TIL MODELLARIDA IRONIYANI AVTOMATIK ANIQLASH UCHUN O‘ZBEK IRONIK KORPUSI

Ergashev Xumoyun Baratali o‘g‘li,
Magistrant
ergashevhumoyun9@gmail.com
Montreal universiteti

Annotatsiya. Ushbu tadqiqot doirasida biz o‘zbek tilida ironiyani o‘rganish va til modellarida avtomatik aniqlash uchun mo‘ljallangan birinchi O‘zbek ironik korpusini (O‘IK) ishlab chiqdik. Ushbu murakkab lingvistik hodisaning pragmatik va semantik nozikliklarini anglash uchun ma’lumotlarning qat’iy arxitekturasini yaratish talab etiladi. Mazkur maqolada ushbu korpusning tuzilishi va uni shakllantirish metodologiyasi batafsil bayon etilgan bo‘lib, unda ma’lumotlarni yig‘ish bosqichlari, ironik va ironik bo‘lmagan misollarni tanlab olish mezonlari hamda korpus tematikasi aniqlashtirib o‘tilgan.

Kalit so‘zlar: *ironiya, korpus, ko‘chma ma’no, hissiyotlar tahlili, til modellari, NLP.*

Abstract. Within the framework of this research, we have developed the first Uzbek Irony Corpus (UIC), specifically designed for the study and automatic detection of irony in language models. Comprehending the pragmatic and semantic nuances of this complex linguistic phenomenon requires the creation of a rigorous data architecture. This article details the structure and methodology of this corpus, specifying the stages of data collection, the selection criteria for ironic and non-ironic examples, and the thematic distribution of the corpus.

Keywords: *irony, corpus, figurative language, sentiment analysis, language models, NLP.*

Kirish. Dasturiy vositalar yordamida ironiyani avtomatik aniqlash ilmiy hamjamiyatda tobora katta qiziqish uyg‘otmoqda, ayniqsa, Twitter (X) kabi mikrobloglarni o‘rganishda bu yaqqol ko‘zga tashlanadi. Boshqa tillarda, jumladan

ingliz [1]; [2]; [9] va fransuz [4]; [5] tillarida mavzuga oid korpuslar yetarlicha topiladi. Ammo, o‘zbek tilida hali ironik korpus mavjud emas.

Mazkur yo‘nalishda hissiyotlar tahlili (sentiment analysis) bo‘yicha o‘zbek tilida ilk keng ko‘lamli ishlar Kuriyozov va boshqalarga (2019 [6]) tegishlidir. Shuningdek, tildagi annotatsiya qilingan ishonchli ma’lumotlar taqchilligi sharoitida, Saidov va boshqalarning (2026 [8]) ishlari metodologik muqobil yo‘lni – hissiyotlar tahlili va atoqli otlarni aniqlash (NER) uchun to‘liq sintetik korpuslar yaratishni tadqiq etadi.

Hernández-Farías va b. (2016 [3])ga ko‘ra, ironik korpuslarni yaratishda ikkita asosiy usul qo‘llaniladi. Birinchi usul, avtomatik belgilash. Bu usul metadiskursiv belgilar, ya’ni heshteglar (masalan: #ironiya, #sarkazm) yordamida ma’lumotlarni avtomatik yig‘ishdan iborat. Bunda ijtimoiy tarmoq foydalanuvchilari post yozish jarayonida o‘z xabarlarining tabiatini o‘zlari aniqlaydilar. Mazkur yondashuv “muallifning fikri uning kommunikativ niyatini ko‘rsatuvchi eng ishonchli ko‘rsatkich” degan farazga tayanadi. Ushbu metodning asosiy afzalligi – keyinchalik qo‘lda annotatsiya qilishga ehtiyoj sezmasdan, qisqa vaqt ichida juda katta hajmdagi misollarni to‘plash imkoniyatidir.

Ikkinchi usul, qo‘lda belgilash va kraudsorsing. Qo‘lda belgilash ma’lum bir xabarda ironiya bor-yo‘qligini baholash uchun uchinchi tomon annotatorlarini jalb qilishga tayanadi. Avtomatik belgilashdan farqli o‘laroq, bu yondashuvda qaror qabul qilish muallifdan tashqari tashqi kuzatuvchiga ham o‘tadi va bu jarayon ko‘pincha qat’iy annotatsiya protokoli asosida boshqariladi. Kraudsorsing (ommaviy autsorsing) ushbu metodning o‘ziga xos shakli bo‘lib, maxsus platformalar orqali ko‘plab ishtirokchilarni jalb qilgan holda keng ko‘lamli annotatsiyalarni olish imkonini beradi. Bu strategiya ironiyani idrok etishdagi subyektivlikni kamaytirish va fikrlar mosligi (konsensus) orqali annotatsiya jarayonini barqarorlashtirishga qaratilgan.

Yuqorida qayd etilgan korpuslar hissiyotlar tahlili uchun mustahkam poydevor yaratsa-da, ular ironiya masalasini bevosita qamrab olmaydi. Bu esa o‘z navbatida O‘IK korpusini yaratish zaruratini ko‘rsatib beradi. Ushbu bo‘shliqni to‘ldirish maqsadida biz yuqorida ko‘rsatilgan korpus yaratishning ikkinchi metodi, ya’ni qo‘lda belgilash va kraudsorsing metodiga asoslangan O‘zbek ironik korpusi (O‘IK) ni ishlab chiqdik. Ushbu korpus ikki bosqichli annotatsiya tuzilishi bilan ajralib turadi. Birinchidan, binar tasniflash: gapning tabiatini binar (ikki tomonlama) aniqlash (ironik yoki ironik emas); ikkinchidan, ironiyadagi lokusni aniqlash: gap tarkibida ironik effektning yuzaga kelishiga yoki uning kuchayishiga sabab bo‘lgan qismni (lokus) korpusda aniq belgilash. Keyingi usul modelga nafaqat natijani berishni, balki tushuntirishni (explainability) o‘rgatishga imkon beradi.

Metodologiya. Ushbu korpus olti oy davomida shakllantirildi. U jami 499 ta matn birliklaridan iborat bo‘lib, shundan 208 tasi ironik, 291 tasi esa ironik bo‘lmagan matnlardir. Ironiyani turli nutqiy vaziyatlarda va turli matn janrlarida kuzatish imkonini yaratish maqsadida, ma’lumotlar uchta alohida manbadan olindi. Ma’lumotlarning bunday rang-barangligi ironiyaning rasmiydan tortib so‘zlashuv uslubigacha bo‘lgan turli nutq qatlamlarida namoyon bo‘lishini tahlil qilishga imkon beradi. Ma’lumotlar quyidagi manbalardan jamlangan:

1. Badiiy asarlar: romanlar, hikoyalar va she’riyat, shuningdek, askiya janri.
2. Ijtimoiy tarmoqlar: Facebook va Telegram platformalaridan olingan postlar va sharhlar.
3. Kundalik muloqotlar: kundalik hayotdagi muallif tomonidan kuzatilgan jonli suhbatlar yoki teleko‘rsatuvlardan olingan parchalar.

Quyidagi 1-jadvalda korpus matnlarining manbalar bo‘yicha taqsimoti keltirilgan.

1-jadval. O‘IK korpusining matn manbalari va guruhlari (ironik/ironik emas)ga ko‘ra tuzilishi

Manba/Guruh	Ironik	Ironik emas	Jami
Badiiy asarlar	101	275	376
Ijtimoiy tarmoqlar	86	16	102
Kundalik muloqotlar	21	0	21
Jami	208	291	499

Annotatsiya jarayoni. Har bir matn qo‘lda ikki bosqichli annotatsiyadan o‘tkazildi. Dastlabki bosqichda gaplar ikki toifaga ajratildi: ironik yoki ironik emas. Shundan so‘ng, ikki nafar tashqi annotator (mos ravishda tilshunoslik va tarjima sohalari mutaxassislari) ham xuddi shu vazifani bajardilar. Annotatorlar o‘rtasidagi moslik darajasini Koen Kappasi koeffitsienti yordamida o‘lchadik. Muallif va ikki annotator o‘rtasidagi moslik ko‘rsatkichlari tegishli ravishda 0,41 va 0,45 ni tashkil etdi. Bundan tashqari, ikki tashqi annotatorning o‘zaro moslik darajasi 0,44 ga teng bo‘ldi. Landis va Koch shkalasiga ko‘ra, ushbu natijalar *“mo‘tadil moslik”* darajasini ko‘rsatadi. Ironiyani aniqlash kabi murakkab va subyektiv vazifa uchun bu natija qoniqarli deb hisoblanadi.

Uch baholovchi o‘rtasidagi kelishmovchiliklarni yuzaga keltirgan to‘rtta asosiy omil aniqlandi. Bu esa ironiya va boshqa uslubiy vositalar yoki nutq qatlamlari o‘rtasidagi chegaralar yaqin va aniqlashga murakkab ekanligini ko‘rsatadi.

1. **Samimiy mubog‘ala va ironiya o‘rtasidagi chalkashlik**

Birinchi qiyinchilik ironik maqsaddagi va samimiy mubog‘ala (giperbola) o‘rtasidagi farqni aniqlash bilan bog‘liq. Ba’zan mubolag‘adan ironiya yuzaga keltirish uchun foydalanilsa, ba’zan esa ironik maqsadda emas, balki shunchaki gapni tasdiqlash va kuchaytirish uchun ishlatiladi. 1-misolda *jala shiddati* samimiy ifodalangan bo‘lsa-da, undagi kuchli mubog‘ala tufayli ikki annotator ushbu gapni ironik deb belgilagan.

1-misol. *Uni bag‘rimga bosib, beg‘ubor ko‘zlariga tikildim. Nimalardandir umidlanib, chuqur xo‘rsindim... Oqshomga yaqin jala ham tindi. Asfaltni ariqcha topolmagan loyqa suvlar egallab, “dengiz” hosil bo‘ldi*.

2. **Ma’no ikkiyoqlamaligi yoki kontekst yetishmasligi**

Ba’zi gaplarda kontekstual ma’lumot yetarli emasligi sababli, ularning talqini butunlay annotatorning subyektiv idrokiga bog‘liq bo‘lib qoladi. 2-misol ko‘rsatib berganidek, so‘zma-so‘z bayon va ironik tanqid o‘rtasidagi farq faqat so‘zlovchining niyatini anglashga (inferensiya) tayanadi. Yetarli kontekstual ishoralar (umumiy kontekst/common ground) bo‘lmaganda, annotatorlar o‘rtasidagi moslikni barqarorlashtirish qiyinlashadi.

2-misol. *Uch-to‘rt kun bo‘yinbog‘ni hilpiratib qatnab, xat hujjatlarga qarab ko‘zimni pishitib, kun o‘tar qilib yurdim. O‘sha paytgacha partiya deganda, xalq manfaati yo‘lida fikrlar xilma-xilligi, bahs-munozara uyi, umuman siyosiy qarashlar markazi, deb yurar ekanman.*

3. **Ritorik so‘roq gaplar**

Ritorik so‘roq gaplar ham avtomatik, ham insoniy aniqlash jarayoni uchun tez-tez uchraydigan “tuzoq” hisoblanadi. Garchi ritorik so‘roq ironiyaning asosiy vositalaridan biri bo‘lsa-da, u boshqa pragmatik vazifalarni (ta’kidlash, g‘azab, hayajon) ham bajarishi mumkin. 3-misoldagi ritorik so‘roq strukturasi, tarkibida hech qanday ironik ziddiyat bo‘lmasa ham, chalkashlik keltirib chiqargan.

3-misol. *O‘ylab ko‘rsam Yozuvchilar uyushmasida qizlarning otalari uchun to‘kkan qanchadan-qancha ko‘z yoshlariga guvoh bo‘libman, tarix guvoh bo‘lganlari qancha, guvohlarsiz to‘kilgan ko‘zyoshlar-chi?!*

4. **Samimiy yumor va idiomatik iboralarning murakkabligi**

Nihoyat, samimiy yumor va turg‘un birikmalarning qo‘llanilishi katta qiyinchilik tug‘diradi. 4-misoldagi To‘ybekaning gaplari ushbu qiyinchilikni yaqqol ko‘rsatadi: “Teng tengi bilan, tezak qopi bilan” maqolining qo‘llanilishi ham

qiyoslashni, ham idiomani o‘z ichiga oladi. Bu yerda samimiy hazil va ironiya o‘rtasidagi chegara shunchalik nozikki, bu baholovchilar o‘rtasida muntazam tushunmovchilik manbasiga aylanadi.

4-misol. - *Ey... singlim, hali sen bilmaysan, - dedi, - u yigitni bir ko‘rgin-da, hu, ded ketabergin... sen tugil, shu yoshim bilan menim ham unga tekkim keldi, - dedi va xoxalab yubordi. Kumushbibi chirt etib yuzini To‘ybekaga o‘girdi. -Tezroq tegib qoling. -Koshki edi tegalsam, - dedi To‘ybeka, - men uning bir tukiga ham arzimayman. Ammo sen bo‘lsang uning bilan tenglashar eding. Teng-tengi bilan, tezak qopi bian. Xa-xa-xa!...*

Korpusning tematik tahlili

Korpusning ikki guruhidan (ironik va ironik bo‘lmagan) biriga xos bo‘lgan mavzuviy kamchiliklar (thematical bias) bor-yo‘qligini baholash maqsadida biz mavzuviy modellashtirish (topic modeling) jarayonini amalga oshirdik. O‘zbek tili uchun optimallashtirilgan matnni avtomatik qayta ishlash (TAL/NLP) vositalarining taqchilligi sababli, ushbu tahlil Gensim kutubxonasining [7] LDA (Latent Dirichlet Allocation) algoritmi yordamida korpusning inglizcha versiyasi ustida olib borildi. Ushbu bosqichdan ko‘zlangan maqsad, har bir guruh ichidagi mavzular xilma-xilligini tekshirish; hamda ironik va ironik bo‘lmagan matnlar o‘rtasida mavzuviy jihatdan nisbiy tenglik mavjudligiga ishonch hosil qilishdir. Buning uchun uchtadan o‘ntagacha bo‘lgan turli konfiguratsiyalarni sinab ko‘rdik va yakunda yetti mavzuli modelni (k=7) tanlab oldik. Ushbu tanlov mavzular o‘rtasidagi aniq farqni saqlab qolgan holda, korpusning semantik xilma-xilligini aks ettirish uchun maqbul (optimal) deb topildi. Quyidagi 2-jadvalda ushbu yetti toifa doirasida har bir guruh (ironik va ironik bo‘lmagan) bo‘yicha matnlarning taqsimoti keltirilgan:

2-jadval. Korpusning tematik bo‘linishi

Mavzu	Ironik matnlar (n=208)	Ironik bo‘lmagan matnlar (n=291)
Mavzu 0	36	49

Mavzu 1	27	33
Mavzu 2	33	47
Mavzu 3	22	35
Mavzu 4	34	54
Mavzu 5	25	36
Mavzu 6	31	37

Ushbu jadvaldagi 4-mavzudagi juz’iy miqdoriy tafovutlarga qaramay, mavzular har ikki sinf o‘rtasida umuman olganda muvozanatlashgan degan xulosaga kelishimiz mumkin. Faqatgina ironiya uchun ajratilgan alohida mavzu mavjud emas, bu esa bizning korpusimiz ironiyani suhbat mavzusidan qat’i nazar, umumiy hodisa sifatida o‘rganish imkonini berishini tasdiqlaydi.

Xulosa. Garchi O‘zbek ironik korpusi (O‘IK) o‘zbek tili uchun tabiiy tilni qayta ishlash (NLP), xususan, hissiyotlar tahlili sohasida oldinga tashlangan qadam bo‘lsa-da, ushbu tadqiqot loyihasiga xos bo‘lgan metodologik cheklovlarni ham ta’kidlab o‘tish joiz. Birinchidan, ma’lumotlar to‘plamimizning hajmi (499 ta matn) nisbatan kichik. Qo‘lda annotatsiya qilingan korpusni shakllantirish ko‘p vaqt talab qiladigan vazifa bo‘lib, u katta insoniy va moliyaviy resurslarga tayanadi; mazkur magistrlik dissertatsiyasi doirasida bizda bunday imkoniyatlar to‘liq mavjud emas edi. Garchi ushbu hajm chuqur sifat tahlili (qualitative analysis) o‘tkazish imkonini bersa-da, optimal darajadagi umumlashtirishga erishish uchun, odatda ulkan ma’lumotlar hajmini talab qiladigan chuqur o‘rganish (deep learning) modellarini o‘qitishda qiyinchilik tug‘diradi. Shuningdek, yuqorida sanalganidek, mavzuga ko‘ra modellashtirish (LDA) tahlili ironik va ironik bo‘lmagan guruhlar o‘rtasida 4-mavzu taqsimotida juz’iy nomutanosiblikni aniqladi. Bu holat qisman ikki sinf o‘rtasidagi miqdoriy nomutanosiblik bilan bog‘liq bo‘lsa-da, korpusdagi ayrim mavzular ironik ifoda uchun qulayroq bo‘lishi mumkin. Mavzularning bunday notekis taqsimlanishi kelgusida modelning ironiyani oddiy mavzu mazmunidan ajratib olish qobiliyatiga ta’sir ko‘rsatishi ehtimoldan xoli emas.

Foydalanilgan adabiyotlar ro‘yxati

1. Beals C. A linguistic analysis of verbal irony. PhD thesis, University of Chicago, Department of Linguistics, 1995.

2. Ghosh A., Li G., Veale T., Rosso P., Shutova E., Barnden J., Reyes A. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015) (pp. 470–478). Association for Computational Linguistics, 2015. https://doi.org/10.18653/v1/S15-2089

3. Hernández-Farías D. I., Patti V., Rosso P. Irony detection in Twitter: The role of affective content. ACM Transactions on Internet Technology, 16(3), Article 19, 2016. Pp. 1–24. https://doi.org/10.1145/2930663

4. Karoui J., Aussenac-Gilles N., Benamara F., Hadrich Belguith L., (cinquième auteur si requis). Détection automatique de l’ironie dans les tweets en français. In Actes de la 22e Conférence sur le Traitement Automatique des Langues Naturelles. TALN, 2015.

5. Karoui J. FrIC : Un corpus et un schéma d’annotation multi-niveaux pour l’ironie dans les tweets. In Actes de la conférence conjointe JEP-TALN-RECITAL 2016. Vol. 7, COLTAL.

6. Kuriyozov E., Matlatipov S., Alonso M. A., Gómez-Rodríguez C. Deep learning vs. classic models on a new Uzbek sentiment analysis dataset. In Human Language Technologies as a Challenge for Computer Science and Linguistics – 2019. Pp. 258–262. Wydawnictwo Nauka i Innowacje, 2019.

7. Rehurek R., Sojka P. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2). 2011.

8. Saidov B., Barakhnin V., Madirimov S., Ibragimov U., Meylikulov S., Normamatov S., Bahodirova F., Matnazarov J., Fayzullaeva Z. Dual-source

synthetic Uzbek corpora for sentiment analysis and NER with controlled emoji signals. Data, 11(2), 28. 2026. https://doi.org/10.3390/data11020028

9. Van Hee C., Lefever E., Hoste V. SemEval-2018 Task 3: Irony detection in English tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018) (pp. 39–50). Association for Computational Linguistics, 2018. https://www.aclweb.org/anthology/S18-1005