УДК 374

Jurayeva Maftuna Bahtiyor qizi

Namangan, Uzbekistan

WORK WITH ENGLISH CORPUS AND ANALYSIS OF THE INFORMATION PROVIDED BY THEM

Annotation: This article reflects the views of scientists involved in the development of corpus linguistics - the general principles and linguistic features of the corpus, information that the linguistic corpus was assembled in accordance with certain norms and standards and equipped with a special search engine.

Key words: corpus linguistics, word frequency, research, corpus linguistics, WordSmith 5.0 software.

INGLIZ TILI KORPUSLARI BILAN ISHLASH VA ULARNI TAQDIM ETADIGAN MA’LUMOTLARNI TAHLIL QILISH

Аннотация: Ushbu maqolada korpus lingvistikasi - korpusning umumiy printsiplari va til xususiyatlarini ishlab chiqish bilan shug’ullangan olimlar fikri, lisoniy korpus muayyan me’yorlar va standartlarga muvofiq to’plangan va maxsus qidiruv tizimi bilan ta’minlanganligi to’g’risida ma’lumotlar aks ettirilgan.

Kalit so’zlar: korpus lingvistikasi, so’z chastotasi, ilmiy tadqiqot, korpus tilshunoslik, WordSmith 5.0 dastur.

РАБОТА С АНГЛИЙСКИМ КОРПУС И АНАЛИЗ ПРЕДОСТАВЛЯЕМОЙ ОНИМИ ИНФОРМАЦИИ

Аннотация: В данной статье отражены взгляды ученых, причастных к развитию корпусной лингвистики - общие принципы и языковые особенности корпуса, сведения о том, что лингвистический корпус был собран в соответствии с определенными нормами и стандартами и снабжен специальной поисковой системой.

Ключевые слова: корпусная лингвистика, частотность слов, исследования, корпусная лингвистика, программное обеспечение WordSmith 5.0.

Korpusdagi eng tez-tez uchraydigan so’zlar ro’yxatini yaratish oson va tez amalga oshiriladi va natijalar har doim qiziqarli. Aksariyat dasturlar natijalarni alifbo tartibida yoki paydo bo’lish chastotasiga qarab ko’rish imkonini beradi.

Natijalarni ko’rishning qaysi usulini tanlamasligingizdan qat'i nazar, ro’yxatlar odatda har bir so’z shakli uchun misollar sonini va so’z shakli ifodalagan butun korpusning foizini ko’rsatadi. Korpus tilshunoslari "turlar" va "tokenlar" haqida gapirishga moyildirlar, bunda turlar soni korpusdagi alohida yoki noyob so’zlar sonini anglatadi va tokenlar soni korpusdagi so’zlarning umumiy sonini bildiradi. Bu yerda biz darhol turni tashkil etuvchi barcha biznesga kirishamiz. Misol uchun, ba'zi tadqiqotlar va ba'zi dasturiy ta'minot dasturlari bir xil, boshqalari esa ikkita "- is" + "n't" sifatida emas. Siz BNCda eng ko‘p uchraydigan 100 ta so‘zni sanab o‘tgan 1-jadvalda ko‘rishingiz mumkinki, dasturiy ta’minot alohida tur sifatida sanabgina qolmay, balki u “ ’s ” “(as in it’s and she’s)” va “ ’s ” “(as in Susan’s)”. Buning sababi, BNC nutqning bir qismi teglangan bo’lib, bu bobning boshqa joylarida muhokama qilinadi, bu dasturiy ta'minotga ularni alohida turlar sifatida hisoblash imkonini beradi. Boshqa joylarda muhokama qilingan lemmatizatsiya masalasi ham bor va so’z chastotasi ro’yxatida, masalan, birlik va ko’plik yoki fe'lning turli xil ta'sirlangan shakllari so’z chastotasi ro’yxatida yoki alohida-alohida ro’yxatga olinishi kerakmi? Ko’pgina so’z chastotasi ro’yxati so’z shakllarini alohida ro’yxatga oladi.

So’z chastotasi ro’yxati yaratilganda, turlar va belgilar soni ham ko’rsatiladi. Bundan tashqari, ba'zan ko’rsatiladi nisbat -type/token nisbati yoki TTR (ya'ni, number of types divided by the number of tokens) - bu korpuslarni solishtirishning bir usuli. TTR qanchalik baland bo’lsa, korpusdagi turlarning xilma-xilligi shunchalik ko’p bo’ladi, bu esa TTR past bo’lgan korpusdan farqli o’laroq, so’zlarning kamroq takrorlanishini bildiradi. Biroq, TTR nisbati muammoli bo’lishi mumkin, chunki korpus qanchalik katta bo’lsa, TTR shunchalik past bo’ladi, bu so’zlarning takrorlanishi ehtimoli ko’proq bo’ladi. Ushbu muammoni hal qilish yo’li korpusning bo’limlari uchun TTRni hisoblash orqali nisbatga keladigan standartlashtirilgan turdagi/token nisbatidan (STTR) foydalanishdir. Buni amalga oshirgandan so’ng, dasturiy ta'minot barcha bo’limlar bo’yicha o’rtacha hosil qiladi va bu o’rtacha STTR hisoblanadi.

Mayk Skottning (2008) taniqli WordSmith 5.0 dasturi har safar so’z chastotasi ro’yxatini yaratganda TTR va STTR ni ta'minlaydi. Ushbu nisbatlarga qo’shimcha ravishda, foydalanuvchiga o’rtacha so’z uzunligi, turli uzunlikdagi so’zlarning chastotasi, shuningdek, jumlalar, paragraflar, sarlavhalar, bo’limlar soni va so’zlarning o’rtacha soni kabi ko’plab boshqa statistik ma'lumotlar ham taqdim etiladi. har birida mavjud. WordSmith shuningdek, so’zning korpusda qayerda paydo bo’lishini ko’rsatadigan vizual dispersiya syujetini taqdim etishi mumkin. Bu foydalidir, chunki har doim tez-tez uchraydigan so’z butun korpusda topilmasligi xavfi mavjud va uning chastotasi uning ma'lum bir matn yoki janrda ko’p ishlatilishi natijasidir. Ushbu funktsiya bitta matnga qo’llanilganda ham foydalidir, chunki u matnning alohida qismlarida ma'lum so’zlar tez-tez uchraydigan hodisani “matn birikmasi(textual colligation)” (Hoey, 2005) ochib berishi mumkin. Misol uchun, Guardian gazetasi maqolalarini o’rganish (O'Donnell va boshqalar, 2008) kechagi voqealar deyarli har doim birinchi xatboshida bo’lishini aniqladi.

1-3-jadvallarda uchta turli korpusdagi eng tez-tez uchraydigan so’zlar ulardagi ma'lumotlarni tasvirlash uchun berilgan.

1-jadval ingliz tilining 100 million so’zdan iborat umumiy korpusi bo’lgan BNCdan, 2 va 3-jadvallar esa ixtisoslashgan korpuslardan, Gonkong moliyaviy xizmatlar korpusidan (HKFSC “the Hong Kong Financial Services Corpus” - 7,3 million so’z) va Gonkong muhandislik korpusidan (HKEC “the Hong Kong Engineering Corpus” - 9,2 million so’z).

Korpus tilshunosligida yangi bo’lganlar uchun har uch korpusda ham eng tez-tez uchraydigan so’zlar funktsiyali so’zlar yoki grammatik so’zlar ekanligi diqqatga sazovordir va bunday so’zlarning ingliz tilidagi hal qiluvchi rolini aniq ko’rsatadi. Bu, shuningdek, har bir ro‘yxatning tepasida joylashgan bir xil oltita so‘z bo‘lib, oxirgi uchtasining reytingi o‘zgarib turadi va bu yozma matnlardan tashkil topgan har qanday korpusga yoki asosan yozma matnlar (BNC 90% yozma matnlardan iborat va ikkita ixtisoslashgan korpus ikkalasi ham 90% dan

ortiq yozma matnlardan iborat). Ahmad (2005) BNCda eng ko’p uchraydigan so’zlarni o’rganib chiqdi va bu oltita asosiy so’z butun korpusning taxminan 20 foizini tashkil etishini aniqladi va bu HKFSC va HKEC uchun ham to’g'ri ekanligini ko’rish mumkin. Ahmad shuningdek, BNCdagi eng yaxshi ellikta so’zning barchasi funktsiyali so’zlar ekanligini va birgalikda korpusning 40 foizga yaqinini tashkil etishini aniqladi. Aynan shu erda biz BNC va ikkita ixtisoslashgan korpusdagi eng tez-tez uchraydigan so’zlar o’rtasidagi qiziqarli farqlarni topamiz.

Birinchisi ikkala ro‘yxatda ham kuchli yigirmatalikda, Gonkong 17 va 19-o‘rinlarni egalladi (here, we conflate Hong and Kong since it is, in essence, one item in these Hong Kong-based corpora, with fewer instances of Kong simply because Kong’s is counted separately). Faqat birinchi ellikta ichida HKFSCda yana o’n uchta leksik so’z va HKECda to’qqizta leksik so’z bor.

Bundan tashqari, leksik so’zlarning o’zi biz (masalan, company, financial, group, shares, million, limited, business, assets and investment) moliyaviy xizmatlar korpusi va (masalan, system, water, building, energy, construction, design, works, environmental and air) muhandislik korpusining ro’yxatini ko’rib chiqayotganimizni osonlik bilan aytadi.

Darhaqiqat, umumiy korpusda leksik so’zlarning juda kamligi va ularning o’ziga xosligi yo’qligi (masalan, said, time, now, just, people, know, very, new and way) bu ro’yxatning umumiy korpusdan ekanligini ko’rsatadi, holbuki eng yaxshi 100 ta so’z ro’yxatida leksik so’zlarning ko’pligi. ikkita ixtisoslashgan korpus bizga bu so’zlar aslida nima bo’lishidan qat'i nazar, ular haqiqatan ham ixtisoslashgan korpus ekanligini aytadi.

1-jadval

BNCda eng ko’p uchraydigan 100 ta so’z

№	*So’z*	*Chastota*	*%*	*№*	*So’z*	*Chastota*	*%*
1	the	6,184,700	6.29	51	can	235,400	0.24

2	of	2,939,100	2.99	52	her	218,300	0.22
3	and	2,681,700	2.73	53	said	208,700	0.21
4	a	2,162,600	2.20	54	who	205,500	0.21
5	in	1,821,400	1.85	55	one	196,200	0.20
6	to	1,628,400	1.66	56	so	189,300	0.19
7	it	1,087,500	1.11	57	up	179,500	0.18
8	is	998,200	1.01	58	as	177,400	0.18
9	to	934,300	0.95	59	them	173,300	0.18
10	was	923,600	0.94	60	some	171,200	0.17
11	I	887,500	0.90	61	when	171,200	0.17
12	for	841,200	0.86	62	could	168,300	0.17
13	that	730,800	0.74	63	him	164,900	0.17
14	you	695,400	0.71	64	into	163,400	0.17
15	he	681,000	0.69	65	its	163,200	0.17
16	be	664,400	0.68	66	then	159,500	0.16
17	with	657,500	0.67	67	two	156,100	0.16
18	on	647,500	0.66	68	out	154,200	0.16
19	by	509,600	0.52	69	time	154,200	0.16
20	at	479,000	0.49	70	my	152,500	0.16

2 - jadval

HKFSCda eng ko’p uchraydigan 100 ta so’z

№	*So’z*	*Chastota*	*%*	*№*	*So’z*	*Chastota*	*%*
1	the	484,439	7.32	51	been	13,205	0.20
2	of	297,781	4.50	52	total	12,790	0.19
3	and	203,418	3.07	53	management	12,743	0.19
4	to	175,283	2.65	54	fund	12,608	0.19
5	in	167,158	2.52	55	all	11,994	0.18
6	a	99,448	1.50	56	exchange	11,968	0.18
7	for	70,878	1.07	57	per	11,953	0.18
8	or	60,299	0.91	58	market	11,450	0.17

9	is	57,610	0.87	59	New	11,438	0.17
10	as	55,712	0.84	60	date	11,416	0.17
11	on	54,750	0.83	61	Net	11,353	0.17
12	by	50,530	0.76	62	directors	11,345	0.17
13	be	41,224	0.62	63	value	11,332	0.17
14	with	38,725	0.58	64	December	10,887	0.16
15	at	38,351	0.58	65	capital	10,861	0.16
16	are	37,198	0.56	66	property	10,656	0.16
17	Hong	37,171	0.56	67	were	10,541	0.16
18	Kong	36,184	0.55	68	I	10,490	0.16
19	that	33,563	0.51	69	US	10,380	0.16
20	from	31,166	0.47	70	services	10,322	0.16

3-jadval

HKECda eng ko’p uchraydigan 100 ta so’z

№	*So’z*	*Chastota*	*%*	*№*	*So’z*	*Chastota*	*%*
1	the	582,437	6.81	51	control	13,266	0.16
2	of	335,785	3.92	52	power	13,218	0.15
3	and	265,945	3.11	53	new	12,606	0.15
4	to	204,117	2.39	54	C	12,399	0.14
5	in	168,798	1.97	55	used	12,360	0.14
6	a	132,246	1.55	56	been	12,348	0.14
7	for	117,325	1.37	57	if	12,327	0.14
8	be	95,221	1.11	58	under	11,779	0.14
9	is	84,323	0.99	59	use	11,698	0.14
10	on	59,914	0.70	60	were	11,684	0.14
11	with	58,878	0.69	61	services	11,625	0.14
12	by	53,059	0.62	62	site	11,439	0.13
13	or	51,863	0.61	63	management	11,174	0.13
14	as	51,413	0.60	64	our	11,164	0.13
15	are	45,190	0.53	65	development	11,020	0.13

16	at	42,836	0.50	66	their	11,018	0.13
17	that	38,881	0.45	67	its	10,993	0.13
18	from	37,521	0.44	68	m	10,969	0.13
19	Hong	33,999	0.40	69	road	10,939	0.13
20	Kong	32,806	0.38	70	more	10,909	0.13

Bu farqning sababi nimada? Sinclair javob beradi. U ixtisoslashgan korpus uchun “tilning umumiy ko’rinishi uchun zarur bo’lganidan ko’ra, odatdagi tadqiqotlar uchun ancha kichikroq korpus kerak bo’ladi” degan qiziqarli kuzatishni amalga oshiradi. Buning sababi shundaki, ixtisoslashtirilgan korpuslarda umumiy korpusga qaraganda “har xil so’z shakllarining soni, ya`ni lug’at hajmining taxminiy bahosi”. Sinklerning ta'kidlashicha, 2 va 3-jadvallarda tasdiqlangan bu tendentsiya “kichik, ehtimol texnik lug’at” ni ta`kidlaydigan ixtisoslashgan korpus bilan bog’liq va shuning uchun “maxsus hududning xarakterli lug`ati chastotalar ro’yxatida sezilarli darajada namoyon bo’ladi”.

Tadqiqot jarayonida korpuslar tasnifini maxsus printsiplarga asoslangan holda keltirish, ularni ma’lum ma’noda qo’llash sohalariga ko’ra turkumlash, shartli ravishda, yakuniy shaklga keltirildi va mavjud terminlardan foydalanildi.

Korpus lingvistikasidan maksimal darajada foydalanish uchun dasturiy ta'minotni ishlab chiqishdagi o’zgarishlardan xabardor bo’lish va ular taqdim etayotgan ko’plab funktsiyalardan to’liq foydalana olishingizga ishonchingiz komil bo’lishi muhimdir. Bundan tashqari, foydalanuvchi dasturiy ta'minot tomonidan yaratilgan natijalarni tahlil qila olishi va ular uchun ishonchli tushuntirishni taklif qilishi muhimdir. Ushbu bobda so’z chastotasi ro’yxati, kalit so’zlar ro’yxati va muvofiqlik kabi asosiy qidiruvlar tushuntirilgan va tasvirlangan. Shuningdek, frazeologiyaga qiziqish ortib borayotgani va frazeologik variatsiyalarni kiritish muhimligi ta'kidlangan. Statistik ko’rsatkichlarning qisqacha sharhi ularning foydaliligini ta'kidladi va birgalikda tanlovlarning ahamiyatini aniqlashda ularning cheklovlarini muhokama qildi.

Foydalanilgan adabiyotlar ro'yxati

1. Ataboev N.B. ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis // International journal of TEST Engineering and Management. – India. Vol.81/ November-December, 2019. – P. 4170-4176.

2. Primov A. Tilining milliy korpusini yaratish korpusini yaratish muammolari // Tilshunoslikning dolzarb muammolari. Actual problems of linguistics. / DOI:10.13140/RG.2.2.31122.89280.

3. Mengliyev B. va Hamroyeva Sh. "Korpus lingvistikasi: korpus tuzish va undan foydalanish": Amaliy mashg`ulot uchun qo`llanma (Uzbek Edition) Paperback – September 5, 2020.

4. Кокорева А. А. Корпус параллельных текстов в обучении иностранному языку // Вестник Тамбовского госуниверситета. Гуманитарные науки. Педагогика и психология, 2013. - С. 57–62.

5. Мосина М. А. Интеграция современных образовательных педагогических и информационно-коммуникационных технологий в процессе лингвометодической подготовки будущего учителя иностранного языка // Фундаментальные исследования. 2013. — № 11–8. — С. 1699–1703.

6. Juraev, E. S. (2018). Foreign experience in conducting financial policies for the development of small business. Russia, Экономика и социум.

7. Xolmirzaev U. A., Juraev E. S. Problems of improvement of debitor debt debt analysis //Мировая наука. – 2020. – №. 1. – С. 100-105.

8. Razzakov S. J., Juraev B. G., Juraev E. S. Sustainability of walls of individual residential houses with a wooden frame //Structural Mechanics of Engineering Constructions and Buildings. – 2018. – Т. 14. – №. 5. – С. 427-435.

9. Mahlberg, Michaela. 2006. Lexical cohesion: Corpus linguistic theory and its application in English language teaching. International Journal of Corpus Linguistics (3): 363–383.