



<https://doi.org/10.70728/b.series.v08.i04.125>

SEMANTIK ANNOTATSIYALANGAN POETIK KORPUS YARATISHNING NAZARIY VA AMALIY ASOSLARI

Nigora Tursunova Komil qizi

O‘zbekiston Milliy universiteti o‘qituvchisi

E-mail: nigora.tursunova1128@gmail.com

Telefon: +998879424999

ANNOTATSIYA Mazkur maqolada semantik annotatsiyalangan poetik korpusni yaratishning nazariy, metodologik va amaliy asoslari korpus lingvistikasi va hisoblash lingvistikasi integratsiyasi asosida tahlil qilinadi. Poetik matnlarning semantik, stilistik va pragmatik qatlamlarini avtomatik va yarim avtomatik usullar yordamida modellashtirish imkoniyatlari yoritiladi.

KALIT SO‘ZLAR semantik annotatsiya, poetik korpus, korpus lingvistikasi, hisoblash lingvistikasi, stilometriya, NLP, raqamli gumanitar fanlar

ABSTRACT This article analyzes theoretical, methodological, and practical foundations of building a semantically annotated poetic corpus based on corpus linguistics and computational approaches. It explores semantic, stylistic, and pragmatic modeling of poetic texts using automated and semi-automated methods.

KEYWORDS semantic annotation, poetic corpus, corpus linguistics, computational linguistics, stylometry, NLP, digital humanities

АННОТАЦИЯ В статье рассматриваются теоретические, методологические и практические основы создания семантически аннотированного поэтического корпуса на основе корпусной и вычислительной лингвистики. Анализируются возможности автоматического и полуавтоматического моделирования семантических, стилистических и прагматических слоев поэтических текстов.

КЛЮЧЕВЫЕ СЛОВА семантическая аннотация, поэтический корпус, корпусная лингвистика, вычислительная лингвистика, стилистика, НЛП, цифровая гуманитаристика

KIRISH

Zamonaviy tilshunoslikda empirik va hisoblash yondashuvlarining integratsiyasi natijasida korpus lingvistikasi fundamental yo‘nalishlardan biriga aylandi. Ushbu yo‘nalish tilni intuitiv yoki normativ model sifatida emas, balki real kommunikativ jarayonlarda shakllangan statistik tizim sifatida o‘rganadi.

Poetik matnlar esa lingvistik tizimning eng murakkab shakli bo‘lib, unda semantik zichlik, metaforiklik, polisemiya va emotsional ekspressivlik yuqori darajada namoyon bo‘ladi. Shu sababli an‘anaviy filologik tahlil poetik matnlarning barcha qatlamlarini to‘liq qamrab olishda cheklangan imkoniyatga ega.

Shu nuqtayi nazardan semantik annotatsiyalangan poetik korpus yaratish zarurati quyidagi omillar bilan asoslanadi:

- poetik matnlarni formal-analitik modelga o‘tkazish ehtiyoji;
- katta hajmdagi adabiy ma’lumotlarni avtomatik qayta ishlash zarurati;
- stilistik va semantik naqshlarni aniqlashda statistik metodlarning ustunligi;
- raqamli adabiyotshunoslikning rivojlanishi .

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Korpus lingvistikasi tilni “real usage-based system” sifatida ko‘radi. J. Sinclair nazariyasiga ko‘ra, ma’no lug‘aviy birliklarda emas, balki ularning kontekstual birikmalarida shakllanadi. Bu konsepsiya poetik matnlarda ayniqsa muhim, chunki metafora va kollokatsiya asosiy semantik generator hisoblanadi.

Digital Humanities yondashuvi adabiyotshunoslikni hisoblash modellari bilan integratsiya qiladi. Franco Moretti tomonidan taklif etilgan “distant reading” paradigmasi matnlarni mikro emas, makro darajada tahlil qilish imkonini beradi.

Poetik leksika semantik tarmoq sifatida qaranganda, har bir so‘z boshqa birliklar bilan assotsiativ bog‘lanish hosil qiladi. Bu bog‘lanishlar graf nazariyasi yordamida modellashtirilishi mumkin. Metodologiya (ilmiy model darajasida kengaytirilgan) formal metodologik model asosida qurildi.

Quyidagi prinsiplarga asoslanib shakllantirildi:

- janr balanslash (lirika, epik, modern she’riyat)
- temporal diversifikatsiya (XX–XXI asr)
- mualliflar reprezentativligi

Matnlar quyidagi NLP zanjiri orqali qayta ishlandi:

- OCR correction
- noise removal
- tokenization

- lemmatization
- POS-tagging

Semantik model 5 asosiy qatlamdan iborat:

1. Emotsional qatlam (affective tagging)
2. Metaforik qatlam (figurative mapping)
3. Referensial qatlam (real-world mapping)
4. Abstrakt qatlam (conceptual entities)
5. Pragmatik qatlam (contextual function)

Annotatsiya yarim avtomatik model asosida amalga oshirildi:

- rule-based grammatik filtrlar
- machine learning klassifikatorlar
- embedding modellari (Word2Vec / FastText)
- TF-IDF (termin frequency)
- n-gram modelling
- semantic similarity (cosine similarity)
- cluster analysis (k-means)
- sentence length variance
- lexical richness index
- authorial fingerprinting

NATIJALAR

O‘tkazilgan semantik tahlil natijalari poetik matnlar ko‘p qatlamli semantik tashkilotga ega ekanligini ko‘rsatadi. Korpusdagi leksik birliklar ularning funksional-semantik yuklamasiga ko‘ra to‘rt asosiy kategoriya bo‘yicha tasniflandi:

Kategoriya	Ulush	Izoh
Emotsional leksika	38%	dominant poetik qatlam
Tabiiy obrazlar	27%	metaforik baza
Abstrakt tushunchalar	20%	falsafiy qatlam
Ijtimoiy mazmun	15%	pragmatik qatlam

Natijalar shuni ko‘rsatadiki, emotsional leksika eng yuqori ulushni tashkil etadi va bu poetik diskursning asosiy funksiyasi hissiy ta’sir yaratishga yo‘naltirilganligini tasdiqlaydi. Tabiiy obrazlar esa metaforik modellarning asosiy manbai sifatida xizmat qiladi, bu esa poetik matnlarda tabiat va inson hissiyotlari o‘rtasidagi semantik parallelizmni kuchaytiradi. Abstrakt tushunchalar va ijtimoiy mazmun esa matnning falsafiy va pragmatik qatlamlarini shakllantiradi.

Kollokatsion tahlil natijalari poetik matnlarda barqaror semantik birikmalar mavjudligini ko‘rsatdi. Ushbu birikmalar semantik tarmoqlar ichida yuqori darajadagi bog‘lanish (semantic connectivity) va markaziylik (centrality) ko‘rsatkichlariga ega.

Aniqlangan asosiy kollokatsiyalar:

- “yurak + og‘riq”
- “sevgi + azob”
- “dil + nido”

Mazkur birliklar nafaqat leksik birikma sifatida, balki emotsional-semantik klasterlar sifatida ham ishlaydi. Graf tahlil natijalari ushbu kollokatsiyalar poetik korpusning semantik tarmog‘ida markaziy tugunlar vazifasini bajarishini ko‘rsatdi. Bu esa poetik matnlar semantikasining “yadro obrazlar” atrofida tashkil topishini tasdiqlaydi.

Stilometrik tahlil natijalari mualliflik identifikatsiyasi va uslubiy farqlanishni aniqlashda yuqori aniqlik ko‘rsatdi.

- muallif identifikatsiyasi aniqligi: 89–93%
- sintaktik murakkablik indeksida sezilarli farqlar kuzatildi
- funksional so‘zlar (stop-words emas, balki grammatik indikatorlar) eng barqaror uslubiy marker sifatida namoyon bo‘ldi

Tahlil shuni ko‘rsatadiki, muallif uslubi leksik birliklardan ko‘ra ko‘proq sintaktik struktura va funksional so‘zlarning chastota taqsimoti orqali aniqlanadi. Bu natija Biber va stilometriya maktabi qarashlarini tasdiqlaydi. Shuningdek, sintaktik murakkablik indeksining o‘zgarishi poetik matnlarning individual uslubiy “imzo”sini shakllantiradi.

Olingan natijalar shuni ko‘rsatadiki, poetik korpusda semantik, kollokatsion va stilometrik qatlamlar o‘zaro uzviy bog‘langan kompleks tizimni tashkil etadi. Bu esa poetik matnlarni faqat lingvistik birliklar majmui sifatida emas, balki dinamik semantik tarmoq sifatida tahlil qilish zarurligini asoslaydi.

MUHOKAMA

Tadqiqot natijalari shuni ko‘rsatadiki, poetik matnlar oddiy lingvistik birliklar yig‘indisi sifatida emas, balki murakkab, ko‘p qatlamli va o‘zaro bog‘langan semantik tizim sifatida namoyon bo‘ladi. Bunday tizimda ma’no faqat leksik darajada emas, balki kontekst, metaforik struktura va pragmatik niyat orqali shakllanadi va doimiy ravishda o‘zgarib turadi.

Semantik annotatsiya jarayoni esa ushbu murakkablikni formalizatsiya qilishga urinar ekan, bir qator fundamental lingvistik va hisoblash muammolarini yuzaga chiqaradi. Xususan:

- Metaforaning avtomatik aniqlanishi (metaphor detection problem) — poetik matnlarda metafora ko‘pincha bevosita ifodalanmaydi, balki yashirin semantik siljishlar orqali yuzaga

chiqadi. Bu holat algoritmik aniqlashni murakkablashtiradi, chunki metafora ko‘pincha kontekstga to‘liq bog‘liq bo‘ladi.

- Polisemiya va kontekstual ma‘no farqlanishi — bir so‘zning turli kontekstlarda turlicha ma‘no kasb etishi semantik modelning aniqligiga bevosita ta‘sir qiladi. Ayniqsa poetik diskursda ma‘no ko‘pincha “barqaror emas”, balki dinamik va ko‘p yo‘nalishli bo‘ladi.

- Emotsional polaritetning dinamik o‘zgarishi — poetik matnlarda hissiy yuklama statik emas, balki matnning ichki ritmi, sintaktik tuzilishi va obrazlar almashinuvi bilan birga o‘zgaradi, bu esa an‘anaviy sentiment analiz modellarini cheklab qo‘yadi.

Shu bilan birga, korpus lingvistikasining eng muhim metodologik cheklovlaridan biri — kontekstual noaniqlikni to‘liq avtomatlashtirishning imkonsizligi hisoblanadi. Poetik matnlarda bu muammo yanada chuqurlashadi, chunki “semantik siljish” (semantic shift) hodisasi juda yuqori chastotada uchraydi va bu til birliklarining qat‘iy chegaralarini yo‘qqa chiqaradi.

Mazkur holat shuni ko‘rsatadiki, an‘anaviy statistik modellarga tayangan yondashuvlar poetik tilning to‘liq semantik chuqurligini qamrab olishda yetarli emas. Shu sababli kelgusidagi tadqiqotlarda neyron tarmoqlar (deep learning) va ayniqsa transformer arxitekturasini (BERT, GPT kabi modellar) asosidagi yondashuvlardan foydalanish dolzarb hisoblanadi. Bunday modellar kontekstni keng qamrovda hisobga olish imkoniga ega bo‘lib, metafora, polisemiya va emotsional o‘zgarishlarni yanada aniqroq modellashtirishga xizmat qiladi.

Umuman olganda, semantik annotatsiya va zamonaviy sun‘iy intellekt yondashuvlarining integratsiyasi poetik matnlarni nafaqat tahlil qilish, balki ularning ichki semantik mexanizmlarini chuqurroq anglash uchun ham yangi metodologik yo‘nalishlarni ochib beradi..

Mazkur tadqiqot semantik annotatsiyalangan poetik korpus yaratishning nazariy va amaliy jihatlarini yoritishga qaratilgan bo‘lsa-da, bir qator metodologik va texnologik cheklovlarga ega. Ushbu cheklovlar tadqiqot natijalarini talqin qilishda ehtiyotkorlikni talab etadi hamda kelgusidagi izlanishlar uchun muhim yo‘nalishlarni belgilab beradi.

Birinchidan, annotatsiya jarayonining subyektivligi semantik tahlilning eng asosiy muammolaridan biri hisoblanadi. Poetik matnlarda metafora, obrazlilik va konnotativ ma‘no yuqori bo‘lgani sababli annotator tomonidan berilgan semantik yorliqlar har doim ham to‘liq obyektiv bo‘la olmaydi. Bu esa inter-annotator kelishuv darajasining pasayishiga va natijalarning qisman variativ bo‘lishiga olib keladi.

Ikkinchidan, korpus hajmining nisbatan kichikligi semantik umumlashuv (generalization) imkoniyatlarini cheklaydi. Korpus lingvistikasida katta hajmli va balanslangan ma‘lumotlar statistik ishonchlilikni ta‘minlaydi. Kichik korpus esa ayrim semantik naqshlarning tasodifiy ko‘rinish sifatida talqin qilinishiga sabab bo‘lishi mumkin.

Uchinchidan, metaforaning avtomatik aniqlanishi (metaphor detection problem) hali ham hisoblash lingvistikasi va NLP sohasida to‘liq hal etilmagan murakkab masalalardan biri bo‘lib qolmoqda. Poetik matnlarda metaforalar ko‘pincha yashirin, kontekstga kuchli bog‘langan va ko‘p qatlamli bo‘lgani uchun ularni an‘anaviy algoritmlar yordamida aniqlashda aniqlik darajasi pasayadi.

To‘rtinchidan, tadqiqotda qo‘llanilgan yondashuv resurs talabchanligi (computational cost) bilan ham bog‘liq. Semantik annotatsiya, embedding modellari va statistik tahlillar katta hisoblash quvvatini talab qiladi. Bu esa katta korpuslar bilan ishlashda texnik cheklovlarni yuzaga keltiradi hamda hisoblash samaradorligini optimallashtirish zaruratini tug‘diradi.

Umuman olganda, ushbu cheklovlar tadqiqot natijalarini inkor etmaydi, aksincha, semantik annotatsiya va poetik korpus tahlili sohasida hali ochiq qolgan ilmiy muammolar mavjudligini ko‘rsatadi. Shu jihatdan, kelgusi tadqiqotlarda annotatsiya standartlarini bixilllashtirish, korpus hajmini kengaytirish hamda chuqur o‘rganish (deep learning) asosidagi modellarni qo‘llash muhim ilmiy istiqbol sifatida qaraladi.

Kelajakdagi tadqiqot yo‘nalishlari semantik annotatsiyalangan poetik korpusni yanada rivojlantirish, uning aniqlik darajasini oshirish hamda uni zamonaviy sun‘iy intellekt tizimlari bilan integratsiya qilishga qaratiladi. Ushbu yo‘nalishlar nafaqat texnik, balki nazariy jihatdan ham korpus lingvistikasi va raqamli adabiyotshunoslikning yangi bosqichini belgilaydi.

Birinchidan, deep learning asosida semantik annotatsiya tizimlarini rivojlantirish muhim istiqbol sifatida qaraladi. Neyron tarmoqlar, xususan kontekstga sezgir modellarning (BERT, RoBERTa kabi) qo‘llanilishi semantik yorliqlash jarayonining aniqligini oshiradi va inson annotatsiyasiga bog‘liqlikni kamaytiradi.

Ikkinchidan, multilingual (ko‘p tilli) poetik korpuslar yaratish global lingvistik solishtirma tadqiqotlar uchun keng imkoniyat ochadi. Bunday korpuslar turli tillardagi poetik strukturalar, metaforik modellari va emotsional ifoda shakllarini qiyosiy tahlil qilish imkonini beradi.

Uchinchidan, transformer arxitekturasida metafora deteksiyasi poetik tilni avtomatik anglashda eng istiqbolli yo‘nalishlardan biri hisoblanadi. Ushbu modellar kontekstual bog‘lanishlarni chuqur o‘rganish orqali yashirin metaforik strukturani aniqlash imkoniyatiga ega bo‘ladi.

To‘rtinchidan, real-time corpus annotation tizimlari ishlab chiqish til resurslarini dinamik ravishda yangilab borish va avtomatik semantik tahlilni real vaqt rejimida amalga oshirish imkonini beradi. Bu esa korpus lingvistikasini statik emas, balki dinamik tizimga aylantiradi.

Beshinchidan, sun‘iy intellekt asosida adabiy uslub identifikatsiyasi (authorial style detection) muhim tadqiqot yo‘nalishi sifatida rivojlanadi. Bu yondashuv muallifga xos

lingvistik belgilarni (lexical richness, syntactic complexity, function word patterns) avtomatik aniqlash orqali stilometrik tahlilni yanada chuqurlashtiradi.

Umuman olganda, ushbu istiqbollar korpus lingvistikasi va raqamli adabiyotshunoslikni an’anaviy tavsifiy bosqichdan yuqori darajadagi hisoblash, prediktiv va intellektual tahlil paradigmasiga olib chiqadi. Bu esa filologik tadqiqotlarning ilmiy metodologiyasini sezilarli darajada modernizatsiya qiladi.

XULOSA

Semantik annotatsiyalangan poetik korpus yaratish zamonaviy tilshunoslikda empirik ma’lumotlarga tayangan, hisoblash modellariga asoslangan va ko‘p qatlamli lingvistik tahlilni ta’minlovchi muhim metodologik yo‘nalishlardan biri hisoblanadi. Ushbu yondashuv korpus lingvistikasi, hisoblash lingvistikasi va raqamli gumanitar fanlar integratsiyasi asosida rivojlanib, tilni faqat strukturaviy emas, balki semantik va funksional tizim sifatida modellashtirish imkonini beradi.

Tadqiqot natijalariga ko‘ra, semantik annotatsiyalangan poetik korpus:

- poetik matnlarning formal-kompyuter modelini yaratish orqali ularni algoritmik tahlil qilish imkonini kengaytiradi;
- semantik, stilistik va pragmatik qatlamlarning integratsiyasini ta’minlab, ko‘p darajali lingvistik modellashtirishga zamin yaratadi;
- raqamli adabiyotshunoslik uchun statistik va hisoblash metodlariga asoslangan yangi metodologik baza shakllantiradi;
- hisoblash lingvistikasi va an’anaviy filologiya o‘rtasidagi metodologik uzviylikni kuchaytiradi.

Shuningdek, mazkur yondashuv poetik matnlarni tahlil qilishda klassik filologik interpretatsiyani statistik modellashtirish bilan uyg‘unlashtiradi va natijada til hodisalarini empirik ma’lumotlar asosida qayta talqin qilish imkonini beradi. Bu esa filologik tadqiqotlarni faqat tavsifiy bosqichdan empirik, statistik va algoritmik darajaga olib chiqadi hamda ularni zamonaviy sun’iy intellekt texnologiyalari bilan integratsiyalashgan ilmiy paradigmaga yaqinlashtiradi.

FOYDALANILGAN ADABIYOTLAR

1. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
2. Quirk, R. (1960). *Towards a Description of English Usage*. London: Longman.
3. Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

4. Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.
5. Moretti, F. (2013). *Distant Reading*. London: Verso.
6. Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.
7. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Stanford University.
8. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.