

O‘ZBEK VA INGLIZ TILLARIDAGI PARALLEL KORPUSDA FRAZEMALARNI MORFOLOGIK VA SEMANTIK TEGGLASH MODELIMuallif:

Nurullayeva Oydin Nazirjon qizi

TAFU

E-mail: oydinnurullayeva85@gmail.com

Annotatsiya. Maqolada o‘zbek–ingliz frazemalar parallel korpusi uchun morfologik va semantik teglash modeli taqdim etiladi. Tadqiqotda 420 ta frazema juftligi asosida ikki qatlamli annotatsiya tizimi ishlab chiqilgan bo‘lib, bu tizim frazemaning semantik yaxlitligini saqlagan holda uning morfologik tuzilishini ham ko‘rsatadi. Model frazemalarni to‘g‘ri segmentlash va semantik tahlil qilish imkonini beradi, bu esa tarjima sifatini yaxshilashda va parallel korpuslarda frazemaning aniqligini oshirishda muhim ahamiyatga ega.

Kalit so‘zlar: o‘zbek–ingliz parallel korpusi, frazema, morfologik teglash, semantik teglash, iyerarxik tagset, frazemani lemmalash, n-gramm, PMI, JSON-schema, annotatorlar kelishuvi.

Аннотация. В статье представлен метод морфологической и семантической разметки узбекско-английского параллельного корпуса фразем. На основе 420 пар фразем разработана двухуровневая система аннотации, которая сохраняет семантическую целостность фразем и отображает их морфологическую структуру. Модель позволяет корректно сегментировать фраземы и проводить их семантический анализ, что важно для улучшения качества перевода и повышения точности фразем в параллельных корпусах.

Ключевые слова: узбекско-английский параллельный корпус, фразема, морфологическая разметка, семантическая разметка, иерархический tagset, лемматизация фразем, n-gramm, PMI, JSON-схема, согласие аннотаторов.

Abstract. This paper presents a morphological and semantic tagging model for an Uzbek–English parallel corpus of phrasemes. Based on 420 aligned phraseme pairs, the study develops a two-layer annotation system that preserves the semantic integrity of phrasemes while also reflecting their morphological structure. The model facilitates accurate segmentation and semantic analysis, which is crucial for improving translation quality and enhancing the accuracy of phraseme identification in parallel corpora.

Keywords: Uzbek–English parallel corpus, phraseme, morphological tagging, semantic tagging, hierarchical tagset, phraseme lemmatization, n-grams, PMI, JSON schema, inter-annotator agreement.

Kirish

Korpus lingvistikasi va tabiiy tilni qayta ishlash (NLP) amaliyotlarida eng “nozik” obyektlardan biri – frazema. U matnda bir nechta soʻz koʻrinishida kelsa-da, maʼno va funksiyada koʻpincha yaxlit leksemaga tenglashadi; shu bois uni oddiy soʻzlar ketma-ketligi sifatida teglash frazemaning semantik yaxlitligini yoʻqotadi. Bu muammo parallel korpuslarda yanada kuchliroq seziladi: ekvivalent frazemalar koʻpincha soʻzma-soʻz mos tushmaydi, obraz va struktura farqi tarjimani murakkablashtiradi.

Parallel korpuslar boʻyicha anʼanaviy yondashuvlarda material yigʻish va qayta ishlash, odatda xom maʼlumotni olish, gap moslash, segmentlash, tokenlash/normalizatsiya va yakuniy moslash bosqichlarini oʻz ichiga oladi. [2] Ochiq parallel resurslar infrastrukturasida OPUS turli formatlarda parallel maʼlumotlar va interfeyslarni taqdim etadi. Biroq yirik parallel korpuslarning koʻpchiligida frazemalar “tayyor” belgilangan qatlam sifatida doim ham mavjud emas; natijada frazemalarni qidirish va tahlil qilish koʻpincha qoʻlda yoki cheklangan heuristikalar orqali bajariladi.[5]

Morfologik annotatsiya amaliyotida ingliz tili uchun P.Marcus [3] va hamkorlar yaratgan Penn Treebank tajribasi POS teglashning metodik tayanchi sifatida koʻriladi; u millionlab soʻzli korpusda POS va sintaktik skelet annotatsiyasini standartlashtiradi. Lekin idiomatik birliklar koʻpincha komponentlarga parchalanib ketishi muammosi saqlanadi. Semantik teglash boʻyicha yirik amaliy yondashuvlardan biri – Paul Rayson va hamkorlar ishlab chiqqan USAS; u iyerarxik tagset bilan avtomatik semantik belgilashning amaliy samaradorligini koʻrsatgan. [4] Bu tajriba iyerarxik semantik tagset foydasini koʻrsatsa-da, frazemalar uchun “dominant sema”ni ajratish va tillarar moslashni metodik jihatdan qatʼiy belgilash alohida vazifa boʻlib qoladi.

Mazkur maqolaning maqsadi – oʻzbek-ingliz frazemalar parallel korpusi uchun morfologik va semantik teglash modelini metodik jihatdan aniq bayon qilish, tagsetni real misollar bilan koʻrsatish va eng muhimi, resurs sifatini himoya qiladigan baholash taklif qilishdir. Bunday resurslar davlat tilini raqamli muhitga integratsiya qilish va sunʼiy intellekt imkoniyatlaridan foydalanishni kuchaytirish kabi vazifalar bilan hamohang. [6]

Masalaning qoʻyilishi. Tadqiqot materiali 420 ta oʻzbek–ingliz frazema juftligidan iborat. Har bir yozuvda: (1) oʻzbekcha frazema, (2) inglizcha muqobil, (3) lugʻaviy izoh, (4) kontekst misoli saqlanishi va baza JSON formatida tashkil qilingani koʻrsatiladi. Korpusni toʻldirishda lugʻat manbalari asosiy oʻrin tutadi: Oʻzbek tili frazeologik lugʻati (Rahmatullayev va boshq., 2022), Inglizcha-oʻzbekcha idiomalar lugʻati (Yusupov, 2014), hamda Oʻzbek tilining izohli lugʻati (2020) kabi manbalar, shuningdek Oxford Dictionary of Idioms va Cambridge International Dictionary of Idioms keltiriladi.

Annotatsiya sxemasi: ikki qatlamli yondashuv. Dissertatsiyada frazemalarni belgilash uchun ikki qatlamli annotatsiya konsepsiyasi asoslanadi: (a) Leksik-semantik qatlam frazemani yaxlit

segment sifatida ajratadi; (b) Morfologik (morfosintaktik) qatlam frazema ichidagi komponentlarga grammatik teglarni berishni davom ettiradi. Bu yondashuv frazemaning semantik yaxlitligini saqlagan holda, uning ichki grammatik qurilishini ham ko'rsatadi. Segmentlash uchun XMLga o'xshash belgilash ishlatilishi mumkin: <idiom>... </idiom> . Metodik jihatdan bu "majburiy" qadam: aks holda POS-teglagich idiomani komponentlarga parchalaydi va qidiruv/struktur tahlil "frazema" obyektini yo'qotadi

Morfologik tagset. Morfologik teglashning asosiy g'oyasi: frazema butunligicha qaysi so'z turkumidagi leksik birlikka teng bo'lsa, shu turkum bilan teglanadi. Dissertatsiyada to'rt asosiy teg joriy qilinadi: *ot-frazema*, *sifat-frazema*, *fe'l-frazema*, *ravish-frazema*. Lemmaga keltirish ham shu prinsipga tayangan: masalan, qo'li ochiq birikmasi "qo'l+ochiq" kabi bitta segment/lemma sifatida birlashtiriladi. Bu yondashuv frazemaning barcha grammatik variantlarini "bitta obyekt" sifatida qidiruvda jamlash muammosini yumshatishga xizmat qiladi .

UDning morfologik tavsifi lemma – POS-features qatlamlaridan iborat; u korpuslararo moslash uchun qulay universal qatlam beradi. Shu sabab frazema sinfini XPOS darajasida, komponentlarning grammatik xususiyatlarini esa FEATS darajasida kodlash UD bilan moslikni saqlagan holda frazema qatlamini ajratishga yordam beradi. [3]

Semantik tagset va iyerarxiya. Semantik teglash "mushtarak sema" va "dominant sema" tamoyiliga tayanib, 4 asosiy kategoriya bilan boshlanadi: (1) shaxs – Fper, (2) makon/fazo –Flok, (3) vaqt – Ftemp, (4) miqdor/daraja – Fkvnt. Teglar iyerarxik kengaytiriladi: masalan, shaxs kategoriyasi doirasida mikroteglar (Fper hlt iqt, Fper hlt ijt, Fper hlt rh, Fper xrktr va h.k.) keltiriladi Flok, Ftemp va Fkvnt kategoriyalari ham dominant sema asosida misollar bilan asoslanadi: "yuzma-yuz kelmoq" – Flok (yaqin masofadagi uchrashuv), "ko'z ochib yumguncha" – Ftemp (juda qisqa fursat), "to'rt ko'zi bilan kutmoq" – Fkvnt (sonli metafora).

Ma'lumotlar modeli va sifat nazorati. Tadqiqotda korpus yozuvlari JSON formatida saqlanishi, JSON-schema orqali formal validatsiya va Node.js muhitida audit skripti yordamida xatolarni tekshirish imkoniyati qayd etiladi. Bu struktur xatolar (maydon yo'qligi, format mos kelmasligi) va mantiqiy xatolarni (ID takrorlanishi, teg formatining buzilishi) kamaytirishga xizmat qiladi. Foydalanuvchi interfeysi React.js asosida tuzilgani, shakl/ma'no/kontekst bo'yicha bir vaqtda qidirish imkonini berishi ko'rsatilgan.

Avtomatik komponentlar: n-gramm va PMI orqali frazema nomzodlari. N-gramm ko'p so'zli birliklarni konkordans statistikasi orqali "nomzod" sifatida chiqaradi; amaliyotda minimal chastota filtri va kollokatsion o'lehovlar, jumladan PMI ishlatiladi. PMI va kollokatsiyalarning "kutilganidan ko'ra tez-tez birga chiqishi" g'oyasini K.Church, P.Hanks klassik maqolasida asoslagan. [7]

flowchart TD

A[Manbalar: frazeologik lugʻatlar + matnlar] --> B[Normalizatsiya: variantlarni birlashtirish]

B --> C[Identifikator: oʻzbek–ingliz juftlikni bogʻlash]

C --> D[Segmentlash: <idiom> ... </idiom>]

D --> E[Morfologik teglash: ot/sifat/feʼl/ravish-frazema + komponent POS]

E --> F[Semantik teglash: Fper/Flok/Ftemp/Fkvnt + mikroteg]

F --> G[JSON modeliga yozish]

G --> H[JSON-schema validatsiya + audit skript]

H --> I[Qidiruv/vizualizatsiya: veb interfeys]

Natijalar va namunalar.

Korpusning ishlanma natijasi va strukturasi. 420 juftlikli baza har bir frazema uchun “turgʻun birlik + maʼno + kontekst + tarjima” toʻrtligini beradi. Teglashning ikki qatlamli modeli korpus qidiruvini amaliy jihatdan soddalashtiradi: frazema <idiom> bilan segmentlanganda, qidiruv natijasida komponentlar boʻyicha parchalanib ketmay, aynan frazema obyekt sifatida qaytadi. Shu bilan birga, tadqiqot natijasi tokenlash/lemmalash amaliyoti toʻliq avtomatlashtirilmagani va roʻyxatning qoʻlda tuzilgani sabab kelgusida bu bosqichlar rejalashtirilganini taʼkidlaydi.

1-jadval. Korpusning minimal statistik pasporti.

Koʻrsatkich	Qiymat	Izoh / manba
Frazema juftliklari soni	420	JSON formatidagi baza
Har bir yozuv maydonlari	4+	UZ frazema, EN muqobil, izoh, kontekst
Morfologik teglar soni	4	ot-/sifat-/feʼl-/ravish-frazema
Semantik 1-daraja teglar	4	Fper/Flok/Ftemp/Fkvnt
Ekvivalentlik turlari	4	toʻliq, qisman, obrazi boshqa, noekvivalent
QA mexanizmi	mavjud	JSON-schema + Node.js audit
Interfeys	mavjud	React.js qidiruv interfeysi

2-jadval. Tagset misollari.

Qatlam	Teg	Qisqa ta’rif	Dissertatsiyadagi tipik misol
Morfologik	fe’l-frazema	harakat/amalni bildiruvchi frazema	yerga urmoq
Morfologik	sifat-frazema	sifat funksiyasidagi frazema	qo’li ochiq
Semantik	Fper hlt iqt	shaxsning iqtisodiy holati	qo’li yupqa; o’ziga to’q
Semantik	Flok	makon/fazoviy obraz	yuzma-yuz kelmoq
Semantik	Ftemp	vaqt/davomiylik/tezlik	ko’z ochib yumguncha
Semantik	Fkvnt	miqdor/daraja/ko’lam	to’rt ko’zi bilan kutmoq

3-jadval. Teglangan parallel juftliklar.

ID	UZ (kontekst parchasi)	Teglar	EN (kontekst parchasi)	Teglar
FZ_0001	<idiom>aloqasini uzmoq</idiom>	morph=fe’l-frazema; sem=Fper	<idiom>cut ties</idiom>	morph=verb-idiom; sem=Fper
FZ_0002	<idiom>alg’ov-dalg’ov qilmoq</idiom> / ostin-ustin qilish	morph=fe’l-frazema; sem=Fper	<idiom>throw into chaos</idiom>	morph=verb-idiom; sem=Fper

Modelning kuchli tomoni – frazemani korpusda “mustaqil obyekt” sifatida qayd etish va uni morfologik-semantik tagsetga bog’lashdir. Bu qidiruvdagi noaniqliklarni kamaytiradi, statistik hisoblarni tozalaydi va tarjimada ekvivalentlik darajalarini sistemali ko’rsatishga yordam beradi. Konseptual jihatdan yondashuv frazemalarni maxsus qatlam sifatida ajratadigan MAGPIE kabi resurslar bilan hamohang: MAGPIE BNC asosida idiomatik ifodalarni belgilash va korpus vositalari orqali qidirish imkonini beradi. [1]

Amaliy qo‘llanishlar uch yo‘nalishda ayniqsa kuchli. Birinchisi, tarjima amaliyoti va tarjima texnologiyalarida frazema muqobilini tanlash: semantik teglar “qidiruv filtri” sifatida ishlaydi. Ikkinchisi, lingvodidaktika: frazemalarni kontekst va semantik guruhlar bo‘yicha o‘rgatish, misollarni “Fper/Flok/Ftemp/Fkvnt” doirasida qurish mumkin. Uchinchi yo‘nalish esa NLP/AI: n-gramm+PMI bilan nomzod frazemalarni topish, so‘ng teglangan baza bilan validatsiya qilib, mashina tarjimasi va qidiruv tizimlarida idiomalar bilan bog‘liq xatolarni kamaytirish. Bu yo‘nalishlar davlat tilini raqamli muhitga faol integratsiya qilish va AI imkoniyatlaridan samarali foydalanish talablariga ham mos keladi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

- [1] Haagsma H., Boss J., Nissim M. (2020). MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). – P. 279–287.
- [2] Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation // Proceedings of the 10th Machine Translation Summit. – Phuket, Thailand: AAMT, 2005. – P. 79–86.
- [3] Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330.
- [4] Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. LREC’04 Workshop, 7–12.
- [5] Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. LREC 2012, 2214–2218.
- [6] O‘zbekiston Respublikasi Prezidentining “Davlat tilini yanada rivojlantirishga oid qo‘shimcha chora-tadbirlar to‘g‘risida”gi PQ-308-son qarori. 20.10.2025. <https://lex.uz/uz/docs/-7776360>
- [7] Church K.W., Hanks P. Word Association Norms, Mutual Information, and Lexicography // Computational Linguistics. 1990. – Vol. 16. – № 1. – P. 22–29.