УДК 81’33, 81.512.1

Б.Э. Хәкимов,
М.Р. Шәехов

РУСЧА-ТАТАРЧА МАШИНА ТӘРҖЕМӘЧЕСЕ ӨЧЕН ПАРАЛЛЕЛЬ ТЕСТ КОРПУСЫНДА ҖӨМЛӘЛӘР ТӘҢГӘЛЛЕГЕ

При подготовке обучающих и тестовых корпусов для систем машинного перевода осуществляется поиск готовых параллельных текстов на двух языках в различных доступных источниках. Обычно считается, что пары предложений из оригинальных текстов и их переводов, если они были опубликованы и имеют официальный статус, являются эквивалентными. Однако в результате анализа таких предложений установлено, что степень их эквивалентности в определенных случаях может не соответствовать задачам разработки систем машинного перевода. В данной статье изучаются типичные случаи, в которых проявляются различия в требованиях к эквивалентности перевода предложений в исходных текстах и в качестве отдельных единиц параллельного тестового корпуса, используемого для оценки качества результатов машинного перевода. В результате исследования установлено, что предложения, извлекаемые из параллельных текстов на двух языках, и их переводы даже при полной контекстной эквивалентности могут потребовать значительного количества лексических, грамматических и других трансформаций.

Ключевые слова: машинный перевод, качество перевода, параллельный корпус, эквивалентность перевода, татарский язык, русский язык.

The current practice of preparing training and test corpora for machine translation systems usually involves searching for parallel texts in real use. Pairs of sentences from the original texts and their published translations are considered a priori equivalent. However, a detailed qualitative analysis of such sentences shows that this does not always meet the requirements of machine translation systems. This paper discusses typical differences between the equivalence of texts and the equivalence of parallel corpus units. The results show that sentences from the authentic texts and their translations, despite of their complete parallelism and contextual equivalence, may require a significant amount of lexical, grammatical and other transformations.

Keywords: machine translation, translation quality, parallel corpus, translation equivalence, Tatar language, Russian language.

Нейрочелтәрле машина тәрҗемәсе системалары өчен өйрәтүче корпуслар һәм тест корпусларын әзерләүнең төп ысулы мөмкин кадәр күбрәк чыганаклардан ике телдә әзер параллель җөмләләр туплаудан гыйбарәт. Русча-татарча машина тәрҗемәсе сыйфатын бәяләү өчен, махсус параллель корпус төзелде [Хакимов, Шаехов, с. 283–292]. Гадәттә мондый төр корпусларны әзерләгәндә оригиналь текстлардан һәм рәсми, басылып чыккан тәрҗемәләрдән алынган җөмлә парлары тулысынча эквивалент, тәңгәл дип санала. Әмма андый җөмләләргә анализ ясау нәтиҗәсендә аларның тәңгәллек дәрәҗәсе машина тәрҗемәсе системалары таләпләренә җавап бирмәскә дә мөмкин икәнлеге ачыклана.

Эквивалентлык – тәрҗемә гыйлеменең төп мәсьәләләреннән берсе. Тәрҗемә теориясендә ул төрле лингвистик һәм экстралингвистик факторларга бәйле булган чагыштырма характеристика буларак карала. В.Н. Комиссаров эквивалентлыкны тәрҗемә белән төп нөсхәнең максималь мөмкин булган лингвистик якынлыгы дип билгели һәм төп нөсхә белән тәрҗемә арасында аның берничә дәрәҗәсен аерып күрсәтә: коммуникация максаты дәрәҗәсе, ситуацияне тасвирлау дәрәҗәсе, ситуацияне тасвирлау ысулы дәрәҗәсе, әйтелгәннәрнең структур төзелеше дәрәҗәсе, лексик берәмлекләр семантикасы дәрәҗәсе [Комиссаров, б. 167]. Л.С. Бархударов фикеренчә, эквивалентлык – ул төп нөсхә һәм тәрҗемә телендә текстларның мәгънәви яктан туры килүендә чагыла торган семантик категория [Бархударов, б. 150]. Шулай итеп, тәрҗемәнең реципиентка йогынтысы төп нөсхәгә тиңдәш булсын өчен, аларның тәңгәллеге лексик, грамматик һәм синтаксик дәрәҗәдә генә түгел, ә коммуникатив ситуация һәм максат дәрәҗәсендә дә тәэмин ителергә тиеш. Тагын да катлаулырак дәрәҗәләрдәге эквивалентлыкка ирешү өчен, тел чараларының гади тәңгәллегенә игътибар итмәскә мөмкин дип санала.

Әлеге мәсьәлә машина тәрҗемәсе максатларында параллель корпуслар әзерләү процессында үзенчәлекле рәвештә чагылыш таба. Машина тәрҗемәсенең теләсә кайсы конкрет нәтиҗәсенә карата төрле дәрәҗәләрдәге тәңгәллек күзлегеннән барлык таләпләр кулланырлык булса да, мондый прагматик юнәлешле параллель корпусларда урын алган җөмлә парлары аерым үзенчәлекләргә ия.

Төп үзенчәлек контекст күләме белән бәйле. Тәрҗемә ителә торган эчтәлекнең контекстуаль мәгънәсе бөтен текст яки аның бер өзеге дәрәҗәсендә аерылып тора. Кагыйдә буларак, машина тәрҗемәсе максатларына хезмәт итә торган параллель корпусларга аерым җөмләләр контекст бәйләнешләрен сакламыйча гына урнаштырыла, чөнки текстлардан барлык җөмләләр түгел, ә билгеле бер критерийларга туры килгәннәре генә сайлап алына. Димәк, чыганак текстлардагы параллель җөмләләр ике телдә дә коммуникатив ситуация һәм контекст дәрәҗәсендә эквивалент булырга мөмкин (мәсәлән, әдәби әсәрләрдә төп эчтәлек – образлар аша, ә матбугаттагы яңалык текстларында фактлар аша бирелә). Параллель корпуска кертелгәч исә, контексттан максималь бәйсезлек шартларында шул ук җөмләләр структур төзелеш яисә лексик берәмлекләр дәрәҗәләрендә тәңгәллек таләпләренә инде туры килмәскә дә мөмкин. Башкача әйткәндә, тәрҗемә ителгән җөмлә эквивалентлыгының кайбер факторлары әлеге җөмлә конкрет контексттан аерылгач юкка чыга. Шуңа бәйле рәвештә мондый корпусларда параллель җөмләләрне бәяләү һәм аларга төзәтмәләр кертү өчен, билгеле бер кагыйдәләр билгеләү зарурлыгы туа. Мәсәлән, җөмләләр тәңгәллегенә ирешү тагын да төгәлрәк дәрәҗәдә мөмкин булган очракта, тиешле төзәтмәләр кертеп, мөмкин кадәр күбрәк контекстларга туры килерлек вариант барлыкка китерү максатка ярашлы була.

Җөмлә парларының контексттан бәйсез эквивалентлыгы аларның төп нөсхәдәнме, тәрҗемәдәнме булуына караганда мөһимрәк, шуңа күрә дә төзәтмәләр параллель җөмләләрнең икесенә дә кертелә ала. Без аерым очракларда хәтта әдәби әсәрләрдән алынган оригиналь җөмләләрне дә үзгәртергә мөмкин дип саныйбыз, чөнки матур әдәбият тәрҗемәсендә гадәттә югарырак дәрәҗәдәге эквивалентлык турында сүз бара. Бер яктан, корпуста реаль тексттан булган оригиналь җөмләне саклап калу өчен, тәрҗемәгә үзгәрешләр кертү кулайрак. Икенче яктан, әгәр корпус лингвистик күренешләрнең төрлелеге (шул исәптән грамматик формалар һәм лексик берәмлекләр) ягыннан репрезентативлырак булса, машина тәрҗемәсе сыйфатын бәяләүнең нәтиҗәлелеге дә арта. Шулай итеп, төзәтү варианты һәр очракта аерым сайлана. Моннан тыш, һәр вариантны редакцияләүгә сарыф ителә торган вакыт һәм хезмәт күләме дә игътибарга алынырга мөмкин.

Гомумән алганда, машина тәрҗемәсе нәтиҗәләрен бәяләү өчен параллель җөмләләр корпусын әзерләгәндә, экспертка параллель җөмлә парларының максималь эквивалентлыгы һәм тәрҗемәнең табигыйлеге арасында баланс саклау бурычы йөкләнә.

Русча-татарча машина тәрҗемәсе сыйфатын анализлау өчен хезмәт итә торган параллель җөмләләр корпусын төзү барышында, текст составындагы һәм корпусның аерым берәмлекләре буларак җөмләләр эквивалентлыгына карата булган таләпләрдәге аермаларны истә тотып, түбәндәге төр төзәтмәләр кертелде. Әлеге мәкаләдә параллель русча-татарча тест корпусыннан [Хусаинов, Хакимов, Шаехов, 2022] татарча җөмләләрне төзәтү мисаллары китерелә: башта чыганак текстлардан үзгәрешсез алынган җөмләләр, җәя эчендә татарча җөмләнең төзәтелгән варианты бирелә.

1. Контексттан бәйсез вариантка үзгәртү: *Бушуева **сидела** в кабине второй машины. – Бушуева икенче машинаның кабинасында **бара иде** (Бушуева икенче машинаның кабинасында **утыра иде**)*.

Әлеге мисалда транспорт чарасы пассажирына карата *сидеть – барырга* фигыльләренең контекст эквивалентлыгы күзәтелә. Татарча җөмләгә төзәтмә кертү барышында фигыль туры тәңгәллеккә алмаштырыла, югарырак дәрәҗәдәге эквивалентлыкка да зыян килми, җөмләнең табигыйлеге югалмый. Димәк, бу очракта төзәтмә кертү үзен аклый. Тагын бер мисалда тәрҗемә төгәлрәк вариантка үзгәртелә: *Все знают – и никто **не понимает**. – Барысы да белә – һәм беркем **аңлата алмый** (Барысы да белә – һәм беркем дә **аңламый**)*.

2. Параллель җөмләдә туры тәңгәллеге булмаган сүзләрне бетерү: *Вспомнила Суок из сказки «Три толстяка». – «Три толстяка» әкиятендәге Суокны **да** искә төшерде.* (*«Три толстяка» әкиятендәге Суокны искә төшерде).*

Мисаллардан күренгәнчә, параллель җөмләләр еш кына аныклаучы яки көчәйтүче функция башкара торган тел чаралары йөрткән мәгънә төсмерләре белән аерылып торырга мөмкин: *По словам Тяминова Артура, картофель хорошо хранится. – Тяминов Артур*

әйтүенчә, бәрәңге ***бик*** *әйбәт саклана (Тяминов Артур әйтүенчә, бәрәңге әйбәт саклана).*

3. Грамматик форманы үзгәртү: *Минздрав ответил на предложение вернуть курилки в* ***аэропорты****. – Сәламәтлек саклау министрлыгы тәмәке тарту урыннарын* ***аэропортка*** *кайтару тәкъдименә җавап бирде. (Сәламәтлек саклау министрлыгы тәмәке тарту урыннарын* ***аэропортларга*** *кайтару тәкъдименә җавап бирде).*

Китерелгән мисалда, җөмләләр эквивалентлыгын арттыру өчен, төп нөсхәдә булган күплек сан формасы тәрҗемәдә дә торгызыла.

4. Стилистик хаталарны төзәтү һәм тәрҗемәне стилистик яктан үзгәртү: ***Убежище*** *в январе в ЕС получили только 13 граждан Грузии, а 1288* ***гражданам дали отказ****. – Гыйнварда ЕСта* ***сыенуны*** *Грузиянең 13 гражданы гына алган, ә 1288* ***гражданга кире кагылган****. (Гыйнварда ЕСта* ***сыену мөмкинлеген*** *Грузиянең 13 гражданы гына алган, ә 1288* ***гражданның үтенече кире кагылган****).*

Бу төрдәге мисаллар күрсәткәнчә, чыганак буларак рәсми нәшер ителгән текстлар һәм аларның тәрҗемәләре кулланылуга карамастан, җөмләләрдә хаталар да булырга мөмкин (ешрак стилистик характерда). Тәңгәллек дәрәҗәләре күзлегеннән караганда, мондый хаталарны төзәтү югарыдагы мисаллар белән чагыштырганда капма-каршы нәтиҗәгә китерә, ягъни төзәтмәләр керткәч, бигрәк тә билгеле бер контекст белән турыдан-туры бәйле булмаган очракта, тәңгәллекнең түбәнрәк дәрәҗәсенә генә түгел, ә югарырагына да ирешергә мөмкин.

Түбәндәге мисалда исә төзәтүнең төрле юнәлештәге ике төрен күрергә була: бер яктан, *бумага* сүзенең тәрҗемәсе контекстка бәйле булмаган вариантка алмаштырыла һәм, шулай итеп, тәңгәллек түбәнрәк дәрәҗәгә күчә. Икенче яктан, сүзгә-сүз тәрҗемәне үзгәртү исәбенә (*со словами – мондый сүзләр белән*), сүзтезмәнең гомуми эквивалентлык дәрәҗәсе күтәрелә: *Они должны за подписью руководителя присылать в Роскомнадзор* ***бумагу со словами****:… – Алар җитәкче имзасы белән «Роскомнадзор»га* ***мондый сүзләр белән хәбәр*** *җибәрергә тиешләр:… (Алар җитәкче имзасы белән Роскомнадзорга* ***мондый эчтәлекле кәгазь*** *җибәрергә тиешләр:…)*

5. Телләрнең берсендә синонимик лексик берәмлек яки грамматик форма булганда вариант өстәү: *Приглашать в школы* ***героев*** *войны. – Мәктәпкә сугыш* ***геройларын*** *чакырырга (1) Мәктәпкә сугыш* ***геройларын*** *чакырырга. 2) Мәктәпкә сугыш* ***каһарманнарын*** *чакырырга).*

Әлеге мисалдагы русча *герой* сүзенең татар телендә ике варианты бар. Димәк, әгәр тест корпусында синонимнарның берсе генә теркәлгән булса, автомат тикшерү вакытында тәрҗемәнең тулысынча эквивалент булган икенче, синонимик варианты төгәл түгел дип бәяләнәчәк. Шул сәбәпле параллель корпуска ике пар җөмлә өстәлә, аларда бер телдә бер генә вариант кулланыла, ә икенчесендә аерыла. Мондый хәл лексика дәрәҗәсендә дә, грамматика дәрәҗәсендә дә барлыкка килергә мөмкин. Түбәндәге мисалдан күренгәнчә, рус телендәге фигыльнең инфинитив формасының татар телендә ике эк-

виваленты бар: *В теории его* ***сосчитать*** *несложно. – Теориядә аны* ***санарга*** *авыр түгел (1) Теориядә аны* ***санарга*** *авыр түгел. 2) Теориядә аны* ***санау*** *авыр түгел).*

Телләрнең берсендә теге яки бу грамматик мәгънәнең формаль күрсәткече булмаган очракларда да параллель җөмлә вариантларын өстәргә кирәк була. Мәсәлән, рус телендә үткән заманда фигыльнең зат күрсәткечләре юк, татар телендә исә бу грамматик мәгънә ачык белдерелә: *Так и* ***скупила*** *все в этом маленьком магазине. – Шулай итеп, бу кечкенә кибеттәге бөтен әйберне* ***сатып алып та бетердем*** *(1) Шулай итеп, бу кечкенә кибеттәге бөтен әйберне* ***сатып алып та бетердем****. 2) Шулай итеп, бу кечкенә кибеттәге бөтен әйберне* ***сатып алып та бетерде****).*

Тагын да төгәлрәк караганда, соңгы мисалда *сатып алып та бетердең* дигән тагын бер вариантны өстәргә кирәк. Ә татар телендә үткән заманның төп ике формасы булуын да исәпкә алсак, мөмкин булган вариантлар саны икеләтә арта. Бу очракта ике стратегия булырга мөмкин: мөмкин булган барлык вариантларны кертү яисә статистик яктан ешрак очрый торганнарын сайлау.

Тикшеренү нәтиҗәләре күрсәткәнчә, еш кына чыганак текстлар һәм аларның тәрҗемәләре арасында тулы контекстуаль эквивалентлык булганда да, аерым җөмлә парлары буларак махсус корпуска өстәгәндә, шактый күләмдә лексик, грамматик һәм башка трансформацияләр таләп ителә. Шундый лингвистик мәгълүмат җыелмаларының гади тел эквивалентлыгы күзлегеннән караганда төгәлрәк булуы машина тәрҗемәсенең сыйфатын арттырырга мөмкинлек бирә.

Әдәбият

Бархударов Л.С. Язык и перевод. М., 1975. 230 с.

Комиссаров В.Н. Общая теория перевода (лингвистические аспекты): Учеб. для ин-тов и фак. иностр. яз. М.: Высшая школа, 1990. 253 с.

Федоров А.В. Основы общей теории перевода. М.: Высшая школа, 1983. 300 с.

Хакимов Б.Э., Шаехов М.Р. К вопросу создания параллельного тестового корпуса для задачи машинного перевода в русско-татарской паре // Восьмая Международная конференция по компьютерной обработке тюркских языков «Turk-Lang-2020». (Труды конференции). Уфа: ИИЯЛ УФИЦ РАН, 2020. С. 283–292.

Хусаинов А.Ф., Хакимов Б.Э., Шаехов М.Р. База данных тестового параллельного корпуса на русском и татарском языках. Свидетельство о регистрации базы данных 2022620573, 17.03.2022.

Koller W. Equivalence in Translation Theory // Ed. A. Chesterman, Readings in Translation Theory. Helsinki: Oy Finn Lectura Ab., 1989. P. 99–104.

Хәкимов Булат Эрнст улы,
филология фәннәре кандидаты, Казан федераль университетының билингваль һәм цифрлы белем бирү кафедрасы доценты,
ТР ФА Гамәли семиотика институтының әйдәп баручы фәнни хезмәткәре

Шәехов Марат Рәшит улы,
ТР ФА Гамәли семиотика институтының әйдәп баручы фәнни хезмәткәре