ლია ოთინაშვილი
სოხუმის სახელმწიფო უნივერსიტეტი

კორპუსის ლინგვისტიკა და ლექსიკოგრაფია

კორპუსი წარმოადგენს ლექსიკონის შედგენის ერთ-ერთ უმნიშვნელოვანეს ფაქტორს. სწორედ იგი გახდა ამ ბოლო ხანებში შექმნილი არაერთი ლექსიკონის საფუძველი. ასეთი მჭიდრო დამოკიდებულება გამოწვეულია იმით, რომ სალექსიკონო მასალის შერჩევისას ლექსიკოგრაფი ეყრდნობა კორპუსის მონაცემთა ბაზას, რომელიც ამ შემთხვევაში გვევლინება, როგორც რაოდენობრივ ასევე ხარისხობრივ კრიტერიუმად. შესაბამისად, საპასუხისმგებლო საქმეა კორპუსის ტიპის შერჩევა, რომლის მონაცემთა ბაზა და ენობრივი იარაღი საჭირო მასალა იქნება ლექსიკონისათვის.

როგორც ჩანს, სამივე ასპექტი ნიშანდობლივი მხარდამჭერია კორპუსის ლინგვისტიკის მნიშვნელოვანი როლისა ლექსიკოგრაფიაში. თუმცა, ბოლო ხანებში პროდუქტი, რომელიც აერთიანებს ლექსიკონსა და კორპუსის მონაცემებს ფორმულირდა და ბუნებრივი ენის დამუშავება (Natural Language Processing) ხასიათდება მთელი რიგი ორმხრივი ურთიერთობის სფეროებით: ლექსიკური მონაცემები გამოიყენება კორპუსის ანოტირებისას, ხოლო ლექსიკური აღწერილობები პირიქით, აღებულია უკვე ანოტირებული კორპუსიდან.

ზემოაღნიშნული საკითხების განხილვისას ყურადღებას გავამახვილებთ დანიელი ლინგვისტისა და ლექსიკოგრაფის ჯონ ბერგენჰოლცის შეხედულებებზე. იგი მიიჩნევს, რომ ლექსიკონი არის პროდუქტი, რომელიც ადამიანს აწვდის მთელ რიგ მასალას ლინგვისტური ობიექტების შესახებ, როგორიცაა სიტყვა, მორფემა, სიტყვათა ჯგუფი და ა.შ. ლექსიკონის შედგენისას ვითვალისწინებთ ან უნდა ვითვალისწინებდეთ მომხმარებლის საბაზისო ცოდნას. ეს უკანასკნელი კი განარჩევს ორი ტიპის საჭიროებას: ურთიერთობაზე ორიენტირებულ საჭიროებას და ცოდნაზე ორიენტირებულ საჭიროებას.

პირველი ტიპის შემთხვევაში საქმე გვაქვს ერთ ან მაქსიმუმ ორ ენასთან, რომელთაგან ერთ-ერთი აუცილებლად მშობლიურია. ამ შემთხვევაში ხდება წაკითხვა-გაგება და თარგმნა - მშობლიური ენიდან უცხო ენაზე და პირიქით.

მეორე ტიპის შემთხვევაში საქმე გვაქვს შემეცნებითი ინფორმაციის მიღებასთან, რომელიც გულისხმობს სპეციალიზებულ სფეროს - კულტურულ-ენციკლოპედიურ ფაქტებს ან თავად ენის შესწავლას.

ზემოაღნიშნული ორივე ტიპის ლექსიკონის შექმნასა თუ გამოყენებაში დიდი როლს თამაშობს კორპუსი, თუმცა, როგორც ვთქვით, ეს არ არის ცალმხრივი პროცესი. ხდება შერჩევა, თუ რა სალექსიკონო მასალა იქნება გამოყენებული კორპუსში და, პირიქით, რა იქნება გამოყენებული ლექსიკონში კორპუსის ბაზიდან.

ყველა კორპუსის ლინგვისტური ფუნდამენტი ეყრდნობა ლექსიკოგრაფიულ სამუშაოებს. ეს მოიცვას მონაცემთა შერჩევას და კორპუსში წარმოჩნას. კორპუსთა უმეტესობა შეიცავს წერილობით მასალას - წიგნებს, ჟურნალებს, გაზეთებს. კამათის საგანს წარმოადგენს ის ფაქტი, რომ ზემოაღნიშნული მასალა უნდა დაბალანსდეს ზეპირსიტყვიერი მასალით. ბრიტანული ნაციონალური კორპუსის (BNC) 10%-ს შეადგენს სწორედ ზეპირსიტყვიერი მასალა. ეს პროცენტულობა იშვიათია მსგავსი ტიპის დიდი კორპუსებისათვის. სწორედ ლექსიკოგრაფიის პრეროგატივაა ამ საკითხის გადაწყვეტა, თუ რომელმა ენამ უნდა იმუშაოს კორპუსის შიგნით და როგორ.

აშკარაა, რომ ლექსიკონის ზომას განსაკუთრებული გავლენა აქვს კორპუსის შემადგენლობასა და სიდიდეზე. მაგ., გერმანული სალექსიკონო ელექტრონული პროექტი ელეხიკო შეიცავს 1 400 000 სიტყვას, რომელთაგან 1 000 000 სიტყვა შესულია (Digitales Wörterbuch der deutschen Sprache) ტექსტის კორპუსის ბაზაში.

ამ ფაქტის სინამდვილეში კიდევ ერთხელ გავრწმუნებს ბრიტანული ნაციონალური კორპუსი (BNC). ამ კორპუსის მაგალითზე ვრწმუნდებით, რომ ლექსიკოგრაფიული მონაცემების სიდიდე კიდევ უფრო მნიშვნელოვან გავლენას ახდენს არა მხოლოდ ტექსტის კორპუსის, არამედ თვით ნაციონალური კორპუსის შექმნასა და ზომაზე. სწორედ ბრიტანული ნაციონალური კორპუსი იყო პირველი, რომელიც 1993 წელს შეიქმნა ლექსიკოგრაფიული კონსორციუმისა და რამდენიმე ლექსიკონის გამომცემელთა ინიციატივით. იგივე მხარდაჭერა ჰქონდა 2004 წელს ამერიკული ნაციონალური კორპუსის (ANC) შექმნის მცდელობას. მონცემთა ბაზას წარმოადგენდა COBUILD-ის ლექსიკონი.

თუმცა, უნდა აღინიშნოს ის ფაქტიც, რომ ლექსიკოგრაფთა უმრავლესობა თანხმდება, რომ კორპუსი, რომელიც შეიცავს 60-დან 100 მილიონამდე სიტყვას, წარმოადგენს შესანიშნავ ბაზას ლექსიკონისათვის, რომლის შემადგენლობაშიც იქნება 50 000-დან 60 000 სიტყვამდე.

კორპუსი აქტიურად მონაწილეობს სპეციალიზებული ლექსიკონების ე.წ. დარგობრივი დომენების შექმნაში (Specialized Domains), რისთვისაც კორპუსი იყენებს განსაკუთრებულ მეთოდებს. ამ ტიპის ლექსიკონები შეიცავს რამდენიმე ასეულ ან ათასეულ სიტყვას. მათთვის საკმარისია კორპუსი, რომლის ბაზაც მილიონ სიტყვაზე ნაკლებია და წყაროს წარმოადგენს სპეციალური ლიტერატურა, სახელმძღვანელოები, ბროშურები და ა.შ. სხვა შემთხვევაში კი, როცა საქმე გვაქვს უფრო ვრცელ დომენებთან, როგორიცაა ბიოტექნოლოგია, ბიოლოგია და ა.შ., რა თქმა უნდა, საჭიროა უფრო დიდი კორპუსი.

ლექსიკოგრაფიის ამ სფეროსთვის შექმნილი კორპუსის მთავარი საკითხია მასში დაცული შესაბამისი ტექსტების შერჩევა. სირთულეს წარმოადგენს იმის გადაწყვეტა, შეესაბამება თუ არა მოცემული ტექსტი არჩეულ სპეციალიზებული დომენის სფეროს. მაგ., ბიოლოგია არის არის ჩაკეტილი სფერო, მას აქვს ნაწილობრივი თანხვედრები ფიზიკასთან და ქიმიასთან. ასე რომ, ამ ტიპის კორპუსი უნდა შეიცავდეს ტექსტებს, რომლებიც მჭიდრო კავშირში იქნება ყველა მოცემულ სფეროსთან.

სპეციალიზებული ლექსიკოგრაფიისათვის შექმნილი კორპუსის დროს მომხმარებელთათვის დამატებით პრობლემად იჩენს თავს კორპუსში შემავალი ტექსტების ტიპოლოგია. აუცილებელია ტექსტები იყოს ლექსიკონის პარალელური. ეს გამოწვეულია იმით, რომ საჭირო ტერმინოლოგია, რომელიც დაცულია კორპუსის ბაზაში, მომხმარებელს სჭირდება არა როგორც ტერმინთა ნუსხა ექვივალენტებითურთ, არამედ, ის, თუ როგორ არის საჭირო მონაცემები დაცული ტექსტებში. ეს კი მოიცავს კოლოკაციას (შესიტყვებებს), ტერმინოლოგიურ და მორფოსინტაქტიკურ ვარიანტებს.

საყურადღებოა ის ფაქტი, რომ არსებობს ლექსიკონთა ტიპი, რომლებიც უკვე თავად შეიცავს ტექსტურ მაგალითებს. მათი ყველაზე გამორჩეული თვისებაა მოახდინოს დასამუშავებელი ერთეულების ილუსტრირება. COBUILD-ის ლექსიკონი იყო ერთერთი პირველი ტექსტებდართული ლექსიკონი, რომელშიც მაგა-

ლითებად გამოყენებულია სწორედ ტექსტები. განმარტებების სახით ვხვდებით მთელ რიგ სრულ წინადადებებს ლიტერატურული ნამუშევრებიდან ან სწორედ კორპუსული მონაცემებიდან, წყაროს დეტალური მითითებით. აქვე უნდა დავძინოთ, რომ ენის შემსწავლელთა ლექსიკონებში ხშირად გამოყენებულია მხოლოდ სინტაგმები ან გამოგონილი მაგალითები.

აშკარაა, რომ ლექსიკოგრაფიული მონაცემების აღწერა ეფუძნება კორპუსის მონაცემებს. თუ საქმე გვაქვს კომპიუტერულ ლინგვისტიკასთან, მაშინ აქ ვხვდებით უფრო ფართო სპექტრს და ტექსტები არის მარკირებული (Tokenized), ანოტირებული (Lemmatized) მითითებულია მეტყველების ნაწილები (Tagged part of speech), შესაძლებელია ახლდეს მოზრდილი გრამატიკული ანალიზი.

დასასრულს, ზემოთ განხილული საკითხების შესაჯამებლად გვინდა აღვნიშნოთ, რომ კორპუსის ლინგვისტიკასა და ლექსიკოგრაფიას შორის კავშირს ისტორიული ფაქტორი განაპირობებს. აშკარაა, რომ შესაბამისობის დაძებნა ბიბლიასა და შუასაუკუნეების ხანის ავტორებს შორის, ასევე თარგმანები ორ ენაზე, როგორც პარალელური კორპუსის წყარო, ლათინური ტექსტების ანოტირება სტრიქონებრივი თარგმანებითურთ (Interlinear translation), როცა ეს ყველაფერი გამოიყენებოდა, როგორც ენის შემსწავლელი მასალა, შეიძლება ჩაითვალოს პრეისტორიული ლექსიკოგრაფიული კორპუსის შექმნის გზად.

მე-19 საუკუნის ბოლოს კადინგის ნაშრომი გერმანულენოვანი ლექსიკონის შესახებ, რომელიც ემყარება კორპუსს, რომლის ბაზასაც 11 მილიონი სიტყვა წარმოადგენს, არის პირველი რეალური კორპუსზე დაყრდნობით შექმნილი ლექსიკოგრაფიული შრომა, რასაც მოჰყვა 1920-30-იანი წლების ლექსიკონები და სიტყვათა ფუნდამენტური ნუსხები, ჩ. მიულერისა და ა. ჯულიანდის შრომები ლექსიკურ სტატისტიკაზე დაფუძნებით.

სხვა მხრივ, ადრეული კომპიუტერული კორპუსები, დაახლოებით სიტყვათა 1 მილიონიანი ბაზით, როგორებიც იყვნენ ბრაუნის კორპუსი და ლანკასტერ-ოსლო-ბერჯინის კორპუსი, რომლებიც გამოიყენებოდა გრამატიკული შრომებისათვის, მასიურად არ გამოიყენებოდა ლექსიკოგრაფიაში ზომის გამო.

პირველი ლექსიკონები, რომლებიც შეიქმნა ელექტრონული ტექსტების კორპუსის ბაზაზე არის: "ოქსფორდის ინგლისური ენის ლექსიკონი" (Oxford English Dictionary) და COBUILD-ის ლექსიკონი.

LIA OTINASHVILI
Sokhumi State University

Corpus Linguistics and Lexicography

Summary

The work "Corpus Linguistics and Lexicography" represents what kind of connection is between the Corpus Linguistics and Lexicography.

The work reveals that intercommunication among these two disciplines is obvious and very necessary. The existence of the Corpora made possible to overcome a lot of difficulties in dictionaries and Lexicography and vice versa.

This kind of the relation is caused by that during the choosing the dictionary entries the lexicographer relies on the Corpus database as the lexical databases are used during the Corpus Annotation.