O‘ZBEK TILI UCHUN TARJIMA TEXNOLOGIYASINI AVTOMATLASHTIRISHNING LINGVISTIK ASOSLARI

Shamsiyeva Gulshoda Asliddin qizi
O‘zbekiston Milliy universiteti, 1-kurs magistranti
e-mail: gulshoda.xayrullayeva.98@mail.ru

ANNOTATSIYA

Ushbu maqolada o‘zbek va ingliz tillaridagi ilmiy va rasmiy matnlarni tarjima qila oladigan mashina tarjimasi texnologiyasini yaratishning lingvistik asoslari tahlil qilingan. Bunday tarjima dasturining yaratilishi ilmiy va rasmiy matnlarni tarjima qilishdagi muammo va kamchiliklarni bartaraf etadi va ilmiy izlanish olib boruvchilar uchun katta qulaylik yaratadi. Maqolada avtomat tarjima texnologiyasini yaratishda ijobiy natijaga erishish uchun bir qancha takliflar ham berib o‘tilgan. Shuningdek, “Google Translate” tarjima dasturi haqida ham to‘xtalib o‘tilgan bo‘lib, uning yutuq va kamchiliklari ko‘rsatib berilgan. Mashina tarjimasi uchun zarur bo‘lgan gap modellari haqida ham to‘xtalingan bo‘lib, bu borada qilingan tadqiqot ishlari tahlilga tortilgan. Tadqiqotimizning asosiy maqsadi ilmiy izlanishlar yo‘lida keng imkoniyat yaratish uchun terminologik parallel korpuslar asosida ilmiy va rasmiy matnlarni (o‘zbek va ingliz tilida) tarjima qiladigan mashina tarjimasi texnologiyasining lingvistik asoslarini tahlil qilishdan iborat.

Kalit so‘zlar: Mashina tarjimasi, Parallel korpus, Morfoanalizator, Sintaktik analizator, Tarjima lug‘atlar, gap modellari, Modellashtirish.

ABSTRACT

This article analyzes the linguistic basis of the development of machine translation technology that can translate scientific and official texts in Uzbek and English. The creation of such a translation program eliminates the problems and shortcomings in the translation of scientific and official texts and creates great convenience for researchers. The article also offers a number of suggestions for achieving positive results in the development of automatic translation technology. Google Translate translation program was also mentioned its advantages and disadvantages. The models of speech required for machine translation were also discussed, and research work in this area was analyzed. The main purpose of our research is to analyze the linguistic basis of machine translation technology, which translates scientific and official texts (in Uzbek and English) on the basis of

terminological parallel corpus to create a wide range of opportunities for scientific research.

Key words: Machine Translation, Parallel Corpus, Morphoanalyzer, Syntactic Analyzer, Translation Dictionaries, Speech Models, Modeling.

АННОТАЦИЯ

В данной статье анализируются лингвистические основы развития технологий машинного перевода, которые могут переводить научные и официальные тексты на узбекский и английский языки. Создание такой программы-переводчика устраняет проблемы и недостатки при переводе научных и официальных текстов и создает большие удобства для исследователей. Также в статье предлагается ряд предложений по достижению положительных результатов в развитии технологии автоматического перевода. Также была упомянута программа-переводчик Google Translate, ее преимущества и недостатки. Также обсуждались модели речи, необходимые для машинного перевода, и анализировались исследовательские работы в этой области. Основной целью нашего исследования является анализ лингвистических основ технологии машинного перевода, которая переводит научные и официальные тексты (на узбекский и английский языки) на основе терминологического параллельного корпуса для создания широких возможностей для научных исследований.

Ключевые слова: Машинный перевод, параллельный корпус, морфоанализатор, синтаксический анализатор, переводческие словари, модели речи, моделирование.

Bugunga qadar mashina tarjimasi sohasida bir qancha olimlar tomonidan ilmiy izlanishlar olib borilgan. N. Abduraxmonova “Mashina tarjimasining linvistik ta’minoti” nomli monografiyasi A.But, R. Richans, J. Xatchins, P. Braun (AQSh); G.G.Belonogov, J.Allen, Z.Shalyapina, N.D. Andreev, I.A. Melchuk, V.Yu. Rozensveyg, Yu.N. Marchuk, R.G. Piotrovskiy, Yu.A. Motorin, K.B. Bektaev, A.N. Belyaev, I.K. Belskiy, A.V. Zubov, G.E. Miram, L.L. Nelyubin, V.I. Perebiynos, V.A. Chijakovskiy, Ye.A.Shingarev, R.G.Kotov, Babushkina N.V, O.Yu. Mansurova, A.S. Panina, A.A. Xoroshilov (Rossiya); M. Nagao (Yaponiya); A.Vaxer (Estoniya); Fransiyada J.Astrouni; R.Sinha, A.Jain (Hindiston); B.Bleyzer, U.Shvol, A.Storrer (Germaniya) singari tadqiqotchilarning ishlarini sana o‘tadi.[68] Massachusetts Texnologiya Institutining olimi Philipp Koehn “Pharaoh” (“Fir’avn”) mashina

[68]Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).

tarjimasi mexanizmi orqali frazalarga asoslangan statistic mashina tarjimasi modellarini tavsiflab beradi. [69]

XVIII asrda faylasuflar G. V. Leybnis va Deskart jumla va so‘zlarning o‘zaro bog‘liqligini kodlash nazariyasini fanga olib kirishadi. Mashinaning tarjima qila olish imkoniyati haqidagi g‘oyani Ch. Bebbidj (1791-1871) 1836-1848 yillarda o‘sha davrdan 100 yil keyin paydo bo‘lgan elektron raqamli mashinalarning mexanik prototipi - raqamli analitik mashinalar bo‘yicha olib borgan loyihasida qayd etgan. XX asrning 30 yillarida “tarjimon mashinalar” amalda qo‘llanilgan bo‘lsa, Estoniyada mexanik tarjimani ro‘yobga chiqarishda A.Vaxerning nazariy qarashlari “Vaba Maa” nomli ro‘znomada e’lon qilindi (1924). Fransiyada J.Astrouni tomonidan avtomatik bilingval lug‘atlardan foydalanish taklif etilib, u “Mexanik miya” deb nomlangan mashina tarjimasi ishlanmasi uchun patentga ega bo‘ldi. 1933 yilda Rossiyada lingvistik arifmometr muallifi P.P.Smirnov Troyanskiy tomonidan esperanto tili asosida tillar o‘rtasidagi grammatik boshqaruv usuli bilan taqsimlash metodi va bilingual lug‘atni o‘z ichiga oluvchi tizim yaratildi.[70] Mashina tarjimasi sohasida birinchi ilmiy-amaliy konferensiya 1952-yilda Massachutes texnika universitetida o‘tkaziladi va 1954-yilda NyuYorkdagi Jorjtaun universiteti bilan IBM kompaniyasi hamkorligida birinchi IBM II rus tilidan ingliz tiliga tarjima qilishga mo‘ljallangan mashina tarjimasi dasturi yaratiladi. Bu dastur L.Dorster rahbarligida o‘tkazilib, kimyo sohasi doirasida 250 ta leksik birlik hamda 6 ta grammatik qoidalar bilan chegaralgan edi. (“Dorster tajribasi” yoki “Jorjtaun sinovi”). Turkiy tillar bo‘yicha ilk mashina tarjimasi ham 1961-yilda Dorster boshchiligida amalga oshirilgan. Ingliz-turk tili mashina tarjimasi tizimida 700 ta so‘zshakldan iborat bo‘lib, lug‘at turk tili hamda ingliz tilidagi ekvivalentliklarining asos va suffikslari ro‘yxatidan iborat bo‘lgan.[71]

O‘zbek tili uchun mashina tarjimasiga doir deyarli fundamental izlanish olib borilmagan edi. N. Abduraxmonova “Mashina tarjimasining linvistik ta’minoti” monografiyasida M.Xakimov ishlarida kengaytirilgan matematik modelga asoslanilgan mashina tarjimasi texnologiyasi bo‘yicha tadqiqotlar amalga oshirilganligi, biroq o‘zbek tilidagi ma’lumotlarning yetarli emasligi yaratiladigan dasturiy ta’minotlarga ta’sir ko‘rsatganligi, shuningdek S.Muhamedov R.R.Piotrovskiy bilan hammualliflikda yozgan “Injenernaya lingvistika i опыт

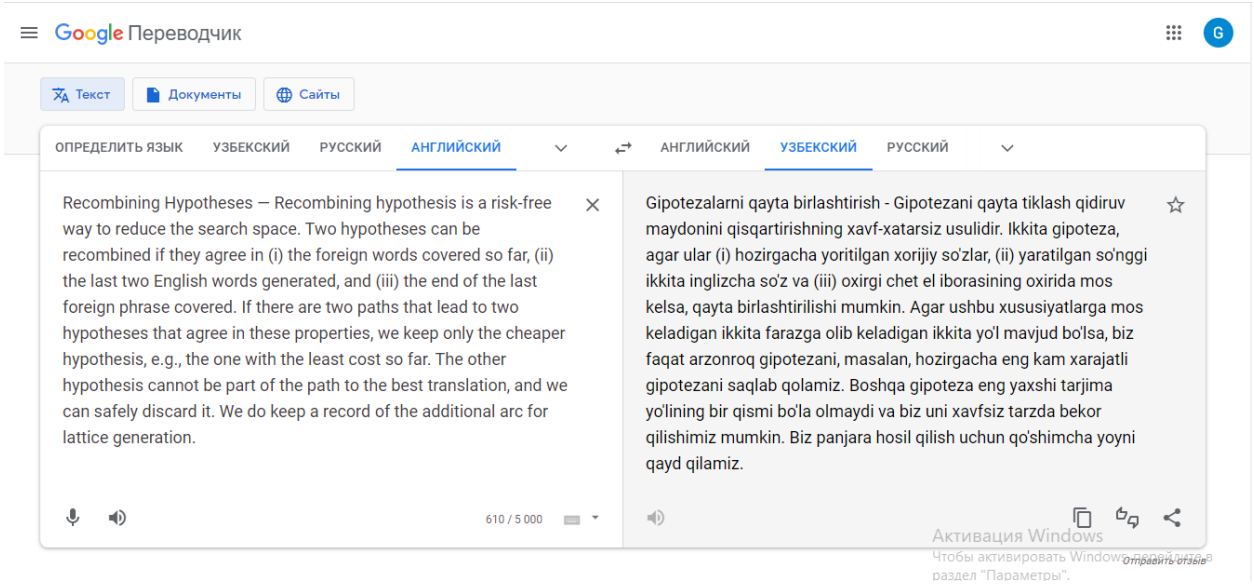
[69] Philipp Koehn. A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. Machine Translation: From Real Users to Research 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004. Washington, DC, USA, September/October 2004

[70] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).

[71] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).// М.Г.Мамедова, З.Ю.Мамедова Машинный перевод: эволюция и основный аспекты моделирования. – Баку, 2005. – С. 34.

sistemno-statisticheskogo issledovaniya uzbekskix tekstov” nomli kitobida lingvistik modellar, modellashtirish va uning umumiy tamoyillari haqida fikr yuritilib, o‘zbekcha matnlarning kvantativ modellari keltirilganligini qayd etib o‘tgan.[72] Nilufar Abduraxmonova tomonidan “Inglizcha matnlarni o‘zbek tiliga tarjima qilish dasturining lingvistik ta’minoti (sodda gaplar misolida)” mavzusidagi falsafa doktori (PhD) ilmiy darajasini olish uchun yozilgan (2018) dissertatsiyasida o‘zbek tilining lingvistik modellari, sintaktik strukturalar orqali tarjima qilish, qolaversa, transfer usuli qardosh bo‘lmagan tillar uchun mos ekanligi ilmiy jihatdan o‘rganilgan.[73]

Yuqorida aytib o‘tilgan tadqiqotlarga qaramasdan o‘zbek tilining avtomatik tarjima dasturi mavjud emas. Bir tildagi matnlarni o‘zbek tiliga tarjima qilish uchun “Google Translation” dasturidan keng foydalaniladi. Bu dastur bir qancha qulayliklarga ega bo‘lish bilan birga, tarjimada xatoliklar ham uchraydi. Bu, ayniqsa, ilmiy va rasmiy matnlar tarjimasida ko‘p kuzatiladi. Misol tariqasida Philipp Koehnning **“A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models”** maqolasidan parcha tarjima qilib ko‘rdik va dastur bizga quyidagi natijani berdi:



[72] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018). // Мухамедов С.А., Пиотровский Г.Г. Инженерная лингвистика и опыт системно – статистического исследования узбекских текстов. –Т.: Фан, 1986; Махмудов М.А., Пиотровская А.А., Садыков Т. Система машинного анализа и синтеза тюркской словоформы // Переработка текста методами инженерной лингвистики. –Минск, 1982.

[73] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).

Recombining Hypotheses — Recombining hypothesis is **a risk-free way** to reduce the search space. Two hypotheses can be recombined if they agree in (i) the foreign words covered so far, (ii) the last two English words generated, and (iii) the end of the last foreign phrase covered. If there are two paths that lead to two hypotheses that agree in these properties, we keep only **the cheaper hypothesis**, e.g., the one with the least cost so far. The other hypothesis cannot be part of the path to the best translation, and we can **safely discard** it. We do keep a record of the **additional arc** for **lattice generation**.

*Gipotezalarni qayta birlashtirish - Gipotezani qayta birlashtirish qidiruv maydonini qisqartirishning **xavf-xatarsiz** usulidir. Ikkita gipoteza, agar ular (i) hozirgacha yoritilgan xorijiy so‘zlar, (ii) yaratilgan so‘nggi ikkita inglizcha so‘z va (iii) oxirgi chet el iborasining oxirida mos kelsa, qayta birlashtirilishi mumkin. Agar ushbu xususiyatlarga mos keladigan ikkita farazga olib keladigan ikkita yo‘l mavjud bo‘lsa, biz faqat **arzonroq gipotezani**, masalan, hozirgacha eng **kam xarajatli gipotezani** saqlab qolamiz. Boshqa faraz eng yaxshi tarjima yo‘lining bir qismi bo‘lishi mumkin emas, va biz uni **xavfsiz tashlab yuborishimiz** mumkin. Biz **panjara hosil qilish uchun** qo‘shimcha **yoyni** qayd qilamiz.*

Tarjima natijasida ko‘rinib turibdiki, ayrim so‘zlar va terminlar matn ma’nosiga va gap mazmuniga to‘g‘ri kelmaydigan holda tarjima qilinyapti. Bu esa kontekstdagi ma’noning anglashilmasligiga olib keladi. Buning asosiy sabablaridan biri “Google Translate” dasturida o‘zbek tili ma’lumotlar bazasining to‘liq emasligidadir.

Philipp Koehn tavsiflab bergan statistik mashina tarjimasi modellarini va “Google Translate” ni analiz qilgan holda O‘zbek tilshunosligida ilk bor korpusga asoslangan mashina tarjimasi texnologiyasini joriy etmoqchimiz. Buning uchun bizga katta hajmdagi korpus kerak bo‘ladi. Buning uchun Nilufar Abduraxmonova rahbarligida yaratilgan https://uzbekcorpus.uz/ dan foydalanish maqsadga muvofiq deb topdik. Bizning asosiy maqsadimiz terminologik parallel korpuslar asosida ilmiy va rasmiy matnlarni tarjima qiladigan avtomatik mashina tarjima dasturini yaratishdir.

Matnlarni terminologik, struktur, morfologik, sintaktik tomondan to‘g‘ri tarjima qiladigan avtomat tarjima dasturi texnologiyasini yaratish uchun katta hajmdagi korpus bilan bir qatorda quyidagilar kerak bo‘ladi:

1. Parallel korpus
2. Morfo analizator
3. Sintaktik analizator
4. Tarjima lug‘atlar

Tarjima dasturi to‘g‘ri ishlashi uchun matnlarni o‘xshashlik jihatidan ehtimollilik modellarini yaratib chiqishimiz zarur bo‘ladi. Modellashtirishning quyidagi turlari mavjud:

- So‘z asosida modellashtirish, ya’ni so‘z segmentlari bo‘yicha muqobil variantlarni yaratish;
- So‘z birikmalari yoki frazalar bo‘yicha muqobillashtirish;
- Gaplarni o‘zaro muqobillashtirish.

Avvalo, yaratilayotgan mashina tarjimasi uchun taxminan nechta gap modeli kerak bo‘lishini aniqlab olish lozim. O‘zbek tilshunosligida gap modellari bir qancha olimlar tomonidan o‘rganilgan bo‘lib, xususan, Nilufar Abduraxmonova aynan tarjima dastur uchun ingliz va o‘zbek tillaridagi gap modellarini tadqiq etish maqsadida mashina tarjimasi nazariyasi bo‘yicha ilmiy izlanishlar olib borgan va “Mashina tarjimasining linvistik ta’minoti” nomli monografiyasining to‘rtinchi bobini aynan mashina tarjimasi tizimida gap modellarini yaratish nazariyasiga bag‘ishlagan.[74] Monografiyada 50 dan ortiq o‘zbek-ingliz tillaridagi gap modellari keltirilgan bo‘lib, ular ingliz tilidagi ot kesimli sodda gap va fe’l kesimli sodda gap modellari, shuningdek, inglizcha matnlardagi ega va kesim moslashuvining o‘zbek tilidagi modellari hisoblanadi. Tadqiqot ishida kesimi o‘timisiz fe’ldan iborat bo‘lgan asosiy gap modellarining 10 ta, kesimi yordamchi fe’l bilan hosil bo‘lgan asosiy gap modellarining 14 ta, kesimi o‘timli fe’l bo‘lgan asosiy gap modellarining 13ta, *to be* yordamchi fe’li ishtirok etadigan gap shakllarining 20 dan ortiq turi ko‘rsatib berilgan. Mashina tarjimasi to‘g‘ri ishlashi uchun qo‘shma gap modellari ham kerak bo‘ladi. Ingliz va o‘zbek tillaridagi qo‘shma gap modellarini avtomatik tarjima uchun ishlab chiqish oldimizda turgan vazifalardan biri bo‘lib hisoblanadi.

Mashina tarjimasi uchun matn so‘zlarni mazmuniy birliklarga ajratib chiqish muhim bosqich sananladi. Biz WordFast texnologiyasida sigmentatsiya jarayonini amalga oshiramiz. Bunda tadqiqot obyektimiz o‘zbek va ingliz tillaridagi avtorefetlar bo‘ladi.

O‘zbek-ingliz tillaridagi ilmiy va rasmiy matnlarni to‘g‘ri tarjima qila oladigan mashina tarjimasi texnologiyasini yaratish o‘zbek olimlar, umuman, ilimiy tadqiqot ishini olib boruvchilar uchun keng imkoniyat yaratadi. Yuqoridagilarni hisobga olib, “Google Translate” tarjima dasturining yutuq va kamchiliklarini tahlil qilgan holda o‘z tadqiqot ishimizda ilmiy terminlar va ilmiy matnlar bazasini kengaytirib, ilmiy parallel matnlarni to‘ldirish hisobiga boyitib o‘zbek tilshunosligida ilk bor mashin atrjimasi texnologiyasini yaratishni maqsad qildik.

[74] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).

FOYDALANILGAN ADABIYOTLAR

1. Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VIInternational Conference “Modern Problems of Applied Mathematics and Information Technology -Al-Khorezmiy 2018”, pp. 37–38, Tashkent, Uzbekistan (2018)

2. Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) ildis. aftoref.” (2018).

3. Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2 (38):12-7.

4. Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.

5. Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.

6. Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.

7. Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.

8. A. Ismailov, M. M. A. Jalil, Z. Abdullah and N. H. A. Rahim, "A comparative study of stemming algorithms for use withthe Uzbek language," 2016 " 3rd International Conference on Computer and Information Sciences (ICCOINS), 2016, pp. 7-12, doi: 10.1109/ICCOINS.2016.7783180.

9. Philipp Koehn. A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. Machine Translation: From Real Users to Research 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004. Washington, DC, USA, September/October 2004

10. Jalil, Masita & Ismailov, Alisher & Abd Rahim, Noor Hafhizah & Abdullah, Zailani. (2017). The Development of the Uzbek Stemming Algorithm. Advanced Science Letters. 23. 4171-4174. 10.1166/asl.2017.8332