O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2026, [14]
ISSN 2181-7324(Print)

FILOLOGIYA
http://journals.nuu.uz
Social sciences
ISSN 3093-8996(Onlayn)

UDK: 811.111’373.46+512.133’373.46

Dilafro‘z XODJAYEVA,
DSc, dotsent, Buxoro davlat universiteti, Buxoro, O‘zbekiston
E-mail: d.i.xodjaeva@buxdu.uz, ORCID: 0000-0003-0099-6138

PhD M.Ahmedova taqrizi asosida

KORPUS TURLARI VA UNI TERMINOLOGIK TADQIQOTLARDA QO‘LLASH TAMOYILLARI

Annotatsiya

Mazkur maqolada lingvistik korpuslarning asosiy turlari, ularning tasniflanish mezonlari hamda terminologik tadqiqotlarda qo‘llanilish imkoniyatlari yoritilgan. Og‘zaki va yozma, umumiy va maxsus, annotatsiyalangan va annotatsiyalanmagan, shuningdek, parallel korpuslarning o‘ziga xos xususiyatlari tahlil qilingan. Korpus ma’lumotlarini qayta ishlashning muhim bosqichlari hisoblangan annotatsiyalash, lemmalash, tokenizatsiya va konkordanslar yaratish jarayonlarining lingvistik tahlildagi ahamiyati ochib berilgan. Shuningdek, korpus texnologiyalarining terminologik birliklarni aniqlash, ularning qo‘llanish chastotasi, semantik xususiyatlari va kontekstual funksiyalarini o‘rganishdagi imkoniyatlari ko‘rsatib berilgan. Korpus turini tadqiqot maqsadiga muvofiq tanlash hamda lingvistik ma’lumotlarning reprezentativligi va ishonchliligini ta’minlash terminologik tadqiqotlarning muhim metodologik tamoyillari sifatida asoslangan.

Kalit so‘zlar: korpus lingvistikasi, lingvistik korpus, korpus tipologiyasi, og‘zaki korpus, yozma korpus, umumiy korpus, maxsus korpus, parallel korpus, korpus annotatsiyasi, morfologik annotatsiya, sintaktik annotatsiya, semantik annotatsiya, lemmalash, tokenlash, konkordans.

TYPES OF LINGUISTIC CORPORA AND PRINCIPLES OF THEIR APPLICATION IN TERMINOLOGICAL RESEARCH

Annotation

This article examines the major types of linguistic corpora, their classification criteria, and their potential applications in terminological research. The distinctive features of spoken and written, general and specialized, annotated and unannotated, as well as parallel corpora are analyzed. Particular attention is given to the main procedures of corpus data processing, including annotation, lemmatization, tokenization, and concordance generation, highlighting their significance for linguistic analysis. The article also discusses the potential of corpus-based approaches for identifying terminological units, examining their frequency of occurrence, and investigating their semantic characteristics and contextual functions. The importance of selecting appropriate corpora according to research objectives and ensuring the representativeness and reliability of linguistic data is emphasized as a fundamental methodological principle of terminological research.

Keywords: corpus linguistics, linguistic corpus, corpus typology, spoken corpus, written corpus, general corpus, specialized corpus, parallel corpus, corpus annotation, morphological annotation, syntactic annotation, semantic annotation, lemmatization, tokenization, concordance.

ТИПЫ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ И ПРИНЦИПЫ ИХ ПРИМЕНЕНИЯ В ТЕРМИНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Аннотация

В статье рассматриваются основные типы лингвистических корпусов, критерии их классификации и возможности применения в терминологических исследованиях. Анализируются особенности устных и письменных, общих и специализированных, аннотированных и неаннотированных, а также параллельных корпусов. Особое внимание уделяется основным процессам обработки корпусных данных, включая аннотирование, лемматизацию, токенизацию и составление конкордансов, раскрывается их значение для лингвистичского анализа. Рассматриваются возможности корпусного подхода при выявлении терминологических единиц, определении частотности их употребления, изучении семантических характеристик и контекстуальных функций. Подчёркивается важность выбора корпуса в соответствии с целями исследования, а также обеспечения репрезентативности и достоверности языковых данных как основных методологических принципов терминологических исследований.

Ключевые слова: корпусная лингвистика, лингвистический корпус, типология корпусов, устный корпус, письменный корпус, общий корпус, специализированный корпус, параллельный корпус, корпусная аннотация, морфологическая разметка, синтаксическая разметка, семантическая разметка, лемматизация, токенизация, конкорданс.

Kirish. Linguistik tadqiqotlarda korpus – bu ma’lum bir tizim asosida to‘plangan matnlar yoki nutq yozuvlari to‘plami bo‘lib, u tilni qo‘llanishi, tuzilishi va o‘zgaruvchanligini o‘rganish uchun asosiy manba sifatida xizmat qiladi. Tadqiqotchilar til hodisalarini ularning tabiiy muhitida o‘rganishga intilar ekan, korpusning ahamiyati yanada ravshanlashadi. Korpuslar nafaqat real hayotdagi til ishlatilishini aks ettiradi, balki tilga oid xosliklar, til o‘zgarishi va kognitiv jarayonlar haqida muhim xulosalar chiqarishga yordam beradigan katta hajmdagi ma’lumotni ham taqdim etadi.

Korpuslar turli xususiyatlari, maqsad va ifodalovchi kontekstlariga qarab bir necha toifaga bo‘linadi. Eng asosiy toifalarga og‘zaki va yozma, annotatsiyalangan va annotatsiyalanmagan, sinxron va diaxron korpuslar kiradi. Og‘zaki korpuslar – bu nutq transkripsiyalari yoki yozuvlari, yozma korpuslar esa adabiyot, ommaviy axborot vositalari va boshqa yozma manbalarni o‘z ichiga oladi. Shu bilan birga, korpuslar maxsus yoki umumiy bo‘lishi ham mumkin. Masalan, maxsus korpuslar tibbiyot yoki huquqiy til kabi aniq bir sohani o‘rganishga qaratilgan bo‘lsa, umumiy korpuslar turli mavzu va janrlarni o‘z ichiga oladi. Bundan tashqari, o‘quvchilar korpusi kabi kichik toifalar ham mavjud bo‘lib, ular chet tili so‘zlovchilarining nutqini tahlil qilishga imkon beradi. Yoki parallel korpuslar yordamida bir nechta tillardagi tarjimalar taqqoslanadi. Ushbu korpus turlarining qo‘llanilish doirasi juda keng: leksikografiya, tilni diagnostik test qilish, sotsiolingvistik tadqiqotlar va hatto kompyuter lingvistikasi sohalarida ham foydalaniladi.

Akademik tadqiqotlarda muhokamalarni ilmiy dalillar bilan asoslash va manbalarga to‘g‘ri havola qilish juda muhim.

Ayniqsa, empirik ma’lumotlarga tayanadigan korpus lingvistikasi sohasida bu masala juda dolzarb hisoblanadi. Bu sohada korpusni yaratish va natijalarni tahlil qilishda qat’iy metodlarga rioya qilish talab etiladi. Ishonchli va nufuzli manbalarga tayanish tadqiqotchilarga oldingi ilmiy ishlardan foydalanishga, xulosalarini asoslashga va tadqiqot jarayonini shaffof qilishga imkon beradi. Aslida, korpus asosida chiqarilgan xulosalarning to‘g‘ri va ishonchli bo‘lishi ko‘p jihatdan o‘sha korpuslarning sifatiga bog‘liq. Shu sababli, tadqiqot olib borishda korpuslarning afzalliklari va cheklovlarini chuqur tushunish, ularning aniq tadqiqot savollariga mos kelishini aniqlash juda muhimdir.

Og‘zaki korpuslar – bu odamlarning tabiiy nutqiy kommunikatsiyalari yozma shaklidan iborat bo‘lib, ular tahlil qilish uchun odatda transkripsiya qilinadi. Ushbu turdagi korpuslar fonetik, fonologik va nutqni o‘rganish uchun muhim ahamiyatga ega, chunki ular real vaqt rejimida ishlatiladigan tilni aks ettiradi. Masalan, Switchboard korpusi telefon orqali muloqotlarni yozib olgan bo‘lsa, Cambridge va Birmingham Corpus of Spoken English turli hayotiy vaziyatlardagi og‘zaki suhbatlarni qamrab olad [1].

Yozma korpuslar esa kitoblar, maqolalar va internetdagi materiallar kabi yozma shaklda yaratilgan matnlardan iborat. Bunday korpuslar, ayniqsa, grammatik va semantik tahlillar uchun juda muhim bo‘lib, ular yozuv uslubi va matn tuzilishini saqlab qoladi. Masalan, British National Corpus (BNC) turli janrlarni o‘z ichiga olgan katta yozma korpus bo‘lib, lug‘at boyligi, grammatik tuzilmalar va matn izchilligi ustida tadqiqotlar o‘tkazishda keng qo‘llaniladi[2]. Yozma korpuslardan shuningdek, zamonaviy Natural Language Processing (NLP) dasturlarida til modellarini yaratishda ham faol foydalaniladi. Bu esa ularning ilm-fanda va texnologiyalarda ahamiyatini yanada oshiradi.

Agar korpuslar annotatsiyalanmagan bo‘lsa, ular oddiy matn shaklida, hech qanday o‘zgartirishsiz, qo‘shimcha ma’lumotlarsiz asl shaklida bo‘ladi.

Annotatsiyalangan korpuslar tilshunoslikdagi tadqiqotlar uchun juda foydali bo‘lib, ular asosiy matn yoki nutqdan tashqari qo‘shimcha ma’lumot qatlamlarini o‘z ichiga oladi. Bunday annotatsiyalar, masalan, so‘z turkumlarini belgilash (POS), sintaktik tahlil qilish, semantik rollarni aniqlash yoki diskurs markerlarini qo‘shishni o‘z ichiga oladi. Masalan, Penn Treebank kabi resurslar, sintaktik jihatdan annotatsiyalangan korpuslar sifatida, sintaktik parserlar uchun muhim asos bo‘lib xizmat qiladi[3]. Ushbu turdagi annotatsiyalar tadqiqotchilarga tilning turli jihatlarini yanada chuqurroq o‘rganish va murakkab bog‘liqliklarni aniqlashga imkon beradi.

Korpusni annotatsiyalash unga qo‘shimcha tilshunoslik ma’lumotlarini qo‘shish jarayonidir, ya’ni matndagi so‘zlarga tegishli ma’lumotlarni, masalan, ularning qaysi so‘z turkumiga tegishliligini, qanday sintaktik xususiyatlarga egaligini ko‘rsatadigan belgilashlarni qo‘shish demakdir. Masalan, bir xil yozilgan, lekin ma’nosi yoki talaffuzi turlicha bo‘lgan so‘zlarni aniqlashda so‘z turkumi bo‘yicha belgilash (POS – Part of speech – belsgilash) juda qo‘l keladi: agar matnda “present” so‘zi uchrasa, u bir nechta ma’noni anglatishi mumkin: ot sifatida – sovg‘a; fe’l sifatida – sovg‘a berish; sifat sifatida – hozirgi, mavjud. Bularning ma’nolari bir-biridan keskin farq qiladi. Bundan tashqari, talaffuzda ham o‘zgarish bor: masalan, fe’l bo‘lganda pre’sent so‘zidagi urg‘u oxirgi bo‘g‘inga tushadi.

Matnni annotatsiyalashning oddiy usullaridan biri bu so‘zlarga tegishli ma’lumotlarni pastki chiziq yordamida qo‘shib yozishdir. Masalan: present_NN1 (singular common noun), present_VVB (base form of a lexical verb), present_JA (general adjective).

So‘z turkumi bo‘yicha belgilashdan tashqari, korpuslarni annotatsiyalashning boshqa usullari ham bor. Ular tilni tahlil qilishning turli darajalariga mos keladi. Masalan:

Fonetik annotatsiya – bu og‘zaki korpusda so‘zlarning qanday talaffuz qilinganligini ko‘rsatadigan ma’lumotlarni qo‘shishni anglatadi.

Prosodik annotatsiya – og‘zaki nutq korpusida urg‘u, intonatsiya yoki pauzalar kabi prosodik xususiyatlar haqida ma’lumot qo‘shiladi.

Sintaktik annotatsiya – bu gapning tahlilini ibora va gap qismlari kabi birliklarga bo‘lib ko‘rsatish haqida ma’lumot kiritishdir.

Semantik annotatsiya – so‘zlarning ma’nosi bo‘yicha farqlarni ko‘rsatadigan ma’lumotlar qo‘shiladi. Masalan, funksiya so‘zi matematikada ham, tilshunoslikda ishlatilishi mumkin, lekin ularning ma’nosi turlicha bo‘ladi, garchi talaffuzi va yozilishi bir xil bo‘lsa ham.

Pragmatik annotatsiya – og‘zaki muloqotda ishlatiladigan iboralar yoki so‘zlarning nutq aktlari bo‘yicha qanday vazifani bajarishini ko‘rsatadi. Misol uchun, ingliz tilidagi okay so‘zi yoki o‘zbek tilidagi xo‘p so‘zi ba’zan tasdiq, ba’zan fikr bildirishni so‘rash yoki yangi mavzuni boshlash uchun ishlatilishi mumkin.

Diskurs annotatsiyasi – matndagi bog‘lanishlarni aniqlash uchun ishlatiladi. Misol uchun, “Men talabalarni baholab, ularning bahosini e’lon qilaman” gapida “ularning” olmoshi “talabalar” so‘ziga bog‘langan.

Stilistik annotatsiya - matnda nutq yoki fikr qanday ifoda etilganini ko‘rsatish uchun ma’lumot qo‘shishni anglatadi. Misol uchun, to‘g‘ridan-to‘g‘ri nutq, bilvosita nutq yoki erkin bilvosita fikr turlari ajratilishi mumkin.

Leksik annotatsiya – matndagi har bir so‘zning asosiy shaklini, ya’ni lug‘atdagi bosh shaklini belgilashni o‘z ichiga oladi. Masalan, teaching so‘zining asosiy shakli teach bo‘ladi.

Agar korpus oldindan annotatsiyalangan bo‘lsa, bu turli avtomatlashtirilgan jarayon va tahlillarda yordam beradi. Masalan, so‘z turkumi bo‘yicha belgilangan korpuslar yordamida avtomatik ravishda so‘z chastotasi ro‘yxatlarini yoki grammatik tasnifga ega lug‘atlarni tuzish mumkin. Bunday holatlarda plans (fe’l) va plans (ot) alohida-alohida so‘zlar sifatida ko‘rib chiqiladi va alohida sanaladi, chunki ko‘pchilik tahlillar uchun bu zarur hisoblanadi[4].

Shuningdek, avtomatik sintaktik tahlil qilish, ya’ni matn yoki korpusning tuzilishini tahlil qilish jarayonini ko‘rib chiqaylik. Oldindan belgilangan matn sintaktik tahlilning birinchi bosqichi bo‘lib, bu bosqichdan keyingi jarayonlarni ancha muvaffaqiyatli o‘tkazishga imkon beradi.

Bundan tashqari, nutq sintezi misolini olaylik. Masalan, agar biror matn “ovozli kitob” orqali ovozli o‘qiladigan bo‘lsa, sintezator conduct, project, survey kabi so‘zlarning qaysi so‘z turkumi sifatida ishlatilganini bilishi kerak: bu ot yoki fe’lmi? Chunki bu qo‘llanish talaffuzga ta’sir qiladi va matnni to‘g‘ri tushunish uchun muhim hisoblanadi.

Lemmalash – zamonaviy korpuslarda uchraydigan yana bir keng tarqalgan annotatsiya turi hisoblanadi. Lemmatlash jarayonida korpusdagi har bir so‘zga uning lemma, ya’ni asosiy shakli beriladi. Bu jarayon so‘zning asl grammatik shakllaridan, masalan, ko‘plik yoki fe’lning turli nisbatlaridan abstraktsiya hosil qiladi.

Lemma (yunoncha lĕmma —‘mavzu, bosh so‘z) lug‘at maqolasidagi bosh so‘zdir. Ushbu termin, shuningdek, lug‘at maqolasi deb ham talqin qilinishi mumkin, ya’ni:

– lug‘at maqolasining bosh so‘zi (lemma);

– lug‘at maqolasining to‘liq matni, ya’ni bir bosh so‘zdan boshlanib, keyingi bosh so‘zgacha davom etadigan qism.

Bu yerda biz lemma terminini birinchi ma’noda, ya’ni lug‘at bosh so‘zi sifatida ishlatamiz, maqola (entry) terminini esa ikkinchi ma’noda, ya’ni lug‘at maqolasining to‘liq matni sifatida tushunamiz. Shunday qilib, lug‘at maqolasi ikki asosiy qismdan iborat:

– lemma (bosh so‘z);

– lemma haqida ma’lumot (masalan, talaffuz, ma’no, qo‘llanilishi).

Demak, lemmalar ko‘pincha lug‘atlardagi bosh so‘zlarga o‘xshaydi va lemmatizatsiya qilingan korpuslar tadqiqotchilarga so‘zning barcha shakllarini izlashda osonlik yaratadi, chunki infleksiyalangan shakllarni alohida qidirish ko‘pincha qiyin bo‘lishi mumkin.

Lemmalash, ayniqsa, morfologiyasi yaxshi rivojlanagan til korpuslarni tahlil qilishda juda foydali. Bu tillarda bir ot yoki fe’lning o‘zi o‘nlab turli shakllarga ega bo‘lishi mumkin. Lemmalash bunday turli shakllarni bir joyga jamlashni osonlashtiradi va zarur shakllarni e’tibordan chetda qoldirish xavfini kamaytiradi. Ingliz tili grammatik jihatdan boy til

bo‘lmasa ham, masalan, bring, sing, drive, send va stand kabi fe’llarning turli shakllari mavjud (masalan, brings, brought, singing, drove, stood). Lemmalar asosida izlash bunday shakllarning barchasini bir joyda tez va oson topish imkonini beradi.

Lemmalash jarayonida ba’zi qiyinchiliklar ham uchraydi. Masalan, ilgari uchramagan yangi so‘zlarni tahlil qilish yoki bir xil ko‘rinishga ega bo‘lsa-da, turli ma’noga ega bo‘lgan so‘zlarni farqlash zarur bo‘ladi. Kontekstga qarab bunday so‘zlar turli asosiy shakllarga ega bo‘lishi mumkin.

Lemmalash jarayoni ikki bosqichdan iborat. Birinchi bosqichda, matndagi so‘z shakllari ularning asosiy shakllari bilan bog‘lanadi. Bu asosiy shakllar leksik birliklarni, ya’ni lemmalarni ifodalaydi. Ikkinchi bosqichda, aniqlangan asosiy shakllardan bazaga kiritiladiganlari tanlanadi.

Matnni lemmalash jarayoni odatda quyidagicha amalga oshiriladi:

Matnni so‘zlar ro‘yxatiga aylantirish (tokenizatsiya jarayoni).

Har bir so‘zning korpusdagi joylashuvini ko‘rsatadigan ro‘yxat, ya’ni konkordansini yaratish.

Ushbu ro‘yxat asosida so‘zlarni ularning lemmalari bilan bog‘lash.

Stemlash va lemmalash, ayniqsa, qidiruv tizimlari kabi axborotlarni izlash tizimlarida foydalidir, bunda foydalanuvchilar biror so‘z (masalan, teach) bilan so‘rov yuborishi mumkin, lekin ular so‘zning turli shakllari (masalan, teaches, teacher, teaching va boshqalar) keltirilgan natijalarni kutishadi. Stemlash va lemmatizatsiya, shuningdek, avtomatik o‘rganish algoritmlarida matnni qayta ishlash samaradorligini oshirishga ham muhim ahamiyatga ega.

Annotatsiyalash va lemmalash, agar sifatli va standartlarga muvofiq amalga oshirilsa, korpusni tadqiqot uchun ancha qulay va qimmatli manbaga aylantiradi. Ular qidiruv tizimlarini optimallashtirish, lingvistik tahlillarni chuqurlashtirish va turli ma’lumotlarni tizimlashtirish uchun asos bo‘lib xizmat qiladi.

Misol uchun, agar matnni lemmatizatsiya qilsak, ko‘plikdagi otlar birlik shaklida, fe’llar esa ega bilan kelishuvlarsiz asosiy shaklda ifodalanadi. Buni quyida misol orqali tahlil qilamiz. Quyida biz lemmalash jarayonidan o‘tmagan asl matn va lemmalanmagan matnni ko‘ramiz. Bunda asl so‘zlar o‘rniga ularning lemmalari keltirilgan bo‘ladi.

Lemmalash jarayonidan o‘tmagan asl matn: lingvistik korpus uchun semantik teglarni yaratish ma’lumotlarni izohlash va lingvistik tadqiqotlarda muhim qadamdir. Semantik teglar tilning turli semantik jihatlari va hodisalarini ajratib ko‘rsatib, matn ma’lumotlarini tasniflash va tavsiflashga yordam beradi.

Lemmlash jarayonidan o‘tgan matn: lingvistik korpus uchun semantik teg yaratish ma’lumot izohlash va lingvistik tadqiqot muhim qadam. Semantik teg til turli semantik jihat va hodisa ajrat ko‘rsat, matn ma’lumot tasnif va tavsif yordam ber.

Lemmalashdan oldin matnni tokenizatsiyalash zarur. Tokenizatsiya – bu matnni token deb ataluvchi kichik birliklarga bo‘lish jarayonidir. Tokenlar so‘zlar, iboralar yoki tinish belgilari bo‘lishi mumkin. Masalan, zamonaviy-kun iborasi alohida qismlarga – zamonaviy va kun – bo‘linishi mumkin, bu esa har bir qismni alohida tahlil qilish imkonini beradi[5]. Yuqoridagi matnning so‘z darajasida tokenizatsiyasini amalga oshiramiz, ya’ni berilgan matndagi so‘zlarni ajratamiz: “Lingvistik”, “korpus”, “uchun”, “semantik”, “teglarni”, “yaratish”, “ma’lumotlarni”, “izohlash”, va “lingvistik”, “tadqiqotlarda”, “muhim”, “qadamdir”, “Semantik”, “teglar”, “tilning”, “turli”, “semantik”, “jihatlari”, va “hodisalarini”, “ajratib”, “ko‘rsatib”, “matn”, “ma’lumotlarini”, “tasniflash” va “tavsiflashga”, “yordam”, “beradi”.

Maxsus va umumiy korpuslar haqida gap ketganda, har bir tur ma’lum tadqiqot maqsadlariga xizmat qilishini tushunish muhimdir. Umumiy korpuslar, masalan, British National Corpus (BNC), turli janrlar va uslublarni qamrab olgan matnlar va nutqlardan iborat bo‘lib, til o‘zlashuvi va o‘zgarishini o‘rganish uchun qulay vosita hisoblanadi[6]. Bunga qarama-qarshi ravishda, maxsus korpuslar ma’lum bir mavzu yoki soha bo‘yicha chuqurroq ma’lumot beradi. Masalan, Corpus of Academic English faqat akademik yozuvlarni o‘rganish uchun yaratilgan bo‘lib, bu sohadagi tilning o‘ziga xos xususiyatlarini o‘rganishga imkon beradi[7].

Maxsus korpuslarga yana bir misol sifatida suhbatlarga oid ma’lumotlar to‘plamini keltirish mumkin. Masalan, Santa Barbara Corpus of Spoken American English norasmiy suhbatlar va muloqotlarni qamrab oladi[8]. Ushbu korpus tabiiy muhitda tilni o‘rganish uchun muhim bo‘lib, sotsiolingvistik tahlillar, jumladan, ijtimoiy guruhlar, gender farqlari va kontekstga bog‘liq nutq strategiyalarini o‘rganishga imkon beradi.

Umuman olganda, korpuslarning xilma-xilligi tilshunoslik tadqiqotlarining qanchalik keng va qamrovli ekanligini ko‘rsatadi. Og‘zaki, yozma, annotatsiyalangan, umumiy yoki maxsus korpuslarning har biri o‘ziga xos tadqiqot ehtiyojlarini qondiradi. Masalan, tadqiqotning maqsadi va gipotezalari qanday bo‘lishiga qarab, mos korpusni tanlash kerak bo‘ladi. D.Biber va boshqalar bu haqda shunday degan: tadqiqotchilar o‘z maqsadlariga mos korpusni ehtiyotkorlik bilan tanlashlari kerak, chunki bu tanlov tadqiqot natijalariga va ularni talqin qilishga katta ta’sir qiladi[9].

Shu sababli, korpus turlari va tadqiqot maqsadlari o‘rtasidagi bog‘liqlikni yaxshi tushunish nafaqat tadqiqotlarning sifatini oshiradi, balki lingvistika sohasida chuqurroq bilim hosil qilishga ham yordam beradi.

ADABIYOTLAR

1. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge University Press, 1998. – P. 148.
2. Burnard L. Reference guide for the British National Corpus (XML edition). – Oxford University Computing Services, 2007. – P. 229.
3. Du Bois J. W., Chafe W. L., Meyer C., Thompson S. A. Santa Barbara Corpus of Spoken American English. – University of California, Santa Barbara, 2000.
4. Hyland K. Disciplinary Discourses: Social Interactions in Academic Writing. – University of Michigan Press, 2004. – P. 126.
5. Jenna Kanerva, Filip Ginter, Tapio Salakoski. Universal Lemmatizer: A Sequence to Sequence Model for Lemmatizing Universal Dependencies Treebanks // Natural Language Engineering 27 (5), 2020. – P. 27.
6. Leech Geoffrey. Developing Linguistic Corpora: a Guide to Good Practice. – Lancaster University, 2004. – P. 79.
7. Leech Geoffrey. Corpora and Theories of Linguistic Performance. // J. Svartvik (Ed.), Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1992. – P. 105-126.
8. Marcus M.P., Marcinkiewicz M.A., Santorini B. Building a Large Annotated Corpus of English: The Penn Treebank. – Computational Linguistics, 1993. – P. 313-330.
9. McEnery T., Wilson A. Corpus Linguistics: An Introduction. – Edinburgh University Press, 2001. – P. 175.