کارپس ژبپوهنه
Corpus Linguistics

Bushra Khatoon

Abstract

Corpus Linguistics is an offshoot of Linguistics which emerged in 1970s and gained momentum in the present century. Computers have dramatically changed every aspect of our life and same is the case with linguistics. Like other fields of life interfaces between linguistics and computer science are also growing and are likely to be of high interest to linguists and this interest is accelerating more and more in future. Nowadays computer corpora are used for a number of purposes such as dictionary writing, lexical analysis, translation of different languages, study of semantic changes, neologisms, and learning other languages. This paper is an attempt to introduce corpus linguistics, its scope and types to the readers so as to make them familiar with this branch of linguistics and to motivate them to collect computer corpora for Pashto language.

ژبپوهنه د یوې ژبې سائنسي مطالعې ته وئیلے شي ـ د ژبې نه مراد څۀ خصوصي ژبه نۀ ده بلکې په عمومي توګه د ژبې سائنسي مطالعه او څېړنه ـ ژبپوهنې اوس د یوې غټې ونې شکل اختیار کړ ے دے او ډېرې څانګې او څړګې ئې سپړدې دي چې یوه نوې څانګه په کښې د کارپس ژبپوهنې په نوم پېژندے شي ـ په دے مقاله کښې به د کارپس ژبپوهنې په اړه خبره کولے شي ـ

د کارپس ژبپوهنې تعریف مونږ داسې کولے شو:

کارپس ژبپوهنه د کمپیوټري کار پورا په مدد د ژبې تجزیه ده ـ دا تجزیه د کمپیوټر په مدد کولے شي ـ او مختلف قسمه خصوصي سافټ وئیرز په کښې استعمالېږي ـ[1]

۱ کارپس څۀ ته وائي

د انګريزۍ کارپس لفظ د لاطينۍ ژبې لفظ Corpus نه اخذ شوے دے چې مطلب ئې دے Body يعني بدن، جسم ـ لاطيني لفظ Corpus انګريزۍ ته د فرانسيسۍ دغاړې په دوه شکلونو کښې راغلے دے ـ

اولنے Corpse ـ دا د زړې فرانسيسۍ Cors نه جوړ شوے دے ـ

او دويم Corps دا لفظ په اتلسمه پېړۍ کښې د جديدې فرانسيسۍ په لار راغلے دے ـ

> دا اولنے لفظ Corps په انګريزۍ کښې په ديارلسمه پېړۍ کښې د Cors په شکل کښې راغے ـ په څوارلسمه پېړۍ کښې هغه خپل اصلي حرف P په کښې دوباره اولګيدو۔ اول اول کښې ددې مطلب صرف body وۀ او د څوارلسمې پېړۍ په اخره کښې دا صرف Dead body يعني مړ وجود د پاره په مکمله توګه استعمالېدل شروع شو۔ او بل طرف ته اصلي لفظ Corpus په خپله هم په څوارلسمه پېړۍ کښې انګريزۍ ته رانوتو۔ [2]

۲ د کارپس اصطلاحي معنې

Corpus واحد لفظ دے او ددې جمع Corpora يا Corpuses دے ـ په ژېپوهنه کښې مونږ د کارپس تعريف داسې کولے شو چې کارپس د يوې ژبې د داسې ټکړو مجموعې ته وئيلے شي چې د تحقيق د پاره په يوه منظمه طريقه چونه کړې شوې او راغونډې شوې وي او د يوې ژبې د نمونې په توګه استعماليدے شي بلکې په جديده ژېپوهنه کښې به مونږ د کارپس تعريف داسې کوؤ چې د نمونه جاتي وېنا وي يا ليکلو شوؤ متنونو مجموعه چې په مشيني توګه لوستے شي او د ګڼو ژبني معلوماتو د پاره ئې توضيع او تشريح کېدے شي ـ[3]

هسې کارپس هر يو متن ته وئيلے شي ـ ولې په اصطلاحي توګه هغه متن ته وئيلے شي چې په مشين باندې لوستے شي او په خصوصي توګه د متنونو د معلوم شمېر مجموعه چې په مشين لوستل کېدے شي ـ او په اُصولي توګه يو کارپس د صحيح ژبنو خصوصياتو د مطالعه کولو د پاره جوړولے شي ـ

په ژېپوهنه کښې د کارپورا نه مراد هم کمپيوټري کارپورا دے اګر چې د کارپس ژېپوهنې اصطلاح په ۱۹۸۰ء کښې استعمال شوه خو د کارپس په بنياد د ژبې مطالعه د کا مسکي نه اول شروع شوې وه ـ Field linguist باؤز (Baos) ۱۹۴۰ء کښې هم دا طريقه استعمال کړې

وه او د structuralism روایت سپیر، نیومین، بلوم فیلډ او د پائک هم دا طریقه په استعمال کښې وه هغه وخت به ژبپوهانو د کاغذونو په ټکړو د یوې ژبې ټکي لیکل او د پیزارونو په ډبو کښې به ئې جمع کول ـ وخت سره سره دا ټکي په کمپیوټرو کښې خوندي کېدل شروع شول او د مختلفو سافټ وئیرو په مدد د ژبې هر اړخیزه څېړنې شروع شوې ـ [4]

په ۱۹۶۰ء کښې دغه شان د لاس لیکلے او د ټائپ لیکلے کارپورا په استعمال کښې وے او د کمپیوټر لیکلے کارپورا رواج د ۱۹۷۰ء نه شروع شو په حقیقت کښې د جدیدې کارپورا غټه فائده په مشیني توګه د کارپورا لوستل دي ـ بل د کمپیوټر د ټولو نه غټه فائده دا ده چې ددې د سرچ، چوڼ کولو ترتیبولو او formatting سپیډ ډېر زیات دے ـ ددې خرچ ډېر کم دے ـ او انساني محنت په کښې هم ډېر کم لګي ـ د وخت ډېر غټ بچت دے او نتیجه ئې ډېره صحیح او قابل اعتباره وي ـ د کمپیوټري کارپورا ژبنی تجزیه نن سبا د کمپیوټري سافټ وئیرو په وجه ډېره اسانه شوې ده ـ تېرو شلو کالونو کښې د ژبې داسې جدیدې او باریکې تجزیې د کمپیوټري کارپورا نه بغېر ناممکنه وي کمپیوټري کارپورا په ژبه باندې څېړنو او تجزیو کښې ډېر غټ بدلون راوستے دے ـ په کارپس ژبپوهنه کښې د کمپیوټر د بې کچه استعمال د وجې دې ته نن سبا Computer Corpus linguistics هم وئیلے شي چې CCL ئې مخفف دے ـ

۳ دکارپس ژبیوهنې ساحه

د کارپس ژبپوهنې ساحه اوس ډېره پراخه شوې ده او د ژبپوهنې په ګڼو څانګو کښې استعمالېدے شي او د ژبپوهانو کار ئې ډېر اسان کړې دے ـ دنظریاتي او تطبیقي ژبپوهنې ماهران د کارپورانه د خپل خپل مطلب کار اخیستے شي ـ د ټولنیزې ژبپوهنې ماهر د یوې ژبې د مقامي ویونکو د خبرواترو اډیو ریکارډنګ د طبقاتي فرق او لهجو مطالعه کولو د پاره استعمالولے شي ـ د نفسیاتي ژبپوهنې ماهر هم دغه ریکارډنګ د ژبې د لغزشونو د مطالعه کولو د پاره استعمالولے شي او د لغت لیکونکي ئې د مختلفو فقرو د تکرارېدو یعني فریکوېنسي د پاره استعمالولے شي ـ ګرامر پوهان ئې د ژبې د ګرامري ساخت مطالعه کولو د پاره استعمالولے شي ـ دغه شان د ژبې اُستاذ او د عدالتي ژبپوهنې ماهر ئې د خپل مقصد د پاره استعمالولے شي ـ [5]

کارپورا د دۀو ژبو د ترجمو د مطالعې د پاره هم استعمالولے شي ـ د ۱۹۹۰ء نه پس د کارپس په بنیاد د دوۀ ژبو ډکشنرو ډېره زر وده اوموندہ ـ دغه شان د اسلوبیاتو په څېړنه کښې هم کارپورا ډېره ښه ذریعه ده ـ د یو ثقافت ځانګړي اصطلاحات هم د کارپورا په ذریعه مطالعه کولے شي ـ [6]

کله چې یوکارپس جوړولے شي ـ نو راغونډ شوي متنونه په سافټ Version کښې محفوظول، ترتیبول او ذخیره کول وي ـ د وېنا وي نمونو په نسبت لیکلي شوي متنونه محفوظول زیات اسان وي ـ د لیکلو شوۍ متنونو د پاره په عامه توګه اپټکل کرېکټر ریکګنیشن سافټ وئیرSoftware Optical Character Recognition ضروري دے ـ دا سافټ وئیر الیکټرانک متن فائیلو نه سکېن کوي ـ په داسې کارپورا کښې د ژوند د مختلفو شعبو نه متنونه شاملولے شي لکه د اخبارونو، رسالو، د سرکاري کاغذونو د تجارت د کاغذونو د خبرو اترونه نمونې اخستې شي ـ [7]

په یو کارپس کښې لاندیني خصوصیات په کار دي ـ

- په یو کارپس کښې د هغه ژبې عمومي او خصوصي ټول خصوصیات په کار دي د کومې نه چې دا راټول شوي او ډیزائن شوي وي ـ یو کارپس کښې د ژبې ټول اجزا لکه لفظونه، فقرې او جملې شاملې وي خو د یو کارپس سائیز د کارپس د لفظونو په شمېره انحصار کوي ـ
- په یو کارپس کښې د ژوند د هرې شعبې نمائنده ګي په کار ده ـ لکه د سائنس، طب، ارټس، انجنیرنګ او د ټولو فنونونه اصطلاحات او لفظونه په کارپس کښې راغونډول په کار دي ـ
- پکار ده چې یو کارپس د چاپ شوۍ ټولوذریعو نقل وي ـ په دې کښې د تلفظ، هجو، او املائي ټولو اُصولو خیال ساتل په کار دي ـ
- یو کارپس په سافټ فارم کښې محفوظول په کار دي چې د ژبې څېړن کاران دا ډېټا ببس د ژبې د تجزیې، د شماریاتي تجزیې او ترجمو ژباړو د پاره استعمالوي ـ [2]

٤ د کارپس ژبپوهنې یوه لنډه تاریخچه

کارپس جمع کول د ۱۹۶۰ء راهسې شروع شوي وو۔ په انګریزی ژبه کښې د تجرباتي ګرېمري څېړنو د پاره اولنے کارپس د Randolph Quirk Survey of English usage وۀ چې وروستو د ډیرو عشرو د پاره یو معیاري انګریزي ګرامر جوړ شو په ۱۹۸۵ء کښې دا Comprehensive Grammar of the English language په نوم چاپ شو او وروستو په ژبه کښې د تجرباتي څېړنو د پاره یو reference جوړ شو تر ډېره وخته دا کارپس کمپیوټري شوے نۀ وۀ ـ دا ډېر وروستو د ۱۹۸۰ء د عشرې په اخره کښې کمپیوټري شو۔ دا اوس د International Corpus

of English یعني ICE په نامه پېژندلے شي ۔ د Quirk سروې د وېناوي اولیکلې ژبې ګډون وۀ ۔ په یو ملین لفظونو کښې د وېناوي انګریزۍ پینځه لاکهه ټکي شامل وو ۔ [8]

د ۱۹۷۰ء اخره کښې Jan Svartvik د وېناوي انګریزۍ کارپس کمپیوټر کښې واچوۀ۔ چې د London Lund په نوم پېژندلے شي ۔ په دې کښې فونو لوجیکل او فونېټک معلومات هم ورکړي شوي دي ۔ په ۱۹۶۰ء کښې دویم کارپس پراجېکټ د براؤن کارپس دے چې د Rhode جزیرې په Providence ښار کښې د براؤن یونیورسټۍ پروفېسر نیلسن فرانسز او هینري کاؤسیرا compile کړو ۔ دا کارپس په یو ملین لفظونو مشتمل وۀ چې د امریکې پینځلس قسمونو پینځه سوه متنونو باندې مشتمل وۀ ۔ دا کارپس ډېر احتیاط سره په منظمه طریقه را ټول شوے وۀ او د پروف ریډنګ غلطۍ ترې هم ویستلې شوې وې ۔[9]

دې پسې بل کمپیوټري کارپس په یو ملین لفظونو مشتمل لینکاسټراوسلو۔ برګن کارپس Lancaster/oslo-Bergen corpus وۀ چې په LOB پېژندلے شي ۔ د ۱۹۶۵ء د کال چاپ شوي مواد د پینځه سوه متنونو نه دوه زره لفظونه راجمع کړي شوي وو۔ دا کارپس د Lancaster یونیورسټۍ Prof. Geoffrey Leech د نګرانۍ لاندې راټول شوے وۀ ۔ دا په اصل کښې د امریکې د براؤن کارپس په مقابله کښې د برطانوي انګریزۍ د پاره راټول شوے وۀ ۔ دا کارپس په دې وېب سائټ باندې سرچ کولے شي ۔

http:// khrit.hit.uib.no/icame/manuals/lob/index.htm[10]

دریم غټ کارپس پراجېکټ د انګریزۍ English Lexical Studies وۀ چې Edinburgh ښارکښې په ۱۹۶۳ء کښې شروع شو او په برمنګهم کښې ختم شو۔ ددې کارپس Principal Investigator ، John Sinclair وۀ ، دے اولنے ماهر وۀ چې کارپس ئې د Lexical Investigation د پاره استعمال کړو۔ دا پراجېکټ د وېناوي اولیکلي متن تقریباً په یو ملین لفظونو مشتمل وۀ د John Sinclair اخرنے رپورټ په ۱۹۷۰ء کښې د ټائپ لیکلے یو څو کاپۍ تقسیم شوې ۔[11]

د ډکشنرۍ جوړولو د پاره غټې غټې کارپورا هم جوړې شوې دي په ۱۹۸۰ء کښې Collins Publishers او برمنګهم یونیورسټۍ یو مشترکه پراجیکټ وۀ چې د پروفېسر John Sinclair د نګرانۍ لاندې جوړ شو چې د COBUILD په نامه پېژندے شي او اوس تقریباً پینځه سوه ملینه لفظونو باندې مشتمل دے او په http:www.mycobuild.com/about-collins-corpus.aspx وېب سائټ باندې سرچ کولے شي ۔

دغه شان اکسفورډ یونیورسټۍ پریس د خپلې ډکشنرۍ د مرتب کولو د پاره اکسفورډ انګلش کارپس جوړ کړے دے چې تقریباً په دوه بلینه لفظونو مشتمل دے ۔

د ۱۹۹۰ء په شروع کښې برټش نیشنل کارپس جوړ شو۔ دا په سلو ملینو لفظونو مشتمل وۀ چې نوي في صده ئې په لیکلي متن مشتمل دے او لس في صده ئې په وېناوي خبرواترو مشتمل دے ۔ دا د ژبې د ډېټا یوه ښه ذریعه ده چې په انټرنیټ په www.natcorp.ox.ac.uk وېب سائټ باندې سرچ کېدے شي ۔ [10]

۵ د کارپس قسمونه

د کارپس لاندیني قسمونه دي ۔

۱.۵ Reference Corpus

ریفرنس کارپورا د ګڼو مقصدونو د پاره استعمالولے شي ۔ په دې کښې د یوې ژبې معیاري لغات شامل وي ۔ دا کارپس د ژبپوهانو د پاره په معنو پوهېدو یوه اهمه ذریعه ده ۔ دې ته General corpora هم وئیلے شي ۔ لکه برټش نیشنل کارپورا BNC یا Bank of English.

۲.۵ Monitor Corpora

مانیټر کارپس یو داسې کارپس دے چې ژبه مانیټر کوي ۔ دا په ژبه کښې بدلونو نه مانیټر کوي ۔ دا وخت تروخته اپ ډېټ کولے شي ۔ دا کارپس د لغاتو په بدلون کښې دلچسپي لري ۔ لکه په ژبه کښې نوي لغات راننوتل او زاړۀ لغات نوې معنو کښې استعمالېدل یا د لفظونو، متلونو او اقوالو بدلون یا دوي کښې معنوي بدلون ۔ دا قسم کارپورا زیات تر د اخبارونو او رسالو په ژبه مشتمل وي ۔ د اخبارونو د ژبې خپل ځانګړے سټائل وي ۔ د اخبارونو نه علاوه په دې کار پورا کښې کال په کال د ادب، تاریخ او سائنس نوي چاپ شوي کتابونه هم شاملولے شي ۔ [12]

۳.۵ Parallel Corpora

Parallel کارپس ته ټرانسلېشن کارپس هم وئیلے شي ۔ دا کارپس د یوې ژبې په اصلي متنونو مشتمل وي چې هم ددې متن په بله ژبه کښې ترجمه هم وي ۔ کله کله دا ترجمه په ډېرو ژبو کښې هم وي ۔

٤.٥ Historical Corpora

تاريخي کارپورا په تاريخي موادو مشتمله وي ـ لکه دمثال په توګه Helsinki کارپس يا Archer ـ

٥.٥ Regional Corpora

دا د يوې علاقې د ژبې په لغاتو مشتمل وي ـ لکه WCNZE يعني Willington Corpus of written new Zealand English.

٦.٥ Learner Corpora

د يوې ژبې د زده کړې د پاره چې غېر مقامي خلقو ته کوم ضروري لغات زده کول په کار وي ـ دا قسم کارپورا په ټولو هغه لغاتو مشتمله وي لکه ICLE ، International Corpus of Learner English.

٧.٥ Multilingual Corpora

دې قسم کارپورا کښې د ګڼو ژبو او کم ازکم د دؤو ژبو لغات شامل وي ـ دې کښې اکثر د پرتليزه څېړنو د پاره يو متن په دوؤ يا ډېرو ژبو کښې ورکړے شوے وي ـ

٨.٥ Spoken Corpora

عام کارپورا کښې وېناوي او ليکلي متنونه دواړه شامل وي ـ خو په Spoken کارپورا کښې صرف وېناوي مواد شامل وي ـ لکه London-Lund Corpus of Spoken English

٩.٥ Internet Corpora

نن سبا په ژېپوهانو کښې دا Trend جوړ شوے دے چې انټرنېټ هم د کارپورا په توګه استعمالوي ـ دا په خصوصي توګه هغه وخت فائده مند وي چې کله مونږ د يو لفظ په حقله کتل غواړو چې ايا دا لفظ شته او کنه او په کومو کومو مطلبونو کښې استعمالېږي ـ

١٠.٥ Annotated Corpora

په Annotated کارپورا کښې څۀ ژبنی تجزيې د مخکښې نه شوې وي، لکه د لفظونو وېش يا د جملو تجزيه ـ

لنډه دا چې کارپس ژبپوهنې ډېره وده موندې ده او نن سبا په ژبپوهنه کښې په ګڼو اړخونو کښې استعمالېږي ـ په کارپس ژبپوهنه کښې نړیوالې تحقیقي مجلې چاپ کېدل هم شروع شوي دي لکه Corpus Linguistics & Linguistic theory International Journal of، Corpus Linguistics Corpora او ددې نه علاوه د کتابونو ډېر سیریز د کارپس ژبپوهنې اوځي چې په کښې د ژبې او کېموټر په تعلق بحث شوے وي لکه studies in corpus linguistics او English corpus linguistics ددے نه علاوه څۀ External links هم دې چې مختلف کارپورا پرې مطالعه کېدے شي، لکه Penn parsed corpora of historical English Linguistic Data consortium او corpora discussion list وغېره ـ

حوالې

[۱] http://www.as.uni.heidelberg.de/personen/nesselhauf/files/corpus-linguistcs, practical introduction. pdf (accessed June 2, 207)

[۲] http://www.eolss.net/sawple-chapter/co4/e6-91-17 PDF (accessed May 19, 017)

[۳] Tony MCEnery et all, corpus-Based Language Studies. An Advance Resource Book. (London & Newyork: Rotledge Taylor & Francis Group 2006) P. 4

[٤] Ibid- P-3

[٥] Michael stubbs, " Language Corpora" in The Handbook of Applied Linguistics, Eds: Alan Davies & Catherine Elder (Malden USA: Blackwell Publishers, 2004) P. 107

[٦] Ibid- P. 124

[۷] Randi Reppen & Rita Simpson-Vlach, "Corpus Linguistics" in An Introduction to Applied Linguistics (uk: Holder education An Hachettu Uk Company, 2011) P. 95

[۸] Wolfgang Teubert & Anna Cermakora, Corpus Linguistics A short Introduction (London: Continuum Publishers, Reprint 2008) P. 51

[۹] Ibid P- 52

[۱۰] Howard Jackson & Peter Stock well, An Introduction to the Nature & Function of Language (London & Newyork: Continuum International Publishing Group 20) P. 12

[۱۱] Wolfgang Teubert & Anna Cermakora, Corpus Linguistics A short Introduction (London: Continuum Publishers, Reprint 2008) P. 54

[۱۲] Ibid, P. 72