

Dr. sc. Memnuna Hasanica

KARAKTERISTIKE I PRAKTIČNI PRIMJERI KORPUSNE LINGVISTIKE SA POSEBNIM OSVRTOM NA NJEMAČKI JEZIK

Sažetak

Korpus je sistematski zbir autentičnih tekstova ili dijelova tekstova. Svrha korpusne lingvistike predstavlja opisivanje sadržaja i odnosa unutar određenog jezika i izučavanje diskursa kao posrednika znanja i društvenih normi. Korpus predstavlja reprezentativan izvod iz jednog jezika ili njihove različitosti. Korpore omogućavaju empirijske izjave o jeziku. Korištenje korpore ima smisla svuda, gdje su neophodne informacije u vezi sa upotrebom jezika. Primjenljivost korpusnog istraživanja je moguća na nekoliko različitih polja (u leksikografiji, morfologiji, semantici, fonologiji, sintaksi itd.).

Ključne riječi: *sadržaj, korpus, veličina korpusa, primarni podaci, reprezentativnost, tekstualne probe*

Korpusna lingvistika – šta je to?

Korpusna lingvistika je, pored ispitivanja govornika i eksperimenata, jedna od metoda, uz pomoć koje se upotreba jezika istražuje na osnovu autentičnih jezičkih podataka. Kako bismo imali solidnu osnovu podataka za empirijsko istraživanje upotrebe jezika, nudi se rad sa tekstnim korpusom. **Korpus** je zbir tekstova ili dijelova tekstova, koji su svjesno odabrani i podijeljeni prema određenim lingvističkim kriterijima. Pod tekstovima se u tom kontekstu ne podrazumijevaju samo proizvodi pisanog jezika kao novinski članci, romani, kuharske knjige, e-mailovi, pisma ili dnevnici, nego i usmene izjave, koje mogu biti u obliku ugovora, radioemisija, telefonskih razgovora ili neformalnih razgovora za stolom na ručku. Tekstovi, koji su sadržani u korpusu, nazivaju se **primarni podaci**.

Korpus ima svrhu, da služi kao dio jezika, koji se treba istražiti. Pri tome je bitno, objasniti, da li se jedan jezik želi općenito istražiti (npr. jedan jezik u cjelosti), ili samo određenu raznolikost datog jezika. Pod pojmom raznolikost se podrazumijeva određeni izraz jezika, koji je definisan kroz vanjezične faktore, kao što su vrijeme, prostor, grupa

govornika ili komunikativna situaciju. Raznolikosti zavise od grupe govornika (npr. jezik omladine ili stručni jezik medicine), dok su npr. švicarski njemački ili švapski regionalno definisane raznolikosti.

Korpus, koji se treba koristiti da bi se istražio jedan jezik, izgleda potpuno drugačije nego korpus, koji se npr. koristi za analizu jezika omladine. Bitna razlika je da korpus jezika omladine sadrži samo tekstove omladine, razgovore među prijateljima, školske referate i SMS-ove. Onda bi bilo besmisleno, kada bi se korpus, koji treba da prikaže jedan dio cjelokupnog njemačkog jezika, sastojao samo od tekstova mladih govornika.

Danas se korpore – množina od korpusa – često nalaze u elektronskoj formi. Takve korpore se isčitavaju kompjuterski i mogu se uz pomoć kompjutera analizirati i obrađivati. To predstavlja veliki napredak u odnosu na početke korpusne lingvistike, kada su korpore postojale samo u čistom papirnom obliku, a poslije su se sakupljale i obrađivale na karticama. Ipak su korpore, koje postoje samo u čistom papirnom obliku, do danas veoma rasprostranjene. Tu ima više razloga. Sa jedne strane određene vrste tekstova, kao ručno pisana pisma ili srednjovjekovno ručno pismo, ne mogu se jednostavno prebaciti u format za kompjutersko isčitavanje. Isti se mogu, naravno, skenirati, ali se ne mogu jednostavno prebaciti u tekstualni dokument.

Kvalitet jednog istraživanja ne ovisi od toga da li je korpus moguće kompjuterski isčitavati ili ne. Mogućnosti i metode su kod korpora na papiru ili obliku dokumenta principijelno iste, iako je kod kompjuterski isčitavajućih korpora pretraživanje znatno manje i vrednovanje je često jednostavnije. Nedostaci korpusa na papiru se često karakteriziraju time da je preobrazba u tekstualni dokument neprimjenjiva. Kao posljedica, posebno kod manjih analiza korpusa, kao u okviru studentskog domaćeg ili seminarskog rada, često ima više smisla da se izbjegne kompjuterski isčitana verzija korpusa i da se umjesto toga neophodne informacije ručno vade iz korpusa.

Šta je korpus i kako je građen?

Svaka zbirka tekstova ne predstavlja korpus, iako se u praksi zbirke tekstova često koriste kao korpore, koje ne odgovaraju definisanim kriterijima. U idealnom slučaju korpus se gradi na osnovu određenih lingvističkih posmatranja. Tekstovi, iz kojih se sastoji korpus, odabiru se na osnovu unaprijed definisanih kriterija.

Bitno za koncepciju korpusa su veličina i sadržaj korpusa, kao i njihova otpornost i reprezentativnost.

Korpore su principijelno izdvojene. One su određene za to da istražuju primjenu jezika u dijelovima ili u njegovoj kompletnoj širini. One služe kao dio jedne jezičke cjeline, koju treba istražiti. To znači da korpus omladinskog jezika mora biti tako sastavljen da omogući generalnu izjavu o njemačkom omladinskom jeziku, korpus standardnog jezika treba da da izjavu o svim područjima standardnog jezika. Korpus treba biti reprezentativan za temeljne jezične cjeline.

Reprezentativnost korpusa je najviši cilj prilikom izgradnje korpusa. Kako bismo dostigli reprezentativnost, neophodno je da osnovna cjelina koja će se istraživati bude jasna. Ako je to npr. omladinski jezik, onda se mora razjasniti ko sve spada u omladinu. Pripada li djevojčica od devet godina ili dječak od jedanaest godina već u grupu omladine? Broje li se mladi odrasli sa osamnaest ili devetnaest još kao omladina? Također, treba razjasniti da li se svaki iskaz omladinca može pridodati jeziku omladine. Radi li se npr. kod školskog referata jednog trinaestogodišnjaka o jeziku omladine? Priča li jedna šenaestogodišnjakinja sa svojim djedom i nanom jezikom omladine? Da li se radi možda kod tekstova odraslih, koji su kao tekstovi adresirani na omladinske časopise za omladinske čitatelje, o jeziku omladine?

Intuitivna ideja jezika omladine ili standardnog jezika nije dovoljna da se napravi reprezentativan korpus. Neophodnije je da se intuitivno znanje sprovede u objektivne, obavezne kriterije, koji će prilikom odabira tekstova, pogotovo kod slučajeva koji su na granici, pomoći pri odluci da li tekst pripada definisanoj osnovnoj cjelini ili ne. Pored toga je bitno da je jasno da se sve vrste tekstova u jeziku ne javljaju često. Tako se uputstva za upotrebu, javni govori ili sudske presude mogu sa sigurnošću dodijeliti standardnom jeziku, jer većina govornika se sa ovom vrstom teksta znatno rjeđe susretne nego sa novinskim tekstovima, televizijskim intervjuima ili poslovnim pismima. Osim toga je bitno razmotriti da li postoji vidna neuravnoteženost između vrste i broja tekstova, koje govornik aktivno proizvodi, te vrste i broja tekstova, koje on pasivno bilježi.

Ovdje je jasno da nije samo bitno da se postavljanje pitanja za analizu korpusa tačno ograniči, nego da je, također, bitno da se utvrdi koje vrste tekstova se trebaju ubaciti u koji dio korpusa. U pojedinačnom slučaju to će ovisiti od predmeta istraživanja.

Patentni recept ne postoji. Ako je izgradnja korpusa završena, onda u normalnom slučaju ne slijede nikakve dodatne izmjene. Sadržaj, obim i struktura ostaju nepromijenjeni. Korpus time ispunjava kriterij **otpornosti**. Ovaj kriterij ne odgovara takozvanoj monitornoj korpori.

Što se tiče **veličine korpusa**, često se čuje – što je korpus veći, to je bolje. Veličina korpusa se mjeri tekstualnim riječima. Na samim počecima korpusne lingvistike i sve do osamdesetih godina 20. stoljeća najveće kompjuterski iščitane korpore imale su oko million tekstualnih riječi (npr. američki Smeđi korpus i njegov britanski duplikat Lancaster Oslo/Bergen-korpus, LOB, istovjetni kao i njemački LIMAS korpus). Dugo vremena je važno pravilo da je standardna veličina korpore milion tekstualnih riječi. Tehničkim napretkom evidentan je konstantan rast u smislu obima tekstualnih korpora. Tako npr. britanski nacionalni korpus (BNC) obuhvata 100 miliona tekstualnih riječi i tzv. banka engleskog jezika više od 550 miliona tekstualnih riječi. Institut za njemački jezik (IDS) u Mannheimu arhivira danas veliki broj korpora na njemačkom jeziku u opsegu od preko dvije milijarde tekstualnih riječi.

Veličina, međutim nije sve. Ovisno o predmetu istraživanja korpus od deset ili dvadeset hiljada tekstualnih riječi može dati pouzdane informacije o istraživanim pitanjima. Tako korpus standardnog njemačkog jezika sa dvadeset hiljada tekstualnih riječi daje pouzdanu sliku o tome koje riječi se najčešće javljaju u njemačkom jeziku, jer se većinom radi o funkcionalnim riječima kao članovima, zamjениčkim veznicima i pomoćnim glagolima. Te vrste riječi nalaze se u većem broju u svakom tekstu. Za utvrđivanje stranih riječi, kao npr. rječnik fotografije ili pedijatrije može biti dovoljan jedan mali korpus.

Pored veličine korpusa još jedna dodatna veličina je od značaja za izgradnju korpusa, a to je veličina dobivenih tekstova odnosno tekstualnih isječaka. U korpusu sadržani tekstovi i dijelovi tekstova označavaju se kao probe, odnosno **tekstualne probe**, jer se one biraju iz osnovne jezičke cjeline kao stručnog jezika fotografije, kako bi služile kao predstavnici tih raznolikosti. Općenito, postoje dvije mogućnosti, ili da se odabrani tekstovi u cjelokupnoj dužini integrišu u korpus ili se ovisno od teksta unaprijed definisani broj tekstualnih riječi ubaci u korpus. Ako se kao u slučaju Bonnskog LIMAS korpusa izgradi korpus sa ukupnom veličinom od milion tekstualnih riječi, onda se mora kod čvrste veličine proba od 2.000 tekstualnih riječi

uraditi probe od 500 tekstova. Ukoliko će se u korpus preuzimati cjelokupni tekstovi, onda zvaničan broj proba u korpusu zavisi od dužine pojedinačnih tekstova. Ukoliko se radi kod odabranih proba o veoma dugim tekstovima kao romanima, onda će korpus sadržavati znatno manje različitih proba, nego kod odabranih kratkih tekstova, tipa telefonskih razgovora, kuharskih recepata ili viceva.

Ovim dolazimo do sljedećeg bitnog pitanja, a to je **sadržaj** korpusa. Uglavnom bi bilo moguće da se u korpus omladinskog jezika jednostavno ubaci svaki tekst na omladinskom jeziku, sve dok se ne postigne određena veličina korpusa. U tom slučaju je sastav korpusa čisti slučajni proizvod, jer se tekstovi biraju samo na osnovu raspoloživosti. Prema slučajnom principu izgrađeni elektronski korpus omladinskog jezika može sadržavati npr. do 80% e-mailova, SMS-ove, chatove i članke iz učeničkih časopisa, koji prikladno već postoje u elektronskom obliku. Ostalih 20% korpusa bi bili tekstovi, kao ručno pisani referati, pisma i bilješke koji bi morali biti otkucani ili skenirani. Govorni jezik, koji se prvo snima i poslije prepisuje, tj, mora se zabilježiti, u ovom korpusu bi vjerovatno nedostajao. Veliki nedostatak kod ove metode jeste, da je slučajni odabir rijetko reprezentativan. Tako da u upravo opisanom korpusu nedostaju široka područja omladinskog jezika, naime cjelokupan govorni jezik, koji predstavlja veliki dio naše svakodnevne komunikacije. Druga područja kao komunikacija putem interneta bila bi prezastupljena. Pošto slučajni odabir nema obzira prema autoru teksta, može se desiti da tekstovi starije i mlađe populacije mladih, djevojčica i dječaka nisu podjednako vrednovani.

Naposlijetku se preporučuje da se prije početka odabira teksta definiše sa kojim težištem treba uzimati u obzir određene vrste tekstova i grupe govornika. Ako smo omladinski jezik definisali kao zbir svih tekstova koji se produciraju od omladine starosti između 11 i 19 godina, onda moramo u pogledu na te tekstove razmisliti o slijedećim pitanjima: u kojem omjeru treba evidentirati govorni i pisani jezik, te kako trebaju biti zastupljene pojedinačne vrste tekstova, npr. školski referati, lična pisma, e-mailovi, bilješke, dnevници kao pisani tekstovi, lični razgovori, telefonski razgovori između omladine ili sa odraslima kao usmena komunikacija. U odnosu na lične karakteristike autora teksta postavljaju se sljedeća pitanja: koji udio treba da zauzmu tekstovi mlađih i starijih govornika, te koji je udio tekstova dječaka odnosno djevojčica. Ako se uzme u

obzir regionalno porijeklo i nivo obrazovanja, onda treba razmisliti o tome koji udio treba da zauzimaju tekstovi govornika iz pojedinačnih regija odnosno iz različitih škola.

Osim toga, korpus treba da raspolaze opširnom dokumentacijom dobivenog materijala, tzv. **metapodacima**, tako da se u uvijek može dokazati iz kojeg teksta, te od kojeg autora (tzv. autora ili govornika), proizilaze određeni podaci. Korpus treba da sadrži informacije o imenu, starosti, te spolu autora, kao i podatke o vremenu i situaciji prilikom izjašnjavanja odnosno objave teksta. Postojanje podataka autora omogućava da se sprovedu sociolingvističke studije, kao npr. da se istraži utjecaj starosti, spola ili pripadnosti jednoj socijalnoj grupi na jezičkoj primjeni.

Nije samo bitno, da se uzmu u obzir lične karakteristike autora teksta i karakteristike teksta, bitno je, također, da je jasno da jezik nije bezvremenski fenomen nego da je izložen promjenama. To je odmah vidljivo ako se bavite jezicima iz ranijih stoljeća, npr. jezikom srednjeg vijeka, vremenom reformacije ili njemačke klasike. Ukoliko čitate tekstove Waltera von der Vogelweide, Lutherovu Bibliju ili dramu od Goethea, onda neizbježno primjećujete jezičke razlike između tih tekstova i današnjeg njemačkog jezika. Tako kod Waltera von der Vogelweide riječ *frouwe* označava plemenitu damu a riječ *wîp* ženu općenito. Međutim, u današnjem njemačkom jeziku koriste se odgovarajuće riječi, žena (*Frau*) odnosno ženska (*Weib*), vrijednosno neutralno (*žena*) ili pejorativno (*ženska*) kada se priča o ženskim osobama. Dok se kod Goethea pronalaze konjunktivni oblici kao *begänne*, *empfähle* i *schüfe*, odgovarajući oblici u sadašnjem njemačkom jeziku glase *sie würde beginnen* (*ona bi počela*), *ich würde empfehlen* (*ja bih preporučila*) odnosno *er würde schaffen* (*on bi radio*). Niko nije došao na ideju da koristi korpus sa Goetheovim tekstovima kako bi istražio konjunktivne oblike u sadašnjem njemačkom jeziku. Isto tako se mora biti oprezan, kada se koriste tekstovi koji su stari deset, dvadeset ili samo pet godina. Tako se danas jezik omladine znatno razlikuje od jezika koji su govorili njihovi roditelji ili njihove starije sestre ili braća u istoj dobi. Ako želite istražiti jezičke novine, prvenstveno u vokabularu, mora se voditi računa, da korišteni korpus sadrži aktualne podatke.

Za šta se koriste korpora?

Velika prednost korpora jeste, da ne sadrže samo autentične govorne materijale, nego i pored toga daju informacije o učestalosti i korištenju riječi, gramatičkim kategorijama i drugim jezičkim jedinicama. Rad sa korporama je koristan svuda gdje iz bilo kojeg razloga jezička kompetencija pojedinačnog govornika nije dovoljna. To se odnosi na germanističku lingvistiku, koja svoje teorije provjerava empirijski, isto kao i izdavač rječnika ili gramatika, historičar ili sociolog, koji traži kulturalne ključne riječi, kao i tehničarka, koja radi na akustičkom sistemu za prepoznavanje govora, prevodioci stručnih tekstova kao i studenti, koji pišu rad na stranom jeziku.

Korpora se stoga može koristiti za veliki broj različitih namjena. U najbitnija lingvistička područja primjene korpora spadaju:

- istraživanje jezičkih struktura i njihove različitosti,
- kreiranje rječnika (leksikografija),
- kreiranje gramatika (gramatikografija),
- nastava stranog jezika,
- prevođenje, te
- kompjuterska lingvistika.

Već duže vrijeme korpora se koriste prilikom kreiranja rječnika i gramatika. U međuvremenu korpusno bazirani rječnici i gramatike šire se sve više i više. Prvenstveno, elektronski rječnici kao Elexiko IDS-a ili Digitalni rječnik njemačkog jezika 20. stoljeća (DWDS) dobivaju svoj materijal na osnovu tekstualnih korpora. To ima prednost da leksikografu na raspolaganju stoje autentični govorni podaci. Dodatna prednost jeste da korpus nudi dodatne informacije za pojedinačne riječi. Tako se npr. može ustanoviti, u kojim različitim značenjima se može pojaviti riječ *kuća* – naime, ne samo u značenju *„građevine“* u (2a), nego i u značenju *„porodice“* u (2b) odnosno *„puževe kućice“* u (2c).

(2) a) *Ja živim u četvrtoj **kući** na lijevoj strani ulice.*

a1) *Ich lebe im vierten **Haus** auf der linken Seite der Straße.*

b) *Ona dolazi iz dobre **kuće**.*

b1) *Sie kommt aus dem wohlhabenden **Haus**.*

c) *Puž se povukao u svoju **kuću**.*

c1) *Die Schnecke zog in sein **Haus**.*

Pored toga, u korpusu se može ustanoviti u vezi sa kojim drugim riječima ili u kojim idiomatskim frazama se pojavljuje neka riječ. Za riječ *kuća* pronašli bi *Bijela kuća*, *isporuka na kućnu adresu* ili *ispred rasprodane kuće*, itd.

Slično kao kod rječnika, korpus može biti od pomoći prilikom kreiranja gramatika. Najpoznatiji primjer za korpusno baziranu gramatiku je, pretpostavlja se, trodjelna gramatika IDS-a (Zifonun *et al.* 1997). Odgovarajuće pripremljeni korpus i odgovarajući program omogućavaju da se identifikuju i izdvoje određene gramatičke strukture u korpusu. Na osnovu rezultata postojeći gramatički opisi se mogu provjeriti i mogu se dobiti nove spoznaje. Tako možemo dokazati da se prijedlog *zbog* u novinama redovno koristi sa *genitivom* a samo kod izuzetaka sa *dativom*. Ali u korpusima se može naići i na gramatičke strukture, koje do sada nisu bile, ili su neznatno, opisane u gramatikama. To važi za dupli perfekt *sam već pročitao* (*hab gelesen gehabt*) u (3) gdje je stvarni perfekt oblik *sam pročitao* (*habe gelesen*) proširen za drugi particip perfekta (*gehabt*). Slično važi i za takozvani Rheinski progresivni oblik u (4).

(3) *Knjigu sam pročitao u roku od 1 ½ sat.*

(4) a. *Pišem.*

b. *Trenutno kuham.*

Korpore su, također, bitne da se istraže raličitosti njemačkoga jezika. Različitosti određenih grupa, kao što su stručni jezik medicine, stručni jezik lingvistike ili ekonomski njemački jezik, karakterišu da se oni u gramatici u velikoj mjeri preklapaju sa standardnim njemačkim jezikom, ali i da dijelom imaju veliki stručni vokabular, koji se izbjegava kod govornika standardnog jezika. Nazivi kao *glaukom* za povišen očni pritisak, *genus* za gramatički rod i *tratte* za prenos mjenice pripadaju stručnim jezicima. Stručne tekstualne korpore ovdje se mogu primjenjivati, kako bi se izradili specijalni stručni rječnici.

Usporedba stručnog tekstualnog korpusa sa korpusom standardnog jezika omogućava da se prepoznaju razlike između stručnog jezika i standardnog jezika. Znanje o razlikama olakšava izučavanje stručnog jezika. Slično važi i za druge grupne specifične ili regionalne govorne različitosti, kao što su omladinski jezik, zatvorski jezik, saksonski ili kelnski. I ovdje se mogu, kroz usporedbu korpusa standardnog jezika i specijalnog korpusa, prepoznati karakteristike govornih različitosti.

Posebno bitne su korpore u području stranih jezika, sa jedne strane za sticanje znanja tog stranog jezika i kod pisanja tekstova na stranom jeziku, a sa druge strane za prevođenje stranih tekstova na njemački jezik. U području nastave stranog jezika korpore se mogu

koristiti za razvoj nastavnog materijala. To se odnosi na udžbenike izdavačkih kuća, kao i nastavni materijal koji se razvija od strane predavača. Razumno je npr. da se osnovni vokabular i osnovna gramatika, koju izučavaju učenici, definiše uz pomoć korpusa, jer bi učenici trebali poznavati strane riječi i gramatičke strukture sa kojima će se najčešće susretati prilikom čitanja i u vođenju razgovora. Korpore se, osim toga, mogu koristiti, kako bi se proizveli tekstovi sa prazninama, u koje bi učenici trebali ubaciti odgovarajuće riječi ili gramatičke oblike.

Specijalne korpore za izučavanje, koje dobivaju tekstove od stručnjaka stranog jezika, omogućavaju predavaču da prepozna tipične greške i poteškoće učenika. Nadalje, korpore daju predavačima kao i stručnjacima u ličnoj studiji vrijedne informacije o korištenju riječi i gramatičkim strukturama – npr. u kojem kontekstu, odnosno u kojoj situaciji, se koristi zamjenica *ti* odnosno *Vi*; koji glagoli se tvore sa pomoćnim glagolom *sein* / *biti*, a koji sa *haben* / *imati*; kada iza prijedloga *u* stoji dativ (*u kući*), a kada akuzativ (*u kuću*), itd.

Baš zbog takvih tekstualnih informacija su korpore i za prevodioce vrijedan izvor informacija. Jednojezične korpore mogu pored konteksta davati i frekventnost informacija i time dozvoljavaju identifikaciju rijetkih riječi i stilističkih poremećaja u stranom jeziku koje se moraju pronaći za njemačku korespondenciju.

Korpore ne nude samo informacije o vokabularu i gramatici nego i informacije za ustrojavanje i strukturiranje tekstova – npr. kako su izrađene određene vrste tekstova; koje dijelove sadrže; kako se u jednom tekstu stvara smisao; kako se prezentuju argumenti za postavljanje neke hipoteze; kako se uvodi zamjena govornika, kako se ista realizuje. Analiza tekstualnih korpora ovdje može dati bitne upute i smjernice za sve koji proizvode sami određene vrste tekstova, posebno one sa specijalnim formalnim zahtjevima kao znanstvenim radovima ili stručnim tekstovima, na maternjem ili na stranom jeziku. To mogu biti autori stručnih tekstova, ambiciozni novinari, recenzenti, znanstvenici ili studenti, ali i zaposlenici u nekoj gradskoj upravi.

Zaključak

Skup teksta, bilo pisanog ili govornog jezika u svom osnovnom značenju predstavlja korpus. Poznato je da postoje opšti i specijalizirani korpusi koji su namijenjeni različitim vrstama lingvističke analize. Pravilno upotrijebljen korpus odnosno njegova

reprezentativnost, a posljedično i rezultati koje taj korpus pruža, postižu se ne veličinom analiziranog korpusa nego prvenstveno raznolikošću, odnosno pravilnim i kvalitetnim odabirom izvora prilikom analize korpusa.

Danas su korpore neizbježne kod lingvističkih istraživanja, jer dozvoljavaju neposredan pristup autentičnom jeziku. Korpore omogućavaju lingvistima da postojeću teoriju istraže empirijski i da na osnovu korpusnog materijala razviju novu hipotezu.

Literatura

1. Lemnitzer, Lothar / Zinsmeister, Heike (2010): Korpuslinguistik. Eine Einführung. 2., durchges. und aktualisierte Auflage, Tübingen: Narr.
2. McEnery, Tony / Xiao, Richard / Tono, Yukio (2006): Corpus based language studies, London: Routledge
- Scherer, Carmen (2006): Korpuslinguistik. Heidelberg: Winter.
2. Zifonun, Gisela, Ludger Hoffmann, Bruno Strecker et.al. (1997): Grammatik der deutschen Sprache, Berlin & New York: de Gruyter

CHARACTERISTICS AND PRACTICAL EXAMPLES OF CORPUS LINGUISTICS WITH A SPECIAL REFERENCE TO GERMAN LANGUAGE

Abstract

Corpus is a systematic collection of authentic texts or parts of texts. The purpose of corpus linguistics is to describe contents and relations within a particular language, and to discuss discourse as an intermediary of knowledge and social norms. The corpus represents a representative excerpt from a single language or diversity of languages. Corpora allow empirical statements about language. The use of corpora makes sense wherever information regarding use the language is needed. The applicability of corpus research is possible in several different fields (in lexicography, morphology, semantics, phonology, syntax, etc.).

Key words: *content, corpus, corpus size, primary data, representativeness, text examples*