karakteristikama metapodataka.[225]

Iako je TEITOK i dalje relativno nov alat, njime se koristi nekoliko projekata sa Sveučilišta u Lisabonu, kao i nekoliko projekata iz drugih država, ali je i mnogo projekata u razvoju. Na službenoj stranici alata, navedeno je 10 projekata koji su u tolikoj mjeri razvijeni da mogu biti dostupni javnosti, a to su učenički korpus portugalskog jezika COPLE2, povijesni korpus judeošpanjolskog jezika CoDiAJe, povijesni korpus portugalskog jezika Corpus de Textos Antigos, referentni korpus portugalskog jezika Corpus África, učenički korpus portugalskog jezika EFFE-On, učenički korpus latvijskog i litvanskog jezika ESAM, povijesni korpus galicijskog jezika Gondomar, povijesni korpus slovenskog jezika IMP, korpus govornog portugalskog jezika MADISON i povijesni korpus španjolskog i portugalskog jezika Post Scriptum. Kako je navedeno ranije, alat je vrlo prilagodljiv, stoga sučelja i njihove mogućnosti nisu za sve korpuse jednaki. [226]

Među prvim je projektima upravo korpus COPLE2. To je učenički korpus portugalskog jezika kao drugog ili stranog jezika, a sastoji se od pisanih i izgovorenih tekstova čiji su autori studenti tečajeva portugalskog jezika na Institutu portugalske kulture i jezika (port. *Instituto de Cultura e Língua Portuguesa ICLP*) i Centru za evaluaciju portugalskog kao stranog jezika (port. *Centro de Avaliação de Português Língua Estrangeira, CAPLE*) na Fakultetu humanističkih znanosti Sveučilišta u Lisabonu (port. *Faculdade de Letras da Universidade de Lisboa, FLUL*) .[227] Cilj je ovog projekta omogućiti pomagalo koje će koristiti pri istraživanju, obuci nastavnika te didaktici, ali i odrediti lingvistički profil učenika koji portugalski jezik uče kao drugi ili strani jezik u svrhu nastave ili procjene kompetencija.[228] Sučelje korpusa COPLE2 prilično je jednostavno i vrlo je intuitivno. S lijeve se strane nalazi stupac s odjeljcima, a s desne

[225] Janssen, 2016.
[226] CLUL. TEITOK. URL: http://alfclul.clul.ul.pt/teitok/site/index.php?action=projects (5.6.2017.)
[227] Ibid.
[228] CLUL. COPLE2. URL: http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=home (5.6.2017.)

strane sadržaj koji se prikazuje odabirom pojedinog odjeljka. Zanimljiva značajka sučelja je odjeljak „XML Files“ u kojemu se nalaze XML dokumenti podijeljeni prema jezicima. Odabirom jezika, a zatim i dokumenta, korisnik dobiva uvid u fotografiju izvornog dokumenta, materinji jezik studenta, drugi strani jezik kojeg student poznaje, razinu poznavanja jezika u pitanju, žanr i temu dokumenta, broj pojavnica u dokumentu i trajanje učenja portugalskog jezika u mjesecima. Pokraj fotografije izvornog dokumenta nalazi se digitalni tekst, a iznad njega nalaze se opcije prikaza: a) s obzirom na tekst (transkripcija, studentski oblik, nastavnikov oblik i ortografski ispravljen oblik); b) opcije prikaza lingvističkih oznaka (POS, lema) i c) opcije prikaza boja i slikovnih sadržaja.

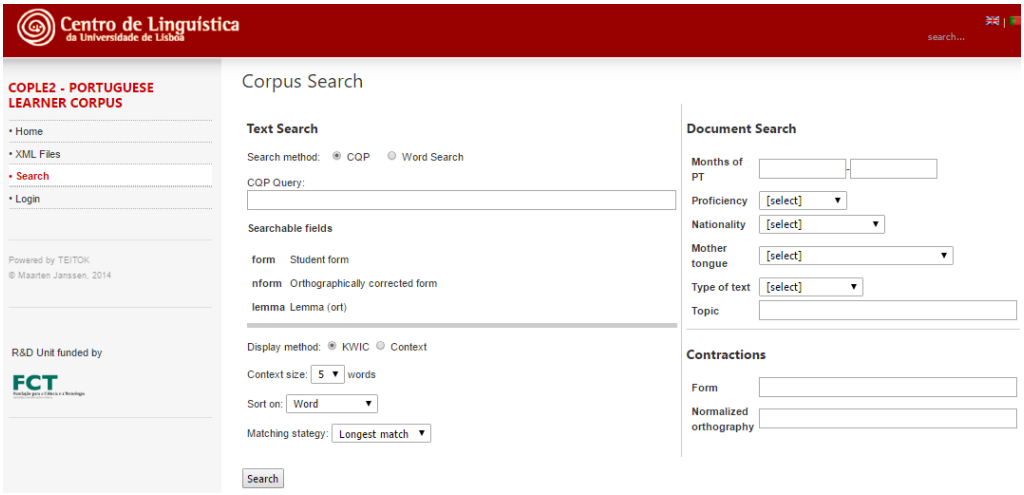


Slika 25. *Prikaz digitalnog i izvornog teksta dokumenta u korpusu COPLE2*[229]

Odjeljak „Search“ služi za pretraživanje korpusa. Odabirom odjeljka, automatski se otvara stranica koja sadrži prazno polje za pretragu pomoću jezika upita CQP Query Language, uz kratku uputu upotrebe. Tu je i opcija „advanced“, koja korisnika vodi ka složenijim, odnosno naprednijim mogućnostima pretraživanja korpusa. Ovdje korisnik može odabrati želi li pretraživati korpus pomoću jezika CQP Query Language ili pretraživanjem tražene riječi. Ukoliko korisnik odabere pretraživanje unosom tražene riječi, moguće je dodatno definirati parametre studentski oblik, ortografski ispravljen oblik i lemu. Moguće je odabrati prikaz rezultata između formata KWIC i kontekstualnog prikaza, kao i daljnje podešavanje navedene dvije vrste prikaza u smislu veličine tekstne okoline, parametra sortiranja i strategije pronalaska odgovarajućeg rezultata. S desne se strane nalaze polja za upis podataka za pretraživanje

[229] CLUL. COPLE2. URL: http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=file&id=English/en001CVMTD.xml (7.6.2017.)

dokumenata i kontrakcija. Za dokumente je moguće unijeti trajanje učenja jezika u mjesecima, razinu poznavanja, nacionalnost učenika, materinski jezik učenika, vrstu teksta i temu teksta. Za kontrakcije je moguće unijeti izvorni oblik te oblik nakon postupka normalizacije ortografije. S lijeve je strane također dostupan i odjeljak „Login“, gdje se korisnik može prijaviti u sustav.

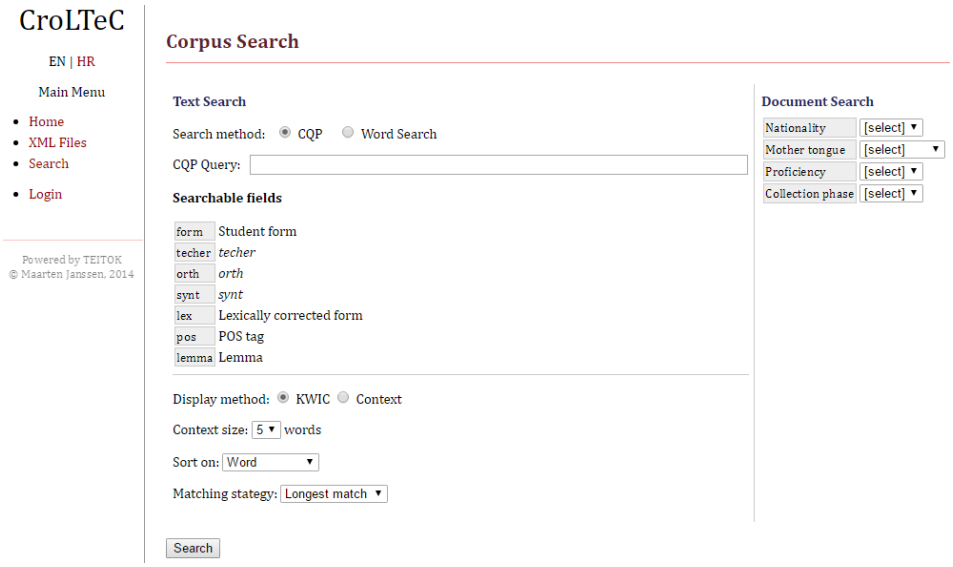


Slika 26. *Prikaz sučelja za napredno pretraživanje korpusa COPLE2 pri Odabiru CQP metode pretraživanja*[230]

Jedan od projekata kojeg je važno spomenuti u ovom kontekstu jest i ranije spomenuti Hrvatski učenički korpus (engl. *Croatian Learner Corpus*), koji se sastoji od dva potkorpusa – Hrvatskoga učeničkog korpusa pisanog jezika, CROLTEC (engl. *CROatian Learner TExt Corpus*) i Hrvatskoga učeničkog korpusa govornog jezika, CROLSEC (engl. *CROatian Learner SpEech Corpus*). To su korpusi čija je građa sačinjena od tekstova učenika hrvatskoga jezika kao drugog ili stranog jezika različitih razina, od početne razine A1 i više, prema šest razina Zajedničkog europskog referentnog okvira za jezike.[231] Cilj im je opisati jezik učenika i mjeru u kojoj on odstupa od standardnog jezika, omogućiti dubinsku analizu jezika učenika te izvlačenje važnih lingvističkih uzoraka, kontrastivnu analizu međujezika te analizu grešaka potpomognutu računalom.[232] CROLTEC je trenutačno i dalje u razvojnoj fazi s alatom TEITOK. Prijepisi su dovršeni, prebačeni su u TEI XML format te su unesene sve učeničke samoispravke. Prema najrecentnijim informacijama, CROLTEC sadrži 1,054.287 pojavnica i

[230] CLUL. COPLE2. URL: http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=cqp&act=advanced (7.6.2017.)
[231] Mikelić Preradović; Berać; Boras, 2015.
[232] Ibid.

razvojni tim pod vodstvom prof. dr. sc. Nives Mikelić Preradović trenutačno radi na MSD označavanju (engl. *morphosyntactic description, MSD*), nakon čega će korpus biti dostupan za pretraživanje. Implementacija sheme za označavanje pogrešaka započet će tijekom jeseni 2017. godine.



Slika 27. *Prikaz sučelja za napredno pretraživanje korpusa CROLTEC pri Odabiru CQP metode pretraživanja*[233]

5. Anketno istraživanje

U sklopu rada provedeno je i anketno istraživanje pod naslovom „Poznavanje alata i tražilica korpusne lingvistike“. Nakon pregleda nekih od dostupnih alata i tražilica, nastoji se dati uvid u to koliko zapravo ljudi, koji se jezikom bave u struci ili u obrazovanju, znaju o korpusima i korištenjem korpusima. Pitanja su u potpunosti preuzeta, a neka u određenoj mjeri i preinačena, iz anketnog istraživanja koje je provela je Kristina Posavec u sklopu svog doktorskog rada pod naslovom „Uloga računalnih korpusa u poučavanja hrvatskoga kao drugoga i inoga jezika“. Ciljni ispitanici anketnog istraživanja K. Posavec bili su lektori Croaticuma – Centra za hrvatski kao drugi i strani, dok je ovo anketno istraživanje imalo nešto širu ciljnu skupinu ispitanika. Istraživanju je cilj bio saznati koliko se jezičara, odnosno osoba koji se u svojem obrazovanju ili struci bave jezikom, dakle, prevoditelji, lektori, nastavnici, studenti filoloških smjerova, novinari i slično, zna služiti korpusnim alatima i tražilicama i jesu

[233] Mikelić Preradović, N. CroLTeC. URL: http://teitok.iltec.pt/croltec/index.php?action=cqp&act=advanced (9.6.2017.)

li tijekom svojeg obrazovanja naišli na bilo kakav sat, tečaj ili kolegij čija je tema bila proučavanje jezika pomoću računalnih korpusa. Budući da je ovaj rad svojevrsni pregled različitog softvera za pretraživanje korpusa i samim time služi kao afirmativni i poticajni rad u pogledu služenja računalnom korpusnom lingvistikom, cilj upitnika bio je saznati status korpusne lingvistike među ispitanicima. Anketno je istraživanje bilo anonimno i distribuirano je, s naglaskom na ciljnu skupinu ispitanika, privatnim kontaktima autora rada, studentima i profesorima Filozofskog fakulteta te putem Facebook grupa „Prevoditelji“ i „Jezičari“.

5.1. Sastav anketnog upitnika

Anketni upitnik izrađen je pomoću internetske usluge Google Forms. Sastoji se od ukupno 24 pitanja koja su podijeljena u tri dijela. Prvi se dio anketnog upitnika zove „Osobne informacije“ i cilj je bio prikupiti osobne informacije ispitanika, točnije: a) spol; b) dob; c) trenutačnu razinu obrazovanja; d) vrijeme završetka formalnog obrazovanja; e) područje specijalizacije najvišeg završenog stupnja obrazovanja; f) trenutačno mjesto zaposlenja; g) strane jezike koje ispitanici poznaju i h) kojim se jezičnim izvorima ili materijalima najčešće koriste u poslu ili obrazovanju.

Drugi se dio anketnog upitnika zove „Upoznatost s korpusnom lingvistikom“ i cilj je bio provjeriti koliko je ispitanika upoznato s konceptom korpusne lingvistike i do koje razine. Pitanjima je cilj bio utvrditi: a) jesu li ispitanici sudjelovali na bilo kakvom kongresu, seminaru, radionici ili tečaju vezanom uz korpusnu lingvistiku i što im je na takvoj vrsti edukacije bilo najkorisnije; b) jesu li ikada slušali kolegij iz korpusne lingvistike i kada je to bilo posljednji put, c) koliko su kolegija s elementima korpusne lingvistike slušali i d) koriste li se u obrazovanju ili u poslu metodama i alatima iz korpusne lingvistike i kojima.

Posljednji se dio anketnog upitnika zove „Vaše poznavanje korpusnih alata“ i cilj je bio provjeriti koliko su ispitanici upoznati s korpusnim alatima i u kojoj se mjeri njima koriste u različite jezične svrhe. Pitanjima je bio cilj utvrditi: a) koliko su ispitanici upoznati s korištenjem korpusnih alata u obrazovanju i poslu, analizom korpusa, stvaranjem korpusa i korištenjem korpusa za poboljšanje vlastitog znanja o jeziku; b) u kojoj se mjeri analizom i konkordancijama koriste pri učenju o jeziku, pomoći u poslu/obrazovanju te radu s učenicima/studentima; c) u kojoj se mjeri koriste *online* i *offline* alatima za pretraživanje i obradu korpusa; d) u kojoj se mjeri koriste korpusnim metodama kako bi kod sebe ili kod svojih učenika ili studenata poboljšali vokabular, gramatiku, čitanje, pisanje, slušanje i govor; e) u

kojoj se mjeri slažu s tvrdnjom da će im sudjelovanje u organiziranoj edukaciji o korpusnoj lingvistici koristiti za učenje o stvarnoj uporabi jezika, učenje o stvaranju korpusa, učenje o radu s korpusom, analizu konkordancija, pronalaženje jezičnih uzoraka, stvaranje i razvoj boljih materijala za učenje, učenje o korištenju korpusnih alata izravno u nastavi i razumijevanje nastavne metode izravnog korištenja korpusnih alata u nastavi (engl. *data-driven learning, DLL*); f) kojim su se hrvatskim računalnim korpusima koristili; g) kako bi sebe opisali kao korisnika tehnologije i računala i h) kako bi ocijenili svoj stupanj informatičke pismenosti, odnosno služenja računalom i računalnim alatima.

5.2. Rezultati anketnog istraživanja

Anketni upitnik bio je otvoren sveukupno sedam dana, tijekom kojih ga je riješilo ukupno 110 ispitanika. Niže su navedeni i pobliže opisani rezultati prema navedenim trima dijelovima anketnog upitnika.

5.2.1. Rezultati za dio „Osobne informacije“

Od ukupno 110 ispitanika, 84 (76,4%) su ispitanika bile žene i 26 (23,6%) su ispitanika bili muškarci. Najviše je ispitanika u dobnoj skupini od 20 do 29 godina, njih čak 67 (60,9%). Nadalje, 25 (22,7%) ispitanika u dobnoj je skupini od 30 do 39 godina, 9 (8,2%) ispitanika u dobnoj je skupini od 40 do 49 godina, 4 (3,6%) ispitanika u dobnoj je skupini od 50 do 59 godina, 3 (2,7%) ispitanika u dobnoj je skupini ispod 20 godina i 2 (1,8%) ispitanika u dobnoj su skupini iznad 60 godina.

Što se tiče trenutačne razine obrazovanja, dvije su kategorije izjednačene po broju ispitanika. Ukupno 62 (56,4%) ispitanika izjavilo je kako su po stupnju obrazovanja magistri/magistre ili magistri/magistre u tijeku, dakle, 31 (28,2%) ispitanik u svakoj kategoriji. Njih 17 (15,5%) izjavilo je kako trenutačno studiraju na preddiplomskom studiju, 12 (10,9%) ih je izjavilo kako trenutačno studiraju na postdiplomskom studiju, 10 (9,1%) ih je izjavilo kako imaju prvostupničku titulu i 9 (8,2%) ih je izjavilo kako imaju doktorsku titulu. Njih čak 54 (49,1%) studira trenutačno, 30 (27,3%) ih je završilo formalno obrazovanje prije više od 5 godina, 15 (13,6%) ih je formalno obrazovanje završilo unutar protekle godine dana, 7 (6,4%) ih je završilo formalno obrazovanje prije 4 do 5 godina i njih 4 (3,6%) završilo je formalno obrazovanje prije 2 do 3 godine.

Područje specijalizacije najvišeg završenog stupnja obrazovanja ispitanika je raznovrsno. Budući da svi ispitanici pitanje nisu shvatili na jednak način, tako ni njegovi odgovorni nisu jednako specifični. Nakon pregleda svih odgovora, najčešća područja specijalizacije ispitanika,

a koja su relevantna za anketu, bila su prevoditeljstvo, lingvistika, nastavništvo i književno-kulturološki smjerovi. U manjini su novinarstvo, arhivistika i krovni termin „filologija“. Vezano uz područja specijalizacije, najčešći jezici u pitanju su engleski, francuski, ruski, talijanski i hrvatski. Sljedeće pitanje bilo je vezano uz trenutačno mjesto zaposlenja. Čak 41 (37,3%) ispitanik zaposlen je u ustanovi visokog, višeg, srednjeg ili osnovnog obrazovanja, 19 (17,3%) ispitanika su studenti, 18 (16,4%) ispitanika zaposleno je u prevoditeljskim agencijama i školama stranih jezika, 16 (14,5%) ispitanika je nezaposleno, 5 (4,5%) ispitanika su *freelanceri* ili vlasnici vlastitih obrta ili tvrtki, te je 11 (10%) ispitanika zaposleno u ostalim granama djelatnosti. Sljedeće je pitanje bilo vezano uz strane jezike koje ispitanici govore. Očekivano, najviše ispitanika označilo je da govori engleski, njih čak 109 (99,1%). Nadalje, njemački govori 47 (42,7%) ispitanika, francuski 25 (22,7%), španjolski 21 (19,1%), talijanski 18 (16,4%), ruski 14 (12,7%), slovenski 9 (8,2%), švedski 8 (7,3%), norveški 4 (3,6%). Po 3 (2,7%) ispitanika govore japanski, slovački, češki, poljski i nizozemski, po 2 ispitanika govore makedonski, mađarski, rumunjski, kineski i grčki te po 1 ispitanik govori ukrajinski, korejski, portugalski i turski. Na pitanje koji su najčešći jezični izvori/materijali kojima se koriste u svojem poslu ili obrazovanju, ispitanici su mogli odabrati samo dvije vrste. Najveći je broj ispitanika odabrao rječnike, njih 86 (78,2%). Ukupno 39 (35,5%) ispitanika odabralo je udžbenike i priručnike, 36 (32,7%) ih je odabralo gramatike, 28 ih je odabralo korpuse (25,5%) i ukupno je 14 ispitanika odabralo „ostalo“, pod čime su naveli internetske resurse i repozitorije, prijevodne memorije i terminološke baze, pravopise, znanstvene knjige i članke te usluge automatskog prevođenja poput Google Translatea.

5.2.2. Rezultati za dio „Upoznatost s korpusnom lingvistikom“

Ukupno 44 (40%) ispitanika afirmativno je odgovorilo na pitanje jesu li ikada sudjelovali na kongresu, seminaru, radionici ili tečaju iz korpusne lingvistike ili jesu li imali određeni doticaj s lingvističkim korpusima. Sljedeće je pitanje bilo vezano uz prethodno, i cilj mu je bio saznati koji su naslovi ili teme ispitanicima bili najkorisniji u slučaju da su sudjelovali na navedenim oblicima edukacije. Na pitanje je odgovorilo 33 (30%) ispitanika. Ukupno 3 korisnika odgovorilo je da se ne sjećaju ili ne znaju. Tablica niže prikazuje odgovore ostalih 30 ispitanika.

Popis korpusa koje mogu koristiti	Dostupni korpusi, vrste korpusa, pretraživanje korpusa, specijalizirani korpusi
Pretraživanje korpusa i obrada korpusnih podataka	Pretraživanje korpusa, korpusni alati i tehnike
Računalni korpusi	E-adrese različitih računalnih korpusa
Programiranje	Kako se koristiti korpusima
Općenito upoznavanje s načinom uporabe korpusa	Postojeći korpusi na hrvatskom jeziku
Pretraživanje korpusa	Uporaba korpusa
British National Corpus (BNC)	Pretraživanje korpusa
Korpusna analiza	*Tips and tricks* za što učinkovitije pretraživanje korpusa
Koncept *word sketches*	Kako upotrebljavati korpus, kako napisati rad zasnovan na materijalu iz korpusa
Sastavljanje korpusa	Dizajn korpusa, obilježavanje, paralelni korpusi, statistika u korpusnoj lingvistici
Programiranje u svrhu pretraživanja korpusa	Služim se korpusom, no nisam bio na radionicama
Uvod u indoeuropsku lingvistiku	Pretraživanje korpusa, sastavljanje korpusa
Kako prikupljati korpus; CHILDES, CHAT	Upotreba jezičnih korpusa u prevoditeljskoj djelatnosti; Strojno prevođenje i jezični korpusi EU
Učila sam uglavnom samostalno, a ne na tečajevima	Nepostojanje sveobuhvatnog korpusa suvremenog srpskog jezika
Regularni izrazi, statistika	Sketch Engine, Termex

Tablica 1. *Prikaz odgovora ispitanika na pitanje koji su im naslovi ili teme bili najkorisniji na radionici ili tečaju*

Nadalje, samo je 22 (20%) ispitanika afirmativno odgovorilo na pitanje jesu li ikada slušali kolegij iz korpusne lingvistike. Sljedeće je pitanje bilo namijenjeno ispitanicima koji su na prethodno pitanje odgovorili afirmativno, gdje je cilj bio saznati kada su odslušali zadnji kolegij iz korpusne lingvistike. Međutim, ovdje se odazvao veći broj ispitanika, njih 46 (41,8%), od kojih je 23 (50%) ispitanika izjavilo da nikada nisu pohađali takav kolegij, 11 (23,9%) ispitanika izjavilo je da su ga odslušali prije 2 do 3 godine, 6 (13%) ispitanika izjavilo je da ga je odslušalo unutar protekle godine dana, 5 (10,9%) ispitanika izjavilo je da je odslušalo prije više od 5 godina i 1 (2,2%) ispitanik izjavio je kako ga je odslušao prije 4 do 5 godina.

Na pitanje koliko su kolegija koji posjeduju elemente korpusne lingvistike odslušali, 46 (41,8%) ispitanika odgovorilo je da nisu odslušali nijedan takav kolegij, 33 (30%) ispitanika

odgovorilo je da je odslušalo 1 takav kolegij, 17 (15,5%) ispitanika odgovorilo je da su odslušali 2 takva kolegija, 9 (8,2%) ispitanika odgovorilo je da su odslušali 3 takva kolegija i 5 (4,5%) ispitanika odgovorilo je da su odslušali više od 3 takva kolegija. Sljedeće pitanje bilo je jesu li se ispitanici u obrazovanju ili poslu ikada koristili metodama i alatima iz korpusne lingvistike. Greškom, u razdoblju od jednog dana u postavkama ankete nije bilo podešeno da je pitanje obvezno, što je rezultiralo manjim odazivom od ukupno 106 ispitanika. Od navedenih 106 ispitanika, afirmativno ih je odgovorilo 62 (58,5%). Sljedeće je pitanje bilo vezano uz prethodno, a cilj mu je bio saznati kojim su se metodama i alatima iz korpusne lingvistike ispitanici koristili u slučaju da su afirmativno odgovorili. Na pitanje je ukupno odgovorilo 55 (50%) ispitanika, od kojih su 2 ispitanika izjavila kako se ne sjećaju i da ih je previše, a odgovori ostalih 53 ispitanika navedeni su u tablici niže.

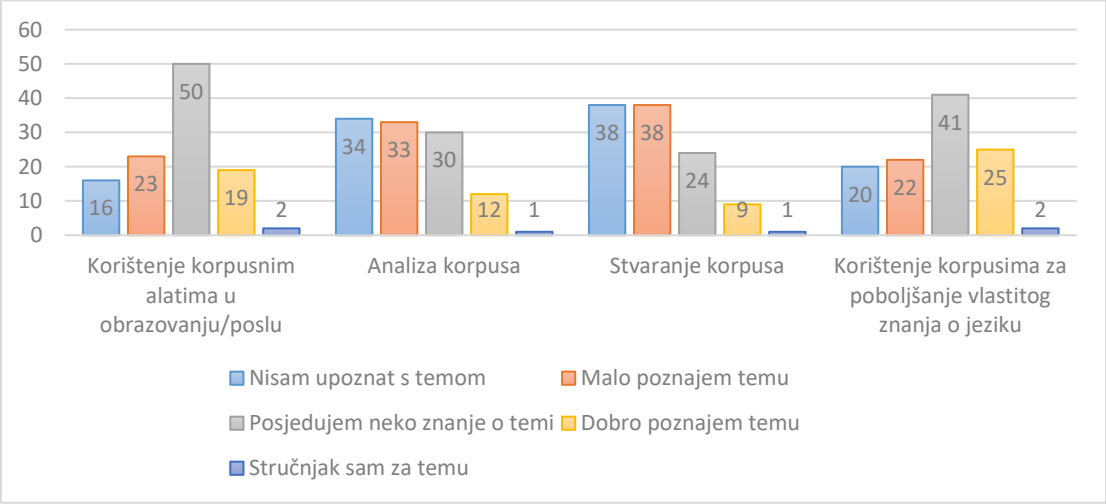
Korpusom engleskog jezika	MSD označavanje, normalizacija, lematizacija, parseri, NoSketch Engine, Nooj
Corpus.byu.edu	Uspoređivanje termina
BootCat, Intelitext...	Leksikologija
Fran.si, Gigafida	Sketch Engine, AntConc
Konkordancije, kolokacijske baze	Analiza stanja na korpusima u odnosu na odabranu temu, tj. određen specijaliziran temat određene znanstvene grane ili djela
Riznica IHJJ; Linguee.fr, Eur-Lex	Dostupnim materijalom u sveučilišnoj knjižnici
Hrvatski nacionalni korpus, Riznica IHJJ, British National Corpus, Corpus of Contemporary American English, Corpus of Middle English Prose and Verse	Hrvatskim računalnim korpusima
Sketch Engine, Wordsmith...	hrWaC i Hrvatski nacionalni korpus
Pretraživanje korpusa (kao pomoć pri prevođenju)	Pretraživanje korpusa
Sketch Engine, Bonito	Analiza korpusa, stvaranje korpusa
Korpus hrvatskog jezika i korpus poljskog jezika	Pretraživanjem korpusa radi provjere da koristi li se neki izraz i u kojoj mjeri
Pretraga korpusa po kolokacijama i prema vrstama riječi	Malo sam se koristila alatima iz korpusne lingvistike. Većinom je to bilo pretraživanje korpusa za razne zadatke iz jezičnih vježbi na fakultetu
FIDA, Gigafida	Sketch Engine, LF Aligner
Wordsmith	Pretraživanje korpusa - traženje primjera za ispite/vježbe ili za prevođenje (kolokacije, kombinacije riječi i prepozicija npr.)
Sketch Engine, AntConc	Crawlanje, statistička analiza podataka
Traženje kolokacija i provjeravanje učestalosti pojavljivanja riječi i fraza	Pretraživanje korpusa, označavanje za korpus

Treebanking	Pretraživanje korpusa za određivanje sličnosti između antonima u engleskom
Online rječnici (Cambridge; Oxford; Theasurus.com; Urban dictionary; hrvatski jezični portal)	British National Corpus, Corpus of Contemporary American English, IATE, Eurovoc, Hrvatski nacionalni korpus, Struna, Corpuscle
COCA - Corpus of Contemporary American English	Korpusima
AntConc, WordNet	Wordsmith
HJK, Slovenský národný korpus	COCA, Språkbanken
Bonito, Hrvatski milijunski korpus, Čestotni rječnik	Online korpusima
CHILDEST, CHAT	KWIL, KWIC
Primjena u istraživanju aspektualnih vrijednosti	Korpus američkog govornog jezika COCA, srpski korpus Matematičkog fakulteta
COCA - Corpus of Contemporary American English	Sketch Engine, Hrvatski nacionalni korpus
Kolokacije	HNK
Najviše baza podataka CHILDES i CLAN programi, zatim Linguee.fr	

Tablica 2. *Prikaz odgovora ispitanika na pitanje kojim su se metodama i alatima iz korpusne lingvistike koristili u obrazovanju ili poslu*

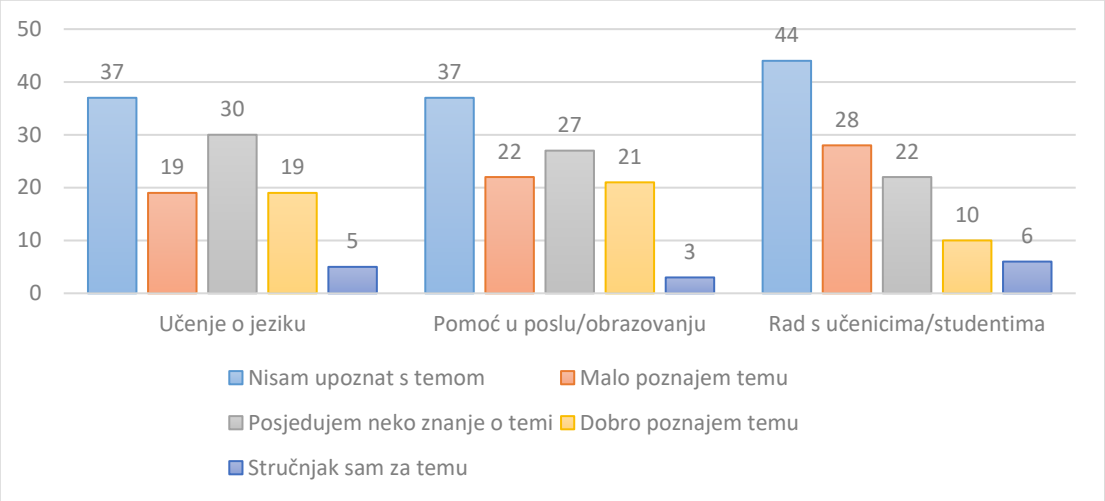
5.2.3. Rezultati za dio „Vaše poznavanje korpusnih alata“

Cilj prvog pitanja bio je provjeriti koje znanje ispitanici posjeduju o određenim temama vezanim uz korpusnu lingvistiku. Rezultati su prikazani u Dijagramu 1.



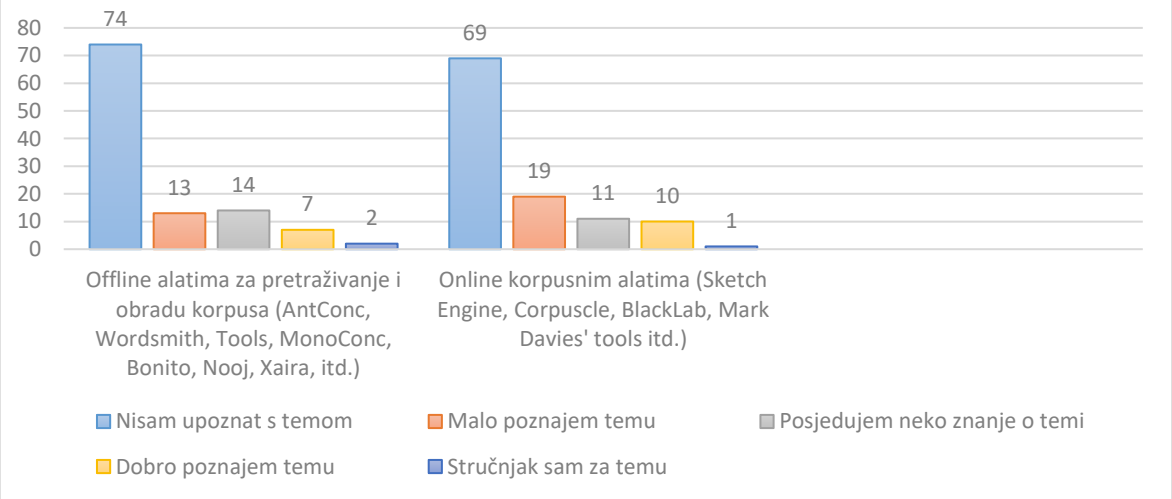
Dijagram 1.

Cilj drugog pitanja bio je saznati koliko ispitanici znaju o analizi i korištenju konkordancija u pojedinim područjima. Rezultati su prikazani u Dijagramu 2.



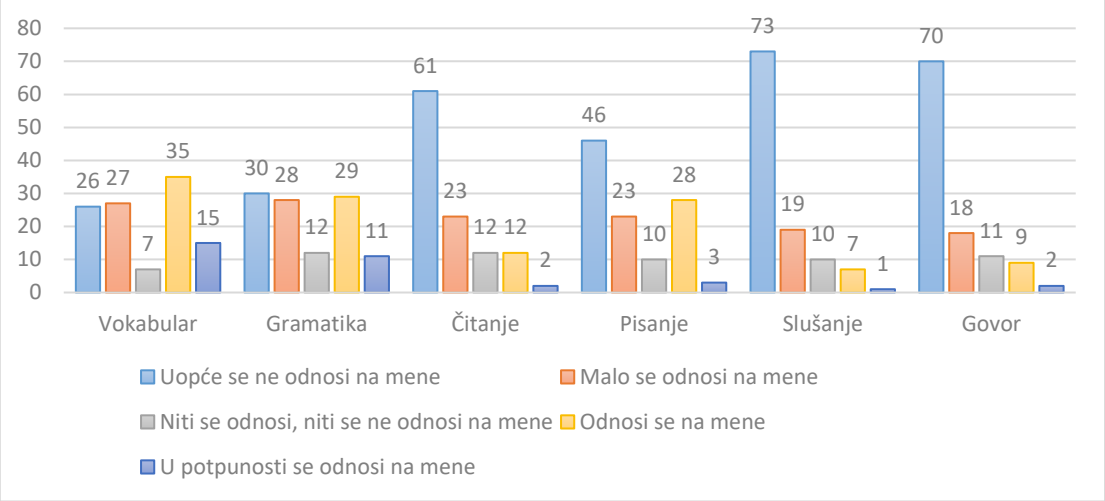
Dijagram 2.

Cilj trećeg pitanja u ovom dijelu ankete bio je saznati koliko ispitanici znaju o korištenju *offline* i *online* alatima za pretraživanje i obradu korpusa te učenju o njima. Rezultati su prikazani u Dijagramu 3.



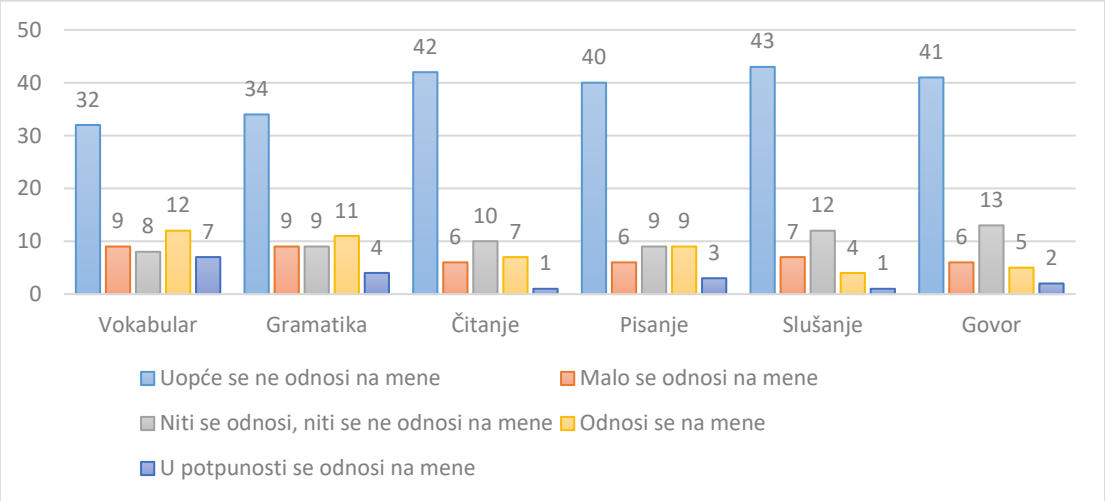
Dijagram 3.

Cilj četvrtog pitanja bio je saznati koriste li se ispitanici korpusnim metodama kako bi kod sebe popravili vokabular, gramatiku, čitanje, pisanje, slušanje i govor. Rezultati su prikazani u Dijagramu 4.



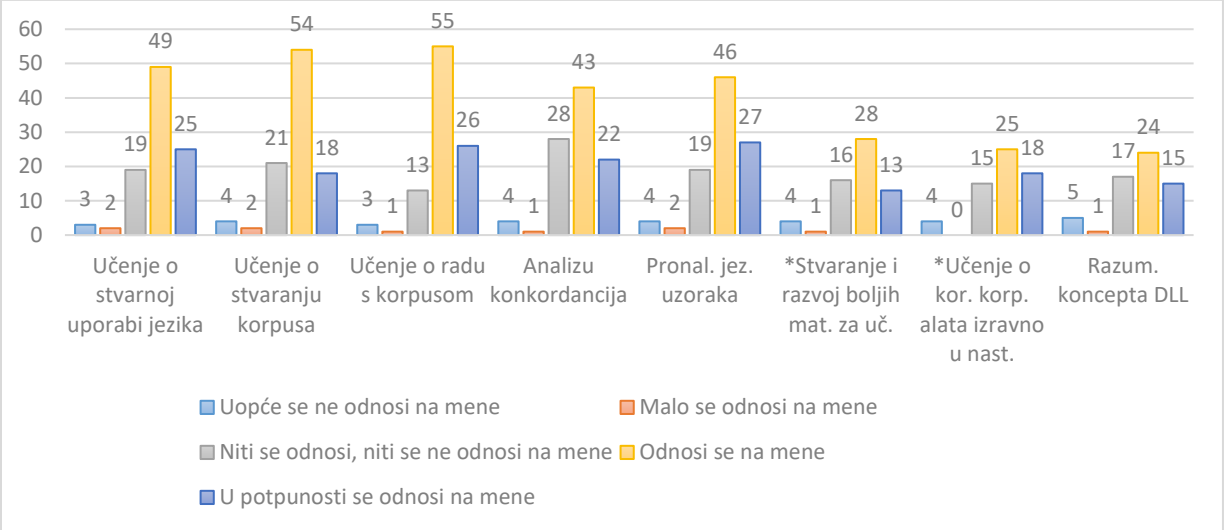
Dijagram 4.

Peto pitanje bilo je naznačeno kao pitanje isključivo namijenjeno prosvjetnim djelatnicima. Cilj ovog pitanja bio je saznati koriste li se ispitanici korpusnim metodama kako bi kod svojih učenika ili studenata popravili vokabular, gramatiku, čitanje, pisanje, slušanje i govor. Rezultati su prikazani u Dijagramu 5.



Dijagram 5.

Svrha šestog pitanja bila je saznati slažu li se ispitanici s tvrdnjom da im sudjelovanje u organiziranoj edukaciji o korpusnoj lingvistici koristi za određene istraživačke i obrazovne aktivnosti. Posljednje su tri aktivnosti bile namijenjene isključivo prosvjetnim djelatnicima. Rezultati su prikazani u Dijagramu 6. Važno je napomenuti kako je zbog posljednje 3 aktivnosti bilo potrebno isključiti zahtijevanje odgovora u svakom retku. Uslijed toga, neki ispitanici nehotice nisu riješili i prvih 6 aktivnosti koje su se odnosile na sve ispitanike, što je rezultiralo manjim brojem odgovora.

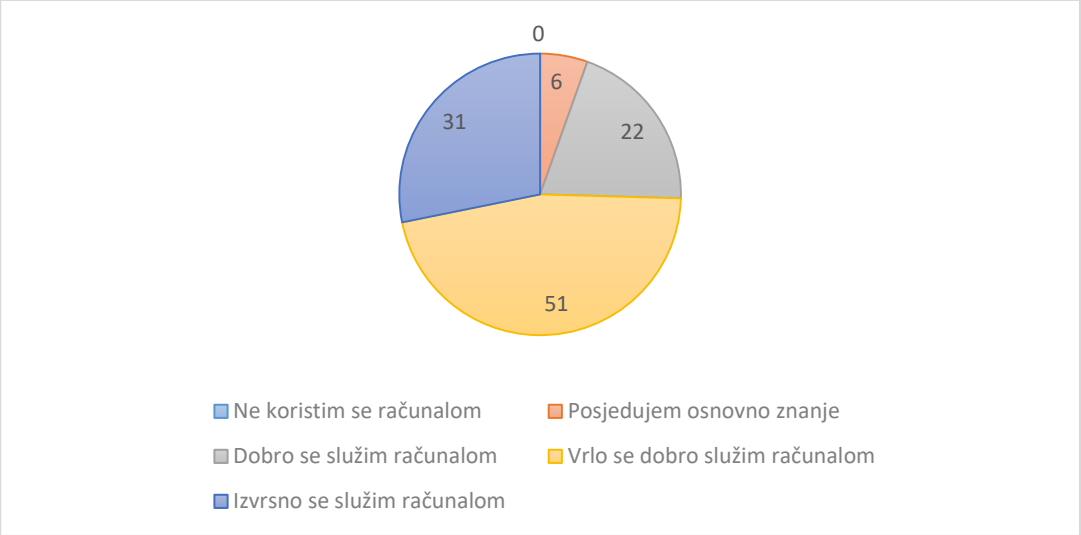


Dijagram 6.

Na pitanje kojim su se od hrvatskih računalnih korpusa koristili, korisnici su imali mogućnost višestrukog odabira. Najviše je ispitanika odabralo Hrvatski nacionalni korpus (HNK), njih čak 54 (49,1%). Zatim, 39 (35,5%) ispitanika odabralo je Riznicu IHJJ-a, 35 (31,8%) ispitanika izjavilo je kako se ne koristi nijednim računalnim korpusom hrvatskog jezika, 30 (27,3%) ispitanika odabralo je Korpus leksikografskog zavoda Miroslav Krleža i 27 (24,5%) ispitanika odabralo je *web* korpus hrWaC. Pod opcijom „Ostalo“, navedena su dodatna tri izvora. Po jednom se spominju Corpus of Contemporary American English (COCA), Korpus savremenog srpskog jezika na Matematičkom fakultetu Univerziteta u Beogradu i Hrvatski jezični portal. Iako od posljednja dva izvora jedan nije hrvatski, a drugi se ne smatra korpusom, navedeni su svi odgovori radi vjernosti anketi.

Posljednja dva pitanja anketnog upitnika od ispitanika su tražili da opišu, odnosno ocijene sebe u tehnološko-računalnoj okolini. Na pitanje kako bi se opisali kao korisnici tehnologije i računala, 42 (38,2%) ispitanika opisalo se izjavom „kada se ostali počnu koristiti novom

tehnologijom, tada sam je i ja željan/na isprobati“, 39 (35,5%) ispitanika opisalo se izjavom „uvijek isprobavam nove tehnologije“, 22 (20%) ispitanika opisalo se izjavom „kada mnoštvo ljudi sudjeluje, tada se i ja pridružim“, 6 (5,5%) ispitanika opisalo se izjavom „među zadnjima sam od svojih kolega koji isprobavaju nove tehnologije (kada se više ne smatraju „novima“), no na kraju ih isprobam“ i samo 1 korisnik opisao se izjavom „ne volim i ne trebam nove tehnologije te je moj napredak spor, ukoliko uopće i postoji“. Posljednje pitanje zahtijevalo je od ispitanika da ocijene svoj stupanj informatičke pismenosti, odnosno služenja računalom i računalnim alatima. Od 110 ispitanika, nijedan ispitanik nije izjavio kako se ne koristi računalom, 6 (5,5%) ispitanika izjavilo je kako posjeduju osnovno znanje, 22 (20%) ispitanika izjavilo je kako se dobro služi računalom, 51 (46,4%) ispitanik izjavio je kako se vrlo dobro služi računalom i 31 (28,2%) ispitanik izjavio je kako se izvrsno služi računalom. Rezultati su prikazani u Dijagramu 7.



Dijagram 7.

5.3. Osvrt na rezultate anketnog upitnika

Uzevši u obzir sveukupne rezultate ankete, može se doći do nekoliko zanimljivih zaključaka. Prvenstveno je važno napomenuti kako su se, prema anketi, gotovo svi ispitanici opisali kao informatički, odnosno računalno pismene osobe u određenim razinama, što je veliki poticaj za tematiku ovog anketnog upitnika. Iako se sa sigurnošću može ustanoviti da korpusna lingvistika nema marginaliziran položaj među uzorkom jezičara nad kojim je provedeno ovo istraživanje, ona svakako nije zastupljena niti korištena u očekivanoj mjeri. U upitniku je sudjelovalo najviše ispitanika iz dobne skupine od 20-29, njih oko 60%, što znači da pripadaju

generacijama koje su od rane dobi izložene uporabi računala i koje ne bi trebale imati ozbiljnijih problema s usvajanjem novih računalnih alata. Više od polovice ispitanika nije imalo doticaj s lingvističkim korpusima ili ikakvom vrstom organizirane edukacije za korpusnu lingvistiku (kongresi, seminari, tečaji ili radionice), a velika ih većina nije ni slušala kolegij iz korpusne lingvistike. Sudeći prema tome, s obrazovnog aspekta, korpusna lingvistika i dalje je nedovoljno eksponirana, prezentirana ili promicana kao pouzdan i koristan izvor za kojekakve jezične svrhe. Međutim, određeni rezultati ipak upućuju na situaciju koja nije toliko „siva". Od ispitanika koji jesu odslušali kolegij iz korpusne lingvistike, najviše ih pripada u kategorije onih koji su ga odslušali prije 2 do 3 godine ili unutar protekle godine dana. To upućuje da se, u obrazovnom smislu, situacija popravila s obzirom na prijašnje godine, što je vrlo ohrabrujući podatak i svakako je poželjno da se takav trend nastavi. Nemoguće je očekivati da se obrazovne promjene ili novosti dogode naglo – najčešće je slučaj da se uvode postepeno.

Što se tiče kolegija koji posjeduju elemente korpusne lingvistike, više od polovice ispitanika slušalo je barem jedan, što je isto vrlo ohrabrujući podatak. Ovi rezultati ukazuju na blagi porast zastupljenosti korpusne lingvistike u obrazovanju – iako nije na zavidnoj razini, svakako se može smatrati dobrim početkom. Iako veliki broj ispitanika nije sudjelovao na organiziranoj edukaciji ili kolegiju iz korpusne lingvistike, zanimljiv je podatak da se više od pola ispitanika koristi korpusnim alatima i metodama u obrazovanju ili poslu. Vrlo je moguće da su do tih alata ili metoda došli putem osobnih poznanstava ili samostalnom naobrazbom i istraživanjem, što je vrlo pohvalno, iako je mnogo učinkovitije i jednostavnije da osobe mogu pohađati nekakav oblik naobrazbe u tom pogledu. To je posebice istinito kada se u obzir uzme činjenica da sve osobe nisu jednako voljne ili sposobne samostalno istražiti ovo područje. Također je zanimljivo da je većina ispitanika na pitanje koriste li se određenim *offline* i *online* korpusnim alatima odgovorila da nisu upoznati s temom, što, pak, može biti indikacija da su oni ispitanici koji su izjavili da se koriste korpusnim alatima i metodama u poslu ili obrazovanju većinom korisnici s najosnovnijim znanjem o služenju korpusnim alatima, da su vičniji pamćenju imena samih korpusa, a ne alata kojima ih pretražuju i slično.

Trećina ispitanika izjavila je kako se ne služi nijednim hrvatskim korpusom, što nije pretjerano zabrinjavajući podatak. Očekivano, najviše ih se koristilo Hrvatskim nacionalnim korpusom i Riznicom IHJJ-a, a ponešto manji broj ispitanika koristilo se Korpusom leksikografskog zavoda Miroslav Krleža i korpusom hrWaC. Iako je većina ispitanika koristila jedan ili više ovih korpusa, status korpusa u hrvatskoj očito nije ni blizu statusa rječnika, udžbenika i priručnika te gramatika kao primarnih jezičnih pomagala. Na pitanje kojim se

pomagalom najčešće koriste, velika je većina odabrala rječnike kao primarna pomagala. Prema glasovima, korpusi su bili posljednji od četiri ponuđena pomagala – dakle, iza rječnika, udžbenika i priručnika te gramatika, tim redoslijedom. Međutim, tu je najvjerojatnije prisutno više uzroka. Budući da Hrvatska nije mnogoljudna zemlja, korpusna lingvistika počela se razvijati mnogo kasnije nego u zemljama koje predvode u tom području, stoga je i za očekivati znatno rjeđu uporabu korpusa te njihovo promicanje kao lingvističkih izvora. Nadalje, rezultati ukazuju da je u obrazovni kurikulum potrebno implementirati više kolegija iz korpusne lingvistike ili s elementima korpusne lingvistike u svim područjima koje imaju veze s jezikoslovljem. Slično tome, bilo bi poželjno ostvariti da veći broj jezičara pohađa određeni oblik organizirane edukacije iz korpusne lingvistike, što se može ostvariti učestalijim organiziranjem navedenih oblika edukacije, učinkovitijom promidžbom i slično.

6. Zaključak

Svrha rada bila je prikazati položaj korpusne lingvistike unutar područja jezikoslovlja, s posebnim naglaskom na njezine prednosti kao istraživačkog i obrazovnog izvora kroz detaljan pregled nekih od računalnih alata i pomagala pomoću kojih korisnici svih razina obrazovanja i kvalifikacija mogu na određeni način proučavati jezik. Od njezinih začetaka i nijekanja od strane generativnih gramatičara do današnjeg doba, korpusna lingvistika razvila se u značajan lingvistički izvor zbog napretka računalne tehnologije, a računalni korpusi pokazuju se kao sve pouzdaniji izvor za proučavanje raznovrsnih elemenata jezika, od vokabulara, jezičnih uzoraka i obrazaca, do semantičke interpretacije i samog gramatičkog aspekta jezika. Iako se oko klasifikacije korpusne lingvistike još vode rasprave oko okvira unutar kojih se ona može definirati, kao samostalna disciplina, zasebna teorija ili puka metodologija, njezin empirijski pristup jeziku uvelike može pružiti uvid u stvarnu uporabu jezika i, što je još važnije, njegovu promjenu, odnosno, evoluciju kroz vrijeme. Napredak računalne tehnologije do razine na kojoj je danas omogućio je izradu mnoštva alata i tražilica koje služe za izradu, uređivanje i pretraživanje korpusa. Iako ih ima mnogo više, središnja točka ovog rada bilo je ukupno osam alata – Corpuscle, Sketch Engine, BlackLab, IMS Open Corpus Workbench (CWB), Xaira, Poliqarp, PhiloLogic i TEITOK. Iako navedeni softverski paketi imaju primarnu zajedničku svrhu, svaki se razlikuje po pregršt značajki i mogućnosti koje korisniku nude za pretraživanje korpusne građe. U radu su predstavljene značajke i arhitektura svakog od alata i on, kao takav, predstavlja svojevrsni pregled raznovrsnih mogućnosti kojima se korisnici mogu služiti ukoliko se odluče baviti korpusnom lingvistikom u bilo kojem kapacitetu. Završni dio rada jest anketno istraživanje provedeno nad 110 ispitanika koji se na razne načine bave jezikoslovljem, čija je svrha bila utvrditi koliko je jezičara upoznato s korpusima i njihovom uporabom. Prema rezultatima ankete, 60% ispitanika nikada nije imalo doticaj s bilo kojim oblikom organizirane edukacije o korpusnoj lingvistici, a 80% ih nikada nije odslušalo kolegij iz korpusne lingvistike, no gotovo 60% njih služi se korpusnim alatima i metodama u obrazovanju ili poslu. Njih 67,3% nije upoznato s raznim *offline* korpusnim alatima, a 62,7% s *online* korpusnim alatima, iako ih velika većina vjeruje kako bi pohađanjem organizirane edukacije mnogo toga naučili o uporabi korpusnih alata. Kako je bilo očekivano, 95% ispitanika smatra se informatički pismenima i sposobnima u određenim razinama, što je uistinu veliki poticaj da se korpusnu lingvistiku uvede u većoj mjeri u kurikulum ili da se organiziraju tečajevi, seminari ili radionice. Iako ima mnogo prednosti, i dalje je relativno nepoznata među jezičarima i upravo je to razlog da se korpusnu

lingvistiku počne implementirati kao jedan od najkorisnijih izvora za istraživanje fenomena jezika.

7. Literatura

- Adam Kilgarriff. URL: http://kilgarriff.co.uk/cv.htm (19.4.2017.)
- Allen, T.; Gladstone, C.; Whaling, R. PhiloLogic4: An Abstract TEI Query System. // Journal of the Text Encoding Initiative [online], 5(2013). URL: https://jtei.revues.org/817 (20.5.2017.)
- Andersen Francis I.; Dean Forbes A. Biblical Hebrew Grammar Visualized. Warsaw, IN: Eisenbrauns, 2012.
- ARTFL Project Research Blog. General Overview Of PhiloLogic4'S Web Architecture. URL: https://artfl.blogspot.hr/2014/12/general-overview-of-philologic4s-web.html (23.5.2017.)
- ARTFL Project Research Blog. PhiloLogic4: The Big Picture. URL: https://artfl.blogspot.hr/2014/12/philologic4-big-picture.html (20.5.2017.)
- ARTFL Project. PhiloLogic. URL: https://sites.google.com/site/philologic3/manual (18.5.2017.)
- Aston, G.; Burnard L. BNC Handbook: Exploring the British National Corpus With SARA. Edinburg: Einburgh University Press, 1998.
- Bennett, Gena R. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. Ann Arbor, MI: University of Michigan Press: 2010.
- Bratanić M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filolofija. 30-31(1998), str. 171-177
- Brieven als Buit. URL: http://brievenalsbuit.inl.nl/zeebrieven/page/search (25.4.2017.)
- Brown Corpus. URL: https://corpling.uis.georgetown.edu/cqp/brown/ (30.4.2017.)
- Chomsky N. Aspect of the Theory of Syntax. Cambridge, Mass: The M.I.T. Press, 1970.
- CLUL. COPLE2. URL: http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=home (5.6.2017.)
- CLUL. TEITOK. URL: http://alfclul.clul.ul.pt/teitok/site/index.php?action=about (2.6.2017.)
- CroALa. URL: http://croala.ffzg.unizg.hr/ (25.4.2017.)
- Croatica et Tyrolensia. Technical note: installing PhiloLohic 4 on localhost. URL: http://crotyr.hypotheses.org/99 (28.5.2017.)

- Ćavar, D.; Brozović Rončević, D. Riznica: The Croatian Language Corpus. // Prace filologiczne. 62 (2012), str. 51-65
- Daniel Janus. Smyrna: An easy Polish concordancer in Clojure. URL: http://danieljanus.pl/talks/reveal.js/2014-lambdadays.html#/7 (15.5.2017.)
- Dobrić N. Corpus Linguistics – The Basic Form of Linguistic Analysis. // Philolohgia. 7(2009), str. 31-41
- Dukes, K.; Atwell, E.; Habash, N. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. // *Language Resources and Evaluation Journal*. 47, 1(2011), str. 33 – 62
- Evert S.; Hardie, A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. // Proceedings of the Corpus Linguistics 2011 conference. Birmingham: University of Birmingham, 2011.
- Github. BlackLab. URL: http://inl.github.io/BlackLab/index.html (25.4.2017.)
- GitHub. PhiloLogic4. URL: https://github.com/ARTFL-Project/PhiloLogic4/blob/master/README.md (20.5.2017.)
- Hardie. Introduction to Xaira Part One: All about Xaira. URL: http://slideplayer.com/slide/793318/ (4.5.2017.)
- Hrvatski nacionalni korpus. Rječnik korpusne lingvistike. URL: www.hnk.ffzg.hr/bb/definicijekl.doc (4.7.2015.)
- Hrvatski nacionalni korpus. URL: http://www.hnk.ffzg.hr/default.htm
- IMS. IMS Corpus Workbench (CWB). URL: http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html (28.4.2017.)
- Janssen, M. TEITOK: Text-Faithful Annotated Corpora. // Proceedings of the Tenth International Conference on Language Resources and Evaluation. Paris: ELRA, 2016.
- Janus, D.; Przepiórkowski, A. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. // The proceedings of Practical Applications of Linguistic Corpora 2005 / uredili Jacek Walinski, Krzysztof Kredens i Stanisław Goźdź-Roszkowski. Frankfurt am Main: Peter Lang, 2005.
- Kilgarriff, A., Baisa, V., Bušta, J. et al. The Sketch Engine: ten years on. // Lexicography ASIALEX. 1 (2014), str. 7-36

- Klobučar Srbić I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica: časopis za leksikografiju i enciklopedistiku. 2, 2/3 (2008), str. 39-51
- Ljubešić, N.; Klubička, F. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. // Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014. / uredili su Felix Bildhauer i Roland Schäfer. Gothenburg: Association for Computational Linguistics, 2014.
- Malmkjaer, K. The Linguistics Encyclopedia. London; New York: Routledge, 2006.
- Mark Davies. Corpus of American Contemporary English. URL: http://corpus.byu.edu/coca/ (2.4.2017.)
- McEnery T. Corpus Linguistics. // The Oxford Handbook of Computational Linguistics / uredio Ruslan Mitkov. New York: Oxford University Press, 2003. Str. 448-462
- McEnery T.; Wilson A. Corpus Linguistics: An Introduction. 2nd ed. Edinburgh: Edinburgh University Press, 2001.
- McEnery, A.; Xiao, R. Character Encoding in Corpus Construction. // Developing Linguistic Corpora: a Guide to Good Practice / uredio Martin Wynne. Oxford: Oxbow Books, 2005. Str. 59-70
- McEnery, T.; Hardie, A. Corpus linguistics: Method, Theory and Practice. Cambridge, England: Cambridge University Press, 2011.
- Megyesi, Beata B. Corpus usage. Uppsala University. URL: http://stp.lingfil.uu.se/~bea/uv/uv09/dokverktyg/02-corpususage.pdf (5.4.2017.)
- Meurer, P. Corpuscle – a new corpus management platform for annotated corpora. // Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian/ uredio Gisle Andersen. Amsterdam; Philadelphia: John Benjamins Pub. Co., 2012. Str. 29-50
- Mikelić Preradović, N. CroLTeC. URL: http://teitok.iltec.pt/croltec/index.php?action=cqp&act=advanced (9.6.2017.)
- Mikelić Preradović, N.; Berać, M.; Boras, D. Learner corpus of Croatian as a second and foreign language // 2015 Multidisciplinary approaches to multilingualism. Frankfurt am Main: Peter Lang, 2015. Str. 107-126.
- National Corpus of Polish. URL: http://nkjp.pl/poliqarp (15.5.2017.)
- Natural Language Processing group. People. URL: http://nlp.ffzg.hr/people/ (12.4.2017.)

- No Sketch Engine. hrWaC.
  URL: http://nl.ijs.si/noske/all.cgi/first_form?corpname=hrwac;align= (25.4.2017.)
- Przepiórkowski, A.; Krynicki, Z.; Dębowski, Ł.; Woliński, M., Janus, D., Bański, P. A Search Tool for Corpora with Positional Tagsets and Ambiguities. // Proceedings of LREC 2004. Lisabon: ELRA, 2004. Str. 1235 - 1238
- Seuren, Pieter .A.M., Western Linguistics: An Historical Introduction. Malden: Blackwell Publishing. 1998.
- Sinclair, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.
- Sveučilišni računski centar. Hrvatski jezični portal.
  URL: http://hjp.znanje.hr/index.php?show=search_by_id&id=elllXRQ%3D
  (1.4.2017.)
- Tadić M. Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika. // Filologija. 30-31(1998), str. 337-347
- Tadić, M. Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. // Suvremena lingvistika. 43-44, 1-2(1997), str. 387-394
- TEI Wiki. Xaira. URL: https://wiki.tei-c.org/index.php/Xaira (3.5.2017.)
- Teubert W.; Čermáková A. Corpus Linguistics: a short introduction. London; New York: Continuum, 2007.
- Text Encoding Initiative. What is the TEI?
  URL: http://www.tei c.org/Vault/SC/J31/WHAT.htm (2.4.2017.)
- The Sketch Engine. URL: https://www.sketchengine.co.uk/ (19.4.2017.)
- Tognini-Bonelli E. Corpus Linguistics at Work. Amsterdam; Philadelphia: J. Benjamins, 2001.
- Uni Computing. Corpuscle.
  URL: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page#
  (15.4.2017.)
- Uni Research. URL: http://uni.no/en/about-uni-research/ (15.4.2017.)
- University of Chicago Library. PhiloLogic.
  URL: https://www.lib.uchicago.edu/efts/ARTFL/philologic/ (18.5.2017.)
- University of Essex. Corpus Linguistics. URL:
  http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction2.html (5.4.2017.)

- University of Oxford IT Services. Xaira.
  URL: http://projects.oucs.ox.ac.uk/xaira/Doc/refman.xml?ID=X01 (1.5.2017.)
- Wikipedia. Corpus Linguistics.
  URL: https://en.wikipedia.org/wiki/Corpus_linguistics (21.3.2017.)
- Wikipedia. Poliqarp.
  URL: https://en.wikipedia.org/wiki/Corpus_linguistics (9.5.2017.)
- Williams G. Reviews: English Collocation Studies: The OSTI report. // Interntional Journal of Corpus Linguistics. 2, 2(2005), str 257 – 266
- Xiao, R. Well-known and influential corpora. // Corpus Linguistics: An International Handbook / uredili Anke Lüdeling i Merja Kytö. Berlin: De Gruyter Mouton, 2008.
- Xiao, R.; Hu, X. Corpus-Based Studies of Translational Chinese in English-Chinese Translation. New York, Dodrecht, London: Springer Heidelberg, 2015.
- Željka Salopek. Pet godina zbirke Croatiae auctores Latini i prva godina projekta Croatica et Tyrolensia. URL: http://dfest.nsk.hr/wp content/themes/boilerplate/2015/prezentacije/Salopek_Zeljka.pdf (28.5.2017.)

Dodatak 1. – Anketni upitnik

1) OSOBNE INFORMACIJE

1. **Spol**
   - ○ Žensko
   - ○ Muško
2. **Dob**
   - ○ manje od 20 godina
   - ○ 20 – 29 godina
   - ○ 30 – 39 godina
   - ○ 40 – 49 godina
   - ○ 50 – 59 godina
   - ○ 60 godina i više
3. **Koja je Vaša trenutačna razina obrazovanja?**
   - ○ univ.bacc. u tijeku
   - ○ univ.bacc.
   - ○ mag. u tijeku
   - ○ mag.
   - ○ dr.sc. / dr.art. u tijeku
   - ○ dr.sc. / dr.art.
4. **Kada ste završili formalno obrazovanje?**
   - ○ Trenutačno studiram
   - ○ Unutar protekle godine dana
   - ○ Prije 2-3 godine
   - ○ Prije 4-5 godina
   - ○ Prije više od 5 godina

5. **Koje Vam je područje specijalizacije najvišeg završenog stupnja obrazovanja?**

   _____

6. **Gdje ste trenutačno zaposleni (naziv institucije, fakulteta, ustanove i slično)?**

   _____

7. **Koje strane jezike govorite?**

   ○ Engleski

   ○ Njemački

   ○ Španjolski

   ○ Francuski

   Ostalo: _____

8. **Koji su najčešći jezični izvori/materijali kojima se koristite u svojem poslu ili obrazovanju?**

   ○ Udžbenici i priručnici

   ○ Gramatike

   ○ Rječnici

   ○ Korpusi

   Nešto drugo: _____

2) UPOZNATOST S KORPUSNOM LINGVISTIKOM

1. **Jeste li ikada sudjelovali na kongresu/ seminaru / radionici / tečaju iz korpusne lingvistike ili imali određeni doticaj s lingvističkim korpusima?**

   ○ Da
   ○ Ne

2. **Ako da, koji su Vam naslovi ili teme bili najkorisniji na radionici / tečaju?**

   _____

3. **Jeste li ikada slušali kolegij iz korpusne lingvistike?**

   ○ Da
   ○ Ne

4. **Ako da, kada ste odslušali zadnji kolegij iz korpusne lingvistike?**

   ○ Unutar godine dana
   ○ Prije 2-3 godine
   ○ Prije 4-5 godina
   ○ Prije više od 5 godina
   ○ Nikada nisam pohađao/la kolegij

5. **Koliko ste kolegija koji posjeduju elemente korpusne lingvistike odslušali?**

   ○ 1
   ○ 2
   ○ 3
   ○ Više od 3
   ○ Nijedan

6. **Jeste li se u obrazovanju ili poslu ikada koristili metodama i alatima iz korpusne lingvistike?**

   ○ Da
   ○ Ne

7. **Ako da, kojima?**

   _____

3) VAŠE POZNAVANJE KORPUSNIH ALATA

***pitanja označena dvjema zvjezdicama rješavaju isključivo osobe koje se bave prosvjetnom djelatnošću*

1. **Koja izjava najbolje opisuje Vaše znanje o sljedećim temama?**

	Nisam upoznat s temom	Malo poznajem temu	Posjedujem neko znanje o temi	Dobro poznajem temu	Stručnjak sam za temu
Korištenje korpusnim alatima u obrazovanju/ poslu	○	○	○	○	○
Analiza korpusa	○	○	○	○	○
Stvaranje korpusa	○	○	○	○	○
Korištenje korpusima za poboljšanje vlastitog znanja o jeziku	○	○	○	○	○

2. **Koja izjava najbolje opisuje Vaše znanje o sljedećim temama?**

Analiza i korištenje konkordancija za…

	Nisam upoznat s temom	Malo poznajem temu	Posjedujem neko znanje o temi	Dobro poznajem temu	Stručnjak sam za temu
Učenje o jeziku	○	○	○	○	○
Pomoć u poslu/ obrazovanju	○	○	○	○	○
Rad s učenicima/ studentima	○	○	○	○	○

3. **Koja izjava najbolje opisuje Vaše znanje o sljedećim temama?**

Korištenje i učenje o…

	Nisam upoznat s temom	Malo poznajem temu	Posjedujem neko znanje o temi	Dobro poznajem temu	Stručnjak sam za temu
Offline alatima za pretraživanje i obradu korpusa (Antconc, WordSmith Tools, MonoConc Pro, Bonito, NooJ, Xaira itd.)	○	○	○	○	○
Online korpusnim alatima (Sketch Engine, Corpuscle, BlackLab, Mark Davies' tools, etc.)	○	○	○	○	○

4. **Koristim se korpusnim metodama kako bih kod sebe poboljšala/poboljšao:**

	Uopće se ne odnosi na mene	Malo se odnosi na mene	Niti se odnosi, niti se ne odnosi na mene	Odnosi se na mene	U potpunosti se odnosi na mene
Vokabular	○	○	○	○	○
Gramatiku	○	○	○	○	○
Čitanje	○	○	○	○	○
Pisanje	○	○	○	○	○
Slušanje	○	○	○	○	○
Govor	○	○	○	○	○

5. **Koristim se korpusnim metodama kako bih kod svojih učenika ili studenata poboljšala/poboljšao:****

	Uopće se ne odnosi na mene	Malo se odnosi na mene	Niti se odnosi, niti se ne odnosi na mene	Odnosi se na mene	U potpunosti se odnosi na mene
Vokabular	○	○	○	○	○
Gramatiku	○	○	○	○	○
Čitanje	○	○	○	○	○
Pisanje	○	○	○	○	○
Slušanje	○	○	○	○	○
Govor	○	○	○	○	○

6. **Slažete li se sa sljedećim izjavama?**
Sudjelovanje u organiziranoj edukaciji o korpusnoj lingvistici koristit će mi za:

	U potpunosti se ne slažem	Ne slažem se	Nisam siguran/na	Slažem se	U potpunosti se slažem
Učenje o stvarnoj uporabi jezika	○	○	○	○	○
Učenje o stvaranju korpusa	○	○	○	○	○
Učenje o radu s korpusom	○	○	○	○	○
Analizu konkordancija	○	○	○	○	○
Pronalaženje jezičnih uzoraka	○	○	○	○	○
Stvaranje i razvoj boljih materijala za učenje**	○	○	○	○	○

Učenje o korištenju korpusnih alata izravno u nastavi**	○	○	○	○	○
Razumijevanje nastavne metode izravnog korištenja korpusnih alata u nastavi (data-driven learning-DDL)**	○	○	○	○	○

7. Kojim ste se od hrvatskih računalnih korpusa koristili?

○ hrWaC
○ Hrvatski nacionalni korpus
○ Riznica Instituta za hrvatski jezik i jezikoslovlje
○ Korpus leksikografskog zavoda Miroslav Krleža
○ Ostalo:_____

8. Kako biste se opisali kao korisnik tehnologije i računala?

○ Uvijek isprobavam nove tehnologije
○ Kada ostali započinju koristiti novu tehnologiju tada sam i ja željan/na ju isprobati
○ Kada mnoštvo ljudi sudjeluje tada se i ja pridružim
○ Među zadnjima sam od svojih kolega koji isprobavaju nove tehnologije (kada se više ne smatraju „novima“) no na kraju ih isprobam
○ Ne volim i ne trebam nove tehnologije te je moj napredak spor, ukoliko uopće i postoji

9. Ocijenite svoj stupanj informatičke pismenosti/služenja računalom i računalnim alatima?

Ne koristim računalo	Posjedujem osnovno znanje	Dobro se služim računalom	Vrlo dobro se služim računalom	Izvrsno se služim računalom
○	○	○	○	○