

ქეთევან მჭედლიშვილი
ილიას სახელმწიფო უნივერსიტეტი, საქართველო
ketevani.mchedlishvili.3@iliauni.edu.ge
DOI: 10.32859/kadmos/16/237-253

შედარებითი კორპუსები: შედგენის მეთოდოლოგია და გამოყენების სფეროები

შესავალი

კორპუსული ლინგვისტიკა გამოყენებითი ენათმეცნიერების ერთ-ერთი დარგია, რომელიც მოიაზრებს ბუნებრივი ენის შესწავლას კომპიუტერისა და პროგრამული ინსტრუმენტების მიერ დამუშავებული ვრცელი ენობრივი მონაცემების საფუძველზე (Hunston 2022, 1). კორპუსულმა რეგოლუციამ დიდი ცვლილებები გამოიწვია ენათმეცნიერებასა და მომიჯნავე დარგებში, მათ შორის ზოგად და სპეციალიზებულ ლექსიკოგრაფიაში. სწორედ კორპუსები იქცა ზოგადი ენისა და დარგობრივი ლექსიკონების შედგენის პროცესში უმნიშვნელოვანეს წყაროდ. მაკენერისა და უილსონის თანახმად, „კორპუსი არის სასრული სიდიდის მქონე ტექსტების ერთობლიობა, რომელიც იკითხება მანქანის/კომპიუტერის მიერ და შედგენილია იმგვარად, რომ ასახავს საკვლევი ენის ან ენის ვარიანტის მაქსიმალურად რეპრეზენტაციულ სურათს“ (McEnery and Wilson 2001, 32).

კორპუსები შესაძლებელია იყოს ერთენოვანი ან ორენოვანი/მრავალენოვანი. არსებობს ორი სახის ორენოვანი/მრავალენოვანი კორპუსი: პარალელური და შედარებითი (Bowker and Pearson 2002, 92). პარალელური კორპუსებისგან განსხვავებით, რომლებიც შეიცავენ პირდაპირ თარგმანებს ორ ან მეტ ენაზე (Hunston 2022, 27), შედარებითი კორპუსი შედგება ორ/რამდენიმე ენაზე ან ერთი ენის რამდენიმე ვარიანტისთვის შედგენილი ერთენოვანი ქვეკორპუსებისგან, რომელთა შესადგენად გამოყენებულია მსგავსი სტრუქტურა და დაცულია ბალანსისა და რეპრეზენტაციულობის პრინციპი (ერთი ჟანრის ტექსტები, ერთი დარგის ტექსტები, ერთი პერიოდის ტექსტები) (Hunston 2002, 15; McEnery and Xiao 2007, 2). შედარებითი კორპუსების გამოყენების კვლევისადმი ინტერესი ჯერ კიდევ გასული საუკუნის 90-იან წლებში იღებს სათავეს (Fung and McKeown 1997; Rapp 1999) და დღესდღეობით, ერთ-ერთ პრაქტიკულ ალტერნატივად განიხილება, განსაკუთრებით დარგობრივ ორენოვან/მრავალენოვან პროექტებში,

როდესაც რთულია პარალელური ტექსტების მოძიება (Utka et al. 2022; Furtado and Teixeira 2022).

უნდა აღინიშნოს, რომ ქართულ ენასთან მიმართებით შექმნილია ზოგადი ენის შედარებითი კორპუსი, რომელიც შეიქმნა „არანეას“ ვებზე-დაფუძნებული შედარებითი კორპუსების ოჯახის პროექტის ფარგლებში (Benko 2024). მიუხედავად ამისა, ამ ეტაპზე დარგობრივი შედარებითი კორპუსების შედგენა და გამოყენების ეფექტურობის შესწავლა ლექსიკოგრაფიულ პრაქტიკაში სიახლეა ქართული ორენოვანი/მრავალენოვანი სპეციალიზებული ლექსიკოგრაფიისთვის.

წინამდებარე სტატიაში მიმოვიხილავთ შედარებითი კორპუსების რაობასთან, მათი გამოყენების სფეროებსა და უშუალოდ დარგობრივი შედარებითი კორპუსების შედგენის მეთოდებთან დაკავშირებულ საკითხებს, რაც მოგვცემს საშუალებას, შევიმუშაოთ აუცილებელი პრინციპები და კრიტერიუმები ინგლისური და ქართული დარგობრივი შედარებითი კორპუსების შედგენისათვის.

შედარებითი კორპუსების რაობა

როგორც ზემოთ აღვნიშნეთ, შედარებითი კორპუსების შექმნისათვის აუცილებელ წინაპირობას სწორედ მათი შედგენის პრინციპებისადმი მსგავსი მიდგომა წარმოადგენს. კენინგის თანახმად, შედარებითი კორპუსებიც განსხვავდება სხვადასხვა პარამეტრისა და მათში შემავალი ინფორმაციის მიხედვით (Kenning 2010, 448). შედარებითი კორპუსების ნაირსახეობებიც შეგვიძლია რამდენიმე ძირითადი პარამეტრის მეშვეობით დავახასიათოთ, როგორიცაა შემავალი ენების რაოდენობა (ერთენოვანი/მრავალენოვანი), სპეციალიზაციის ხარისხი (ზოგადი ენის/დარგობრივი) და შემავალი ტექსტების ტიპი (წერილობითი/ზეპირმეტყველების/ორივე/მულტიმოდალური).

პირველი დიქტომია, რომელიც გვაქვს კორპუსებში, ეყრდნობა მასში შემავალი ენების რაოდენობას. შედარებითი კორპუსები, თავის მხრივ, შესაძლოა, იყოს ერთენოვანი, ორენოვანი ან მრავალენოვანი. ერთენოვანი შედარებითი კორპუსების შემთხვევაში გვაქვს ერთი ენის ტექსტები, მაგრამ შესაძლოა, გვქონდეს სხვაობა სხვა კრიტერიუმების მიხედვით (Sharoff et al. 2013, 5). მაგალითად, სამიზნე აუდიტორია – დარგის სპეციალისტებისა და არასპეციალისტებისთვის შედგენილი ტექსტები (Deléger and Zweigenbaum 2009), დროის მონაკვეთი – მე-19 საუკუნის პრესის შედარება თანამედროვე პრესასთან (Rosset et al. 2012), წყარო – ერთი და იგივე მოვლენა გადმოცემული სხვადასხვა საინფორმაციო სააგენტოს მიერ (Wang and Callison-Burch 2011) ან ერთი ენის რამდენიმე ვარიანტზე შედგენილი სპეციალიზებული შე-

დარებითი კორპუსები (Wissik 2010). ამავე ტიპს შეგვიძლია მივაკუთვნოთ ბრიტანული ეროვნული კორპუსის 1994 და უახლესი 2014 წლის ვერსიები, რომლებიც მსგავსი პრინციპების მიხედვითაა შედგენილი (Brezina, Hawtin and McEnery 2014; Love et al. 2017).

ორენოვანი შედარებითი კორპუსების მაგალითია ინგლისურ-შვედური ახალი ამბების კორპუსი (Talvensaari et al. 2007), აგრეთვე, სხვადასხვა დარგობრივი შედარებითი კორპუსი (Utka et al. 2022; Yu 2024). უნდა აღინიშნოს, რომ ზანეტინის აზრით, ცალკეული ენების ეროვნული კორპუსები, რომლებიც მსგავსი სტრუქტურის მქონეა ფართო გაგებით, შეგვიძლია ჩავთვალოთ დაბალანსებულ ზოგადი ენის შედარებით კორპუსებად (Zanettin 2012, 150).

მრავალენოვანი კორპუსების მაგალითია „პარლამინტის“ (ParlaMint) კორპუსი, რომელიც შედგება ევროკავშირის წევრი 17 სახელმწიფოს ენაზე წარმართული საპარლამენტო დებატებისგან კოვიდ 19-ის თემაზე (Erjavec et al. 2023). შედარებითი კორპუსები შექმნილია მსგავსი სტრუქტურისა და ანოტაციების დართვის წესების მიხედვით და წარმოადგენს სასაუბრო შედარებითი კორპუსების მაგალითს.

მეორე პარამეტრი, რომლის მიხედვითაც შესაძლებელია, შედარებითი კორპუსების დიქტომია გულისხმობს კორპუსების სპეციალიზაციის ხარისხის მახასიათებელს, კერძოდ, კორპუსების დაჯგუფებას ზოგადი ენის და დარგობრივ/სპეციალიზებულ კორპუსებად. ზოგადი ენის კორპუსებს განეკუთვნება ერთენოვანი კორპუსები, როგორიცაა „ბრაუნი“ (Brown) და „ლობი“ (LOB – Lancaster-Oslo-Bergen), რომლებიც მოიცავს სხვადასხვა ჟანრის ტექსტებს: პრესა, მხატვრული ლიტერატურა და სხვა (Francis and Kucera 1979; Johansson et al. 1986). რაც შეეხება დარგობრივ, ანუ სპეციალიზებულ კორპუსებს, რიგი მკვლევრებისა მხოლოდ ამ ტიპის კორპუსებს მიიჩნევს შედარებით კორპუსებად (Kenning 2010, 488; Munday, Finto and Blakesley 2022, 86). შედარებითი დარგობრივი კორპუსი მოიაზრებს ტექსტების ერთობლიობას, რომლებიც რეპრეზენტაციულია და ასახავს კონკრეტულ დარგობრივ ენას/ვარიანტს. როგორც წესი, ამ ტიპის კორპუსები იქმნება ერთი დარგის (მაგალითად, მედიცინა) ტექსტების გამოყენებით, ხოლო სხვა პარამეტრები, როგორიცაა ჟანრი (მაგალითად, სახელმძღვანელოები, სტატიები), ზომა და ენა, განსხვავდება კვლევის მიზნისა და კორპუსის შედგენის კრიტერიუმების მიხედვით (Zanettin 2012, 151; Fantinuoli 2018, 139). მაგალითად, ამ ტიპის კორპუსს მიეკუთვნება „შედარებითი კორპუსებისა და კომპიუტერის მიერ მხარდაჭერილი (CAT) თარგმნის“ კვლევის ფარგლებში შექმნილი სამენოვანი მკერდის კიბოს დარგობრივი კორპუსი, რომელიც შედარებითი კორპუსებიდან მორ-

ფო-კომპოზიციური თარგმანის მეთოდის გამოყენებით ინგლისური, ფრანგული და გერმანულენოვანი სამედიცინო ტერმინოლოგიის მოძიებასა და ამოღებას ისახავდა მიზნად (Delpéch 2014).

მესამე მნიშვნელოვანი პარამეტრი, რომლის მიხედვითაც შეგვიძლია შედარებითი კორპუსების დახასიათება, გულისხმობს მათში შემავალი ტექსტების ტიპებს, კერძოდ შედარებითი კორპუსები შესაძლოა, იყოს წერილობითი (Furtado and Teixeira 2022; Kupietz et al. 2024) ზეპირმეტყველების (Erjavec et al. 2023), ორივე სახის ტექსტის შემცველი (Čermáková et al. 2021) ან მულტიმოდალური (Segouat and Braffort 2009). როგორც ერთენოვანი კორპუსების შემთხვევაში, შედარებით კორპუსებშიც ჭარბობს წერილობითი სახის კორპუსები, როგორც ზოგადი ენის, აგრეთვე დარგობრივი ენის შემთხვევაში.

შედარებითი კორპუსების გამოყენების სფეროები

მართალია, აკადემიურ სივრცეში შედარებითი კორპუსებისადმი ინტერესი ჯერ კიდევ 1990-იან წლებში ჩნდება, მაგრამ 2008 წლიდან შედარებითი კორპუსების შექმნასა და გამოყენებასთან დაკავშირებული სამუშაო შეხვედრების ორგანიზებით (Building and using comparable corpora, BUCC) საფუძველი ეყრება ამ ტიპის კორპუსების შესახებ არსებული საკითხების სისტემურ კვლევას. სამუშაო შეხვედრების ყოველწლიური ხასიათი კი სამეცნიერო წრეებში შედარებითი კორპუსების მიმართ სულ უფრო მეტად მზარდ ინტერესზე მეტყველებს.

შედარებითი კორპუსებისადმი ინტერესის ზრდა რამდენიმე მნიშვნელოვანი ფაქტორით არის განპირობებული. პირველ რიგში, მათი ფართო ხელმისაწვდომობით, ვინაიდან ინტერნეტსივრცეში ერთი და იმავე საკითხის ან დარგის შესახებ არსებული ტექსტების მოძიება მაღალი ხარისხის პარალელური მასალების მოძიებაზე ბევრად ადვილია (Granger 2010, 4; Alonso et al. 2012, 390). გარდა ამისა, უნდა აღინიშნოს, რომ პარალელური კორპუსების რიცხვის ეტაპობრივი ზრდის მიუხედავად, კვლავაც პრობლემურ საკითხად რჩება ყველა დარგისათვის საკმარისი რაოდენობით პარალელური ტექსტების შეგროვება და ამ გზით, კორპუსების რეპრეზენტაციულობის გაზრდა. განსაკუთრებით იმ შემთხვევაში, როცა საქმე ეხება ახლად აღმოცენებულ ან განვითარების პროცესში მყოფ დარგებს, როგორიცაა განახლებადი ენერგია, კიბერუსაფრთხოება (Blancafort et al. 2010, 263; Utka et al. 2022) ან კონკრეტულ საკითხებსა და თემებს, როგორიცაა, მაგალითად, კოვიდ-19 და ვაქცინაცია, სადაც მზარდი ინტერესი ამ თემისადმი უკუპროპორციულია იმ ორენოვან პარალელურ ხელმისაწვდომ მასალათა რაოდენობისადმი.

დენობასთან, რისი მოძიებაც ამ ეტაპზე არის შესაძლებელი (Filippova, Can and Corpas Pastor 2021, 61).

გარდა ამისა, შედარებითი კორპუსების გამოყენების სფეროების ჩამონათვალი საკმაოდ მრავალფეროვანია და მოიცავს ისეთ მიმართულებებს, როგორიცაა დისკურსის ანალიზი, პედაგოგიკა, სოციოლინგვისტიკა, პრაგმატიკა, ტიპოლოგიური კვლევები, შეპირისპირებითი ლინგვისტიკა / ტერმინოგრაფია, თარგმანმცოდნეობა, ბუნებრივი ენის დამუშავება და მანქანური თარგმნის მოდელები, ტერმინებისა და დარგობრივი ლექსიკონების ავტომატური გენერირება, ენათაშორისი ინფორმაციის მოძიება და სხვა (Maia 2003, 28; Sharoff, Rapp and Zweigenbaum 2023, 125).

შედარებითი კორპუსების შედგენისა და გამოყენების საკითხის შესახებ აქტუალური თემების დასადგენად გავაანალიზეთ BUCC-ის¹ ოფიციალურ ვებსაიტზე წარმოდგენილი 2008-2024 წლების საკონფერენციო კრებულებიდან 123 სტატია და გამოყავით რამდენიმე ძირითადი თემატური კატეგორია, რომელშიც გამოყენებული იყო შედარებითი ტიპის კორპუსები: სხვადასხვა სახის შედარებითი კორპუსების შედგენა (ზოგადი, სასწავლო, დარგობრივი), ინფორმაციის ამოღების ავტომატური და არაავტომატური მეთოდები (პარალელური წინადადებების ამოღება შედარებითი კორპუსიდან, ორენოვანი ტერმინოლოგიის ამოღება, ორენოვანი/მრავალენოვანი ლექსიკონების ავტომატურად გენერირება, თარგმნითი ეკვივალენტების მოძიება და სხვა), მანქანური თარგმანი, შეპირისპირებითი ლინგვისტური კვლევები და შედარებითი კორპუსების შედგენისა და/ან გამოყენებისთვის პროგრამული ინსტრუმენტების შექმნა (იხ. ცხრილი 1).

თემატური კატეგორია	სტატიების რაოდენობა
შედარებითი კორპუსების შედგენა	34
ინფორმაციის ამოღების ავტომატური და არაავტომატური მეთოდები	61
მანქანური თარგმნა	11
შეპირისპირებითი ლინგვისტური კვლევები	14
შედარებითი კორპუსების შედგენისა და/ან გამოყენებისთვის პროგრამული ინსტრუმენტების შექმნა	3

ცხრილი 1. შედარებითი კორპუსების შესახებ BUCC-ის საიტზე გამოქვეყნებული სტატიების თემატიკის რაოდენობრივი ანალიზი

1 იხ. <https://comparable.limsi.fr/bucc2024/bucc-introduction.html> [წვდომის თარიღი: 05.12.2024]

ამრიგად, შეიძლება ითქვას, რომ შედარებითი კორპუსების შედგენა, მათი გამოყენება და ინტეგრირება, მათ შორის ორენოვანი/მრავალენოვანი ტერმინოლოგიური და ლექსიკოგრაფიულ პროექტებში კვლევის აქტუალური საგანია. გარდა ამისა, სამეცნიერო ლიტერატურის ანალიზმა გვიჩვენა, რომ დარგობრივი შედარებითი კორპუსები იქმნება როგორც ინგლისურისა და სხვა ენებისთვის, როგორიცაა ინგლისური და იტალიური (Gaballo 2010; Giampieri 2024), ინგლისური და ლიეტუვური (Utka et al. 2022; Usinskiene and Rackeviciene 2022), ინგლისური, ფრანგული, ესპანური, პორტუგალიური და გერმანული (Tagnin and Teixeira 2012; Araceli et. al 2012; Daille 2017; Bartolomé-Díaz and Frontini 2020; Furtado and Teixeira 2022;), აგრეთვე, სხვა ენობრივი წყვილებისთვის, რომელშიც ინგლისური არ ფიგურირებს (Goeuriot, Morin and Daille 2009; Yu 2024). ხოლო დარგების ჩამონათვალი საკმაოდ მრავალფეროვანია და მოიცავს სამართლის (Wissik 2010; Giampieri 2024), საავტომობილო ინდუსტრიის (Yu 2024), კულტურული მემკვიდრეობის (Araceli et. al 2012); არქიტექტურის (Bartolomé-Díaz and Frontini 2020), კიბერუსაფრთხოების (Utka et al. 2022), მიგრაციის (Furtado and Teixeira 2022; Usinskiene and Rackeviciene 2022); სამედიცინო (Goeuriot, Morin and Daille 2009; Delpech 2014, Daille 2017), განახლებადი ენერგიის (Gornostay et al. 2012; Daille 2017), ბუნებრივი კატასტროფების (Buendia Castro 2013), ქვის ჭრის (Gaballo 2010) და სხვა დარგებს.

დარგობრივი შედარებითი კორპუსების შედგენის მეთოდოლოგია

რეპრეზენტაციულობის, ბალანსისა და კორპუსის ზომის განსაზღვრა მნიშვნელოვან როლს ასრულებს დარგობრივი შედარებითი კორპუსების შექმნისათვის საჭირო კრიტერიუმების შემუშავების დროს და ამ პროცესს შესაძლოა, თან ახლდეს სპეციფიკურად დარგობრივი შედარებითი კორპუსების შედგენისათვის დამახასიათებელი პრობლემები და მათი გადაჭრის გზების ძიება.

რეპრეზენტაციულობა გულისხმობს იმას, თუ რამდენად სანდო და მეტ-ნაკლებად ამომწურავია კონკრეტული ენის ან ენის ვარიანტის კორპუსი. ბაიბერის თანახმად, კორპუსის ნიმუში უნდა შეიცავდეს სამიზნე პოპულაციის სრულ მრავალფეროვნებას, რაც შეიძლება ორი პარამეტრის მიხედვით გავზომოთ: სიტუაციური და ლინგვისტური (Biber 1993, 243). პირველი გულისხმობს საკვლევ პოპულაციაში ანუ დარგობრივი კორპუსების შემთხვევაში კონკრეტულ დარგებში არსებული ტექსტების ტიპების, ჟანრებისა და რეგისტრების განსაზღვრას, ხოლო ლინგვისტური მახასიათებელი მოიაზრებს თითოეულ ტექსტში სიტყვათა და თითოეულ ჟანრში შესული ტექსტების რაოდენობას.

რაც შეეხება ბალანსს, იგი მოიაზრებს კორპუსის შინაგან ერთგვაროვნებას, რომელიც განისაზღვრება თითოეული სახის ცვლადის (სიტუაციური ან/და ლინგვისტური) პროპორციით (McEnery and Brookes 2022, 37). დაბალანსებული და რეპრეზენტაციული კორპუსების მაგალითს წარმოადგენს „ბრაუნისა“ და „ლობის“ კორპუსები, სადაც თითოეული 4 მსხვილი ჟანრობრივი კატეგორია დაყოფილია 15 ქვეკატეგორიად და თითოეულში თანაბარი რაოდენობით არის წარმოდგენილი ხუთას-ხუთასი ტექსტი, თითოეული ტექსტი კი 2000 სიტყვაფორმისგან შედგება (Francis and Kucera 1979; Johansson et al. 1986).

რეპრეზენტაციულობისა და ბალანსის საკითხი უფრო მეტად პრობლემურია დარგობრივი კორპუსების, მათ შორის, დარგობრივი შედარებითი კორპუსების შემთხვევაში. კოესტერის თქმით, თუკი კორპუსში წარმოდგენილია მხოლოდ ერთი კონკრეტული ჟანრი, სჯობს, დავეყრდნოთ სიტუაციურ რეპრეზენტაციულობას (Koester 2022, 51). ამ შემთხვევაში საჭიროა, ამ ჟანრის ფარგლებში არსებული ქვეჟანრების ან სხვადასხვა მოსაზრების და დაწესებულების განსაზღვრა, რომლებიც იყენებენ ამ ჟანრის ტექსტებს და ამ გზით, რეპრეზენტაციულობის დაცვა (Usinskiene and Rackeviciene 2022; Giampieri 2024, Yu 2024). მაგალითად, ფურტადო და ტეიქსეირა მიგრაციული დარგობრივი შედარებითი კორპუსის შედგენის პროცესში ეყრდნობიან იმ ავთენტიკურ ჟანრებს, რაც რეალურად გამოიყენება მიგრაციის დარგში და თავშესაფრის მაძიებელთა სოციალურ ჯგუფებში (Furtado and Teixeira 2022, 5). ამ მიდგომას ისინი კორპუსით-განპირობებულს უწოდებენ და თვლიან, რომ ამ თვალსაზრისით, კორპუსს უფრო მეტი შანსი ექნება, იყოს ამ დარგის რეპრეზენტაციული ნიმუში. დარგობრივ ტერმინოლოგიურ კორპუსებში რეპრეზენტაციულობისა და ბალანსის დაცვის კუთხით განსხვავებული მოსაზრება აქვთ სხვა მკვლევრებს. მათი აზრით, რეპრეზენტაციული და დაბალანსებული კორპუსი შეიცავს თანაბარი პროპორციული რაოდენობით სხვადასხვა საკომუნიკაციო დონის ტექსტებს (Bowker 1996, 45). მათი პოზიციის თანახმად, დაბალანსებული დარგობრივი კორპუსის შესაქმნელად საჭიროა როგორც მაღალი სპეციალიზაციის, ასევე ნახევრად-სპეციალიზებული და არასპეციალისტებისათვის გამიზნული ე.წ სამეცნიერო-პოპულარული ჟანრის ტექსტების შეტანა, რაც შექმნის დარგის მეტ-ნაკლებად სრულყოფილ და სანდო ნიმუშს. ამგვარ მიდგომას იყენებენ დარგობრივი შედარებითი კორპუსების შედგენის დროს (Buendia Castro 2013, 348; Utko et al. 2022).

უნდა აღინიშნოს, რომ ფურტადო და ტეიქსეირა (2024, 56) იუ-

რალექსის 2024 წლის კონგრესზე თავის მოხსენებაში ყურადღებას ამახვილებენ იმ ფაქტზე, რომ კორპუსის რეპრეზენტაციულობისა და ბალანსის საკითხი განსაკუთრებულ სირთულეს ქმნის და არ არის გადაჭრილი მრავალენოვან შედარებით სპეციალიზებულ კორპუსებთან მიმართებით, სადაც კულტურულ და ლინგვისტურ სხვაობებს შესაძლოა მოჰყვეს დამატებითი პრობლემების წარმოშობა მრავალენოვანი კორპუსების რეპრეზენტაციულობისა და ბალანსის შეფასებისას. მეტიც, დაშვება, რომ ყველა ენაში გვაქვს მსგავსი ტექსტური ტიპოლოგია და თანაბრად ვხვდებით ყველა ჟანრს, მცდარია და სინამდვილეში, არ შეესაბამება არსებულ რეალობას (Furtado and Teixeira 2024, 56). სამეცნიერო ლიტერატურაში აღნიშნული პრობლემის გადასაჭრელად გონივრულ ალტერნატივად ესახებათ რეპრეზენტაციულობისა და ბალანსის საკითხის გადაჭრა ერთ ენაში, რომელიც ამოსავალ ენას წარმოადგენს (Utka et al. 2022, 131; Furtado and Teixeira 2024, 57), შემდეგ ეტაპზე კი სხვა ენებში საკვანძო სიტყვების/სინთეზური სიების ან ლექსიკური მრავალფეროვნების მაჩვენებლების შედარება ეკვივალენტების მოსაძიებლად, როგორც ამას აღნიშნავენ შემდეგი ავტორები: ბუენდია კასტრო (2013), უსიენსკე და რასკევიჩიანე (2022), ფურტადო და ტეიქსეირა (2022). მსგავს მოსაზრებას ლინგვისტური მახასიათებლების მიხედვით, კორპუსების რეპრეზენტაციულობისა და ბალანსის გასაზომად იზიარებს კოესტერი (2022, 51), რომელიც საუბრობს ლექსიკურ დონეზე „სატურაციის“ ან „სასრულობის“ ხარისხის გაზომვაზე (McEnery and Wilson 2001, 148-67). კორპუსი ითვლება საკმარისი სატურაციის ხარისხის მქონედ ანუ რეპრეზენტაციულად, თუკი ახალი მონაცემების დამატება არ იწვევს ახალი ლექსიკური ერთეულების გამოჩენას (Koester 2022, 52). მაგალითად, ამ მიდგომას ვხვდებით შემდეგ ავტორებთან: ბუენდია კასტრო (2013), გიამპიერი (2024, 46), სადაც, მიუხედავად შედარებითი კორპუსების განსხვავებული სიდიდეებისა, მეტი ტექსტის დამატება არ იწვევს ლექსიკური მრავალფეროვნების გაზრდას და შესაბამისად, კორპუსები საკმარისად რეპრეზენტაციულადაა მიჩნეული.

რაც შეეხება დარგობრივი შედარებითი კორპუსების სიდიდეს, დარგობრივ კონტექსტში გადამწყვეტი მნიშვნელობა ენიჭება მასალების ხარისხს და არა მათ სიდიდეს. ბეიკერის თანახმად, „დარგობრივი კორპუსის შედგენის პროცესში, შესაძლოა, უფრო მეტად შერჩევითი მიდგომა გამოვიყენოთ ტექსტების ამორჩევის პროცესში, რაც გულისხმობს იმას, რომ ტექსტური მონაცემების ხარისხი ან მათი შინაარსი უფრო მნიშვნელოვან როლს ასრულებს, ვიდრე ტექსტური მასალის რაოდენობა“ (Baker 2006, 29).

დარგობრივი შედარებითი კორპუსების შემთხვევაში ვხვდებით

განსხვავებული სიდიდის კორპუსებს, რაც განპირობებულია კვლევის მიზნებითა და ხელმისაწვდომი მასალების რაოდენობით. მაგალითად, ტერმინოლოგიის მოძიებისათვის, მათ შორის ტერმინთა ვარიანტების ამოღებისათვის, კომპოზიციური თარგმნის ავტომატური მეთოდის დამუშავებისა და გლოსარიუმების შედგენისათვის სამიზნე ტოკენების რაოდენობა თითოეული ენისათვის 300,000-800,000 სიტყვაფორმას შეადგენს (Goeuriot, Morin and Daille 2009; Delpech 2014; Daille 2017; Usinskiene and Rackeviciene 2022). სხვა შემთხვევებში კი სიდიდე 1 მლნ. ტოკენიდან 3 მლნ.-მდე აღწევს (Alonso et al. 2012; Utka et al. 2022; Furtado and Teixeira 2022).

შედარებით კორპუსებში მთავარი სირთულე ერთგვაროვანი სიდიდის კორპუსების შედგენაა, რაც ენების განსხვავებული მორფოსინტაქსური მახასიათებლებითა და ხელმისაწვდომი მასალების მოცულობით არის განპირობებული. მიუხედავად იმისა, რომ რიგ კვლევებში კორპუსების სიდიდე მეტ-ნაკლებად იდენტურია ყველა ენისათვის (Utka et al 2022, Furtado and Texteira 2022), სიტყვაფორმათა არაერთგვაროვნება არ წარმოადგენს სიახლეს დარგობრივი შედარებითი კორპუსებისათვის (Gaballo 2010, Buendia Castro 2013), რაც ძირითადად ინგლისურ ენასთან შედარებით მეორე ენაში არასაკმარისი ხელმისაწვდომი მასალების რაოდენობით იყო განპირობებული.

დიდი ხნის განმავლობაში არამსგავსი მოცულობის კორპუსები ე.წ. „არადაბალანსებულ სპეციალიზებულ კორპუსებად“ მიიჩნეოდა, ვინაიდან ლიჩის (2003) მოსაზრებით, მსგავსი რაოდენობის სიტყვაფორმების არსებობა შესაძლებელს ხდის შედარებითი ტიპის კვლევების ჩატარებას სხვადასხვა ენის ვარიანტისათვის. მიუხედავად ამისა, უნდა აღინიშნოს, რომ მორინი და ჰაზემი (2016, 576) თვლიან, რომ იმ შემთხვევაში, თუკი შედარებითი კორპუსის შემადგენელი ერთ-ერთი ენა ინგლისურია, ეს უდავოდ იწვევს სიტყვაფორმათა არაერთგვაროვან რაოდენობას, ვინაიდან ინგლისურ ენაზე შექმნილი სამეცნიერო ლიტერატურისა თუ სტატიების რაოდენობა ჭარბობს სხვა ენებს, თუმცა აღნიშნული ფაქტიც კი არ უშლის მათ ხელს ე.წ. „არადაბალანსებული შედარებითი კორპუსიდან“ ლექსიკონების გენერირების ავტომატური მოდელების დამუშავებაში (Morin and Hazem 2016, 576).

შედარებითი კორპუსების შედგენის პროცესში კიდევ ერთ მნიშვნელოვან ფაქტორს შედარებითობის ხარისხის და კრიტერიუმების განსაზღვრა წარმოადგენს. შაროვის აზრით, შედარებითობის ცნება ეყრდნობა ჩვენს უნარს, შევაფასოთ არსებული განსხვავებები ისეთი ტიპის კორპუსებს შორის, რომლებსაც ჩვენ შედარებითად მივიჩნევთ, მაგრამ ეს პროცესი ჯერ კიდევ უფრო მეტად ჩვენს ცოდნასა და უნა-

რებზეა დამოკიდებული, ვიდრე მეცნიერებაზე (Sharoff 2013, 113). ერთმანეთისგან განარჩევენ მაღალი შედარებითობის, სუსტი შედარებითობის და კავშირის არმქონე ტექსტებისგან შემდგარ შედარებით კორპუსებს (Sharoff et al. 2013, 3). შედარებითობის განსაზღვრისთვის სამეცნიერო ლიტერატურაში შეგვიძლია გამოვყოთ მსგავსება ექსტრალინგვისტური (გარეგანი) პარამეტრების დონეზე, რომელიც მოიაზრებს მსგავსებას დარგის, ჟანრის, პერიოდის, ტექსტების შეგროვების მეთოდის და საკომუნიკაციო კონტექსტების მიხედვით (Maia 2003, 3; McEnery and Xiao 2007; Goeuriot, Morin and Daille 2009, 56; Daille 2017, 87).

გარდა ამისა, ვხვდებით სხვადასხვა სახის სტატისტიკური საზომის გამოყენების მაგალითებს. ძირითადად კორპუსების სინშირული სიების/საკვანძო სიტყვების შედარებას საზიარო ლექსიკის რაოდენობის განსაზღვრისათვის (Utka et al. 2022; Usinskiene and Rackeviciene 2022). აგრეთვე, ვხვდებით ლექსიკური მრავალფეროვნების მსგავსების გაზომვას TTR (type/token ratio) (ტიპების ტოკენებთან შეფარდებისა) ან STTR (Standardized type/token ratio) (ნორმალიზებული ტიპების ტოკენებთან შეფარდების) განსაზღვრის მეშვეობით შემდეგ კვლევებში ბუნდია კასტრო (2013, 346), სეგირი (2014), ფურტადო და ტეიქსეირა (2022; 2024, 57).

ამრიგად, ერთ-ერთ მთავარ გამოწვევას დარგობრივი შედარებითი კორპუსების შედგენის პროცესში სხვადასხვა ენობრივ წყვილში ხელმისაწვდომი მასალისა და ჟანრების არაპროპორციული გადანაწილება ქმნის, რაც მოითხოვს მკვლევრების მხრიდან მკაფიო კრიტერიუმების შემუშავებას კორპუსების შედარებითობის უზრუნველყოფისთვის, რისთვისაც შესაძლებელია ვიხელმძღვანელოთ როგორც მხოლოდ ექსტრალინგვისტური (გარეგანი) პარამეტრების მსგავსებით, აგრეთვე, სხვადასხვა სტატისტიკური საზომის მსგავსებასთან კომბინაციით.

დასკვნა

შეჯამების სახით, უნდა ითქვას, რომ პარალელური კორპუსების სიმჭირის ან არარსებობის შემთხვევაში, შედარებითი კორპუსების გამოყენება, განსაკუთრებით ახლად აღმოცენებული ან მაღალი ინტერესის მქონე დარგებისთვის, კვლევის აქტუალური საგანია, მათ შორის, ორენოვანი/მრავალენოვანი ზოგადი ან დარგობრივი ენის ლექსიკონების შესადგენად. სამეცნიერო ლიტერატურის ანალიზმა გვიჩვენა, რომ ინგლისური და ქართული დარგობრივი შედარებითი კორპუსების შედგენისათვის აუცილებელი წინაპირობა შედარები-

თობის კრიტერიუმების შემუშავებაა, კერძოდ, მსგავსება დარგის, ტექსტების ჟანრების, საკომუნიკაციო კონტექსტისა და პერიოდების მიხედვით. აგრეთვე, იმ ფაქტის გათვალისწინებით, რომ ქართული და ინგლისური ენის დარგობრივ ტექსტებს შორის რაოდენობრივი არაერთგვაროვნება გვაქვს, მიზანშეწონილად მიგვაჩნია რეპრეზენტაციულობის საკითხი გადაიჭრას ამოსავალ, ქართულ ენაში, ხოლო შემდეგ ეტაპზე სიხშირული სიების ან ლექსიკური მრავალფეროვნების მაჩვენებლების შედარების გამოყენებით განისაზღვროს ინგლისური და ქართული კორპუსების მსგავსება.

დამოწმებანი

- Alonso, Araceli, Elena Blancafort, Clement De Groc, Chrystel Million and Geoffrey Williams. 2012. METRICC: Harnessing Comparable Corpora for Multilingual Lexicon Development. In *Proceedings of the 15th EURALEX International Congress* (August 2012), 389-403. Oslo, Norway.
- Baker, Peter. 2006. *Using Corpora in Discourse Analysis*. London & New York: Continuum.
- Bartolomé-Díaz, Zaida and Francesca Frontini. 2020. Building a Domain-specific Bilingual Lexicon Resource with Sketch Engine and Lexonomy: Taking Ownership of the Issues. *Proceedings of the Globalex Workshop on Linked Lexicography* (May), European Language Resources Association, 62-68. Marseille, France. <https://aclanthology.org/2020.globalex-1.11.pdf>.
- Belinda, Maia. 2003. What are Comparable Corpora? *Proceedings of the Corpus Linguistics Workshop on Multilingual Corpora: Linguistic requirements and technical perspectives*. Lancaster, U.K. <https://repositorio-aberto.up.pt/bitstream/10216/14024/2/whatare000072831.pdf>.
- Benko, Vladimir. 2024. The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. In *Text, Speech, and Dialogue. TSD 2024. Lecture Notes in Computer Science*, 1-16. Cham: Springer. https://doi.org/10.1007/978-3-031-70563-2_5.
- Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8 (4): 243-257. https://doi.org/10.1007/978-3-031-70563-2_5.
- Blancafort, Helena, Béatrice Daille, Tatiana Gornostay, Ulrich Heid, Claude Mechoulam, and Serge Sharoff. 2010. TTC: Terminology Extraction, Translation Tools and Comparable Corpora. In the *Proceedings of the 14th EURALEX International Congress*, 263-268. Leeuwarden/Ljouwert, The Netherlands. <https://euralex.org/publications/ttc-terminology-extraction-translation-tools-and-comparable-corpora/>.
- Bowker, Lynne. 1996. *Towards a Corpus-based Approach to Terminography*.

- Terminology* no. 3: 27-52. <https://doi.org/10.1075/term.3.1.03bow>.
- Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routledge. <https://doi.org/10.4324/9780203469255>.
- Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The Written British National Corpus 2014 – Design and Comparability. *Text & Talk* no 41 (5-6): 595-615. <https://doi.org/10.1515/text-2020-0052>.
- Buendia Castro, Miriam. 2013. Phraseology in Specialized Languages and its Representation in Environmental Knowledge Resources. PhD diss., University of Granada.
- Čermáková, Anna, Jarmo Jantunen, Tommi Jauhiainen, John Kirk, Michal Křen, Marc Kupietz and Elaine Uí Dhonnchadha. 2021. The International Comparable Corpus: Challenges in Building Multilingual Spoken and Written Comparable Corpora. *Research in Corpus Linguistics*, no. 9 (1), 89-103. <https://doi.org/10.32714/ricl.09.01.06>.
- Daille, Béatrice. 2017. Primitive Exploration of Variants Using Comparable Corpora. In *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*, 83-108. Terminology and Lexicography Research and Practice. no. 19. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.19>.
- Deléger, Louise and Pierre Zweigenbaum. 2009. Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora*, 2-10. Association for Computational Linguistics, Singapore. <http://aclweb.org/anthology/W/W09/W09-3102>.
- Delpech, Estelle Maryline. 2014. *Comparable Corpora and Computer-assisted Translation*. UK, USA: ISTE, John Wiley & Sons. <https://doi.org/10.1002/9781119002659>.
- Erjavec, Tomaž, Maciej Ogrodniczuk et al. 2023. The ParlaMint Corpora of Parliamentary Proceedings. *Lang Resources & Evaluation* 57, 415-448. <https://doi.org/10.1007/s10579-021-09574-0>.
- Fantinuoli, Claudio. 2018. The Use of Comparable Corpora in Interpreting Practice and Teaching. *The Interpreters' Newsletter* no. 23: 133-149. DOI: 10.13137/2421-714X/22402.
- Filippova, Darya, Burcu Can, and Gloria Corpas Pastor. 2021. Bilingual Terminology Extraction Using Neural Word Embeddings on Comparable Corpora. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, 58-64, Online. INCOMA Ltd. <https://aclanthology>.

- org/2021.ranlp-srw.9.pdf.
- Francis, W. Nelson and Henry Kucera. 1979. *Brown Corpus Manual*. Department of Linguistics, Brown University, Providence, Rhode Island, US. <http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM>.
- Fung, Pascale and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, 192-202. Hong Kong, China. <https://aclanthology.org/W97-0119.pdf>.
- Furtado, Anna and Teixeira, Elisa. 2022. Multilingual Corpus on Migration and Asylum (COMMIRE): planning, compilation, and overall content. *Texts Livre* no. 15: 1-18. DOI: 10.35699/1983-3652.2022.36965.
- Furtado, Anna Beatriz Dimas and Elisa Duarte Teixeira. 2024. Representativeness and Balance in Multilingual Comparable Corpora for Specialized Lexicography Revisiting Available Standards and Measures. Abstract. In *the Book of Abstracts of the XXI EURALEX International Congress* (8-12 October 2024): 56-58. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex_boa_20.pdf.
- Gaballo, Viviana. 2010. Translating Stones: a Corpus-based Linguistic and Lexicographic Study in Specialized Terminology. *Lexis [Online]*, no. 4: 55-64. DOI: 10.4000/lexis.583.
- Giampieri Patrizia. 2024. The Use of Comparable Corpora on (General) Terms and Conditions as a Pedagogical Tool in Translation Training between English and Italian. Doctoral thesis. L-Università ta' Malta, Malta.
- Goeuriot, Lorraine, Emmanuel Morin, and Béatrice Daille. 2009. Compilation of Specialized Comparable Corpora in French and Japanese. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, 55-63, Singapore. Association for Computational Linguistics. <https://aclanthology.org/W09-3110.pdf>.
- Gornostay, Tatiana, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafortd, Serge Sharoffe, and Claude Méchoulam. 2012. Terminology Extraction, Translation Tools and Comparable Corpora: TTC Concept, Midterm Progress and Achieved Results. *LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS)*. (May): 1-4. Istanbul, Turkey. <https://hal.science/hal-00819909/document>.
- Granger, Sylviane. 2010. Comparable and Translation Corpora in Cross-linguistic Research. Design, Analysis and Applications. *Journal of Shanghai Jiaotong University* no. 2: 14-21. https://sites.uclouvain.be/cecl/archives/Granger_Crosslinguistic_research.pdf.
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge

- University Press. <https://doi.org/10.1017/CBO9781139524773>.
- Hunston, Susan. 2022. *Corpora in Applied Linguistics*. 2nd edition. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108616218>.
- Johansson, Stig, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB Corpus: User's Manual*. The Norwegian Centre for the Humanities. <http://helmer.aksis.uib.no/icame/manuals/LOBMAN/INDEX.HTM>.
- Kenning, Marie-Madeleine. 2010. What are Parallel and Comparable Corpora and How Can We Use Them? In *The Routledge Handbook of Corpus Linguistics* 1st edition, ed. Anne O'Keeffe and Michael McCarthy, 487-500. London & New York: Routledge. <https://doi.org/10.4324/9780203856949>.
- Koester, Almut. 2022. Building Small Specialised Corpora. In *The Routledge Handbook of Corpus Linguistics* 2nd edition, ed. Anne O'Keeffe and Michael McCarthy, 48-61. London: Routledge. <https://doi.org/10.4324/9780367076399>.
- Kupietz, Marc, Piotr Banski, Nils Diewald, Beata Trawinski, and Andreas Witt. 2024. EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, 94-103, Torino, Italia: ELRA and ICCL. <https://aclanthology.org/2024.bucc-1.10.pdf>.
- Leech, Geoffrey. 1991. The State of the Art in Corpus Linguistics. In *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, 8-29. New York/ London, NY/England: Routledge.
- Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations. *International Journal of Corpus Linguistics* no. 22(3): 319-344. <https://doi.org/10.1075/ijcl.22.3.02lov>.
- McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics. An Introduction*. 2nd ed. Edinburgh, Scotland: Edinburgh University Press.
- McEnery, Tony and Richard Xiao. 2007. Chapter 2. Parallel and Comparable Corpora: What is Happening? *Incorporating Corpora: The Linguist and the Translator*, edited by Gunilla Anderman and Margaret Rogers, Bristol, 18-31. Blue Ridge Summit: Multilingual Matters. <https://doi.org/10.21832/9781853599873-005>.
- McEnery, Tony and Gavin Brookes. 2022. Building a Written Corpus: What are the Basics? In *The Routledge handbook of Corpus Linguistics* 2nd edition, ed. Anne O'Keeffe and Michael McCarthy, 35-47. London: Routledge. <https://doi.org/10.4324/9780367076399>.
- Morin, Emmanuel and Amir Hazem. 2016. Exploiting Unbalanced Specialized

- Comparable Corpora for Bilingual Lexicon Extraction. *Natural Language Engineering*. No. 22(4):575-601. DOI: 10.1017/S1351324916000140.
- Munday, Jeremy, Sara Ramos Pinto and Jacob Blakesley. 2022. *Introducing Translation Studies Theories and Applications*. London: Routledge <https://doi.org/10.4324/9780429352461>.
- Rapp, Reinhard. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 519-526. Association for Computational Linguistics. <https://doi.org/10.3115/1034678.1034756>.
- Rosset, Sophie, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn and Pierre Zweigenbaum. 2012. Structured Named Entities in Two Distinct Press Corpora: Contemporary Broadcast News and Old Newspapers. In: *Proceedings of the Sixth Linguistic Annotation Workshop*, 40-48. Association for Computational Linguistics, Jeju, Republic of Korea. <http://www.aclweb.org/anthology/W12-3606>.
- Seghiri, Miriam. 2014. Too Big or Not Too Big: Establishing the Minimum Size for a Legal Ad Hoc Corpus. *HERMES – Journal of Language and Communication in Business*, 27(53): 85-98. <https://doi.org/10.7146/hjlc.v27i53.20981>.
- Segouat, Jérémie and Annelies Braffort. 2009. Toward Categorization of Sign Language Corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, 64-67, Singapore. Association for Computational Linguistics. <https://acl-anthology.org/W09-3111.pdf>.
- Sharoff, Serge. 2013. Measuring the Distance Between Comparable Corpora Between Languages. In *Building and Using Comparable Corpora*. Eds Sharoff Serge, Reinhard Rapp, Pierre Zweigenbaum and Pascale Fung, 113-130. Heidelberg, Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-642-20128-8_6.
- Sharoff, Serge, Reinhard Rapp and Pierre Zweigenbaum. 2023. *Building and Using Comparable Corpora for Multilingual Natural Language Processing. Synthesis Lectures on Human Language Technologies*. Cham: Springer. <https://doi.org/10.1007/978-3-031-31384-4>.
- Talvensaari, Tuomas, Jorma Laurikkala, Kalervo Järveli, Martti Juhola and Heikki Keskustalo. 2007. Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems* no 25(1): 1-4. <https://doi.org/10.1145/1198296.119830>.
- Usienske, Olga and Sigita Rackevivienė. 2022. Parallel and Comparable Cor-

- pora for Terminology Analysis in the Domain of Migration. In the *Conference Proceedings Language for International Communication: Linking Interdisciplinary Perspectives: Language for Specific Purposes in the Era of Multilingualism and Technologies*. no. 4: 143-155. Latvia, Riga: University of Latvia Press. <https://doi.org/10.22364/lincs.2023.12>.
- Utka, Andrius Sigita Rackeviciene, Liudmila Mockiene, Aivaras Rokas, Marius Laurinaitis and Agne Bielinskiene. 2022. Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction. In *Selected Papers from the CLARIN Annual Conference 2021*. ed. by Monica Monachini and Maria Eskevich, 126-138. DOI: <https://doi.org/10.3384/9789179294441>.
- Wang, Rui and Chris Callison-Burch, 2011. Paraphrase Fragment Extraction from Monolingual Comparable Corpora. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 52-60. Association for Computational Linguistics, Portland. <http://www.aclweb.org/anthology/W11-1208>.
- Wissik, Tanja. 2010. Development and Use of Comparable Specialized Corpora of National German Varieties. In *Conference paper from Using Corpora in Contrastive & Translation Studies*, 1-18. Ormskirk, United Kingdom. www.lancaster.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/Wissik.pdf.
- Yu, Chenlu. 2024. A German-Chinese E-dictionary of Manufacturing Technology in Automotive Industry Entry Design. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, 661-668. Institute for the Croatian Language. Despot. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf.
- Zanettin, Federico. 2012. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome Publishing, Routledge. <https://doi.org/10.4324/9781315759661>.

Ketevan Mchedlishvili

Ilia State University, Georgia

ketevani.mchedlishvili.3@iliauni.edu.ge

DOI: 10.32859/kadmos/16/237-253

Comparable Corpora: Compilation Methods and Areas of Application

Abstract

Comparable corpora and their application in research have been an object of interest since the 1990s. Following the establishment of the annual workshop series “Building and Using Comparable Corpora” (BUCC) in 2008, there has been an increasing interest in comparable corpora and the study of their effectiveness for bilingual/multilingual projects. Although there is a general comparable corpus of the Georgian language compiled as part of the “Aranea” project, a family of web-crawled comparable corpora, currently, there are no specialized comparable corpora available for the Georgian language. In general, the application of comparable corpora for bilingual/multilingual specialized lexicography in Georgia is a novel research topic that has not been explored before. Therefore, this review paper aims to analyze the concept and types of comparable corpora. It also discusses the advantages of using comparable corpora and the areas of their application. Furthermore, the paper focuses on the methods of compiling specialized comparable corpora, in particular, such issues as representativeness, balance, corpus size, and comparability criteria.