

MÜASİR DİLÇİLİKDƏ KORPUS DİLÇİLİYİNİN YERİ VƏ ƏHƏMİYYƏTİ

Sahilə Mustafayeva

Azərbaycan Dillər Universiteti, Bakı, Azərbaycan

e-mail: sahilamustafa@yahoo.com

Xülasə. Məqalədə korpus dilçiliyinin etimologiyası və korpus dilçiliyi ilə bağlı müxtəlif alimlərin verdiyi təriflər təhlil olunur. Korpus dilçiliyinin müasir dilçilikdə yerinin əhəmiyyəti vurğulanır. D. Kristal, C. Aarts, D. Biber, N. Francis və digər alimlərin korpusla bağlı irəli sürdüyü müddəalar təhlil olunur və onların korpus dilçiliyindəki mövqeləri araşdırılır.

Açar sözlər: dilçilik, korpus, əhəmiyyət, texnologiya, kompüter.

THE PLACE AND THE IMPORTANCE OF CORPUS LINGUISTICS IN MODERN LINGUISTICS

Sahila Mustafayeva

Azerbaijan University of Languages, Baku, Azerbaijan

Abstract. The article analyzes the etymology of corpus linguistics and the definitions given by different scholars regarding corpus linguistics. The importance of the place of corpus linguistics in modern linguistics is emphasized. D. Kristal, J. Aarts, D. Biber, N. Francis and other scholars' propositions about the corpus are analyzed and their positions in corpus linguistics are examined.

Keywords: linguistics, corpus, importance, technology, computer.

МЕСТО И ЗНАЧЕНИЕ КОРПУСНОЙ ЛИНГВИСТИКИ В СОВРЕМЕННОЙ ЛИНГВИСТИКЕ

Сахила Мустафаева

Азербайджанский Университет Языков, Баку, Азербайджан

Резюме. В статье анализируется этимология корпусной лингвистики и определения корпусной лингвистики, данные разными учеными. Подчеркивается важность места корпусной лингвистики в современном языкознании. Анализируются положения Д. Кристалла, Дж. Аарта, Д. Бибера, Н. Фрэнсиса и других ученых о корпусе, а также рассматриваются их позиции в корпусной лингвистике.

Ключевые слова: лингвистика, корпус, важность, технология, компьютер.

1. Giriş

Hazırda korpus dilçiliyi geniş tədqiq olunan sahələrdən hesab olunur.

Bu məqalədə biz korpus dilçiliyini araşdıracağıq və “Korpus nədir?” sualına cavab verməyə çalışacağıq.

İlk əvvəl lüğəvi mənada korpus (ingilis dilində corpus) termininin mənasını araşdırdıq. İngilis dilinin Oksford lüğətində corpus sözünün izahı belə verilir: “Korpus - yazılı mətnlər toplusu, xüsusən də müəyyən bir müəllifin bütün əsərləri və ya müəyyən bir mövzuda yazılar toplusu” [9].

D. Kristal korpus dilçiliyini belə izah edir: “Korpus linqvistikası - yazılı mətnlər və ya qeydə alınmış nitqin yazılı forması olan linqvistik məlumatların toplusudur; o, (korpus) linqvistik təsvirin başlanğıc nöqtəsi və ya dil haqqında fərziyyələri yoxlamaq vasitəsi kimi istifadə edilə bilər”. Uzun müddət “korpus-əsaslı (məhdudiyyətli)” linqvistik təsvirlər, cisimlərin (obyektlərin) məhdudiyyətlərinə işarə edən generativ qrammatiklər tərəfindən tənqid

mövzusu olmuşdur. Onlar belə iddia edirdilər ki, həmin məlumatlar yalnız performans nümunələridir və hələ də korpusdan kənara çıxma vasitəsinə ehtiyac var. Yeni bir dil üzrə hər hansı bir sahədə və ya tarixi araşdırmada həmin dilin korpusundan kənara çıxmaq çox çətin ola bilər. Bununla belə, dilçilərin yerli dilə müntəzəm çıxışı olan dillərdə-natiqlər (və ana dilində danışanların özləri də ola bilər) onların yanaşması korpusla məhdudlaşdırılmış deyil, həmişə “korpus əsaslı” olacaqdır. Korpora (yaxud korpus) bir növ hesablama dilçiliyi üçün əsas yaradır. Kompüter korpusu maşınla oxuna bilən mətnlərin böyük bir toplusudur. 1980-ci illərdən bəri getdikcə daha böyük korpuslar (xüsusilə, ingilis dili) tərtib edilir və həm təbii dillərin işlənməsi proqram təminatının hazırlanmasında, həm də leksikoqrafiya, nitqin tanınması və maşın tərcüməsi kimi tətbiqlərdə istifadə olunur” [4, s.128].

Keçən əsrin ortalarından biz linqvistik tədqiqatın tendensiyasında nəzərə çarpacaq dəyişikliyi müşahidə edirik. Qeyd edək ki, empirik linqvistik tədqiqat və tətbiq prosesləri tədricən dünyanın demək olar ki, bütün linqvistik tədqiqat mərkəzlərində intuisiyaya əsaslanan dil tədqiqatı və tətbiqi yerini tutur.

2. Korpus dilçiliyini müəyyən edən amillər

Linqvistik tədqiqatlara yanaşmada bu cür dəyişikliyin səbəblərini müəyyən etmək asan deyil, çünki kompüter texnologiyasının yaranması və onun dilçilikdə istifadəsi çox sayda ənənəvi dilçilərin intuitiv tədqiqat sahəsindən bu sahəyə yönləndirilməsinin yeganə səbəbi deyil. Yeni nəsil alimlərə davamlı təsir göstərən bu dəyişikliyin arxasında bəzi başqa səbəblər vardı.

Dilçi alimlərin mentalitetindəki, eləcə də tədqiqat ssenarisindəki dəyişikliyin kökünü izləmək üçün aşağıdakı üç amilə əhəmiyyət lazımdır:

- 1) “İstifadə olunan dilin” xüsusiyyətlərinin müəyyən edilməsində ənənəvi nəzəriyyələrin, müşahidələrin və prinsiplərin məhdudlaşdırılması;
- 2) dil korpusunun “real həyatda dil istifadəsinin” ən etibarlı sübutu kimi təqdim edilməsi;
- 3) dilin təsviri, işlənməsi, təhlili və tətbiqi daxil olmaqla, müxtəlif linqvistik fəaliyyətlərdə dil korpusunun geniş istifadəsi [1, s.44].

Bundan əlavə, çoxlu sayda alimlərin korpus dilçiliyi arenasına cəlb edilməsində bəzi digər amillər də mühüm rol oynaya bilər. Məsələn:

- a) Dil korpusu alimlərə təbii dildən normal həyatda faktiki istifadəyə imkan verir;
- b) Korpus dilin əvvəlki dil tədqiqatlarında çatışmayan həqiqilik dərəcəsi ilə təhlili üçün kifayət qədər sübut təqdim edir;
- c) korpora alimlərə induktiv nəticə çıxarmaqla dilin hər hansı aspekti ilə bağlı qəti mövqeyə gəlməyə kömək edir və bu nəticə ilə çoxsaylı fərdi nümunələrin müşahidəsi üzrə yekun qərar verir.

Bu amillərə görə dilçilər daha çox dilin təhlili və təsviri üçün intuitiv şəkildə yığılmış sitat və nümunə dəstlərindən asılı olmağa hazır deyillər. Əksinə, onlar hər hansı müşahidə və ya iddianın yoxlanılması və təsdiqi üçün sərbəst şəkildə mövcud olan böyük həcmdə real həyat dil

məlumatlarını təhlil etməkdə maraqlıdılar. Bu, empirik dilçilik kimi tanınan dil öyrənilməsinə yanaşmada qlobal dəyişikliyə səbəb olmuşdur.

Korpusa əsaslanan dilin təhlili və təsviri dilçiliyin tamamilə yeni bir sahəsi deyil. Sadə dillə desək, bu, dil öyrənməyə yeni yanaşmadır. O, dilçiliyin bütün sahələri üçün nümunələr və dilçilik məlumatları təqdim edir.

Qeyd edək ki, korpus dilçiliyin bütün sahələrinə aid araşdırma, təsvir, tətbiq və təhlil üçün statistik nümunəli dil bazasıdır. Böyük struktura, müxtəlif tərkibə, nəhəng məlumatlara, təsdiq edilmiş istinad orijinallığına, geniş təmsilçiliyə, asan istifadəyə və dəqiq yoxlanıla bilməsinə görə korpus dilçiliyin bütün sahələrində əvəzsiz mənbəyə çevrilmişdir. Əvvəlcədən təklif olunan təriflər və izahatlar məntiqli və intuitiv dəlillər və arqumentlər faktiki istifadənin sübutları ilə bağlı etibarlıdırsa, o halda dilçiliyin istənilən sahəsində alimlər əvvəlki müddəaların və misalların real olub-olmadığını yoxlamaq üçün asanlıqla korpusa müraciət edə bilirlər.

3. Etimologiya

Etimoloji cəhətdən korpus termini latınca “bədən” mənasını verən *corpus* sözündən götürülmüşdür: *corpse* qədim fransız dilində *cors* kimi və *corps* XVIII əsr fransız dilində işlədilmişdi.

Birinci termin ilk dəfə ingilis dilinə XIII əsrdə *kors* kimi daxil olmuşdu. *Corpus* özü XIV əsrdə ingilis dilində işlədilməyə başlandı [2, s. 56].

Termin təsadüfi olaraq insan biliyinin digər sahələrindəki müxtəlif qeyri-lingvistik məlumat kolleksiyalarına və nümunələrə tətbiq olunsada, dilçilikdə və dillə əlaqəli elmlərdə (məsələn, fəlsəfə, psixologiya və s.) dil nümunələrinin böyük kolleksiyası kimi qəbul edilir.

Tarixdə belə fakt var ki, XVI əsrdə imperator Yustinian “Corpus Juris Civilis” (Body of Civil War) (hüquq elmində fundamental əsərlər toplusunun müasir adı) yaratmışdır və bu da “ilk Roma qanunlarının və hüquq prinsiplərinin işlərlə təsvir edilmiş və izahatla birləşdirilmiş məcmuəsindən ibarət idi, yəni həmin toplu qüvvəyə minəcək yeni qanunlar və gələcək qanunvericilik haqqında idi [6, s. 7]. Bununla belə, korpus termininin bu yaxınlarda istifadə edilən mənaya daha yaxın olan XVIII əsrin Latın Korpus Lüğəti hesab edilir. Həmin lüğətdə “əlifba sırası ilə düzülmüş Latın sözləri və ardınca daha asan Latın dilində olan sinonimləri və ya anqlo-saksondakı ekvivalentlər” yığılmışdır [5, s. 197].

Qrammatik təsvirdə ilk cəhdlərdən bəziləri ən azı qismən xüsusi dini və ya mədəni əhəmiyyətə malik olan cisimlərə əsaslanırdı. Məsələn, Pratişaxya ədəbiyyatı Sanskritin səs nümunələrini Vedalarda tapıldığı kimi təsvir etmişdi və Panininin klassik Sanskrit qrammatikası ən azı qismən həmin korpusun təhlilinə əsaslanırdı. Eynilə, ilk ərəb qrammatikləri Qurani Kərimin dilinə xüsusi diqqət yetirmişdilər. Qərbi Avropa ənənəsində alimlər Müqəddəs Kitabın və digər kanonik mətnlərin dilini ətraflı öyrənməyə imkan vermək üçün konkordanslar hazırlamışdılar.

Müasir korpus lingvistikasında əlamətdar hadisə 1967-ci ildə Müasir Amerika ingilis dilinin Hesablama Təhlili kitabının nəşri idi. H. Kuçera və U. Nelson Frensis tərəfindən

yazılmış əsər “Braun Korpusunun” (Brown Corpus) təhlilinə əsaslanırdı.

“Braun Korpusu nədir?” İngilis dilindən hərfi tərcümədə “Qəhvəyi Linqvistik” adlanan bu termini belə izah etmək olar. Braun (yaxud Qəhvəyi) linqvistik Müasir Amerika İngilis dilinin Brown Universitetinin Standart Korpusu (və ya sadəcə Qəhvəyi Korpus) müxtəlif janrların ilk əsas strukturlaşdırılmış korpusu olan Amerika İngilis dilinin mətn nümunələrinin elektron toplusunu əhatə edir. Bu korpus ilk növbədə gündəlik dil istifadəsində söz kateqoriyalarının tezliyi və paylanması elmi tədqiqinin əsasını da qoymuşdu. Dilçiliyin əsası Rod-Aylenddəki Brown Universitetində H. Kuçera və U. Nelson Francis tərəfindən tərtib edilmişdir və 1961-ci ildə Birləşmiş Ştatlarda nəşr olunan əsərlərdən ibarət yaradılmışdı. Toplu təxminən bir milyon sözdən ibarət 500 ingilis dili nümunəsini ehtiva edən ümumi dil korpusudur.

Müxtəlif mənbələrdən diqqətlə seçilmiş bir milyon amerikan ingilis dilindən olan sözləri seçən H. Kuçera və U. Frensis Qəhvəyi Korpusu müxtəlif hesablamalarla təhlil etdilər və sonra zəngin və rəngarəng bir əsər yaratmaq üçün dilçilik, dil tədrisi, psixologiya, statistika və sosiologiyanın elementlərini birləşdirdilər. Digər əsas nəşr 1960-cı ildə R. Kvirkin “İngilis İstifadəsinin təsvirinə doğru” idi və həmin əsərdə İngilis İstifadəsi Sorğusu korpus kimi təqdim edildi.

Qısa müddət sonra Boston nəşiri Houghton-Mifflin, korpus linqvistikasından istifadə edərək tərtib edilmiş ilk lüğət olan yeni Amerikan İrs Lüğəti (The American Heritage Dictionary of the English Language) üçün milyon sözdən ibarət üç sətirlik sitat bazası təqdim etmək üçün Kuçeraya müraciət etdi. AİL (AHD) təlimat elementlərini (dildən necə istifadə edilməlidir) təsviri məlumatla (əslində necə istifadə olunur) birləşdirən yenilikçi addım atdı.

Digər nəşriyyatlar da onun ardınca getdilər. İngilis nəşiri Kollinsin ingilis dilini xarici dil kimi öyrənən istifadəçilər üçün nəzərdə tutulmuş COBUILD monolinqual öyrənənlər lüğəti (Cobuild monolingual learner’s dictionary) Bank of English-dən istifadə etməklə tərtib edilmişdir. R. Kvirk və digərlərinin tərtib etdiyi “The Survey of English Usage Corpus” (İngilis dili istifadəsi korpusunun sorğu kitabı) tərtib edildi və o, ən mühüm Korpus əsaslı Qrammatikalardan birinin əsasını qoydu. Həmin sorğu 1985-ci ildə A Comprehensive Grammar of English Language kimi nəşr edilmişdir.

4. Mövzu ilə bağlı müzakirələr

Korpus dilçiliyində korpus xüsusi konnotativ məna daşıyır. D. Kristala görə, bu, “dil təsvirinin başlanğıc nöqtəsi və ya dil haqqında fərziyyələri yoxlamaq vasitəsi kimi istifadə edilə bilən yazılı mətnlər və ya qeydə alınmış nitqin transkripsiyası olan böyük bir linqvistik məlumat toplusuna” istinad edir. Başqa cür ifadə etsək, o, “həm yazılı, həm də şifahi formada olan dil mətnlərinin məcmusuna aiddir və insan qarşılıqlı əlaqəsinin hər bir sahəsində istifadə olunan dil növlərini təmsil edir. Maşınla oxuna bilən formada saxlanılaraq, hər cür linqvistik təsvir və təhlilə imkan verir” [3, s. 282].

Qeyd edək ki, C. Kennedi bu təriflə razılaşmır, çünki onun fikrincə, belə birölçülü tərif korpusun yaranması prosesində iştirak edən təzadları və çeşidləri təmsil etməyə bilər. Buna görə də, dilçiliyin indiki kontekstində korpus “müəyyən bir dilin, ləhcənin və ya dilin digər alt qrupunun nümayəndəsi olduğu güman edilən və linqvistik təhlil üçün istifadə ediləcək böyük mətnlər toplusu” mənasında istifadə edilməlidir [10, s. 97]. Baxmayaraq ki, yuxarıda qeyd olunan təriflər dil daxilində yaranan sosial-linqvistik komponentləri əhatə etməyə çalışsa da, bu, korpusun dizaynı üçün zəruri hesab edilən linqvistik meyarlara diqqəti yönəldə bilmir.

Bu ehtiyac, korpusun dil nümunəsi kimi istifadə edilmək üçün bəzi açıq linqvistik meyarlara əsasən, seçilən və sıralanan dil “parçalarının” toplusudur [13, s. 99]. Bu fikir C. Sinclair-ə aiddir. Korpus yaratmaq üçün istifadə edilən seçmə üsulları ilə bağlı suala görə, C. Sinclair (1991) “mətnlər” deyil, “parçalar” sözündən istifadə etməyə üstünlük verir. Onun qənaətinə, nümunələr eyni ölçüdədirsə, o zaman mətn deyil, “parçalar” terminindən istifadə etmək olar, yəni həmin *parçalar* məzmun mənbələrindən özbaşına ayrılmış mətn fraqmentləri kimi qəbul edilə bilər [12, s. 79].

Bəzən kolleksiya və arxiv kimi qəbul edilən sözlər, adətən, dil mətnləri toplusuna aiddir. Lakin onlar korpusdan ona görə fərqlənirlər ki, onların seçilməsinə və sifariş edilməsinə ehtiyac yoxdur. Bundan əlavə, daxil olan mətnlərin seçilməsi və sıralanması dil korpusunun dizaynı üçün təklif olunan xətt üzrə olmalıdır. Buna görə də bunlar olduqca çətin bir dil korpusudur. Mətn termini dil məlumatlarının toplusunu ehtiva etdiyi üçün korpusa münasibətdə də istinad edilir. Bu, sadəcə olaraq şifahi və yazılı ünsiyyətdə istifadə olunan çıxarışlara işarə edir.

Korpus, adətən, dilin müəyyən müxtəlifliyini və ya istifadəsini xarakterizə etmək üçün elmi şəkildə toplanmış və maşınla oxuna bilən formada təqdim olunan təbii olaraq yaranan dil mətnlərinin böyük toplusuna istinad edir [12, s.172]. O, dilin, adətən, çoxşaxəli istifadəsi ilə nümayiş etdirdiyi müxtəlifliyi əhatə etmək üçün müxtəlif linqvistik sahələr üzrə müxtəlif mətnlərdən tərtib edilmiş çox milyonlarla sözü ehtiva etmək üçün metodik olaraq hazırlanmışdır.

Korpus yazılı və ya şifahi formada istənilən mətnə istinad edə bilər. Mənşəyi ilə bağlı sənədləşdirilmiş dilin tərkib “parçalarını” ehtiva edən korpus açıq axtarış tapşırıqları üçün standart şəkildə kodlaşdırılır. C. Aarts, D. Biber və digər alimlər korpusu yuxarıda qeyd olunan eyni xətt üzrə müəyyən etmişlər. Bununla belə, bu təriflərin əksəriyyəti konsepsiyanın özünəməxsus fakturasını təfərrüatlarla izah edə bilmir. Əksinə, T. Makeneri və A. Uilson kimi alimlər korpusu korpusun xas xüsusiyyətləri ilə xarakterizə olunan daha incə təsnifat sxemi ilə təsnif etməyə üstünlük verirlər. Onların fikrincə, korpus aşağıdakı üç şeyə istinad edə bilər:

- a) sərbəst şəkildə, o, bir dilin və ya müxtəlifliyin hər hansı bir “mətn hissəsinə” istinad edə bilər;
- b) ən ümumi olaraq, bu, maşın tərəfindən oxuna bilən mətnin əsas hissəsinə aiddir;
- c) daha dəqiq desək, bu, bir dil və ya dil müxtəlifliyini maksimum dərəcədə təmsil etmək üçün seçilən, maşınla oxuna bilən mətnlərin məhdud kolleksiyasına aiddir [11, s. 105].

Korpus barədə S. Hunstonun tərəfi isə digər təriflərdən bir qədər fərqlidir. Hunstona görə, “Dilçilik həmişə bir neçə cümlədən tutmuş yazılı mətnlər və ya lent yazıları toplusuna

qədər hər hansı bir şeydən ibarət olan təbii olaraq meydana gələn dil nümunələrinin toplusunu təsvir etmək üçün “korpus” sözündən istifadə etmişdir. Korpus termini hər hansı bir məlumatın elektron şəkildə saxlanılan və əldə edilən mətnlər toplusu (və ya mətnlərin hissələri) üçün nəzərdə tutulmuşdur [8, s. 141].

Nəticə. Korpusda nəinki şifahi mətn nümunələri, həm də yazılı mətn nümunələri yer ala bilər. Korpusa balanslaşdırılmış və təmsilçi struktur vermək üçün praktiki olaraq mümkün olduğu qədər dil istifadəsinin bütün sahələrindən nümunələr tərtib edilməlidir.

Müasir kompüterin köməyi ilə böyük və çoxölçülü korpus hazırlamaq çətin deyil, baxmayaraq ki, bu, vaxt apara bilər. İnternetə qoşulmuş bir kompüter varsa, müxtəlif ana səhifələrdən, veb-saytlardan və digər saytlardan əldə edilən nümunələrlə istənilən növ yazılı mətnin böyük korpusunu asanlıqla tərtib etmək olar.

Ədəbiyyat

1. Aarts J. Intuition-based and observation-based grammars. In: Aijmer, K. & B. Altenberg (eds.) English Corpus Linguistics: Studies in Honour of J. Svartvik. London: Longman, 1991, pp. 44–62.
2. Biber, D., Conrad S., Reppen R. Corpus linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998, 256 p.
3. Crystal D. The Cambridge Encyclopaedia of the English Language. Cambridge: Cambridge University Press, 1995, 582 p.
4. Crystal D. A Dictionary of Linguistics and Phonetics. 4th Edn. Oxford: Blackwell Publishers, 1997, 628 p.
5. Dash N.S. Corpus Linguistics and Language Technology (With Reference to Indian Languages). New Delhi: Mittal Publications, 2005, 445 p.
6. Francis N. Problems of assembling and computerizing large corpora. In: Johansson, S. (ed.) Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities, 1982, pp. 7–24.
7. Francis N. Language corpora BC. In: Svartvik, J. (ed.) Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Berlin: Mouton de Gruyter, 1992, pp.17–32.
8. Hunston S. Corpora in Applied Linguistics. Cambridge: Cambridge UP, 2002, 241 p.
9. <https://www.oed.com/>
10. Kennedy G. An Introduction to Corpus Linguistics. London: Addison Wesley Longman Inc., 1998, 315 p.
11. McEnery T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996, 209 p.
12. Sinclair J. Corpus, Concordance, Collocation. Oxford: Oxford UP, 1991, 179p.
13. Sinclair J. The empty lexicon. International Journal of Corpus Linguistics. 1(1), 1996, pp. 99–120.