

İlham TAHİROV

**SİSTEMLİ MÜƏLLİF KORPUSU: SABİR RÜSTƏMXANLI
KORPUSUNUN YARADILMASI PRİNSİPLƏRİ
XÜLASƏ**

Məqalədə sistemli müəllif korpusunun yaradılması prinsipləri Sabir Rüstəmxanlı korpusu nümunəsində təhlil edilir. Araşdırmanın məqsədi korpusun formalaşdırılma prinsiplərini, metodoloji əsaslarını, statistik göstəricilərini və elmi-tətbiqi imkanlarını təhlil etməklə sistemli müəllif korpusunun əsas səciyyəvi xüsusiyyətlərini ortaya qoymaqdır. Tədqiqatın materialını yazıçının poetik, publisistik, nəsr, elmi-publisistik və ictimai-siyasi yaradıcılığını əhatə edən əsərləri təşkil edir. Korpusun hazırlanması zamanı materialın reprezentativ seçilməsi, mətnlərin elektronlaşdırılması, çoxmərhələli filtrlənməsi, normallaşdırılması və lemmalaşdırılması vahid metodoloji prinsiplər əsasında həyata keçirilmişdir. Lemmalaşdırma prosesində Azərbaycan dilinin morfoloji sistemi, söz yaradıcılığı xüsusiyyətləri və sözlərin leksik statusu əsas meyar kimi qəbul edilmiş, xüsusilə feil formalarının, şəxssiz feil formalarının və defislə yazılan vahidlərin emalı üzrə ardıcıl metodoloji yanaşma tətbiq edilmişdir.

Araşdırma nəticəsində ilkin mərhələdə 3 220 690 sözləşlətmə (token) və 195 935 müxtəlif söz-formadan ibarət korpus bazası formalaşdırılmış, sonrakı filtrləmə və lemmalaşdırma nəticəsində 2 388 317 sözləşlətməni və 23 433 lemmanı əhatə edən yekun korpus yaradılmışdır. Müəyyən edilmişdir ki, korpusun reprezentativliyi, müəllif mətninin autentikliyinə qorunması, vahid metodoloji prinsiplər əsasında linqvistik emalı və statistik modelləşdirilməsi onun sistemli müəllif korpusu kimi formalaşdırılmasını təmin edən əsas amillərdir. Korpus müəllif idiolektinin, leksik-statistik modelinin, üslubi xüsusiyyətlərinin, açar sözlər sisteminin, söz yaradıcılığı prosesinin və digər korpus əsaslı linqvistik tədqiqatların aparılması üçün etibarlı empirik baza yaradır.

Tədqiqatın elmi yeniliyi Sabir Rüstəmxanlı korpusunun Azərbaycan dilçiliyində sistemli müəllif korpusunun yaradılması prinsipləri əsasında formalaşdırılmış ilk kompleks müəllif korpusu kimi əsaslandırılmasıdır. Təqdim olunan yanaşma Azərbaycan dilində sistemli müəllif korpuslarının yaradılması və korpusəsaslı leksik-statistik tədqiqatların inkişafı üçün metodoloji baza yaradır.

Açar sözlər: *müəllif korpusu, korpus dilçiliyi, Sabir Rüstəmxanlı, lemmalaşdırma, leksik-statistik model, tezlik lüğəti, korpusun reprezentativliyi*

Müasir dilçilikdə korpus yanaşması dil faktlarının empirik əsasda öyrənilməsinin ən məhsuldar istiqamətlərindən biri kimi qəbul edilir. Elektron mətn korpuslarının yaradılması və onların statistik metodlarla təhlili dil vahidlərinin işlənmə tezliyini, leksik-semantik münasibətlərini və funksional davranışını geniş material

əsasında araşdırmağa imkan verir. Bu baxımdan korpus dilçiliyi ənənəvi təsviri metodları kəmiyyət təhlili ilə tamamlayaraq dil hadisələrinin daha obyektiv qiymətləndirilməsi üçün etibarlı metodoloji baza yaradır [Sinclair, 1991; McEnery, Hardie, 2012]. Korpus yanaşmasının əsas üstünlüyü ondan ibarətdir ki, dil haqqında irəli sürülən nəzəri mülahizələr genişhəcmli faktik material əsasında yoxlanılır və empirik dəlillərlə əsaslandırılır.

Korpus dilçiliyinin mühüm istiqamətlərindən birini müəllif korpusları təşkil edir. Belə korpuslar konkret bir müəllifin yaradıcılığını sistemli şəkildə əhatə etməklə onun fərdi leksik fondunu, üslubi, tezlik və diskurs xüsusiyyətlərini öyrənməyə imkan verir. Müəllif korpusları milli korpuslardan fərqli olaraq dilin ümumi mənzərəsini deyil, konkret müəllifin fərdi dil xüsusiyyətlərini əks etdirir. Bu baxımdan müəllif korpusları fərdi üslubun, leksik dominantların və semantik sahələrin empirik-statistik əsasda müəyyənləşdirilməsi üçün mühüm mənbə hesab olunur. Stilometriya sahəsində aparılmış tədqiqatlar da göstərir ki, müəllif üslubunun müəyyənləşdirilməsində yalnız məzmun sözləri deyil, yüksək tezlikli funksional vahidlərin paylanması da mühüm göstərici ola bilər [Burrows, 2002; McEnery, Hardie, 2012].

Son illərdə Azərbaycan dilçiliyində korpus dilçiliyi istiqamətində müəyyən tədqiqatlar aparılmış, milli korpusun yaradılması, leksik bazanın formalaşdırılması, tezlik lüğətlərinin hazırlanması və elektron konkordansların tərtibi sahəsində mühüm addımlar atılmışdır. Bu tədqiqatlar Azərbaycan dilinin korpusəsaslı öyrənilməsi üçün elmi və texnoloji zəmin yaratmışdır. Xüsusilə milli korpus və leksik infrastruktur üzrə aparılan işlər korpus dilçiliyinin Azərbaycan dilçiliyində ayrıca tədqiqat istiqaməti kimi formalaşdığını göstərir [Mammadli, Mahmudov, Tahirov, 2025].

Bu istiqamətdə həyata keçirilən işlərin mühüm davamı kimi Hüseyn Cavid, Mikayıl Müşfiq, “Kitabi-Dədə Qorqud” və Sabir Rüstəmxanlının əsərləri əsasında hazırlanmış elektron konkordanslar Azərbaycan dilində müəllif və mətn korpuslarının yaradılması sahəsində mühüm təcrübə formalaşdırmışdır. Bununla belə, konkret müəllif yaradıcılığını lemmalaşdırılmış, statistik emaldan keçirilmiş və tezlik lüğəti ilə əlaqələndirilmiş sistemli korpus modeli kimi təqdim etmək ayrıca elmi və metodoloji vəzifə kimi meydana çıxır.

Sabir Rüstəmxanlının yaradıcılığı müasir Azərbaycan ədəbiyyatının poetik, publisistik və ictimai-fəlsəfi istiqamətlərini əhatə edən zəngin dil materialı təqdim edir. Onun əsərlərində *dil, söz, xalq, türk, tarix, dövlət, vətən, yaddaş* və *milli kimlik* kimi anlayışlar yalnız mövzu səviyyəsində deyil, leksik seçim və tezlik davranışı səviyyəsində də aparıcı yer tutur. Bu baxımdan Sabir Rüstəmxanlı yaradıcılığının korpus əsasında öyrənilməsi müəllifin fərdi dil sistemini statistik göstəricilər əsasında təhlil etməyə imkan verir.

Təqdim olunan məqalədə Sabir Rüstəmxanlının əsərləri əsasında formalaşdırılmış korpus sistemli müəllif korpusu nümunəsi kimi nəzərdən keçirilir. Məqalədə “sistemli müəllif korpusu” dedikdə müəllif yaradıcılığının mümkün qədər tam

əhatə olunması, vahid metodoloji prinsiplər əsasında emalı və statistik təhlil üçün hazırlanması nəzərdə tutulur. Məqalənin məqsədi bu korpusun formalaşdırılma prinsiplərini, statistik pasportunu, lemmalaşdırma xüsusiyyətlərini və leksik-statistik modelini təhlil etməkdir. Bu baxımdan təqdim olunan tədqiqat yalnız Sabir Rüstəmxanlı yaradıcılığının leksik-statistik xüsusiyyətlərini təsvir etmir, eyni zamanda müəllif korpuslarının Azərbaycan dilçiliyində tətbiqi və inkişaf perspektivlərini nümayiş etdirən nümunəvi araşdırma kimi çıxış edir. Beləliklə, araşdırmada müəllif korpusu həm elektron mətn toplusu, həm də konkret dil şəxsiyyətinin leksik-statistik modelinin qurulmasına imkan verən elmi baza kimi qiymətləndirilir. Bu məqsədlə əvvəlcə müəllif korpusu anlayışının nəzəri əsasları nəzərdən keçirilir, daha sonra isə Sabir Rüstəmxanlı korpusunun formalaşdırılması prinsipləri və metodoloji xüsusiyyətləri təhlil edilir.

Müəllif korpusları müasir korpus dilçiliyinin xüsusi istiqamətlərindən biri olmaqla konkret müəllifin dilinin empirik əsasda öyrənilməsi üçün formalaşdırılan xüsusi dil resurslarıdır. Milli və ya ümumi korpuslardan fərqli olaraq, müəllif korpuslarının əsas məqsədi bütövlükdə dil sistemini deyil, konkret dil şəxsiyyətinin leksik, morfoloji, sintaktik və üslubi xüsusiyyətlərini sistemli şəkildə əks etdirməkdir. Bu baxımdan müəllif korpusu fərdi dil modelinin qurulmasına, müəllif idiolektinin statistik və funksional xüsusiyyətlərinin müəyyənləşdirilməsinə, leksik vahidlərin işlənmə intensivliyinin, semantik əlaqələrinin və kontekstual davranışının obyektiv faktlar əsasında təhlilinə imkan yaradır [Sinclair, 1991; Biber, Conrad & Reppen, 1998; Baker, 2006; McEnery, Hardie, 2012]. Bu məqalədə “müəllif korpusu” dedikdə konkret müəllifə məxsus mətnlərin müəyyən seçim prinsipləri əsasında toplanmış, elektronlaşdırılmış, strukturlaşdırılmış, təmizlənmiş, lemmalaşdırılmış və statistik təhlil üçün hazırlanmış mətnlər toplusu nəzərdə tutulur. Belə başa düşülən müəllif korpusu yalnız elektron mətnlər bazası deyil, həm də konkret müəllifin dil şəxsiyyətinin empirik-statistik modelinin qurulmasına xidmət edən elmi tədqiqat platformasıdır. Sistemli müəllif korpusunun elmi dəyəri yalnız əhatə etdiyi mətnlərin həcmində deyil, ayrı-ayrı dil faktlarını müəllifin bütöv yaradıcılığı daxilində qarşılıqlı əlaqədə və kəmiyyət göstəriciləri əsasında qiymətləndirməyə imkan verməsindədir.

Müəllif korpuslarının yaradılması və onların korpusəsaslı tədqiqi beynəlxalq dilçilikdə artıq uzunmüddətli elmi ənənəyə malikdir. Müxtəlif dillər üzrə həyata keçirilmiş korpus və leksikoqrafik layihələr göstərir ki, müəllif korpusları konkret yazıçının leksik fondunun, üslubi xüsusiyyətlərinin, fərdi dil modelinin və idiolektinin öyrənilməsi üçün etibarlı empirik mənbə hesab olunur. İngilis dilçiliyində Uilyam Şekspirin yaradıcılığı əsasında hazırlanmış lüğətlər, konkordanslar və elektron korpuslar müəllif dilinin sistemli tədqiqinin klassik nümunələrini təşkil edir [Open Source Shakespeare; Shakespeare's Words]. Analoji yanaşmalar Mark Tven, Fyodor Dostoyevski və Aleksandr Puşkin yaradıcılığı əsasında formalaşdırılmış korpus və leksikoqrafik layihələrdə də öz əksini tapmışdır [Mark Twain Project Online; Словарь языка Достоевского, 2001;

Словарь языка Пушкина, 2000; Засорина, 2003; Russian National Corpus]. Həmin layihələr müəllif korpusunun yalnız mətnlərin elektron formada təqdimatı ilə məhdudlaşmadığını, onların elmi prinsiplər əsasında seçilməsi, strukturlaşdırılması, lemmalaşdırılması, konkordanslaşdırılması və statistik təhlilə hazırlanmasının korpusun elmi dəyərini müəyyən edən əsas amillərdən biri olduğunu göstərir. Məhz bu metodoloji yanaşma təqdim olunan tədqiqatın da nəzəri əsasını təşkil edir. Yuxardakı nümunələr müəllif korpuslarının yalnız mətn bazası deyil, müxtəlifistiqamətli linqvistik tədqiqatlar üçün çoxfunksiyalı elmi platforma kimi istifadə olunduğunu göstərir.

Azərbaycan dilçiliyində korpus texnologiyalarının tətbiqi son illərdə ardıcıl inkişaf etmiş, milli korpusun yaradılması, leksik bazanın formalaşdırılması, elektron konkordansların hazırlanması və statistik leksikoqrafiya istiqamətində mühüm nəticələr əldə olunmuşdur [Azərbaycan dilinin tezlik lüğəti, 2010; Mahmudov, 2013; Mahmudov, 2018; Mahmudov, Tahirov, 2025]. Bu istiqamətdə aparılan tədqiqatlar Azərbaycan dilinin korpus infrastrukturunun formalaşdırılması ilə yanaşı, korpusəsaslı dil araşdırmalarının metodoloji bazasını da yaratmışdır. Xüsusilə Azərbaycan dilinin leksik bazasının qurulması, tezlik lüğətlərinin hazırlanması və elektron konkordansların tərtibi sahəsində əldə olunan təcrübə sonrakı müəllif korpuslarının yaradılması üçün mühüm elmi və texnoloji zəmin olmuşdur [Mammadli, Mahmudov, Tahirov, 2025]. Bu baxımdan Sabir Rüstəmханlı korpusu mövcud korpus infrastrukturunun məntiqi davamı kimi çıxış edir və müəllif korpusunun formalaşdırılması sahəsində əldə edilmiş metodoloji prinsiplərin konkret müəllif materialı üzərində sistemli tətbiqini əks etdirir [Tahirov, 2026].

Sabir Rüstəmханlı korpusunun formalaşdırılması zamanı əsas metodoloji prinsip müəllifin dil sistemini mümkün qədər tam və reprezentativ şəkildə əks etdirən mətnlər toplusunun yaradılması olmuşdur. Korpusun əsas mənbə bazasını yazıçının müxtəlif illərdə nəşr olunmuş, poetik, publisistik, nəsr, elmi-publisistik və ictimai-siyasi yaradıcılığını əhatə edən "Seçilmiş əsərləri"nin 15 cildliyi təşkil etmişdir. Bununla yanaşı, müəllifin həmin nəşrə daxil edilməyən, lakin tədqiqatın məqsədləri baxımından əhəmiyyət kəsb edən digər mətnləri də korpusun yaradılması zamanı nəzərə alınmışdır. Belə seçim müəllifin yaradıcılığının ayrı-ayrı mərhələlərini deyil, bütövlükdə idiolektini xarakterizə edən leksik və üslubi xüsusiyyətlərin daha dolğun şəkildə əhatə olunmasına imkan vermişdir. Korpusun reprezentativliyi də məhz materialın janr müxtəlifliyi, xronoloji əhatəliliyi və müəllif yaradıcılığının əsas istiqamətlərini əhatə etməsi ilə təmin olunmuşdur. Seçilmiş material korpus dilçiliyində qəbul olunmuş reprezentativlik prinsipinə uyğun olaraq müəllif dilinin empirik-statistik modelinin qurulması üçün etibarlı faktik baza təşkil edir.

Korpusun formalaşdırılmasının növbəti mərhələsi mətnlərin elektronlaşdırılması, texniki cəhətdən emalı və korpus mühitində işləmə biləcək formaya gətirilməsi olmuşdur. Bu proses yalnız çap mətnlərinin rəqəmsal formata keçirilməsi ilə

məhdudlaşmamış, eyni zamanda onların korpus dilçiliyinin tələblərinə uyğun şəkildə normallaşdırılmasını və strukturlaşdırılmasını da əhatə etmişdir. Mətnlərin elektronlaşdırılması və ilkin emalı zamanı prioritet müəllif mətninin autentikliyinə maksimum dərəcədə qorunması olmuşdur. Bu səbəbdən mətnlər üzərində yalnız korpusun texniki bütövlüyünü və statistik emalını təmin etmək üçün zəruri olan müdaxilələr həyata keçirilmiş, müəllifin fərdi dil xüsusiyyətlərini əks etdirən leksik, qrammatik və üslubi xüsusiyyətlər olduğu kimi saxlanılmışdır. Mətnlər üzərində aparılan ilkin emal zamanı müxtəlif texniki elementlər, tərtibat xüsusiyyətləri və dil sisteminin statistik təsvirinə təsir göstərməyən qeyri-mətn komponentləri təmizlənmişdir. Beləliklə, Sabir Rüstəmxanlı korpusunun formalaşdırılması representativlik və autentiklik prinsiplərinin qarşılıqlı vəhdəti əsasında həyata keçirilmişdir. Bu metodoloji yanaşma sonrakı mərhələlərdə aparılan lemmalaşdırma, konkordanslaşdırma və statistik təhlillərin bilavasitə müəllif dilinin özünəməxsus xüsusiyyətlərini əks etdirməsini təmin etmişdir.

Korpusun linqvistik emalına başlanmadan əvvəl mətnlər tam söz-forma səviyyəsində statistik işlənmişdir. Bu mərhələdə korpusda işlənən bütün söz-formalar ayrıca müəyyənləşdirilmiş, onların işlənmə tezliyi hesablanmış və ilkin statistik baza formalaşdırılmışdır. İlkin emalın nəticələrinə görə korpus 3 220 690 sözləşmədən və 195 935 müxtəlif söz-formadan ibarət olmuşdur. Bu statistik göstəricilər müəllif mətnlərində işlənmiş bütün dil vahidlərini, o cümlədən appelyativ və xüsusi adları, rəqəmləri və texniki elementləri əhatə etmişdir. Lakin müəllif korpusunun əsas məqsədi Sabir Rüstəmxanlı dilinin leksik sistemini və tezlik modelini müəyyənləşdirmək olduğundan, ilkin söz-forma bazası üzərində çoxmərhələli filtrləmə və normallaşdırma prosesi həyata keçirilmişdir. Bu mərhələdə rəqəmlər, rəqəmli birləşmələr, texniki qırıntılar, durğu işarələri, eləcə də xüsusi adlar (şəxs, soyad, ölkə, şəhər, kənd, dağ, çay və s.) korpusun leksik-statistik bazasından çıxarılmışdır. Bununla yanaşı, defislə yazılan vahidlər ümumi söz-forma bazasından ayrılmış və onların leksik statusu müstəqil şəkildə qiymətləndirilmişdir. Belə yanaşma filtrləmə prosesinin yalnız texniki təmizləmə əməliyyatı deyil, eyni zamanda sonrakı lemmalaşdırmanın metodoloji əsasını təşkil edən mərhələ olduğunu göstərir.

Filtrləmə və normallaşdırmadan sonra korpus materialı lemmalaşdırma mərhələsinə keçirilmişdir. Bu mərhələdə müxtəlif qrammatik formalarda işlənən sözlər vahid leksik əsas ətrafında birləşdirilmiş, lemmalaşdırma zamanı qəbul edilmiş metodoloji prinsiplər korpusun sonrakı linqvistik emalı və statistik təhlili üçün metodoloji əsas yaratmışdır.

Korpusun texniki baxımdan hazırlanması onun linqvistik emalının yalnız ilkin mərhələsini təşkil edir. Müəllif korpusunun elmi dəyəri mətnlərin dil vahidləri səviyyəsində sistemli işlənməsindən, müxtəlif qrammatik və leksik variantların vahid prinsiplər əsasında təsnif edilməsindən və statistik təhlilə hazırlanmasından asılıdır. Buna görə də Sabir Rüstəmxanlı korpusunun formalaşdırılmasında əsas metodoloji mərhələlərdən biri lemmalaşdırma prosesi olmuşdur. Məhz bu

mərhələdə müxtəlif qrammatik formalarda işlənən sözlər vahid lemma altında birləşdirilmiş və müəllifin leksik sisteminin obyektiv statistik modelinin formalaşdırılması üçün etibarlı əsas yaradılmışdır.

Lemmalaşdırma Sabir Rüstəmxanlı korpusunun formalaşdırılmasının əsas mərhələlərindən birini təşkil etmişdir. Bu prosesin məqsədi mətnlərdə müxtəlif qrammatik formalarda işlənən sözləri vahid lemma ətrafında birləşdirərək onların real işlənmə tezliyini müəyyənləşdirmək olmuşdur. Lemmalaşdırma zamanı Azərbaycan dilinin morfoloji quruluşu, söz yaradıcılığı xüsusiyyətləri və korpus dilçiliyində qəbul olunmuş ümumi metodoloji prinsiplər əsas götürülmüşdür. Bununla yanaşı, statistik nəticələrin etibarlılığını və sonrakı leksikoqrafik işlənməni təmin etmək məqsədilə bir sıra xüsusi metodoloji qərarlar qəbul edilmişdir. Lemmalaşdırma aşağıdakı əsas prinsiplər əsasında həyata keçirilmişdir.

Lemmalaşdırmanın ilk və əsas prinsiplərindən biri feil lemmalarının müəyyənləşdirilməsi ilə bağlı olmuşdur. Korpusun hazırlanması zamanı təsəvvürlənən bütün feil formaları vahid lemma altında birləşdirilmiş və əsas forma kimi məsdər qəbul edilmişdir. Bu yanaşma Azərbaycan leksikoqrafiya ənənəsinə, izahlı və orfoqrafiya lüğətlərində qəbul olunmuş təqdimetmə prinsipinə, eyni zamanda beynəlxalq korpus dilçiliyində tətbiq olunan lemmalaşdırma metodologiyasına uyğundur. Belə yanaşma feilin zaman, şəxs, kəmiyyət və şəkil kateqoriyalarına aid bütün formalarda işlənməsini vahid leksik vahid çərçivəsində birləşdirməyə və onun faktiki işlənmə tezliyini obyektiv şəkildə müəyyənləşdirməyə imkan verir.

Lemmalaşdırma zamanı xüsusi metodoloji qərarlardan biri feilin inkar formalarının emalı ilə bağlı olmuşdur. Azərbaycan dilində *-ma/-mə* şəkilçisi ilə yaranan inkar formaları müstəqil leksik vahid kimi deyil, müvafiq təsdiq formasının qrammatik variantı kimi qiymətləndirilmişdir. Bu səbəbdən *-mamaq/-məmək* formaları ayrıca lemma kimi qeydə alınmamış, müvafiq olaraq *-maq/-mək* forması altında birləşdirilmişdir. Belə yanaşma tezlik göstəricilərinin süni şəkildə parçalanmasının qarşısını alır, feilin real işlənmə intensivliyini əks etdirir və Azərbaycan dilinin qrammatik quruluşuna uyğun lemmalaşdırma yanaşmasını təmin edir. Bu prinsip korpusun sonrakı mərhələlərdə hazırlanmış tezlik lüğətində bütün feil lemmalarının vahid statistik model üzrə təqdim edilməsini təmin etmişdir.

Lemmalaşdırma zamanı feilin şəxssiz formalarının korpusda təqdim olunması yalnız morfoloji meyarlara deyil, onların leksikləşmə dərəcəsinə əsasən müəyyənləşdirilmişdir. Buna görə də feili sifət və feili bağlama formalarının korpusdakı statusu vahid şəkildə deyil, onların faktiki leksik funksiyası nəzərə alınmaqla müəyyən edilmişdir. Müasir Azərbaycan dilində müstəqil leksik vahid kimi işlənən, nominativ mənə qazanan və lüğət vahidi statusu daşıyan formalar ayrıca lemma kimi qəbul edilmiş, feil paradiqmasının funksional-qrammatik forması kimi işlənən vahidlər isə müvafiq feil lemması daxilində birləşdirilmişdir.

Bu prinsipə uyğun olaraq *yaradan, keçmiş, bölünən, gələcək, qanacaq, görməli, yeməli, çəkinmə* kimi leksikləşmiş formalar ayrıca lemma kimi qeydə alınmışdır.

Eyni zamanda *-a/-ə* şəkilçisi ilə yaranan və təkrar yolu ilə sabitləşmiş *deyə-deyə, öyə-öyə* tipli qoşa feili bağlama konstruksiyaları da ayrıca lemma kimi qəbul edilmişdir. Belə yanaşma morfoloji quruluşla yanaşı, sözün faktiki leksik davranışını və dil sistemində qazandığı funksional statusu da nəzərə almağa imkan vermişdir. Beləliklə, lemmalaşdırma zamanı əsas meyar morfoloji forma deyil, həmin formanın dil sistemində qazandığı leksik status olmuşdur. Bu baxımdan eyni morfoloji model əsasında yaranmış vahidlərin korpusdakı təqdimatı da eyni olmamışdır. Müstəqil nominativ mənə qazanan və lüğət vahidi kimi fəaliyyət göstərən formalar ayrıca lemma kimi qəbul edilmiş, yalnız qrammatik forma funksiyası daşıyan vahidlər isə müvafiq əsas lemma daxilində birləşdirilmişdir. Belə yanaşma lemmalaşdırmanın formal morfoloji təsnifat çərçivəsində məhdudlaşmasının qarşısını almış, müəllif dilinin leksik sisteminin daha obyektiv və funksional təsvirinə imkan yaratmışdır.

Lemmalaşdırma zamanı tətbiq edilmiş metodoloji prinsiplər ayrı-ayrı söz qrupları üzrə fərqli qərarların qəbul edilməsini tələb etmişdir. Bu qərarlar formal morfoloji təsnifata deyil, dil vahidlərinin leksik statusuna, funksional xüsusiyyətlərinə və korpusun statistik məqsədlərinə əsaslanmışdır. Yuxarıda şərh olunan metodoloji yanaşmaları daha sistemli şəkildə təqdim etmək məqsədilə lemmalaşdırmanın əsas prinsipləri aşağıdakı cədvəldə ümumiləşdirilmişdir (bax: Cədvəl 1).

Forma tipi	Korpusdakı status	Əsas meyar
təsriflənən feil	məsdər altında	morfoloji birlik
inkar forması (-mamaq/-məmək)	ayrı lemma deyil	qrammatik inkar
feili isim (-ma/-mə)	ayrı lemma	leksik vahid
leksikləşmiş feili sifət	ayrı lemma	nominativ mənə
leksikləşməmiş feili sifət	feil daxilində	qrammatik forma
leksikləşmiş feili bağlama	ayrı lemma	sabit leksik vahid
digər feili bağlamalar	feil daxilində	qrammatik forma

Cədvəl 1. Sabir Rüstəmxanlı korpusunda tətbiq edilmiş lemmalaşdırma prinsipləri

Cədvəldə təqdim olunan prinsiplər lemmalaşdırmanın yalnız texniki əməliyyat olmadığını, hər bir dil vahidinin korpusdakı statusunun Azərbaycan dilinin leksik, morfoloji və funksional xüsusiyyətləri nəzərə alınmaqla müəyyənləşdirildiyini göstərir. Bu baxımdan lemmalaşdırma prosesində qəbul edilmiş qərarlar korpusun daxili ardıcılığını təmin etməklə yanaşı, tezlik göstəricilərinin obyektivliyini və sonrakı leksikoqrafik təhlillərin etibarlılığını da şərtləndirmişdir. Nəticədə formalaşdırılmış lemmalaşdırma modeli müəllif dilinin statistik təsvirinin daha dəqiq qurulmasına, tezlik lüğətinin hazırlanması və konkordansların tərtibi üçün etibarlı metodoloji zəmin yaratmışdır.

Belə yanaşma korpusun bütün mərhələlərində eyni metodoloji ardıcılığın qorunmasına, müxtəlif leksik və qrammatik hadisələrin vahid meyarlarla qiymətləndirilməsinə imkan vermişdir. Nəticədə lemmalaşdırma prosesi ayrı-ayrı texniki əməliyyatların məcmusu kimi deyil, müəllif dilinin leksik sistemini mümkün qədər

obyektiv əks etdirməyə yönəlmiş kompleks metodoloji mərhələ kimi həyata keçirilmişdir. Məhz bu sistemlilik korpusun sonrakı statistik emalının, tezlik göstəricilərinin və müəllif dilinin leksik-statistik modelinin etibarlılığını təmin edən əsas amillərdən biri olmuşdur (bax: Cədvəl 2).

Emal mərhələsi	Göstərici	Say
İlkin emal	Söz-forma həcmi (token)	3 220 690
	Müxtəlif söz-forma	195 935
Yekun emal	Lemmalaşdırılmış korpus (token)	2 388 317
	Lemma sayı	23 433

Cədvəl 2. Sabir Rüstəmxanlı korpusunun əsas statistik göstəriciləri

Cədvəldə təqdim olunan statistik göstəricilər korpusun çoxmərhələli emal prosesinin nəticələrini əks etdirir. İlkin mərhələdə korpus tam söz-forma səviyyəsində emal edilmiş, bütün söz-formalar və onların işlənmə tezlikləri müəyyənləşdirilmişdir. Sonrakı mərhələlərdə aparılmış filtrləmə, normallaşdırma və lemmalaşdırma nəticəsində korpusun leksik tərkibi dəqiqləşdirilmiş, müəllif dilinin əsas leksik fondunu əks etdirən 23 433 lemma müəyyən edilmişdir. Beləliklə, yekun statistik göstəricilər yalnız korpusun həcmi deyil, eyni zamanda onun sistemli müəllif korpusu kimi formalaşdırılması prosesinin metodoloji nəticələrini də əks etdirir.

Statistik göstəricilər korpusun həcmi əks etdirməklə yanaşı, onun formalaşdırılması zamanı tətbiq edilmiş metodoloji prinsiplərin praktik nəticəsinə də nümayiş etdirir. Beləliklə, çoxmərhələli emal nəticəsində formalaşdırılmış korpus müəllif yaradıcılığının sistemli tədqiqi üçün etibarlı empirik baza yaradır. Bu xüsusiyyətlərinə görə Sabir Rüstəmxanlı korpusu Azərbaycan dilçiliyində sistemli müəllif korpusu nümunəsi kimi dəyərləndirilə bilər.

Sabir Rüstəmxanlı korpusunun ən mühüm xüsusiyyətlərindən biri onun yüksək leksik potensialıdır. Lemmalaşdırma nəticəsində müəyyən edilmiş 23 433 lemma müəllifin fərdi leksikonunun əsas strukturunu əks etdirir və onun dilinin statistik modelini qurmağa imkan verir. Bu leksik baza müəllif yaradıcılığında işlənən ümumişlək və fərdi leksikanın qarşılıqlı münasibətlərini, tematik söz qruplarının paylanmasını, söz yaradıcılığı prosesinin xüsusiyyətlərini, terminoloji qatın və milli-mədəni leksikanın funksional yükünü obyektiv şəkildə öyrənməyə şərait yaradır. Bu baxımdan korpus yalnız tezlik göstəricilərinin əldə edilməsi üçün deyil, həm də müəllifin leksik sisteminin müxtəlif istiqamətlərdən araşdırılması üçün zəngin empirik baza kimi çıxış edir.

Sabir Rüstəmxanlı korpusunun sistemli müəllif korpusu kimi mühüm üstünlüklərindən biri də onun statistik təhlil imkanları ilə bağlıdır. Korpusun vahid metodoloji prinsiplər əsasında formalaşdırılması müəllif dilinin tezlik strukturunu, leksik nüvəsini, müxtəlif tezlik zonalarını və sözlərin funksional paylanmasını obyektiv şəkildə müəyyənləşdirməyə imkan verir. Bu isə müəllif dilinin statistik modelinin qurulmasına, müxtəlif leksik təbəqələrin kəmiyyət göstəricilərinin mü-

qayisəsinə, eləcə də tezlik qanunauyğunluqlarının müəyyən edilməsinə şərait yaradır. Beləliklə, korpus yalnız elektron mətn bazası deyil, həm də müəllif dilinin kəmiyyət baxımından öyrənilməsi üçün etibarlı statistik resurs kimi çıxış edir.

Sabir Rüstəmxanlı korpusu yalnız konkret tezlik lüğətinin hazırlanmasına xidmət edən mətn bazası deyil, çoxşaxəli lingvistik tədqiqatların aparılması üçün geniş imkanlar yaradan sistemli müəllif korpusudur. Korpusun vahid metodoloji prinsiplər əsasında formalaşdırılması müəllif idiolektinin müxtəlif istiqamətlərdən kompleks şəkildə öyrənilməsinə şərait yaradır. Bu baxımdan korpus müəllifin fərdi leksikonunun strukturunun, açar sözlər sisteminin, leksik dominantlarının, tematik sahələrinin, emosional-ekspressiv leksikasının, milli-mədəni leksik qatının və söz yaradıcılığı xüsusiyyətlərinin korpus əsaslı təhlili üçün etibarlı empirik baza rolunu oynayır.

Bununla yanaşı, korpus müəllif üslubunun statistik və funksional-parametrik baxımdan araşdırılmasına, müxtəlif yaradıcılıq mərhələlərinin müqayisəsinə, janrlar üzrə leksik dəyişmələrin müəyyən edilməsinə, eləcə də müəllifin dil şəxsiyyətinin fərdi xüsusiyyətlərinin obyektiv meyarlar əsasında xarakterizə edilməsinə imkan verir. Bu isə ənənəvi filoloji təhlillə müqayisədə müəllif yaradıcılığının daha sistemli və kəmiyyət göstəricilərinə əsaslanan şəkildə öyrənilməsinə təmin edir. Bununla yanaşı, korpus müəllif dilinin müəyyən tarixi mərhələdəki vəziyyətini sistemli şəkildə sənədləşdirməklə gələcək diaxron tədqiqatlar üçün də mühüm empirik mənbə funksiyası daşıya bilər. Bu baxımdan korpus materialı yalnız Sabir Rüstəmxanlının fərdi dil xüsusiyyətlərinin deyil, onun yaradıcılığında əks olunan leksik və üslubi dəyişmələrin zaman dinamikasının izlənilməsinə də imkan verir.

Sabir Rüstəmxanlı korpusu eyni zamanda Azərbaycan dilində hazırlanacaq digər müəllif korpusları ilə müqayisəli tədqiqatların aparılması üçün də əlverişli zəmin yaradır. Vahid metodoloji prinsiplər əsasında formalaşdırılmış müəllif korpuslarının müqayisəsi müxtəlif yazıçıların leksik zənginliyinin, tezlik modelərinin, üslubi dominantlarının, söz yaradıcılığı xüsusiyyətlərinin və fərdi dil modellərinin obyektiv şəkildə qarşılaşdırılmasına imkan verə bilər. Bu istiqamətdə aparılacaq araşdırmalar Azərbaycan korpus stilistikası və müəllif dilinin statistik modelləşdirilməsi sahəsində yeni nəticələrin əldə olunmasına şərait yarada bilər. Eyni metodoloji prinsiplər əsasında yaradılmış müəllif korpuslarının məcmusu ayrı-ayrı idiolektlərin müqayisəsindən daha geniş perspektiv açaraq Azərbaycan ədəbi dilinin müxtəlif dövrlər, janrlar və fərdi yaradıcılıq təcrübələri üzrə dəyişmə və inkişaf meyillərinin empirik əsasda öyrənilməsinə imkan verə bilər.

Nəhayət, Sabir Rüstəmxanlı korpusunun formalaşdırılması zamanı tətbiq edilmiş metodoloji yanaşmalar Azərbaycan dilində sistemli müəllif korpuslarının yaradılması istiqamətində praktik model kimi qəbul edilməlidir. Müxtəlif müəlliflərin yaradıcılıq xüsusiyyətləri və tədqiqatın konkret məqsədləri nəzərə alınmaqla həmin təcrübənin digər müəllif korpuslarının hazırlanmasında da tətbiqinin və inkişaf etdirilməsinin faydalı ola biləcəyini düşünmək olar. Bu baxımdan Sabir Rüstəmxanlı korpusu yalnız konkret müəllif yaradıcılığının öyrənilməsinə xidmət

edən resurs deyil, həm də Azərbaycan korpus dilçiliyində sistemli müəllif korpuslarının formalaşdırılması istiqamətində nümunəvi elmi model kimi dəyərləndirilə bilər.

Aparılmış tədqiqatın nəticələri göstərir ki, Sabir Rüstəmxanlı korpusu Azərbaycan dilçiliyində sistemli müəllif korpusu kimi formalaşdırılmış ilk müəllif korpusudur. Müəllif yaradıcılığını reprezentativ şəkildə əhatə edən mətn bazasının formalaşdırılması, mətnlərin autentikliyinə qorunması, çoxmərhələli filtrləmə, ardıcıl lemmalaşdırma və statistik əməl vahid metodoloji sistem əsasında həyata keçirilmişdir. Bu yanaşma korpusun elmi etibarlılığını və funksional bütövlüyünü təmin etmişdir. Nəticədə formalaşdırılmış korpus müəllif dilinin leksik-statistik modelinin obyektiv şəkildə qurulmasına, onun fərdi leksikonunun, üslub xüsusiyyətlərinin və dil şəxsiyyətinin korpus əsaslı tədqiqinə geniş imkanlar açır. Eyni zamanda bu təcrübə Azərbaycan dilində yeni müəllif korpuslarının hazırlanması, tezlik lüğətlərinin tərtibi və korpus əsaslı linqvistik tədqiqatların inkişafı üçün etibarlı metodoloji və praktik baza yaradır.

ƏDƏBİYYAT

1. Azərbaycan dilinin tezlik lüğəti (Söz kökləri). I cild. / Tərtibçilər: M. Mahmudov, A. Fətullayev və b. – Bakı: Elm, 2010.
2. Baker P. Using Corpora in Discourse Analysis. – London: Continuum, 2006.
3. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge: Cambridge University Press, 1998.
4. Burrows J. F. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. // Literary and Linguistic Computing, 2002, Vol. 17, №3, pp. 267–287.
5. Təhirov İ. Sabir Rüstəmxanlı dilinin tezlik lüğəti. Bakı, “Elm və təhsil”, 2026. – 568 səh.
6. Mahmudov M. Kompüter dilçiliyi. – Bakı: Elm və təhsil, 2013. – 356 s.
7. Mahmudov M. Türk dillərinin milli korpusları. – Bakı: Elm və təhsil, 2018. – 392 s.
8. Mahmudov M., Təhirov İ. Sabir Rüstəmxanlının əsərlərinin 15 cildliyinin konkordansı. (Elektron resurs). – Bakı, 2025 // <https://korpus.azerbaycandili.az/concordance>.
9. Mark Twain Project Online. Edited by the Mark Twain Papers and Project, The Bancroft Library, University of California, Berkeley, and University of California Press, 2007–2024.
10. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012.
11. Məmmədli N., Mahmudov M., Təhirov İ. National Corpus and Lexical Infrastructure: The Experience of Building the Azerbaijani Language Lexical Database. // Türkologiya, 2025, №1, s. 32–45.
12. Open Source Shakespeare: An Experiment in Literary Computing // opensourceshakespeare.org;
13. Russian National Corpus. Authors' Corpus: A. S. Pushkin. // ruscorpora.ru.
14. Shakespeare's Words: A Glossary and Language Companion // shakespeareswords.com.

15. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.
16. Засорина, Л. Н. (ред.). Частотный словарь языка Пушкина. Москва: ИМЛИ РАН, 2003.
17. Словарь языка Достоевского: Идиоглоссарий. Рос. акад. наук, Ин-т рус. яз. им. В. В. Виноградова. Москва: Азбуковник, 2001.
18. Словарь языка Пушкина: в 4 т. Отв. ред. В. В. Виноградов. 2-е изд., доп. Москва: Азбуковник, 2000.

SYSTEMATIC AUTHOR CORPUS: PRINCIPLES FOR BUILDING THE SABIR RUSTAMKHANLI CORPUS

ABSTRACT

The article analyzes the principles for building a systematic author corpus using the Sabir Rustamkhanli Corpus as a case study. The aim of the research is to identify the defining characteristics of a systematic author corpus through an analysis of its construction principles, methodological foundations, statistical parameters, and scientific and applied potential. The study is based on the author's poetic, journalistic, prose, scholarly-publicistic, and socio-political works.

The corpus was developed according to a unified methodological framework, including representative text selection, text digitization, multi-stage filtering, normalization, and lemmatization. During the lemmatization process, the morphological system of the Azerbaijani language, word-formation patterns, and the lexical status of lexical units were adopted as the principal criteria. Particular attention was paid to the consistent processing of verb forms, non-finite verb forms, and hyphenated lexical units.

At the initial stage, a corpus comprising 3,220,690 word tokens and 195,935 distinct word forms was compiled. Following filtering and lemmatization, the final corpus contained 2,388,317 word tokens and 23,433 lemmas. The study demonstrates that the representativeness of the corpus, preservation of textual authenticity, linguistic processing based on unified methodological principles, and statistical modeling constitute the key factors ensuring its development as a systematic author corpus. The corpus provides a reliable empirical basis for investigating the author's idiolect, lexical-statistical model, stylistic features, keyword system, word-formation processes, and other corpus-based linguistic phenomena.

The scientific novelty of the study lies in substantiating the Sabir Rustamkhanli Corpus as the first comprehensive author corpus in Azerbaijani linguistics constructed according to the principles of a systematic author corpus. The proposed approach establishes a methodological framework for the development of systematic author corpora and corpus-based lexical-statistical research in the Azerbaijani language.

Keywords: *author corpus, corpus linguistics, Sabir Rustamkhanli, lemmatization, lexical-statistical model, frequency dictionary, corpus representativeness.*

СИСТЕМНЫЙ АВТОРСКИЙ КОРПУС: ПРИНЦИПЫ СОЗДАНИЯ КОРПУСА САБИРА РУСТАМХАНЛЫ

РЕЗЮМЕ

В статье анализируются принципы создания системного авторского корпуса на примере корпуса Сабира Рустамханлы. Цель исследования заключается в выявлении основных характеристик системного авторского корпуса посредством анализа принципов его формирования, методологических основ, статистических показателей и научно-прикладного потенциала. Материалом исследования послужили поэтические, публицистические, прозаические, научно-публицистические и общественно-политические произведения автора.

Создание корпуса осуществлялось на основе единой методологической системы, включающей репрезентативный отбор материала, оцифровку текстов, многоэтапную фильтрацию, нормализацию и лемматизацию. При лемматизации в качестве основных критериев были приняты морфологическая система азербайджанского языка, особенности словообразования и лексический статус слов. Особое внимание уделено последовательной обработке глагольных форм, неличных форм глагола и единиц, пишущихся через дефис.

На первоначальном этапе был сформирован корпус объёмом 3 220 690 словоупотреблений (токенов) и 195 935 различных словоформ. После проведения фильтрации и лемматизации итоговый корпус составил 2 388 317 словоупотреблений и 23 433 леммы. Установлено, что репрезентативность корпуса, сохранение аутентичности авторского текста, лингвистическая обработка на основе единых методологических принципов и статистическое моделирование являются основными факторами, обеспечивающими его формирование как системного авторского корпуса. Корпус представляет собой надёжную эмпирическую базу для исследования идиолекта автора, его лексико-статистической модели, стилистических особенностей, системы ключевых слов, процессов словообразования и других корпусно-ориентированных лингвистических исследований.

Научная новизна исследования заключается в обосновании корпуса Сабира Рустамханлы как первого комплексного авторского корпуса в азербайджанском языкознании, созданного на основе принципов системного авторского корпуса. Предлагаемый подход формирует методологическую основу для создания системных авторских корпусов и развития корпусно-ориентированных лексико-статистических исследований азербайджанского языка.

Ключевые слова: *авторский корпус, корпусная лингвистика, Сабир Рустамханлы, лемматизация, лексико-статистическая модель, частотный словарь, репрезентативность корпуса.*